Université Paris Ouest
Nanterre La Défense

Ecole Doctorale 139
Connaissance, Langage, Modélisation

THESE DE DOCTORAT
Mathématiques appliquées et applications des mathématiques

présentée par

Saïp CISS

sous le titre

Forêts uniformément aléatoires et détection des irrégularités aux cotisations sociales

Directeurs de Thèse : MM. Patrice BERTAIL et Pierre PICARD

soutenue le 20 juin 2014

Jury :
M. Patrice BERTAIL, Université Paris Ouest, Professeur, directeur
M. Gérard BIAU, Université Pierre et Marie Curie (UPMC), Professeur, examinateur
M. Stephan CLÉMENÇON, Telecom-ParisTech, Professeur, examinateur
M. Pierre PICARD, École Polytechnique, Professeur, directeur
M. Jean PINQUET, Université Paris Ouest et École Polytechnique, Enseignant-Chercheur, examinateur
M. Vincent RAVOUX, Directeur du réseau de la Caisse Nationale d'Allocations Familiales (CNAF), examinateur
M. Fabrice ROSSI, Université Paris 1, Professeur, rapporteur
M. Nicolas VAYATIS, ENS Cachan, Professeur, rapporteur

Résumé

Nous présentons dans cette thèse une application de l'apprentissage statistique à la détection des irrégularités aux cotisations sociales. L'apprentissage statistique a pour but de modéliser des problèmes dans lesquels il existe une relation, généralement non déterministe, entre des variables et le phénomène que l'on cherche à évaluer. Un aspect essentiel de cette modélisation est la prédiction des occurrences inconnues du phénomène, à partir des données déjà observées. Dans le cas des cotisations sociales, la représentation du problème s'exprime par le postulat de l'existence d'une relation entre les déclarations de cotisation des entreprises et les contrôles effectués par les organismes de recouvrement. Les inspecteurs du contrôle certifient le caractère exact ou inexact d'un certain nombre de déclarations et notifient, le cas échéant, un redressement aux entreprises concernées. L'algorithme d'apprentissage *apprend*, grâce à un modèle, la relation entre les déclarations et les résultats des contrôles, puis produit une évaluation de l'ensemble des déclarations non encore contrôlées. La première partie de l'évaluation attribue un caractère régulier ou irrégulier à chaque déclaration, avec une certaine probabilité. La seconde estime les montants de redressement espérés pour chaque déclaration. Au sein de l'URSSAF (Union de Recouvrement des cotisations de Sécurité sociale et d'Allocations Familiales) d'Île-de-France, et dans le cadre d'un contrat CIFRE (Conventions Industrielles de Formation par la Recherche), nous avons développé un modèle de détection des irrégularités aux cotisations sociales que nous présentons et détaillons tout au long de la thèse.
L'algorithme fonctionne sous le logiciel libre R et est disponible sous la forme d'un paquet : *randomUniformForest*. Il est entièrement opérationnel et a été expérimenté en situation réelle durant l'année 2012. Pour garantir ses propriétés et résultats, des outils probabilistes et statistiques sont nécessaires et nous discutons des aspects théoriques ayant accompagné sa conception. Dans la première partie de la thèse, nous effectuons une présentation générale du problème de la détection des irrégularités aux cotisations sociales. Dans la seconde partie, nous abordons la détection spécifiquement, à travers les données utilisées pour définir et évaluer les irrégularités. En particulier, les seules données disponibles suffisent à modéliser la détection. Nous y présentons également un nouvel algorithme de forêts aléatoires, nommé *forêts uniformément aléatoires*, qui constitue le moteur de détection. Dans la troisième partie, nous détaillons les propriétés théoriques des forêts uniformément aléatoires. Dans la quatrième partie, nous présentons un point de vue économique, lorsque les irrégularités aux cotisations sociales ont un caractère volontaire, cela dans le cadre de la lutte contre le travail dissimulé. En particulier, nous nous intéressons au lien entre la situation financière des entreprises et la fraude aux cotisations sociales. La dernière partie est consacrée aux résultats expérimentaux et réels du modèle, dont nous discutons, et à l'évaluation complète de l'ensemble des entreprises d'Île-de-France.

Chacun des chapitres de la thèse peut être lu indépendamment des autres et quelques notions sont redondantes afin de faciliter l'exploration du contenu.

Mots clés : apprentissage statistique, apprentissage automatique, classification, régression, forêts aléatoires, forêts uniformément aléatoires, irrégularités, fraude, cotisations sociales, URSSAF d'Île-de-France.

Abstract

We present in this thesis an application of statistical and machine learning to irregularities in the case of social contributions. These are, in France, all contributions due by employees and companies to the *Sécurité sociale*, the French system of social welfare (alternative incomes in case of unemployment, Medicare, pensions, ...). Social contributions are paid by employees and companies to the URSSAF network, which in charge to recover them. Our main goal was to build a model that would be able to detect irregularities with a little false positive rate. We, first, begin the thesis by presenting the URSSAF and how irregularities can appear, how can we handle them and what are the data we can use. Then, we talk about a new machine learning algorithm we have developed for, *random Uniform Forests* (and their R package *randomUniformForest*) which are a variant of Breiman's *Random Forests* (tm), since they share the same principles but in a different way. We present theoretical background of the model and provide several examples. We use random Uniform Forests to show, when irregularities are fraud, how financial situation of companies can affect their propensity for fraud. In the last chapter, we provide a comprehensive assessment of declarations of social contributions for all companies in Ile-de-France for 2013, by using the model to predict if declarations present irregularities or not, and by estimating the amount of possible recovery for each company.

keywords : machine learning, ensemble learning, classification, regression, random forests, random uniform forests, decision trees, irregularities, fraud, social contributions, URSSAF of Île-de-France.

Laboratoire de Modélisation aléatoire (MODAL'X)
Université Paris Ouest Nanterre La Défense
Modal'X
Bât G, bur. E04
200, avenue de la république - 92000 NANTERRE

Laboratoire d'économétrie
Département d'Économie
Ecole Polytechnique
91128 Palaiseau Cedex, FRANCE

URSSAF d'Île-de-France
22, rue de Lagny - 93100 Montreuil

A Selma Ciss-Lagorce

Remerciements

Je voudrais remercier tous ceux qui ont permis la réalisation de cette thèse et que je ne saurais tous citer. J'avais cru que les remerciements seraient une étape simple, mais éviter de dire merci à chaque phrase est un exercice ardu, tant les rencontres, les idées et discussions avec tous m'ont été utiles.

Mes premiers mots vont à Vincent Ravoux ainsi qu'à mes deux directeurs de thèse Patrice Bertail et Pierre Picard, sans qui cette thèse n'aurait jamais vu le jour. Merci, Patrice, de m'avoir proposé cette collaboration. j'en sais un peu plus sur l'écriture mathématique, un peu plus sur la rigueur nécessaire et je t'en suis grandement redevable. Surtout, Pierre et toi avez été d'une pédagogie rare et d'un soutien sans faille. Merci, Pierre, pour les nombreuses pistes que tu m'as suggérées et pour ton grand savoir-faire économique. La majorité du contenu de cette thèse doit beaucoup à vos idées.

Merci, Vincent, de l'accueil au sein de l'URSSAF d'Île-de-France. Votre capacité de compréhension et d'analyse m'a énormément aidé et je vous suis redevable d'avoir pu proposer une alternative concrète aux problématiques développées dans cette thèse. Yves Rebouillat et vous avez grandement facilité mon intégration au sein d'une entreprise dans laquelle je me suis beaucoup épanoui. Merci, Yves, du cadre de travail que tu m'as accordé. Il a été précieux et des éléments essentiels de ce travail doivent à la grande liberté dont j'ai pu bénéficier.

Je voudrais, bien sûr, chaleureusement remercier l'ensemble des membres du jury.
Merci à Fabrice Rossi et à Nicolas Vayatis d'avoir accepté d'être les rapporteurs de cette thèse. Longue elle est, et je vous formule toute ma gratitude d'en avoir été les rapporteurs exigeants et attentifs.
A Stéphan Clémençon et à Jean Pinquet, j'adresse mes sincères remerciements.
A Gérard Biau, président du jury, j'exprime ma profonde reconnaissance. Tes travaux sur les forêts aléatoires ont imprimé leur marque sur cette thèse et demeurent, pour moi, des références.

L'URSSAF d'Île-de-France a été un des lieux de ce travail de thèse. La diligence et la sympathie de toutes les équipes et personnes que j'y ai rencontrées fut une grande joie. Tout d'abord, merci à Nathalie Hergaux, collègue et camarade. Nos discussions restent un moteur de mes projets et travaux. Merci à toute l'équipe de la (l'ex-) direction de l'Organisation. Je n'oublie surtout pas Gaëlle Jacq. Merci, Gaëlle, pour toute ton aide. Sans toi, point de bases de données et point de modèle. Merci à Serge Mercier et à toute l'équipe du service Statistiques. Merci, Serge, pour ta diligence et tes précieux conseils. Merci Audrey, condisciple de mes débuts à l'URSSAF et déchiffreuse des indicateurs statistiques ; tu m'as évité bien des déconvenues. Merci Virginie, j'espère qu'Audrey et toi pourrez continuer à tracer le sillon. Merci William, Christine, Françoise, Camille. La convivialité et la gentillesse dont vous avez fait preuve m'honorent. Merci, Colette, de votre parfaite maitrise des arcanes administratifs. Le temps nous manque toujours et je vous sais gré de m'en avoir fait gagner à chaque étape de ce travail. Merci à Alain Hurtrel, garant des contrats clairs et concis. Je suis également redevable à l'ANRT, qui a permis l'établissement de ce cadre (CIFRE), entre entreprise et laboratoire.

Je remercie l'ensemble des inspecteurs chargés du contrôle. Votre aide a été essentielle et m'a permis de rapidement comprendre comment aborder le problème de la détection des irrégularités aux cotisations sociales. Je voudrais, en particulier, saluer votre perspicacité qui m'a marqué. Merci à toute l'équipe de la direction du Contrôle pour son implication dans les expérimentations opérationnelles.

Bien que cette thèse soit profondément ancrée dans la volonté d'une application effective et complète, mes laboratoires d'accueil, Modal'X de l'Université Paris Ouest Nanterre et le laboratoire d'Econométrie de l'Ecole Polytechnique, ont été d'une grande disponibilité. A Nathanaël Enriquez, merci de m'avoir accueilli au sein de Modal'X. A Mélanie Zetlaoui, Anna-Karina Fermin, Omar El Dakkak, merci de vos remarques qui m'ont tout particulièrement aidé à formaliser les arbres de décision.

C'est l'occasion ou jamais, et je voudrais remercier l'ensemble des professeurs qui m'ont mené jusque-là. A Mathilde Mougeot, pour votre grande pédagogie, à MM. Sarr (votre cours sur les nombres complexes, 6-ème Bleue, fût un déclic), Sall, Amavi qui m'ont fait découvrir et apprécier les mathématiques, j'adresse mes profonds remerciements. Merci au Cours Sainte-Marie de Hann ; son exceptionnel cadre d'enseignement, sa diversité incroyable et son ouverture culturelle sont des images qui me suivent à tout moment. A Doudou Konaté, Gilles, Georges, Jean, Cheikh, et à tous les autres, nos discussions et nos amitiés d'alors ont, plus que vous ne croyez, contribué à cette thèse.

Je ne pourrais oublier les amis et la famille qui m'ont soutenu dans les moments les moins évidents. Les mathématiques sont une éternelle terre de doutes, faite de chemins arides et sinueux, où les oasis sont rares. Le voyageur, lui, harassé et quelquefois hagard, ne croise sur cette route sans fin que mirage sur mirage dans le désert brûlant. Seule la soif d'idées et de connaissance demeure alors. Vous avez été là dans ces moments et sans vous le chemin eût été autrement plus éprouvant. A Diane, Xavier, Véro, Xav', merci d'avoir été là, vous connaître est un honneur. A Régine Rossi-Lagorce et Bernard Lagorce, merci de tout coeur, pour le soutien que vous m'avez apporté, pour votre amitié et votre enthousiasme. A Jacqueline, Bruno, Pascal, Renaud, Charlotte, Maxime, Elodie, Emilie, Ameth, merci d'avoir contribué, de près ou de loin, à cette thèse.
A Juliette, ton soutien a été décisif.
A Mounirou Ciss, Ndeye Diouf, Colé Seye, merci pour tout.
A mes parents, de tous vos efforts, je vous serai toujours reconnaissant.
A Nelly, amie de tous les doutes et de tous les succès.
A Laurine et à ma Selma préférée, vous êtes là et merci est une expression bien trop courte pour dire combien vous comptez.

Je voudrais, pour finir, saluer les travaux de MM. Vapnik et Chervonenkis, dont la théorie de l'Apprentissage Statistique fascine et façonne encore mon imaginaire. Mon apprentissage dans cette thèse est essentiellement dû au merveilleux ouvrage de Devroye, Györfi et Lugosi (*A Probabilistic theory of Pattern Recognition*), devenu livre de chevet, à l'instar d'un Ellroy ou d'un Simmons. Je ne peux terminer sans saluer la mémoire et l'oeuvre du Professeur Leo Breiman, inventeur des Forêts Aléatoires, dont la philosophie et l'intuition guident, aujourd'hui et demain, l'ensemble de mes travaux.

Ne garder ou n'espérer écrire que l'essentiel dans cette thèse fût une bataille.
Lecteur occasionnel ou lecteur assidu, merci d'avoir pris un peu de ton temps pour parcourir ces lignes.

Table des matières

Chapitre 1

Présentation

1.1 Introduction

Ce premier chapitre introduit les différents points d'ancrage de la détection des irrégularités aux cotisations sociales. Dans un premier temps, nous en distinguons le contexte ainsi que les contributions apportées par ce mémoire de thèse.
Puis, nous présentons l'URSSAF et la mission de recouvrement et de contrôle qui lui est attribuée par la Sécurité sociale.
Dans un troisième temps, nous détaillons les éléments essentiels à l'équilibre entre les cotisations et les prestations sociales. En particulier, nous montrons que la garantie de cet équilibre peut difficilement être établie en l'absence de contrôles précis. Cet aspect constitue le principal argument de ce chapitre. De manière plus explicite, nous suggérons que les contrôles peuvent être une ressource supplémentaire, et très largement supérieure aux montants qu'ils génèrent à ce jour, de financement pour la Sécurité sociale, sans augmentation de leur nombre. Ce résultat est développé tout au long de la thèse et nous montrons comment le mettre en oeuvre.
Dans la quatrième partie de ce chapitre, nous formalisons le problème de la détection et de l'évaluation d'un point de vue opérationnel.
Dans la dernière partie, l'aspect théorique est introduit à travers une présentation de l'apprentissage statistique. Nous en donnons une brève définition et une traduction des problématiques de la détection d'irrégularités dans un cadre probabiliste. Nous discutons essentiellement des méthodes ensemblistes et de la manière dont elles résolvent le problème.

1.2 Contexte de la détection des irrégularités aux cotisations sociales et contributions

Les cotisations sociales sont l'ensemble des versements effectués par les salariés et les employeurs et destinés à financer les mécanismes et dispositifs de la protection sociale (Assurance Maladie, Accidents, Allocations familiales, Retraites, Assurance Chômage,...). Les cotisations sociales sont déclarées par les employeurs et recouvrées par le réseau des URSSAF.

La problématique des irrégularités aux cotisations sociales est avant tout financière et peut s'illustrer par la juxtaposition du déficit de la Sécurité sociale et du montant annuel estimé des irrégularités. Si les deux se chiffrent à plusieurs milliards d'euros, il n'apparaît pas, à ce jour, possible de réintégrer les sommes dues aux irrégularités pour un meilleur équilibre des comptes. Un argument naturel de cette observation est l'hypothèse qu'il faudrait effectuer un bien plus grand nombre de contrôles pour aboutir à des sommes proches de celles estimées pour l'ensemble des irrégularités. Plusieurs éléments y contribuent :
- les montants redressés, chaque année, ne représentent qu'une part minoritaire des montants estimés ;
- toutes les sommes redressées ne sont pas récupérées, en particulier du fait de leur asymétrie (un petit nombre de redressements représentent la majorité des sommes) ;
- le taux de détection des irrégularités n'est pas suffisamment élevé pour permettre une augmentation importante du nombre de contrôles.

L'objectif de cette thèse était donc de proposer un modèle opérationnel, pour toutes les entreprises d'Île-de-France, qui permette :
- d'accroître de manière importante le taux de détection ;
- une augmentation significative des montants redressés sans nécessiter plus de contrôles que ceux réalisés ;
- de rendre récupérables la grande majorité des montants redressés sur la base des recommandations du modèle ;
- de fournir des garanties sur les résultats avant que les contrôles recommandés ne soient effectués.

L'élément central dans sa réalisation a été la possibilité de fournir un processus complet d'évaluation pour l'ensemble des entreprises d'Île-de-France, grâce aux données de l'URSSAF. Une évaluation par un modèle à une telle échelle n'avait, à notre connaissance, jamais été réalisée pour les irrégularités aux cotisations sociales.

Sur le plan opérationnel, notre contribution a été de rendre réalisable cette évaluation. Pour les années 2012 et 2013, toutes les déclarations de toutes les entreprises d'Île-de-France (avec au moins un salarié) ont pu être évaluées afin d'en déterminer le caractère régulier ou irrégulier, ainsi que les estimations des redressements espérés. Le modèle proposé apporte, en particulier, des garanties explicites, aussi bien sur un taux de détection plus important que celui mesuré que sur le montant net des redressements espérés. Il n'a pas vocation à remplacer la politique de contrôle de l'URSSAF car il en est simplement une généralisation.

De manière encore plus précise, il est, à ce jour, possible d'estimer, avec un niveau de confiance élevé, le taux de détection et les montants de redressement qui seraient réalisés à partir des recommandations du modèle. Il n'y a, notamment, aucun prérequis à son application, hormis la disponibilité des bases de données des URSSAF (y compris hors d'Île-de-france).

Du point de vue théorique, la contribution de cette thèse est le développement d'un nouvel algorithme, aux performances similaires à l'état de l'art, inspiré des idées et des forêts aléatoires de Breiman (2001). Nous discutons de ses propriétés, de ses performances et de son application à la détection des irrégularités aux cotisations sociales.

La dernière contribution est une analyse empirique de l'impact de la situation financière des entreprises sur le niveau de fraude (les irrégularités volontaires) dans le cas du travail dissimulé. A partir d'un échantillon de données réelles et synthétiques, nous illustrons la manière dont certaines variables économiques, liées au bilan et au compte de résultats de l'entreprise, influencent la propension à la fraude.

Le point commun entre ces trois éléments fondateurs réside dans les interactions qui les lient. Un impératif de ce travail était, notamment, de rendre opérationnel l'ensemble des thèmes abordés et chacune des contributions concourt à la réalisation des autres.

1.3 L'URSSAF et la mission de recouvrement et de contrôle des cotisations sociales

Le réseau des URSSAF (Union de Recouvrement des cotisations de Sécurité sociale et d'Allocations Familiales) constitue l'organisme chargé du recouvrement et du contrôle des cotisations sociales. A la suite d'une réforme, leur régionalisation est engagée afin de constituer une URSSAF par région. Le cadre de cette thèse s'est déroulé, en grande partie, au sein de l'URSSAF d'Île-de-France, pour laquelle le moteur de détection d'irrégularités a été conçu et expérimenté.

Les URSSAF sont sous la tutelle de l'ACOSS - Agence Centrale des Organismes de Sécurité sociale, laquelle assure également la gestion de trésorerie des différentes branches du Régime général de la Sécurité sociale.

Il existe plusieurs régimes (de cotisations) au sein de la Sécurité sociale, par exemple celui des travailleurs indépendants et des professions libérales. Cependant, dans tout le contenu de la thèse nous ne discutons que du principal, et plus important, le Régime général qui concerne l'ensemble des entreprises (avec au moins un salarié).

Le principe fondateur du recouvrement des cotisations sociales est déclaratif : l'entreprise déclare ses salariés à l'URSSAF et lui verse l'ensemble des cotisations dues pour chaque salaire versé. L'URSSAF enregistre l'entreprise et ses déclarations dans son système d'information et collecte les cotisations de manière périodique - mensuelle ou trimestrielle. Les variations d'effectif sont ainsi prises en compte au fur et à mesure des années d'activité de l'entreprise et les cotisations sont recouvrées d'une manière de plus en plus automatisée. Nous présentons ci-dessous les missions des URSSAF résumant le recouvrement et le traitement des cotisations :

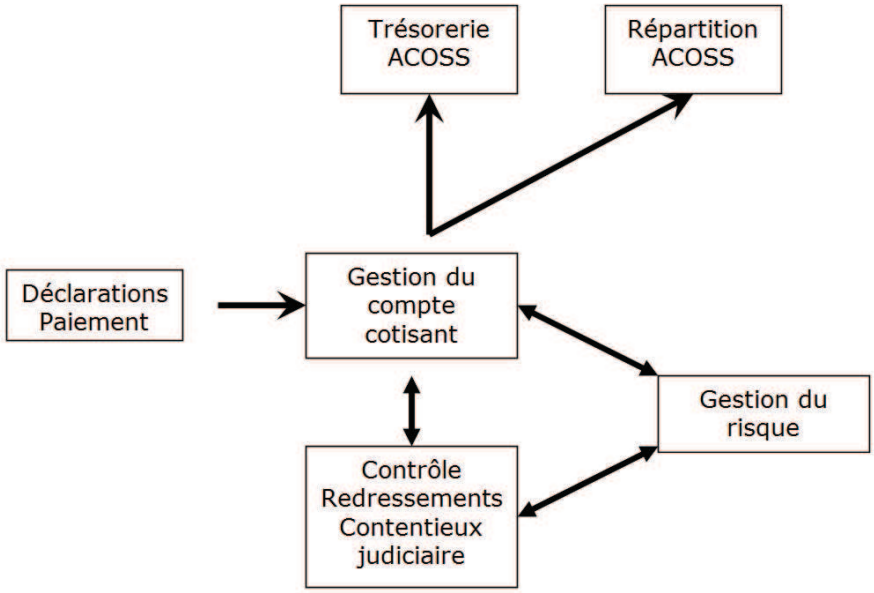

FIGURE 1.1 – Schéma des missions de l'URSSAF d'Île-de-France

Sur le graphique ci-dessus, les missions de recouvrement s'inscrivent dans un cadre dont les éléments principaux sont le lien entre les déclarations et paiements, et la perception qu'en a l'URSSAF, à travers la gestion du compte de cotisations de l'entreprise. Cette dernière est aussi désignée par le terme "cotisant".
Lorsque le compte du cotisant est conforme aux déclarations et paiements attendus, les montants sont redirigés vers l'ACOSS, laquelle centralise le financement des différentes branches. Lorsqu'il y a non-conformité ou un risque associé au compte, différents mécanismes sont mis en oeuvre afin de procéder à sa régularisation. Du point de vue financier, les différences d'interprétation de la législation, notamment après un contrôle, représentent une partie importante des cas de non-conformité.

Les entreprises mentionnent sur des bordereaux dédiés ("Bordereaux Récapitulatifs de Cotisations") le montant des cotisations patronales et salariales et, le cas échéant, la déclaration du nombre d'embauches intervenues depuis le dernier paiement. Chaque année, un tableau récapitulatif de leurs cotisations est établi.

Une fois les montants déclarés, le système informatique les ventile automatiquement dans les catégories appropriées et "vérifie" leur conformité (au moins pour leurs taux d'application) avec la législation. La gestion du compte cotisant est rendue complexe par le fait que l'URSSAF reçoit des paiements fréquents et que, dans le même temps, des prestations sociales sont consommées ou versées de manière encore plus fréquente. Il est donc nécessaire de garantir un équilibre dynamique entre ces deux évènements. Lorsque les cotisations ne peuvent être perçues à la date de paiement échue, ou lorsqu'un redressement a lieu, la fonction de gestion du risque tente de procéder à leur règlement dans les meilleurs délais. Elle propose un ensemble de mesures allant du recouvrement amiable à la procédure judiciaire.

En 2008, seuls 5% des montants déclarés n'étaient pas encaissés à leur échéance. Moins de 2% donnaient lieu à un report sur l'année suivante.

La simplicité du principe déclaratif est altérée par deux processus qui interviennent constamment dans le recouvrement des cotisations :
i) la législation est complexe. Il existe plus de 900 catégories de cotisation possibles et le *Guide du Recouvrement* est constitué de plus de 2 000 pages ;
ii) une cotisation déclarée ne peut être certifiée exacte que si elle est contrôlée.

Ces deux constantes peuvent être illustrées par le paradigme suivant : une entreprise a, virtuellement, la possibilité de déclarer ses cotisations parmi plus de 900 catégories différentes, dont chacune supporte des conditions d'application. Il appartient à l'URSSAF de vérifier que les montants déclarés correspondent à la situation réelle de l'entreprise.

1.4 Les cotisations sociales et leur contrôle

La problématique fondamentale des cotisations sociales est l'équilibre (ou l'excédent) nécessaire entre cotisations versées par les entreprises et prestations reçues par les bénéficiaires. La mission de contrôle est alors un corollaire de la garantie de cet équilibre. Mais, cela suppose de contrôler un grand nombre d'entreprises chaque année, et d'assurer que le coût reste négligeable pour la Sécurité sociale.

En 2013, la Loi de Financement de la Sécurité sociale (LFSS) prévoit un montant total de recettes d'environ 460 Mds d'euros, pour un montant de dépenses de 470 Mds d'euros. Dans ce budget, les cotisations sociales du Régime général représentent environ 329 Mds d'euros, pour des prestations sociales de 340 Mds d'euros. 1 200 000 entreprises constituent le Régime général. Environ 80% emploient 1 à 9 salariés. Et 15% emploient 10 et 49 salariés.

Les cotisations sociales sont constituées d'une part patronale et d'une part salariale. La part patronale est calculée à partir du salaire brut figurant sur le bulletin de paie, mais n'en fait pas partie. La part salariale est calculée sur le même principe et sa déduction du salaire brut donne le montant net payé au salarié.

Afin de rendre plus simple la lecture, nous ne faisons pas de distinction, dans tout le reste de la thèse, entre part patronale et salariale. Chaque cotisation sociale est considérée comme la somme des deux. Les paramètres de calcul sont les mêmes (seules leurs valeurs diffèrent) pour les deux parts et ne modifient pas la comptabilisation en terme de gestion des cotisations. Celles-ci sont principalement affectées aux trois principales branches de la Sécurité sociale : l'*Assurance Maladie*, l'*Assurance Vieillesse* (retraites de base) et les *Allocations Familiales*. L'URSSAF assure également la collecte des cotisations destinées à l'*Assurance Chômage* (depuis 2012), ainsi que celles d'autres organismes de la Sécurité sociale ou de collectivités locales. Le modèle de la Sécurité sociale fonctionne donc comme un système d'assurance des salariés et de la population. Plusieurs éléments le rendent différent d'une assurance privée :

- les cotisations sont obligatoires pour chaque salarié ;
- les bénéficiaires sont couverts sur une durée illimitée ;
- l'ensemble de la population bénéficie des prestations sociales ;
- le montant total des cotisations est très important et dépasse le budget de l'Etat ;
- il n'y a pas d'homogénéité des cotisations.

Ce dernier point rend tout simplement inenvisageable le contrôle de toutes les cotisations. Dans le cas d'une assurance privée, le montant de la "cotisation" est généralement fixe et le cotisant ne détermine ni la nature, ni les conditions d'exécution du contrat. Dans le cas des cotisations sociales, il existe une dépendance avec le montant du salaire brut, du nombre de salariés de l'entreprise, de sa situation géographique, de son secteur d'activité, du type de contrat signé par chaque salarié,...
Ces dépendances sont amplifiées par le nombre de cotisants et créent une multitude de situations particulières. Ces situations se traduisent au sein des URSSAF par l'existence de catégories déclaratives (appelées code-type de personnel) qui définissent explicitement chaque cotisation versée par un taux, une assiette et des conditions d'application. Par exemple, les heures supplémentaires correspondent à une catégorie déclarative précise. Il en est de même pour certains types de contrat ou de professions. De plus, sous certaines conditions, une entreprise peut bénéficier de réductions et mesures dérogatoires. L'ensemble des catégories couvre plus de 900 situations et conditions d'application.

Néanmoins, il demeure possible de synthétiser n'importe quelle cotisation sociale en une formulation explicite. Comme nous l'avons indiqué, la base de calcul d'une cotisation sociale quelconque est le salaire brut. Plus précisément tout ou partie du salaire brut, appelé également *assiette de cotisation*. A l'assiette est appliqué un *taux de cotisation*. Une entreprise emploie généralement plusieurs salariés et, pour chacun, le produit de l'assiette et du taux, pour une catégorie déclarative spécifique, donne le montant de la cotisation. Le nombre total de salariés étant équivalent à la population active, soit plusieurs millions de personnes, dépasser la description des cotisations sociales demeure une tâche non triviale.

Nous formalisons alors le problème de la synthèse ci-après.
On suppose n, le nombre total d'entreprises.
On appelle M_i, la masse salariale (somme des salaires bruts) de l'entreprise $i, 1 \leq i \leq n$, et C_i la somme des cotisations qu'elle verse à l'URSSAF.
Nous supposons également que l'assiette n'est plus tout ou partie du salaire brut, mais un coefficient positif ou nul que l'on applique à la masse salariale.
Alors, pour n'importe quelle cotisation $j, 1 \leq j \leq p$, où p est le nombre total de catégories de cotisation, et pour n'importe quelle entreprise $i, 1 \leq i \leq n$,

$$C_{ij} = M_i a_{ij} u_{ij} t_{ij}, \tag{1.1}$$

avec $a_{ij} > 0, t_{ij} \in [-1, 1], u_{ij} \in [0, 1]$.

C_{ij} est le montant de la cotisation j versée par l'entreprise i,
a_{ij} est le coefficient d'assiette de cette cotisation.
u_{ij} est la proportion d'effectif à laquelle s'applique la cotisation,
t_{ij} est le taux d'application de la cotisation j pour l'entreprise i.

La cotisation totale versée par l'entreprise i s'écrit alors :

$$C_{i.} = \sum_{j=1}^{p} C_{ij}.$$

Pour l'ensemble des entreprises, le montant total des cotisations, C, est défini par :

$$C = \sum_{i=1}^{n} C_{i.} = \sum_{i=1}^{n} \sum_{j=1}^{p} M_i a_{ij} u_{ij} t_{ij}.$$

Chaque cotisation sociale dépend alors uniquement de quatre paramètres et il n'est plus nécessaire de se préoccuper de conditions d'applications, des spécificités des entreprises ou de la situation de chaque salarié. Cette représentation est une réplication du principe déclaratif. Parmi les p catégories de cotisation disponibles, si une entreprise n'est pas concernée par l'une d'entre elles, le coefficient d'assiette vaut simplement 0. En contrepartie, il faut vérifier que chacun des paramètres est bien celui à affecter au calcul de la cotisation spécifiée.

Lorsque L'URSSAF enregistre une cotisation C_{ij}, son système d'information l'impute de manière automatique à la catégorie correspondante. L'assiette est déclarée par l'entreprise, le taux est défini par la législation et le nombre de salariés concernés par la cotisation peut être déduit. Sous réserve que la masse salariale (en l'occurrence chaque salaire brut) déclarée soit exacte, et que l'entreprise n'omette pas une catégorie de cotisation qu'elle devrait déclarer, la cotisation est alors recalculée et comparée avec C_{ij}. Lorsqu'il y a égalité, cela correspond à la relation (1.1). Lorsque ce n'est pas le cas, des écarts entre cotisations attendues et reçues sont générés et le cotisant en est notifié.

Dans la pratique, la problématique de la vérification des cotisations reçues est beaucoup plus complexe. La masse salariale dépend de la déclaration de l'entreprise et une minoration de son montant peut être indétectable. Les primes, frais professionnels, avantages ou autres éléments de rémunération peuvent entraîner de nombreuses omissions ou déclarations erronées, du fait du nombre de salariés. Le coefficient d'assiette (ou l'assiette) en est alors directement affecté. Chaque catégorie de cotisation possède une assiette spécifique. Par exemple, la CSG (Contribution Sociale Généralisée), obligatoire pour tous les salariés, a une assiette représentant 98.25% des revenus d'activité, lesquels sont supérieurs ou égaux à la masse salariale et doivent être précisément déterminés par l'entreprise. Les taux d'application ne sont généralement pas fixes car ils s'appliquent à une part plafonnée (une assiette de cotisation maximale fixée par la Sécurité sociale), et une part déplafonnée. Au-dessus de l'assiette fixée, le taux change. La difficulté la plus importante concerne les conditions d'application et les catégories de cotisation autres que celles obligatoires (ou dont les taux sont fixes). Les premières sont nombreuses (au moins une pour chaque catégorie de cotisation) tandis que les secondes dépendent de spécificités, comme le taux applicable aux Accidents du travail (sa valeur dépend du secteur d'activité), ou les mesures de réduction (dont les taux sont négatifs).

Deux catégories de cotisation, obligatoires, le *Cas général* et la *CSG-CRDS*, constituent l'essentiel des montants recouvrés par l'URSSAF. En 2011, pour les entreprises (du Régime Général) d'Île-de-France, la première représentait environ 42 Mds d'euros et la seconde 11 Mds d'euros, soit, au total, environ 80% des cotisations versées. Toutefois, du fait des montants très importants (pour la seule région Île-de-France, la masse salariale déclarée en 2011 dépassait 195 Mds d'euros) toutes les catégories de cotisation représentent un enjeu financier pour l'URSSAF. Le système d'information, idéalement, doit être en mesure de vérifier, au moins, knp possibilités d'erreurs de cotisation pour garantir les recettes de la Sécurité sociale. $k = 4$, correspond aux quatre paramètres (masse salariale, assiette, taux et proportion d'effectif) de la relation (1.1), $n > 1\ 200\ 000$, est le nombre total d'entreprises en France et $p > 900$, le nombre de toutes les catégories de cotisation. Dans la pratique, les recettes ne sont pas garanties pour une raison : les vérifications (les contrôles) effectuées aboutissent, en moyenne et chaque année, à la découverte d'écarts entre cotisations attendues et versées, supérieurs, au total, à 1 milliard d'euros tandis que, dans le même temps, toutes les vérifications ne peuvent être faites.

`Les irrégularités aux cotisations sociales`

Afin d'éviter un nombre gigantesque de vérifications, l'URSSAF procède à des contrôles sur un certain nombre d'entreprises. Lorsque ces contrôles sont réalisés aléatoirement, il est alors statistiquement possible d'estimer le montant de cotisations omis par les entreprises, avec un niveau de confiance généralement élevé.

Nous définissons les irrégularités aux cotisations sociales comme les montants qui sont omis volontairement ou non dans la déclaration et/ou le paiement des cotisations par les entreprises. Lorsque le caractère est volontaire, l'irrégularité est une fraude, mais alors ce caractère volontaire doit être prouvé par l'URSSAF.

De nombreuses études ont été menées afin de déterminer le montant total des irrégularités aux cotisations sociales. Nous nous référons au rapport du Conseil des prélèvements obligatoires (*La fraude aux prélèvements obligatoires et son contrôle*, 2007, p.71) qui évalue ce montant dans un intervalle de 8 à 15 Mds d'euros annuels. Notons que la fraude (donc le caractère volontaire de l'irrégularité et en particulier le travail dissimulé ou "travail au noir") est, ici, estimée à près de 75% du total. Ce point est très important dans la compréhension des irrégularités car il indique que la plupart d'entre elles le seraient de manière volontaire.

En 2012, le montant total des irrégularités relevées en France (ACOSS, *Le contrôle des cotisants 2012*, p.52) a été d'environ 1 500 millions d'euros, dont au moins 900 millions d'euros n'ont pas été liés au travail dissimulé. La problématique posée par les irrégularités aux cotisations sociales se résume alors à deux aspects :

i) peut-on, dans la même limite relative de ressources que celles allouées à ce jour, détecter, annuellement, des irrégularités pour un montant supérieur, ou égal, à 8 Mds d'euros ?
ii) Ce montant serait-il effectivement récupérable et pourrait-il améliorer les recettes de la Sécurité sociale ?

Le contrôle

Le contrôle des cotisations sociales se présente essentiellement sous deux déclinaisons :
- le contrôle comptable d'assiette (CCA), le type de contrôle le plus répandu, qui consiste à vérifier in situ que l'entreprise a correctement déclaré et payé ses cotisations ;
- le contrôle dans le cadre du travail dissimulé (appelé également contrôle LCTI - lutte contre le travail illégal) qui permet de vérifier que l'entreprise rémunère ses salariés dans le respect de la législation et ne minore ni sa masse salariale, ni ses cotisations. Les contrôles dans le cadre du travail dissimulé représentent, en volume, moins de 5% de l'ensemble et, en valeur, plus de 17% des montants (mais cette part augmente rapidement).

Il existe d'autres types de contrôles (contrôle sur pièces, contrôles partiels d'assiette,...) qui sont des variations des contrôles comptables d'assiette, mais nous discutons essentiellement de ces derniers dans le reste de la thèse, pour plusieurs raisons :
- ce sont les contrôles dont les montants détectés (au total) sont les plus importants ;
- ce sont généralement des contrôles ciblés, bien qu'une partie soit des contrôles aléatoires, avec une probabilité, à priori, supérieure à 50% de trouver des irrégularités ;
- la vérification des cotisations est complète, ou presque, pour ce type de contrôle.
En 2012, un peu plus de 240 000 actions de contrôle ont été menées, dont environ 37% dans le cadre des contrôles comptables d'assiette. Ces derniers ont représenté 60% des montants d'irrégularités relevées. Notons que les actions de contrôle sont aussi bien des contrôles effectifs que des processus visant à mieux informer les entreprises. En pratique, le contrôle a tout autant un objectif de prévention et d'information que de vérification et dissuasion. Les actions explicites de contrôle sont donc moins nombreuses et ont concerné environ 155 000 entreprises en 2012, soit un peu plus d'une entreprise sur 10.

Lorsqu'une irrégularité est détectée, le montant correspondant est appelé redressement. Cependant, il arrive que des sommes soient remboursées aux entreprises, lorsque celles-ci, par exemple par méconnaissance ou par une complexité trop importante de la législation, déclarent et paient un surplus de cotisations. Environ 15% des montants redressés, et plus dans certaines régions comme l'Île-de-France, sont des remboursements aux entreprises. Il existe donc des redressements positifs, lorsque l'URSSAF détecte des montants en sa faveur. Et des redressements négatifs, lorsque les montants sont en faveur de l'entreprise. *Dans la définition que nous avons donnée, une irrégularité, en valeur monétaire, est la différence, lors d'un contrôle, entre la somme des redressements positifs et la somme des redressements négatifs.* Si la différence est négative ou nulle alors il n'y a pas d'irrégularités, afin de maintenir la cohérence de notre contenu, et la valeur, si elle est négative, correspond alors à des erreurs de cotisation.

Formellement, une irrégularité, que nous notons Y, est une variable aléatoire, définie comme non nulle si, et seulement si, pour toute cotisation $j, 1 \leq j \leq p$, de l'entreprise $i, 1 \leq i \leq n$,

$$\sum_{j=1}^{p} C_{ij} < \sum_{j=1}^{p} M_i a_{ij} u_{ij} t_{ij}.$$

On a alors, pour l'entreprise i :

$$Y_i = \mathbf{I}_{\left\{\sum_{j=1}^{p} C_{ij} < \sum_{j=1}^{p} M_i a_{ij} u_{ij} t_{ij}\right\}},$$

soit que $Y \in \{0, 1\}$, et $Y = 1$ si, et seulement si, les montants déclarés et payés sont strictement inférieurs aux montants qui devraient l'être si on avait accès à l'ensemble des informations et paramètres de cotisation de l'entreprise. Notons que du point de vue mathématique, nous imposons que l'absence d'irrégularité est équivalente à dire que l'irrégularité est nulle, soit que $Y = 0$. Le problème posé au contrôle est donc de montrer que l'irrégularité est non nulle. Idéalement, il est tout aussi intéressant de connaitre le montant associé, afin de réserver les ressources, généralement limitées, aux cas les plus significatifs.
Notons R_i, le montant associé à l'irrégularité Y_i, pour l'entreprise i.
Alors,

$$R_i = \left[\sum_{j=1}^{p} \left(M_i a_{ij} u_{ij} t_{ij} - C_{ij}\right)\right] \mathbf{I}_{\{Y_i=1\}}.$$

Du point de vue du contrôle (par l'URSSAF), l'objectif est de vérifier la (non) nullité du couple (Y, R) pour chaque entreprise i, contrôlée et pour un nombre d'entreprises très limité, $i << n$.
Du point de vue mathématique, l'objectif est d'estimer les valeurs prises par le couple (Y, R), avec un niveau de confiance élevé, pour chaque entreprise $i, 1 \leq i \leq n$.

Le montant total des irrégularités que nous avons indiqué (1 500 millions d'euros, en 2012) correspond à la différence entre la somme des redressements positifs et la somme des redressements négatifs. C'est donc un montant net et *toute l'analyse proposée et les chiffres mentionnés ne portent, sauf exception, que sur des montants nets.*
Il serait tentant de doubler le nombre de contrôles (d'une entreprise sur 10 à une sur 5) pour espérer doubler (ou presque) les montants redressés. Cette option ne serait pas réalisable car il faudrait, alors, aussi doubler une partie des ressources dédiées au contrôle. De plus, la probabilité de détection d'une irrégularité est (très) inférieure à 1, ce qui induirait un risque assez grand sur la portée des résultats attendus. Enfin, la distribution des redressements est fortement asymétrique et, comme nous le verrons plus en détail dans le chapitre suivant, l'essentiel des montants redressés provient des grandes entreprises. Néanmoins, ces trois contraintes représentent les limites à lever si l'on souhaite détecter plus d'irrégularités et en augmenter les montants.

Si nous supposons que X, un vecteur aléatoire, est la connaissance, incomplète, que nous avons de $(M_i, a_{ij}, u_{ij}, t_{ij})$ à travers les données déclaratives enregistrées par l'URSSAF, pour chaque entreprise et chaque cotisation, alors on peut définir explicitement la probabilité d'observer une irrégularité :

$$\mathbf{P}(Y = 1 | X = x),$$

et l'espérance conditionnelle du montant associé à l'irrégularité Y et à la connaissance de X :

$$\mathbf{E}(R|(X, Y)).$$

Sur le plan théorique, il s'agit alors de trouver les outils et garanties qui permettent de contrôler et de minimiser les écarts entre les quantités définies ci-dessus et leurs estimateurs, calculés à partir des données observées et des contrôles effectués.

1.5 L'apprentissage statistique

L'apprentissage statistique est une théorie probabiliste et statistique dont l'objectif est la modélisation de la relation entre un phénomène généralement aléatoire et la nature, inconnue, de ce même phénomène. Les outils utilisés sont issus des Probabilités, du fait de l'aspect non déterministe du problème. Le phénomène peut prendre de nombreuses formes, ici les irrégularités aux cotisations sociales, et se traduit généralement par l'existence de données (variables et observations). La nature du phénomène est inconnue car, au mieux, on ne dispose que d'une représentation limitée. Par exemple, ici, nous n'avons qu'un nombre limité de résultats de contrôle pour l'ensemble des entreprises. Si on ne dispose d'aucune information sur cette nature, l'apprentissage est dit *non supervisé*. Dans ce cas, on cherche généralement à constituer différents groupes homogènes dont on essaie d'explorer et de caractériser les propriétés. Si on dispose d'informations sur cette nature, par exemple un échantillon, alors l'apprentissage est *supervisé*. Dans le cas des irrégularités aux cotisations sociales, nous disposons des résultats du contrôle pour un certain nombre d'entreprises et privilégions donc l'apprentissage supervisé pour évaluer les irrégularités. La nature du phénomène est généralement déterminée par le problème que l'on cherche à résoudre. Cela peut être le diagnostic médical, l'attribution de crédits par une banque, le *scoring* (par exemple, la probabilité de défaut d'une entreprise durant son exercice courant), la recherche de nouveaux champs pétrolifères, la prévision des pics de pollution, l'optimisation des ventes d'un produit, l'optimisation d'un processus de production, la reconnaissance de formes, la catégorisation de documents, les résultats renvoyés par un moteur de recherche, l'analyse financière, l'allocation optimale de portefeuille, l'évaluation des spams, la détection de défauts dans un matériau, la détection d'intrusions, d'irrégularités,... Par nature, nous entendons également les résultats associés à chaque observation enregistrée. Par exemple, dans le cas de la détection d'irrégularités, une observation correspond à l'ensemble des informations connues de l'entreprise dans le cadre de sa déclaration de cotisations et le résultat, s'il est connu, est le caractère irrégulier ou non de cette déclaration.

On utilise la notion d'apprentissage, car les modèles utilisés *apprennent* la relation entre les données et les résultats connus. Une fois ces *exemples* (données et résultats) appris, ils n'ont plus besoin de l'être à nouveau, car le modèle *sait* (et dans certains cas de manière exacte) alors la relation entre ces données et leurs résultats. Le résultat fondamental que l'on cherche à atteindre est une bonne capacité de généralisation, c'est-à-dire une grande aptitude à spécifier correctement les résultats sur des données inconnues du modèle, grâce à l'apprentissage déjà effectué. Dans le cas de la détection d'irrégularités, on souhaite que les résultats (le caractère irrégulier ou non de la déclaration) fournis par le modèle sur des entreprises non contrôlées soient aussi proches que possible des résultats effectifs, si ces entreprises étaient contrôlées. La différence entre les résultats estimés et les vrais résultats s'appelle l'erreur de prédiction (ou de généralisation) et on cherche à la rendre minimale.

Mais pour que le modèle soit admissible, il ne peut être restreint à une situation particulière. On doit pouvoir bénéficier de garanties théoriques solides. Pour cela, nous faisons appel à la notion de *consistance*. Un modèle est dit consistant si, lorsque le nombre d'observations augmente, l'erreur de prédiction du modèle tend vers l'erreur la plus petite possible. Il n'est pas garanti que l'erreur minimale soit elle-même petite, mais si l'erreur du modèle y tend, alors cela suffit à apporter des garanties suffisantes. Tout autre modèle ne peut alors, au mieux, que tendre vers l'erreur minimale plus vite (soit, avec moins d'observations) que le modèle choisi.

L'apprentissage statistique, dans le cas supervisé, distingue deux types de problématiques :
- les problèmes dits de classification dans lesquels la nature du phénomène observé prend des valeurs discrètes ou non numériques en nombre fini (par exemple, {0,1}, {"rouge","noir"}, {"malade","non malade"}, {"achat", "sans avis", "vente"}), et tel est le cas lorsqu'on cherche à déterminer le caractère irrégulier ou non de la déclaration de cotisations ;
- Les problèmes de régression, ou cette nature prend des valeurs continues, par exemple lorsqu'on cherche à estimer le montant d'une irrégularité.
Le modèle utilisé pour résoudre ces deux types de problème est appelé, de façon générique, estimateur ou classificateur. Nous lui préférons le néologisme *classifieur*, que nous retenons pour l'ensemble du document. Un classifieur possède généralement une structure algorithmique native en plus de sa représentation mathématique. Notons que la dimension algorithmique de l'apprentissage statistique est fondamentale et intervient très fortement dans les performances des nombreux modèles utilisés.

Pour pouvoir évaluer un modèle, il faut disposer de données d'apprentissage en nombre suffisant. Cet échantillon est noté D_n et défini par $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$.
D_n est donc la représentation d'un certain nombre d'observations, X_i, et de résultats, Y_i, que l'on a déjà sur le problème, par exemple les résultats des contrôles effectués l'année précédente sur les déclarations de cotisation.

`La régression`

Soit Y, une variable aléatoire représentant la nature du phénomène observé et X, les variables qui servent à le décrire.
X est un vecteur aléatoire, c'est-à-dire que X est la représentation de plusieurs variables dont les réalisations (les valeurs) sont aléatoires. Dans le cas des cotisations sociales, chaque catégorie déclarative est une variable et ses valeurs sont aléatoires au sens où il n'existe pas de relation déterministe entre Y et X, ou bien que cette relation est inconnue, et qu'il n'est pas possible de déterminer à l'avance, et exactement, les réalisations conjointes de X et Y. Cependant, X est observé et nous pouvons lui ajouter d'autres variables liées à la déclaration de cotisations, comme les autres informations enregistrées par l'URSSAF. Y est une variable aléatoire, car il n'est pas possible de déterminer à l'avance, et exactement, les réalisations de Y. Dans les cas des cotisations, sociales, le montant R, associé à une irrégularité Y est, par définition, aléatoire. Le problème consiste alors à estimer ce montant en faisant une erreur qui soit la plus petite possible.

$X \in \mathbb{R}^d$, où d est la dimension du problème (son nombre de variables), $R \in \mathbb{R}$, et on peut écrire, par exemple, un modèle de la forme :

$$R = f(X) + \epsilon,$$

où f correspond alors à une relation possible entre R et X, de sorte que $f : \mathbb{R}^d \to \mathbb{R}$, et ϵ est la mesure de l'incertitude de cette relation. ϵ traduit le fait qu'il n'est pas possible de connaître exactement la relation entre R et X.

L'objectif est alors de caractériser $f(X)$ de sorte que l'erreur entre R et $f(X)$ soit minimale. Dans le cas de la régression, l'erreur de prédiction, notée L, est définie par :

$$L(R, f(X)) = (R - f(X))^2.$$

Rechercher f revient alors à minimiser l'espérance de l'erreur de prédiction. Une telle fonction f existe. Elle est définie par :

$$f(X) = \mathbf{E}(R|X)$$

et l'espérance de l'erreur de prédiction s'écrit :

$$\mathbf{E}\{L(R, f(X))\} = \mathbf{E}\{R - \mathbf{E}(R|X)\}^2.$$

Mais, nous ne disposons que d'un échantillon D_n, soit un nombre limité de réalisations du couple (X, R). On ne peut donc pas déterminer f exactement, et à la place on cherche un estimateur de f tel que l'espérance de l'erreur de prédiction entre f et cet estimateur tende vers 0, lorsque le nombre d'observations augmente. Obtenir cette propriété est équivalent à rechercher la consistance (dans le cas de la régression) pour l'estimateur choisi. Notons g cet estimateur. Nous cherchons alors g, de façon à minimiser l'erreur quadratique moyenne donnée par :

$$\hat{L}_n(R, g_n(X)) = \frac{1}{n} \sum_{i=1}^{n} \left(R_i - g_n(X_i)\right)^2.$$

Notons que pour avoir une estimation réaliste de l'erreur de prédiction, il convient d'utiliser un échantillon différent de D_n, appelé échantillon de validation, de sorte que :

$$\hat{L}_N(R, g_N(X)) = \frac{1}{N-n} \sum_{i=n+1}^{N} \left(R_i - g_N(X_i)\right)^2, N > n.$$

Dans le cas de la régression, la difficulté provient du fait que les hypothèses sur la relation entre R et X sont plus nombreuses et plus fortes pour l'obtention de garanties sur le fait que l'espérance de l'erreur de prédiction soit la plus petite possible. A la place, on peut, par exemple, essayer de contrôler cette erreur, c'est-à-dire qu'à chaque valeur de $g(x)$, où x est une réalisation de X, on associe un intervalle dans lequel se trouve la réalisation r de la variable R, avec une grande probabilité. Formellement, à un niveau de probabilité donné et pour une quantité μ, dépendante au moins de g et de x, et pour toutes les réalisations $r \in R$ et $x \in X$,

$$r \in [g(x) - \mu, g(x) + \mu].$$

Dans le problème de la détection des irrégularités aux cotisations sociales, g est un classifieur et nous le construisons de telle sorte que μ soit assez petite et puisse être exprimée simplement. g doit cependant être assez proche de f afin de permettre une certaine précision sans laquelle le montant estimé d'une irrégularité et son intervalle de confiance risquent d'être peu clairs pour l'interprétation.
Dans l'approche que nous avons choisie et pour améliorer les conditions de l'estimation des montants de redressement, nous déterminons d'abord si une irrégularité Y existe puis, si tel est le cas, nous estimons son montant R. Cela permet de nettement simplifier le problème des propriétés opérationnelles de g, car nous nous assurons d'abord de la probabilité d'existence d'une irrégularité. Dans la connexion entre le point de vue théorique et pratique, construire le classifieur g demande de limiter les hypothèses sur la relation entre Y et X et d'analyser les propriétés de l'erreur de prédiction, par exemple, en utilisant la décomposition biais-variance de l'erreur ou directement par les caractéristiques intrinsèques de g.

`La classification`

Dans le cas de la classification, Y prend un nombre de valeurs fini, discrètes ou catégorielles et, afin de simplifier la lecture, nous nous restreignons au cas binaire. Dans ce cas Y ne peut prendre que deux valeurs, par exemple 0 ou 1. En classification, il n'y a pas de relation quantitative entre Y et X et la fonction f définie dans la régression n'existe pas (du moins pas de manière directe) ici. A la place, nous nous référons alors directement à l'erreur de prédiction L, définie par :

$$L(g) = \mathbf{P}(g(X) \neq Y).$$

L'erreur de prédiction est, ici, une probabilité équivalente à un taux d'erreur. On suppose que $Y \in \{0, 1\}$. Idéalement, nous devrions avoir connaissance de

$$\eta(x) = \mathbf{P}(Y = 1|X = x) = \mathbf{E}\{\mathbf{I}_{\{Y=1|X=x\}}\}$$

pour pouvoir estimer si une irrégularité, pour une entreprise quelconque, est suffisamment probable ou non. Comme notre connaissance se limite à D_n, notre classifieur g, tel que $g : \mathbb{R}^d \to \{0, 1\}$, est choisi de sorte à minimiser la quantité

$$L_n = L(g_n)(X) = \mathbf{P}(g_n(X) \neq Y|D_n).$$

Un estimateur de L_n est donné par

$$\hat{L}_n(g_n(X)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{g_n(X_i)\neq Y_i\}}.$$

Dans le cas de la classification, la consistance est plus simple (et moins problématique) à définir. Soit g^*, le classifieur qui minimise

$$L = L(g) = \mathbf{P}(g(X) \neq Y).$$

Alors,

$$g^* = \underset{g}{\texttt{arg min}}\ \mathbf{P}(g(X) \neq Y) = \mathbf{I}_{\{\eta(X) \geq 1/2\}}$$

et

$$L^* = L(g^*) = \mathbf{P}(g^*(X) \neq Y).$$

L^* est appelée l'erreur de Bayes. Elle est la plus petite erreur que l'on puisse commettre en choisissant g^* et ce dernier est appelé *classifieur de Bayes*.

Connaitre g^* revient à connaître la loi de probabilité de (X, Y) afin de pouvoir calculer $\mathbf{P}(Y = 1|X = x)$, mais cette loi est généralement inconnue. A la place, nous devons nous contenter de g_n et estimer L_n. Contrairement à la régression, on ne peut fournir d'intervalle de confiance à une réalisation y de Y car celle-ci ne peut prendre qu'un nombre de valeurs fini. Dans le cas des cotisations sociales, l'irrégularité existe ou n'existe pas. Elle ne peut être entre les deux. Comme elle résulte d'un phénomène aléatoire, on ne peut, au mieux, que lui attribuer une probabilité d'existence.
Pour que L_n soit minimale, elle doit se rapprocher suffisamment de L^*. Construire un classifieur g qui respecte une telle propriété est équivalent à écrire que :

$$\mathbf{E}(L_n) \to L^*,\ lorsque\ n \to \infty.$$

Un tel classifieur est dit consistant et garantit que l'espérance de la probabilité d'erreur faite sur la détection des irrégularités aux cotisations sociales tend vers la plus petite erreur que l'on puisse commettre, lorsque le nombre d'observations augmente. Un des intérêts pratiques majeurs de l'apprentissage statistique réside dans cette capacité de généralisation. Si un classifieur est consistant, plus on augmente la taille de l'échantillon d'apprentissage, ici les déclarations de cotisation et les résultats de contrôle associés, plus l'erreur commise dans la prédiction diminue, jusqu'à atteindre une limite. Cette dernière ne peut pas valoir 0, c'est-à-dire qu'il n'est pas possible de ne pas faire d'erreurs lorsque le classifieur prédit des irrégularités sur des entreprises non contrôlées. Mais pour l'ensemble des entreprises, il est possible d'estimer, avec un niveau de confiance important, le taux d'erreurs. Par exemple, nous pouvons poser :

$$L \in [L^*, L^* + \epsilon], \epsilon > 0.$$

Comme L^* est inconnu, nous cherchons, du point de vue théorique, à connaître ϵ et à le rendre aussi petit que possible, quand $n \to \infty$. Dans un premier temps, nous souhaitons déterminer le meilleur classifieur empirique g_n^* tel que :

$$g_n^* = \underset{g \in \mathcal{C}}{\texttt{arg min}}\ \hat{L}_n(g),$$

où $\mathcal{C}$ représente l'ensemble des classifieurs g susceptibles de minimiser l'erreur de prédiction dans la connaissance de la relation entre X et Y.
Dans un second temps, nous nous intéressons à la différence entre l'erreur de prédiction théorique de g_n^* et la plus petite erreur de prédiction théorique parmi les classifieurs g de $\mathcal{C}$. Elle est donnée par :

$$L(g_n^*) - \underset{g \in \mathcal{C}}{\texttt{inf}}\ L(g),$$

et la connaissance de cette différence permet de contrôler l'erreur minimale commise dans le choix de g_n^*. Comme

$$L(g_n^*) - L^* = (L(g_n^*) - \inf_{g \in \mathcal{C}} L(g)) + (\inf_{g \in \mathcal{C}} L(g) - L^*),$$

la recherche de ϵ, quand $n \to \infty$, revient à étudier les deux parties de l'équation, dont la première est l'*erreur d'estimation*, et la seconde l'*erreur d'approximation*. L'objectif de la consistance est alors le contrôle de ces deux erreurs.

La présentation faite ci-dessus ne suppose pas l'unicité des mesures d'erreur. D'autres types peuvent être utilisés dans le même cadre de minimisation et de nombreux outils théoriques sont disponibles pour le traitement de ces problématiques. Citons, en particulier, la *théorie de Vapnik-Chervonenkis* qui pose les fondements de l'apprentissage statistique et l'ouvrage de Vapnik, *The nature of statistical learning theory* (1995), ainsi que l'ouvrage de référence de Devroye, Györfi et Lugosi, *A probabilistic theory of pattern recognition* (1996). Pour une revue complète des méthodes utilisées en apprentissage statistique, la publication de Hastie, Tibshirani et Friedman, *The elements of statistical learning* (2001) demeure incontournable. Nous discutons d'une de ces méthodes, ci-après, notamment en abordant la structure algorithmique sous-jacente.

`Du classifieur naïf de Bayes aux méthodes ensemblistes`

Pour attribuer de bonnes propriétés à g, il est important de développer des modèles pour lesquels les outils théoriques soient transposables de manière simple dans la pratique. L'avantage réside généralement dans le fait de pouvoir expérimenter rapidement de nouvelles méthodes utiles à la compréhension et à l'inférence. En effet, de nombreux problèmes opérationnels font appel à l'apprentissage statistique, notamment ceux résultant d'autres domaines d'application et dans lesquels des modèles existent déjà. Ce type de problématique nécessite alors souvent une connexion importante entre propriétés théoriques et mise en oeuvre. Le fait de naviguer entre ces deux points de vue se retrouve dans l'ancrage algorithmique des méthodes d'apprentissage statistique, en particulier dans la période récente. L'analyse numérique, grâce à la puissance des ordinateurs, devient également essentielle aux nouveaux développements.

Si nous prenons le cas de la classification, l'évaluation de $\mathbf{P}(Y = 1|X = x)$ (dans le cas des cotisations sociales, la probabilité d'observer une irrégularité conditionnellement à la connaissance de la déclaration de cotisations) est permise explicitement par la formule de Bayes. Elle s'écrit :

$$\mathbf{P}(Y = 1|X = x) = \frac{\mathbf{P}(X = x|Y = 1)\mathbf{P}(Y = 1)}{\sum_{y=0}^{1} \mathbf{P}(X = x|Y = y)\mathbf{P}(Y = y)} \tag{1.2}$$

Cette formulation permet de construire le classifieur g, mais il est nécessaire de faire quelques hypothèses :
- la limitation des observations à D_n ne permet pas de calculer $\mathbf{P}(X = x|Y = 1)$ mais un estimateur, à partir de la distribution empirique des données ;
- on suppose que ses différentes composantes de X sont indépendantes entre elles.

En développant la formule (1.2), on a :

$$\mathbf{P}(Y=1|X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},...,X^{(d)}=x^{(d)}) = \frac{\mathbf{P}(X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},...,X^{(d)}=x^{(d)}|Y=1)\mathbf{P}(Y=1)}{\mathbf{P}(X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},...,X^{(d)}=x^{(d)})}.$$

Le dénominateur ne dépend alors plus de Y et la construction du classifieur peut se poursuivre en explicitant uniquement le numérateur :

$$\begin{aligned}\mathbf{P}(X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},...,X^{(d)}=x^{(d)}|Y=1)\mathbf{P}(Y=1) = \\ \mathbf{P}(Y=1)\mathbf{P}(X^{(1)}=x^{(1)}|Y=1) \\ \mathbf{P}(X^{(2)}=x^{(2)}|Y=1,X^{(1)}=x^{(1)})... \\ \mathbf{P}(X^{(d)}=x^{(d)}|Y=1,X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},...,X^{(d-1)}=x^{(d-1)}).\end{aligned}$$

Si les d composantes de X sont indépendantes alors le produit des probabilités se simplifie et

$$\mathbf{P}(Y=1)\mathbf{P}(X^{(1)}=x^{(1)},...,X^{(d)}=x^{(d)}|Y=1) = \mathbf{P}(Y=1)\prod_{j=1}^{d}\mathbf{P}(X^{(j)}=x^{(j)}|Y=1).$$

En posant, $\mathbf{P}(X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},...,X^{(d)}=x^{(d)}) = Cte$, on obtient finalement

$$\mathbf{P}(Y=1|X^{(1)}=x^{(1)},...,X^{(d)}=x^{(d)}) = \frac{\mathbf{P}(Y=1)\prod_{j=1}^{d}\mathbf{P}(X^{(j)}=x^{(j)}|Y=1)}{Cte}$$

et on peut définir g tel que

$$g(x) = \underset{y\in\{0,1\}}{\texttt{arg max}}\ \mathbf{P}(Y=y)\prod_{j=1}^{d}\mathbf{P}(X^{(j)}=x^{(j)}|Y=y)$$

Un tel classifieur est appelé *classifieur naïf de Bayes* et fournit généralement de bonnes performances, malgré l'hypothèse d'indépendance des variables, laquelle n'est pas vérifiée en pratique. Lorsque le nombre de variables augmente, ou lorsque les données possèdent beaucoup de valeurs en commun, comme la valeur 0, le classifieur naïf de Bayes n'est plus adapté à cause de la complexité du problème. De nombreuses interactions entre variables peuvent exister et la distribution empirique que l'on observe sur D_n peut aussi changer sur de nouvelles données.

Nous nous intéressons alors à des méthodes qui prennent en compte ces interactions et qui soient le moins sensibles à des changements dans la distribution des données. Parmi ces modèles, notre choix va vers les *modèles ensemblistes* qui possèdent les propriétés voulues, et surtout, font très peu, ou pas du tout, d'hypothèses sur les données. L'idée centrale d'une méthode ensembliste (encore appelée *agrégation de modèles*) est la construction de plusieurs modèles, à partir de D_n, et la combinaison de leurs prédictions respectives pour estimer Y. Chaque modèle de base est un classifieur qui possède une règle de décision et la combinaison de ces règles donne la règle de décision du classifieur final. Généralement,

la structure algorithmique intervient à ce niveau, en définissant des étapes, quelque fois nombreuses, dans la construction d'un classifieur de base. La règle de décision est alors l'estimateur qui intervient à la fin de la construction du modèle de base pour évaluer Y. Le classifieur final décide de la manière dont doivent être associés tous les modèles et de la règle de décision qui en résulte.
L'aspect algorithmique est fortement lié au traitement de D_n et à la modélisation des interactions entre variables. Les propriétés théoriques interviennent dans le choix de la règle de décision. Elles président néanmoins aux choix algorithmiques afin de maintenir la cohérence dans la construction du classifieur.

Dans le cas de la régression, $\mathbf{E}(R|X)$ est la quantité primordiale. Si nous considérons un modèle ensembliste qui construit B classifieurs g de paramètre θ, alors on peut par exemple définir le classifieur $\bar{g}^{(B)}$ tel que :

$$\bar{g}_n^{(B)}(x) = \frac{1}{B}\sum_{b=1}^{B} g_n(x, \theta_b),$$

où chaque classifieur $g(X, \theta_b), 1 \leq b \leq B$, estime $\mathbf{E}(R|X, \theta_b)$. La nature et la construction du classifieur sont fondamentales mais l'intérêt de la méthode est qu'il n'est pas nécessaire que chaque modèle de base ait une proximité similaire (ni même une grande) à la quantité à estimer. C'est plutôt la diversité des modèles de base qui régit la performance. La règle de décision finale, dans notre exemple la moyenne, permet la plus petite proximité.

Dans le cas de la classification binaire, la quantité d'intérêt, $\mathbf{P}(Y = 1|X = x)$, s'estime différemment :

$$\mathbf{P}(Y = 1|X = x) = \mathbf{E}\{\mathbf{I}_{\{Y=1|X=x\}}\}$$

et

$$\bar{g}_n^{(B)}(x) = \begin{cases} 1, \ si \ \sum_{b=1}^{B} \mathbf{I}_{\{g_n(x,\theta_b)=1\}} > \sum_{b=1}^{B} \mathbf{I}_{\{g_n(x,\theta_b)=0\}} \\ \\ 0, \ sinon. \end{cases}$$

La règle de décision (ou règle d'agrégation) s'assimile ici à un vote majoritaire parmi tous les modèles de base, et correspond à la même structure probabiliste que celle recherchée. Dans le reste de la thèse, nous nous intéressons exclusivement à des classifieurs de cette forme, pour lesquels nous explorons et proposons une variante dont les principaux arguments sont relatifs à la construction de g.

Les modèles ensemblistes connaissent de nombreux développements depuis l'article initial, et de référence, de Breiman (1996). Au *Bagging* (**b**ootstrap **agg**egat**ing**) s'ajoutent les *forêts aléatoires*, dont la version de référence est également due à Breiman (2001), et le *Boosting* introduit par Freund et Schapire (1997) et amélioré par Friedman (2001, 2002). Ces trois méthodes font partie des modèles ensemblistes les plus utilisés, notamment en raison de leurs performances opérationnelles. Le *Boosting* se distingue du *Bagging* par le fait que la règle de décision est différente de celle que nous avons décrite ci-dessus.
Leur point commun est cependant l'utilisation de la même forme de classifieurs de base (g) en nombre B pour prendre une décision.

i) Le *Bagging* se caractérise par une modification de D_n pour la construction de chaque classifieur g. Cette modification se traduit par un tirage aléatoire, avec remise, de n observations de D_n. C'est ce que l'on appelle le *bootstrap*. Les forêts aléatoires utilisent cette même caractéristique.
ii) A la différence du Bagging, dans lequel toutes les variables interviennent à chaque étape de la construction de g, les forêts aléatoires, du moins dans la version de référence, n'en considèrent qu'un sous-groupe, aléatoire, à chaque étape.

Ces deux techniques constituent un des fondements de ces méthodes et ont comme conséquence de construire des classifieurs de base très différents, bien que leur structure soit la même. L'union de ces différences est plus représentative de la réalité que le choix d'un seul classifieur. L'hypothèse implicite est que le classifieur optimal est celui qui prend en compte tous les autres. Du moins est-il plus simple de le définir ainsi, plutôt que de le chercher parmi le grand nombre de classifieurs possibles. Il convient d'indiquer que ces modèles aléatoires doivent intégrer des mécanismes d'optimisation. L'équilibre entre propriétés théoriques et performances intervient, également, à ce niveau. Parmi les perturbations aléatoires introduites sur les données et les variables, certaines sont plus adaptées que d'autres et il peut être approprié de s'y intéresser.

Nous traitons plus en détail de ces phénomènes, dans le troisième chapitre, en proposant un modèle que nous définissons comme une *forêt uniformément aléatoire*. A la différence des forêts aléatoires de Breiman, les forêts uniformément aléatoires proposent une perturbation plus importante sur les données, et les variables, et limitent les mécanismes d'optimisation. Dans la pratique, cela permet, par exemple, de mieux prendre en compte les données lorsqu'elles ne sont pas toutes disponibles dans un même temps ou bien une moins grande dépendance vis-à-vis de l'échantillon d'apprentissage. D'autres variantes de forêts aléatoires ont été proposées et développées. Pour une revue des versions de forêts aléatoires, nous renvoyons le lecteur vers l'article de Lin et Jeon (2002) qui en donnent un contenu concis et très clair et vers la thèse de Genuer (2010) qui en propose une synthèse.

Nous n'avons cependant pas explicité le classifieur g dans notre présentation des méthodes ensemblistes. g peut être n'importe quel modèle, comme le classifieur naïf de Bayes mais, généralement, il est construit, dans le cas des modèles ensemblistes, grâce à un arbre de décision. Ce dernier constitue une structure algorithmique qui prend D_n en entrée et le partitionne de manière récursive jusqu'à ce que certaines conditions d'arrêt se réalisent. La règle de décision de l'arbre est alors exécutée afin d'évaluer $g(x)$. Les arbres de décision peuvent être très divers, mais nous explorons essentiellement les arbres binaires de type *CART (Classification And Regression Trees)*, introduits par Breiman, Friedman, Olshen et Stone (1984), qui sont les classifieurs de base utilisés, avec une formulation cependant modifiée, dans le Bagging et les forêts aléatoires. Dans un arbre de type CART, à la première étape on divise en *deux sous-partitions* l'ensemble des données grâce à deux procédés :
- en cherchant le meilleur séparateur, pour une variable fixée, par l'évaluation de chacune de ses observations. Celle d'entre elles qui sépare le mieux la variable en deux groupes d'observations, et selon un critère d'optimisation spécifique, est retenue. On l'appelle *point de coupure*.

- Dans un second temps, on choisit la *meilleure variable de séparation* parmi tous les critères d'optimisation calculés sur chaque variable. Par exemple, si $X \in \mathbb{R}^d$, d critères sont évalués et le meilleur d'entre eux est celui qui définit la séparation.
Pour chaque sous-partition créée, on recommence le processus. Lorsqu'on ne peut plus générer de sous-partition, par l'atteinte de conditions d'arrêt, l'arbre de décision est alors totalement construit. Une dernière étape, dite d'élagage, intervient afin de supprimer les régions les moins optimales, issues des sous-partitions créées.

Dans le *Bagging* et les forêts aléatoires, cette étape d'élagage n'existe pas et les mécanismes de perturbation cités plus haut, les points *i* et *ii*, sont ajoutés. Les arbres de décision sont, clairement, au coeur de la construction d'une forêt aléatoire. Cependant, la structure algorithmique s'imbrique fortement dans la définition de leur règle de décision et leurs propriétés, dans le cadre des forêts aléatoires, ne sont pas encore solidement établies.

Dans le troisième chapitre, nous analysons un variante de CART que nous nommons *arbre de décision uniformément aléatoire*. A la différence des arbres de type CART, ses points de coupure sont choisis aléatoirement selon la loi uniforme sur le support de chaque variable. De plus, pour le partitionnement de l'arbre, $\lceil \beta d \rceil$ variables, $\beta \geq 1/d$, sont tirées avec remise à chaque étape.

Mais avant, dans le second chapitre, nous revenons sur le problème de la détection des irrégularités aux cotisations d'un point de vue opérationnel.L'essentiel de la procédure de contrôle des cotisations de l'URSSAF y est détaillé. Nous présentons également les données et le pré-traitement nécessaire à une relation optimale avec le modèle. Nous résumons ce dernier dans la pratique afin de mieux illustrer l'exécution et les contraintes qui peuvent peser.

Dans le quatrième chapitre, nous montrons comment une analyse de type économique, dans le cas de la fraude et lorsque celle-ci s'avère être du travail dissimulé, permet une description complémentaire des données et une réduction des contraintes opérationnelles. En particulier, la situation financière des entreprises explique la majorité des cas les plus importants de fraude et a comme conséquence, inattendue, l'amélioration du processus de détection. Cette propriété est générale, dans le sens où le modèle prédictif n'a pas besoin de connaître la situation financière de toutes les entreprises mais seulement un indicateur synthétique de cette situation pour un petit nombre d'entre elles.

Le dernier chapitre est consacré aux résultats expérimentaux et réels du modèle. Nous y montrons explicitement la démarche pratique de la détection des irrégularités aux cotisations sociales. En particulier, cette dernière ne donne pas lieu à des biais de sélection car le modèle n'a aucune information, à dessein, sur les caractéristiques structurelles (taille, secteur d'activité, masse salariale,...) des entreprises et privilégie la probabilité d'observation d'une irrégularité. Ce point est fondamental, et contre-intuitif, car la politique de contrôle des URSSAF se focalise essentiellement sur des actions ciblées, par exemple un risque plus élevé dans certains secteurs d'activité, mais aussi thématiques, par exemple des catégories de cotisation particulières. Ces actions ne peuvent alors être, par construction, que spécifiques tandis que des modèles, notamment mathématiques, doivent être

capables de généraliser la détection des irrégularités à toutes les entreprises. Cette capacité a plusieurs conséquences :
- le taux de détection des irrégularités mesuré est nettement amélioré car le modèle bénéficie de plus d'exemples d'apprentissage ;
- les montants des redressements effectués sont également accrus ;
- la mise en oeuvre du modèle ne demande pas de ressources autres que celles disponibles.

Nous discutons de ces aspects opérationnels et détaillons les propriétés et les garanties fournies par les forêts uniformément aléatoires.

Chapitre 2

La détection des irrégularités aux cotisations sociales

2.1 Introduction

La problématique des irrégularités aux cotisations sociales émerge, depuis quelques années, comme un sujet économique important. Les montants redressés ou estimés en sont un des principaux enjeux. On peut également citer les distorsions de concurrence que les irrégularités peuvent entraîner, ainsi que l'égalité des cotisants relativement à la législation.
Comme nous l'avons indiqué dans le premier chapitre, et depuis maintenant quelques années, plus d'un milliard d'euros de redressements (tous contrôles confondus) sont notifiés chaque année aux entreprises et plus de 8 Mds d'euros sont estimés être dus par ces dernières, chaque année, à la Sécurité sociale. Il ne paraît pas possible de contrôler 1 200 000 entreprises afin de le vérifier de manière effective.

Néanmoins, pour les sommes redressées, l'URSSAF déploie des moyens de plus en plus importants, notamment dans le cas des irrégularités volontaires (assimilables à de la fraude). Ces moyens vont de l'affectation de ressources supplémentaires au ciblage des entreprises à risque. Dans ce chapitre, nous nous intéressons plus particulièrement à l'URSSAF d'Île-de-France, qui recouvre les cotisations d'environ 400 000 entreprises. Elle enregistre environ 85 Mds d'euros de cotisations (en 2011), soit un peu plus de 20% de l'ensemble des cotisations versées. L'ensemble du travail opérationnel de cette thèse a été effectué grâce aux données de l'organisme et à la synergie avec les équipes.

Afin de bien appréhender les questions posées par la détection des irrégularités, nous commençons par détailler les données liées à la déclaration, les indicateurs et méthodes utilisés par l'URSSAF, les obligations légales, et la structure des résultats des contrôles. Puis, nous résumons la modélisation des données, générée et contrôlée par un algorithme spécifique. En particulier, un aspect important est la *généricité* de ce processus, qui rend les données peu sensibles à des changements dans la législation (taux et assiette de cotisations, conditions d'application,...) ou à des facteurs d'échelle ou de temps.
Dans la dernière partie de ce chapitre, nous donnons un schéma de la construction de l'algorithme de détection, ses capacités et les perspectives offertes par les méthodes de ce

type. Les résultats obtenus sont commentés et analysés dans le dernier chapitre afin de mieux séparer deux niveaux d'analyse :
- le pré-traitement indispensable des variables et données du fait de leur nature mais aussi de leur volume ;
- les résultats, lesquels dépendent fortement du modèle utilisé et dont la discussion doit spécifier de nombreux détails.

D'un point de vue purement opérationnel, la détection des irrégularités aux cotisations sociales pour l'ensemble des entreprises d'Île-de-France ne nécessite alors que les données enregistrées par l'URSSAF, un modèle "indépendant" du volume des données et une unique station de travail.

Sauf mention contraire, nous discutons uniquement des contrôles comptables d'assiette de l'URSSAF d'Île-de-France.

2.2 Le processus et les résultats des contrôles URSSAF

Le recouvrement des cotisations peut se résumer de manière simple. L'entreprise déclare et paie ses cotisations, lesquelles sont enregistrées par le système d'information qui vérifie si :
- pour chaque catégorie de cotisation, le taux et l'assiette sont *compatibles* avec le montant versé ;
- l'entreprise n'a pas d'arriérés de cotisation ou ne présente pas une non-conformité à la législation.

Idéalement, cela permettrait de vérifier toutes les cotisations de manière automatique et d'assurer l'égalité entre montants effectivement dûs et montants versés par les entreprises. Pour les raisons présentées dans le premier chapitre, l'exactitude des montants versés ne peut être assurée par la simple déclaration de cotisations. Une partie essentielle du processus de contrôle est, alors, de corriger les écarts de cotisation, volontaires ou non. Il est régi par trois aspects :
- un caractère légal qui définit la manière dont sont engagées les procédures de contrôle ;
- un caractère administratif, établi par l'Etat, l'ACOSS et l'URSSAF, qui définit les objectifs en terme de politique de contrôle ;
- un caractère opérationnel, exécuté par l'ACOSS et l'URSSAF, qui définit explicitement les critères et les contrôles à effectuer, les indicateurs importants et les résultats attendus.

`La procédure et les objectifs du contrôle`

Les contrôles comptables d'assiette (CCA) portent, par obligation légale, sur les trois dernières années de cotisations de l'entreprise ainsi que celles de l'année en cours au moment du contrôle. Il existe d'autres spécificités (comme les recours des entreprises ou le cas d'éventuelles pénalités) que nous ne détaillons pas car elles n'interviennent pas, ou très peu, dans la modélisation que nous exposons. Précisons cependant qu'une entreprise peut posséder plusieurs établissements. Dans ce cas l'URSSAF est autorisée, mais avec un

certain nombre de contraintes, à échantillonner une partie des déclarations effectuées pour chaque salarié et à extrapoler le résultat d'un éventuel redressement sur cet échantillon à l'effectif total et à tous les établissements de l'entreprise. L'accord de cette dernière est néanmoins requis pour cette procédure, laquelle concerne avant tout les grandes entreprises. Généralement, les contrôles sont ciblés ou thématiques : le ciblage consiste à tenir compte d'un critère particulier (par exemple la taille de l'entreprise, la masse salariale ou le secteur d'activité) pour définir les actions de contrôle. Un thème de contrôle peut être la vérification plus spécifique de certaines catégories de cotisation (comme les mesures de réduction) ou de certains éléments de rémunération. La diversité des missions aboutit à des plans de contrôle, dont chacun peut être caractérisé par un ou plusieurs critères spécifiques. Pour chaque plan de contrôle, il existe des indicateurs qui permettent de mesurer globalement les résultats obtenus. Ce sont la fréquence de redressement, le taux de redressement et dans une moindre mesure, le temps de contrôle moyen par entreprise. Ces indicateurs constituent les principaux outils de mesure de performance du contrôle. Nous pouvons y rajouter le montant total des redressements, lequel permet d'avoir une perspective sur les politiques de contrôle. Notons que pour chaque indicateur, l'URSSAF se voit attribuer un objectif numérique précis, lequel intervient dans la politique de rémunération de l'ensemble de ses salariés (et a une influence sur les plans de contrôle et leur exécution).

`La fréquence et le taux de redressement`

La fréquence de redressement est le rapport entre le nombre de redressements positifs (en faveur de l'URSSAF) et négatifs (en faveur des cotisants), et le nombre total de contrôles.
Le taux global de redressement, est le rapport entre la somme des montants redressés (en faveur de l'URSSAF ou des entreprises, et en valeur absolue) et la somme des montants contrôlés.

Chacun de ces indicateurs peut être calculé sur un plan de contrôle, de façon à pouvoir le comparer à un autre. Afin d'éviter une confusion entre redressements positifs et négatifs, la fréquence et le taux de redressement sont déclinés dans une deuxième version :
La fréquence des redressements positifs, laquelle ne prend au numérateur que le nombre de redressements positifs.
Le taux de redressement débiteur, lequel ne prend au numérateur que la somme des montants de redressement positifs (en faveur de l'URSSAF).

La fréquence des redressements positifs est équivalente à un taux de détection des irrégularités. Elle indique le degré d'efficacité d'une politique de contrôle relativement à la recherche d'irrégularités. Par exemple, une fréquence des redressements positifs de 50% signifie que sur 100 entreprises contrôlées, la moitié des contrôles donneront lieu à un redressement en faveur de l'URSSAF. Le taux de redressement est équivalent au pourcentage de cotisations éludées relativement au montant contrôlé. Par exemple, un taux de redressement débiteur de 2% signifie que sur 100 euros de cotisations contrôlées, le montant des redressements positifs est de 2 euros. Le taux de redressement n'est adapté à la comparaison entre plans de contrôle, ou entre modèles, que lorsque lorsqu'il y a homogénéité des montants contrôlés, de la masse salariale et/ou de la taille des entreprises.

Les variables et les données du contrôle

Plusieurs types d'information peuvent se révéler utiles au contrôle des entreprises. Dans la période récente, la question du croisement des bases de données (et non plus des données uniquement) est posée afin d'améliorer les résultats des contrôles, notamment dans le cas des irrégularités volontaires. Une alternative, ou un procédé complémentaire, est l'utilisation de modèles pour les seules variables disponibles. C'est le cas pour la détection des irrégularités. Toutefois, dans le quatrième chapitre, consacré à la fraude dans le cadre du travail dissimulé, nous montrons comment la présence de variables économiques décrivant la situation financière de l'entreprise, variables habituellement indisponibles, explique la fraude.

Ici, nous résumons les variables utilisées dans le cadre de la modélisation et auxquelles le contrôle peut, à minima, se référer. Deux types de variables sont disponibles au sein de l'URSSAF :
- les codes-type de personnel, lesquels sont les catégories de cotisation, mesurées par la somme versée par l'entreprise pour chaque cotisation ;
- les variables liées à la déclaration de l'entreprise à sa situation courante vis-à-vis de l'URSSAF, et à la politique de recouvrement de cette dernière.

Dans le premier cas, les codes-type de personnel correspondent aux 900 (plus exactement 655 catégories répertoriées pour l'Île-de-France, parmi les 999 virtuellement possibles) catégories de cotisation disponibles. Par exemple, le *Cas général* est la principale catégorie de cotisations, obligatoire pour tous les salariés et abondant les prestations sociales des branches Maladie, Maternité, Invalidité, Décès, Solidarité, Allocations familiales et Vieillesse (en partie). Il correspond au code-type 100. Ce *Cas général* supporte des exceptions, au moins les codes-type 102, 104, 105, 106, 107, 108, 109, 110, 112 et 114. La CSG a pour code-type 260. Elle supporte également des exceptions.
Ces variables ne sont pas évidentes à exploiter et d'autres, plus explicites, sont observées. Elles correspondent, par exemple, au secteur d'activité, aux écarts de cotisation, au nombre de retards de paiements, aux taxations d'office, aux pénalités, au dernier contrôle par l'URSSAF,... Nous récapitulons, un peu plus loin, l'ensemble des variables.

Dans la sélection des entreprises à contrôler, le croisement des données ou la modélisation statistique sont les principaux outils d'évaluation. Le croisement de données, généralement utilisé, définit des critères à partir desquels la sélection s'effectue. Un exemple de plan de contrôle issu d'un croisement de données peut être la recherche de toutes les entreprises, entre 10 et 20 salariés, ayant eu des retards de paiement et n'ayant jamais été contrôlées auparavant. La modélisation statistique utilise l'échantillonnage aléatoire, les modèles linéaires ou encore l'apprentissage statistique. Citons l'approche par un modèle économétrique de Joubert (2009) dans le cadre de la fraude (hors travail dissimulé) aux cotisations sociales dans l'agglomération lyonnaise, grâce à une modélisation du processus de sélection et de détection des entreprises. L'industrialisation des modèles mathématiques est (très) peu courante, à ce jour, et une tendance dans l'évolution des méthodes de contrôle est le *data mining* (fouille de données) qui regroupe à la fois des outils descriptifs, de croisement de (bases de) données pour l'aide à la décision, et, éventuellement, des mo-

dèles prédictifs. Notons que l'apprentissage statistique n'est pas du *data mining*. Il s'en distingue par les outils théoriques, les algorithmes et l'inférence. La partie algorithmique de l'apprentissage statistique est généralement désignée sous le terme *machine learning* ou *apprentissage automatique*.

`Les recours des entreprises`

Une fois le contrôle d'une entreprise effectué, celle-ci dispose de plusieurs voies de recours pour le contester. Le processus de recours peut être très long, en particulier dans le cas des grandes entreprises, et les étapes vont du recours amiable à la Cour de Cassation. Généralement, les contestations concernent peu de cas (5% des entreprises redressées en 2008) mais des sommes élevées (30% des montants redressés cette même année).

2.2.1 Les résultats des contrôles URSSAF

Les montants redressés, par leur importance, donnent généralement l'impression d'une présence massive d'irrégularités dans les déclarations de cotisation des entreprises. Cette impression doit être mise en contraste avec les cotisations sociales versées par ces mêmes entreprises, au minimum, 100 fois plus importantes. La présence massive d'irrégularités ne peut, elle, être déterminée à cause de la nature même des contrôles, thématiques ou ciblés. Plus simplement, plus les contrôles sont ciblés, moins on peut accroître le nombre d'entreprises contrôlées. Lorsque les montants redressés augmentent, la raison principale est le fait de contrôles de plus en plus fréquents sur les entreprises pour lesquelles les enjeux financiers sont élevés et non le fait que le nombre de contrôles augmente (ou que le taux de détection s'améliore).

La principale difficulté dans le contrôle des entreprises est la fréquence des redressements positifs, soit la capacité à détecter des irrégularités en faveur de l'URSSAF relativement au nombre de contrôles effectués. Cette difficulté est masquée par une augmentation quasi-continue des montants redressés depuis, au moins, 2006. Nous montrons dans les lignes qui suivent les raisons de la contradiction entre ces deux événements. Nous développons également l'essentiel des résultats utiles à la compréhension des irrégularités aux cotisations sociales d'un point de vue quantitatif.

`La fréquence des redressements positifs et le taux de redressement`

Nous indiquons l'évolution des principaux indicateurs du contrôle entre 2006 et 2011.
La fréquence des redressements positifs est le rapport entre le nombre d'entreprises redressées, pour un montant net en faveur de l'URSSAF, et le nombre d'entreprises contrôlées.
Le taux global de redressement débiteur est le rapport entre la somme totale des montants redressés en faveur de l'URSSAF et la somme totale des montants contrôlés.

Redressements positifs	2011	2010	2009	2008	2007	2006
Fréquence	55.46%	50.07%	51.04%	47.05%	50%	54.39%
Taux global	2.05%	1.59%	1.86%	1.69%	2.25%	1.87%

TABLE 2.1 – Le pourcentage de redressements positifs relativement au nombre total de contrôles et le montant des redressements relativement aux montants contrôlés, de 2006 à 2011, pour l'URSSAF d'Île-de-France.

En moyenne, parmi les entreprises contrôlées, un peu plus d'une entreprise sur deux fait l'objet d'un redressement en faveur de l'URSSAF lors d'un contrôle comptable d'assiette. Pour les grandes entreprises, le ratio atteint ou dépasse 80%. Les montants redressés représentent, en moyenne, un peu moins de 2% des cotisations contrôlées.

Dans certaines publications, seule la "fréquence de redressement" et le "taux de redressement" sont indiqués. Dans ces deux cas, les redressements négatifs (en faveur de l'entreprise et en valeur absolue pour les montants) sont ajoutés aux redressements positifs. L'équité est alors autant privilégiée que l'efficience. Comme notre propos porte essentiellement sur les irrégularités, et non les erreurs, nous ne retenons que les critères d'efficience. Notons également que le taux de redressement débiteur ne permet pas de faire la distinction entre une augmentation des montants redressés ou une diminution des montants contrôlés.

`Le nombre de contrôles et le montant des irrégularités`

Le nombre de contrôles et la somme des montants redressés permet de préciser une première partie de la progression des résultats.

Contrôles comptables d'assiette	2011	2010	2009	2008	2007	2006
Nombre de contrôles	14472	15604	16872	18141	15981	11470
Montant total net des redressements (milliers d'euros)	130 092	97 034	102 503	59 330	62 936	66 948

TABLE 2.2 – Le nombre total de contrôles et le montant total net des redressements, de 2006 à 2011, pour l'URSSAF d'Île-de-France

Les montants redressés ont doublé entre 2006 et 2011, tandis que, dans le même temps, le nombre de contrôles n'a augmenté que de 25%. Les montants nets redressés pour l'Île-de-France représentaient, en 2011, 17% des sommes redressées au niveau national. Durant cette même année, le montant moyen net de redressement était d'environ 9000 euros. Toutefois, pour trois quarts des entreprises redressées, il est de 2700 euros. Le ciblage accru, notamment celui lié aux enjeux financiers importants, est la principale raison de la forte progression du montant des irrégularités. Cette progression ne s'est pas traduite par une meilleure détection des irrégularités (fréquence des redressements positifs). Pour en comprendre la raison, nous nous intéressons à la répartition des sommes redressées.

`Le poids des enjeux financiers les plus importants`

Afin de mesurer de manière simple, à la fois, la progression des résultats et l'influence des enjeux financiers, nous avons calculé la somme des redressements supérieurs à 100 000 euros et le nombre d'entreprises concernées.

Contrôles comptables d'assiette et redressements supérieurs à 100 000 euros	2011	2010	2009	2008	2007	2006
Somme (milliers d'euros)	94 956	71 801	76 957	43 288	42 863	43 453
Poids dans le montant total net des redressements	73%	74%	75%	73%	68%	65%
Nombre d'entreprises	200	160	146	114	121	112

TABLE 2.3 – Le nombre et le poids des enjeux financiers les plus importants de 2006 à 2011.

De 2006 à 2011, moins de 2% des entreprises ont contribué à plus de 70% du montant total net des redressements pour les contrôles comptables d'assiette. Rappelons que ces derniers représentent, en valeur, deux tiers du montant de tous les types d'irrégularités aux cotisations sociales contrôlées par l'URSSAF.

De façon encore plus précise, c'est le doublement du nombre de contrôles aux enjeux financiers importants (redressements supérieurs à 100 000 euros) qui est à l'origine de la progression des montants redressés. Pour les autres contrôles (98% du nombre total de contrôles), les montants redressés sont passés de 24 (en 2006) à 35 millions d'euros en 2011, soit une moyenne, par entreprise redressée, de 2100 euros (en 2006) à 2500 euros, en 2011.

Pour cette première raison, les irrégularités, telles qu'elles sont détectées à ce jour, ne peuvent être considérées comme une source de financement de la Sécurité sociale.
90% des sommes redressées sont le fait de 10% des entreprises contrôlées, soit environ 1500 cotisants pour chacun desquels le montant redressé dépasse 10 000 euros.
Une deuxième raison, plus primordiale, est la fréquence des redressements positifs trop peu importante pour une détection plus précise des irrégularités.

`La répartition du taux de redressement`

Afin de mieux comprendre les difficultés inhérentes au ciblage des contrôles, nous observons la répartition des entreprises contrôlées en fonction du taux de redressement débiteur (en faveur de l'URSSAF)

Pourcentage des montants de redressement	Pourcentage d'entreprises redressées (période 2006-2011)
inférieur à 1% des cotisations contrôlées	$> 60\%$
entre 1% et 10% des cotisations contrôlées	$< 30\%$
supérieur à 10% des cotisations contrôlées	$< 10\%$

TABLE 2.4 – La répartition relative du nombre d'entreprises redressées relativement au taux de redressement (rapport entre le montant total redressé en faveur de l'URSSAF et le montant total contrôlé).

De 2006 à 2011, plus de 60% des montants redressés représentent moins de 1% des cotisations contrôlés. En valeur monétaire, 4 entreprises redressées sur 10 le sont pour un montant net inférieur à 1000 euros. Du fait de la limitation des ressources, le nombre d'options est alors limité et il est nécessaire de fixer la priorité sur les enjeux financiers importants. Sur les 200 entreprises qui ont constitué 70% des montants redressés en 2011, les grandes entreprises en représentent la majorité.

`L'évolution future des contrôles`

Comparativement au niveau national (+21%), l'évolution entre 2010 et 2011 des montants redressés (+34%) pour l'URSSAF d'Île-de-France s'explique par un contrôle de plus en plus systématique des grandes entreprises. Plus particulièrement, il est très probable que les entreprises aux masses salariales dépassant 1 million d'euros seront toutes contrôlées, tous les 3 ans, dans un futur proche. Même si elles représentent moins de 10% des entreprises (un peu moins de 20 000 en Île-de-France), le niveau de progression quasi-continu des résultats depuis 2006, corrélé avec une augmentation du seul contrôle de ces entreprises dessine une tendance claire. En fait, de manière paradoxale, il est plus avantageux pour une entreprise avec beaucoup de salariés de laisser l'URSSAF corriger ses irrégularités. Si l'entreprise n'est pas contrôlée, les montants éludés se transforment en gain. Si elle l'est, elle s'épargne les coûts supplémentaires de régularisation de ses cotisations, qui sont, de facto, pris en charge par l'URSSAF. Comme le caractère volontaire de l'irrégularité, déclencheur de pénalités, doit être prouvé, l'entreprise peut plaider sa bonne foi et, dans le pire des cas, multiplier les recours.

Au niveau national, les redressements notifiés, dans le cadre des contrôles comptables d'assiette, se sont élevés à plus de 900 millions d'euros en 2012. Le taux de récupération effectif des montants n'est pas connu.

2.2.2 Quelques problématiques de la détection d'irrégularités

L'analyse des résultats des contrôles URSSAF montrent les principales pistes de la détection d'irrégularités. Certaines sont simples et sont des évolutions naturelles des méthodes de contrôle. D'autres relèvent d'une approche massivement générique du problème et ne font aucune hypothèse sur les caractéristiques des entreprises ou l'aspect frauduleux ou non d'une irrégularité. C'est le cas de nombreux modèles en apprentissage statistique. Avant de détailler le modèle proprement dit, nous résumons les principales problématiques liées aux irrégularités aux cotisations sociales :

- les irrégularités détectées ne sont, en général, pas des cas de fraude ;
- la législation, complexe, est une raison, probable, d'un grand nombre d'irrégularités ;
- un peu plus d'un contrôle sur deux aboutit à la détection d'une irrégularité ;
- l'asymétrie des redressements est très importante et un petit nombre d'entreprises constitue l'essentiel des montants redressés ;
- trouver toutes les irrégularités n'est pas opportun. Le critère fondamental est le rapport entre le nombre d'irrégularités correctement prédites par un modèle et le nombre d'entreprises qui vont être contrôlées grâce aux informations produites par le même modèle.

Ces problématiques sont celles dont nous discutons tout au long des lignes qui suivent. La première question à laquelle nous souhaitons répondre est celle de savoir comment détecter, avec une précision aussi élevée que possible, une irrégularité. La seconde question est celle de savoir si l'irrégularité détectée conduit à un montant redressé proche de celui estimé par le modèle. Pour y répondre, nous commençons d'abord par l'ajout de nouveaux indicateurs et une discussion sur le taux de redressement.

`Indicateurs supplémentaires, taux de redressement et propriétés de la détection d'irrégularités`

Nous insistons ici sur le taux de redressement car il représente un indicateur de performance, et de comparaison entre modèles, pour l'URSSAF. Lorsque la mesure de performance doit être effectuée sur deux modèles homogènes en termes de masse salariale, d'effectif et/ou de secteur d'activité considérés, le taux de redressement donne une synthèse de l'impact financier des redressements relativement aux cotisations contrôlées.

Cette synthèse est cependant incomplète car le taux de redressement ne peut pas être utilisé comme indicateur de comparaison : il ne tient pas compte de l'enjeu financier pour la Sécurité sociale, du rendement d'un plan de contrôle ou d'un modèle, ou d'une meilleure répartition des ressources, par exemple pour des contrôles plus approfondis. Afin de préciser notre propos, nous en donnons un paradigme, à travers trois exemples.
1- Supposons deux plans de contrôle qui génèrent la même somme de montants contrôlés, 100 000, pour deux contrôles effectués, dont un qui fournit un résultat en faveur de l'entreprise. Résumons leurs résultats ci-dessous :

	Contrôle 1 (montant redressé)	Contrôle 2 (montant redressé)	Cotisations contrôlées	Taux de redressement débiteur
Plan 1	+10 000	-10 000	100 000	10%
Plan 2	+6 000	-1 000	100 000	6%

Le "Plan 1" a un meilleur taux de redressement mais contribue pour 0 (la somme nette des redressements) aux recettes de la Sécurité sociale. Le "Plan 2" y contribue pour 5000.

2- Dans le deuxième exemple, nous considérons l'absence de prise en compte du nombre de contrôles en faisant varier ce dernier :

	Nombre de contrôles	Montant net redressé	Cotisations contrôlées	Taux de redressement débiteur
Plan 1	20	25 000	100 000	25%
Plan 2	10	20 000	100 000	20%

Le taux de redressement est plus élevé pour le Plan 1, alors que le Plan 2 a un rendement (2000 euros par contrôle) presque deux fois plus important.

3- Dans le troisième exemple, et de notre point de vue le plus problématique, nous faisons varier les cotisations contrôlées, le nombre de contrôles et les montants nets redressés :

	Nombre de contrôles	Montant net redressé	Cotisations contrôlées	Taux de redressement débiteur
Plan 1	40	25 000	50 000	50%
Plan 2	20	50 000	200 000	25%

Le Plan 1 a un rendement quatre fois moins élevé (625 euros), participe moins aux recettes de la Sécurité sociale et mobilise plus de ressources. Son taux de redressement est deux fois plus important que pour le Plan 2.

Pour évaluer un modèle, nous considérons trois critères :
- la fréquence des redressements positifs ou précision ;
- le montant comptable moyen (redressé) ;
- le score d'importance de la détection, qui constitue un nouvel indicateur synthétique.

La fréquence des redressements positifs est équivalente à la *précision*, définie par :

$$Précision = \frac{Nombre\ d'irrégularités\ correctement\ prédites}{Nombre\ d'irrégularités\ prédites}.$$

En d'autres termes, il n'est pas nécessaire de détecter toutes les irrégularités, mais lorsque le modèle en détermine une, la probabilité pour que cela soit réellement le cas doit être la plus grande possible et nécessairement supérieure à 0.5. La limitation des ressources (temps de contrôle, nombre d'inspecteurs,...) rend caduque, et même contre-productive, la recherche de toutes les irrégularités. Plus précisément, le nombre d'entreprises est trop important relativement aux ressources disponibles pour le contrôle. En supposant qu'un modèle puisse détecter toutes les irrégularités, seule une petite partie d'entre elles pourrait être vérifiée. Nous recherchons donc des modèles avec plusieurs propriétés.

i) Pour n'importe quelle entreprise, le modèle doit pouvoir déterminer la probabilité d'existence d'une irrégularité. Ainsi, l'ordre des probabilités détermine la manière dont sont choisies les entreprises à contrôler.

ii) Pour l'ensemble des entreprises détectées, la *précision* doit être aussi proche que possible de 1. Ainsi, le temps passé à contrôler les entreprises n'est plus un temps passé à essayer de détecter les irrégularités (cette étape étant déjà effectuée par le modèle) mais un temps passé à vérifier l'existence effective des irrégularités grâce à la probabilité déterminée par le modèle. Plus cette probabilité est grande, plus l'irrégularité est probable.

iii) Afin d'éviter que ce temps soit perdu en vérifications inutiles, le modèle doit être en mesure de fournir, avant que les contrôles n'aient lieu, une mesure de la précision estimée, strictement inférieure à celle qui sera observée au plan opérationnel. Cette mesure est une fonction du nombre de contrôles qui vont être effectués.

Afin d'uniformiser le cadre de la discussion, nous notons P_r, la *précision*. Les contrôles sont effectués à l'aide des recommandations du modèle et chacun d'eux est associé à une probabilité d'existence d'une irrégularité dans la déclaration de cotisations sociales de l'entreprise concernée. Naturellement, les contrôles associés aux probabilités les plus importantes sont les plus susceptibles de mener à des irrégularités. Plus la probabilité diminue, moins il y a d'évidences sur une irrégularité éventuelle. Ainsi, en dessous d'une probabilité de 0.5, on considère généralement qu'elle n'a pas lieu. Le nombre d'irrégularités, à la fois prédites par le modèle et effectivement observées dans le cadre des contrôles, est le nombre de *vrais positifs*, notés VP. Le nombre d'irrégularités prédites par le modèle, et non observées par les inspecteurs, est le nombre de *faux positifs*, noté FP.
La *précision* se récrit alors sous la forme suivante : $P_r = \frac{VP}{VP+FP}$.

Le dénominateur est la décomposition du *nombre de contrôles* effectués sous la supervision du modèle. Certains aboutiront à l'observation effective d'irrégularités (les vrais positifs, VP), d'autres n'aboutiront pas (les faux-positifs, FP). Savoir trouver les irrégularités ne suffit pas à satisfaire la détection. Il est également nécessaire de connaître (ou d'estimer) les montants découlant de ces irrégularités, afin de proposer la sélection des cas les plus importants. C'est une forme de ciblage, mais il est neutre car un arbitrage doit être réalisé entre la probabilité de détection (probabilité d'existence d'une irrégularité) et le montant de redressement espéré.

iv) De plus, les caractéristiques propres à l'entreprise (effectif, masse salariale, secteur d'activité, zone géographique,...) ne sont pas connues du modèle et n'interviennent pas (du moins pas directement) car les biais de sélection sont à éviter.

v) Le montant comptable moyen, ou montant moyen net redressé, est le rendement du modèle (ou d'un plan de contrôle). Plus le rendement est important, plus le modèle est performant. Nous le définissons ainsi :

$$Montant\ comptable\ moyen = \frac{\sum montants\ redressés\ positifs - \sum montants\ redressés\ négatifs}{Nombre\ de\ contrôles}.$$

Les *montants redressés positifs* sont les montants redressés par l'URSSAF en sa faveur. Les *montants redressés négatifs* sont les montants redressés par l'URSSAF en faveur des entreprises et constituent donc un remboursement à ces dernières.

Le montant comptable moyen pénalise à la fois les redressements négatifs (en faveur de l'entreprise) et les contrôles sans enjeux. Précisons qu'une même entreprise peut, pour un même contrôle, subir un ou plusieurs redressements positifs et/ou un ou plusieurs redressements négatifs.
On note R, la variable aléatoire, à valeurs dans $\mathbb{R}$, correspondant au montant comptable redressé lors d'un contrôle d'une entreprise,
R^+, le montant des redressements positifs,
R^-, le montant des redressements négatifs,
N_C, le nombre de contrôles réalisés.
Pour une entreprise i contrôlée, $1 \leq i \leq N_C$, le montant comptable redressé est alors défini par :

$$R_i = R_i^+ - R_i^-,$$

Le montant comptable moyen pour l'ensemble des entreprises contrôlées se récrit alors sous la forme suivante :

$$\bar{R} = \frac{\sum_{i=1}^{N_C} R_i^+ - \sum_{i=1}^{N_C} R_i^-}{N_C}.$$

Nous avons vu, précédemment, que le taux de redressement n'était pas un indicateur adapté à la comparaison entre modèles. Pour comparer tous les modèles, à la fois par leur capacité à détecter des irrégularités et par leur capacité à estimer les montants espérés de redressement, il nous faut un indicateur universel.
Nous notons M_C, la masse salariale totale des entreprises contrôlées sur les recommandations du modèle. Dans le cas d'un modèle générique, capable d'évaluer la déclaration d'une entreprise quelconque, c'est la masse salariale totale des entreprises enregistrées.
On définit également M_R, la masse salariale de toutes les entreprises redressées grâce au modèle, et N_R, leur nombre.
Le score d'importance de la détection est alors noté S_D, et donné par :

$$S_D = \frac{\left(P_r - \frac{1-P_r}{P_r}\right)\left(1 + \frac{M_C}{M_R N_R}\right)\sum_{i=1}^{N_R} R_i}{M_R} \times 100.$$

Le score d'importance pénalise à la fois les modèles dont les capacités de détection sont faibles et ceux qui ne génèrent que peu de montants redressés. L'hypothèse qui prévaut ici est qu'un modèle qui ne capture qu'une petite partie des irrégularités issues de ses recommandations ne peut pas générer un montant de redressements important, relativement au montant réel des irrégularités, si on le généralise à l'ensemble des entreprises. De la même manière, un modèle possédant une grande précision doit pouvoir être généralisé, autrement dit sa précision ne doit pas s'effondrer quand le nombre de contrôles recommandés augmente. En particulier, un modèle admissible doit avoir un score positif relativement à la masse salariale de toutes les entreprises qu'il est en mesure d'évaluer. Notons que le score d'importance intègre également un équivalent du taux de redressement grâce au terme $\sum_{i=1}^{N_R} R_i/M_R$.

Dans le dernier chapitre, nous revenons plus longuement sur l'évaluation et sur la comparaisons de modèles. La *précision* est le critère intrinsèque et le plus important.
Dans le cas des cotisations sociales, nous cherchons des modèles capables d'évaluer, avec une grande précision, 400 000 entreprises, chaque année, pour l'Île-de-France et 1 200 000 pour la France entière ; cela sans nécessiter autre chose que la seule information disponible et des ressources identiques à celles existantes. Le *rendement* est la capacité à transformer les montants redressés en ressources supplémentaires pour la Sécurité sociale. Plus la précision est importante, plus le rendement augmente. Lorsque le nombre d'entreprises considéré dans le modèle devient assez grand, le *score d'importance* mesure indirectement le montant des irrégularités, relativement à la masse salariale, potentiellement récupérable, si le score est positif.

Ces trois indicateurs ont l'intérêt d'avoir une connexion importante avec la réalité opérationnelle. Plus le nombre de contrôles recommandés par un modèle est exploité, plus les indicateurs transcrivent, en temps réel, sa perte ou son gain d'efficience. Il suffit alors d'appliquer les indicateurs à la politique globale de contrôle de l'URSSAF, à effectif moyen identique, pour savoir si le modèle l'améliore ou non.

`Le modèle générique de la détection d'irrégularités`

Afin de clarifier les objectifs, nous en donnons une formulation explicite :

Un modèle admissible pour la détection des irrégularités aux cotisations sociales est un modèle qui doit remplir plusieurs conditions :
1) il doit sélectionner de manière autonome les entreprises à contrôler ;
2) il ne doit pas exister de biais de sélection ;
3) il doit assurer à l'URSSAF un niveau de détection des irrégularités strictement supérieur à celui observé habituellement sur l'ensemble des contrôles ;
4) le montant comptable moyen redressé (grâce au modèle) doit être supérieur à celui mesuré sur, au moins, 75% des entreprises contrôlées, en l'absence d'utilisation du modèle ;
5) le score d'importance de la détection doit être positif ;
6) le modèle doit apporter des garanties, en matière de fréquence des redressements positifs et de montant comptable, qui doivent être des bornes inférieures aux résultats qui seront observés au niveau opérationnel, sur la base des recommandations du même modèle.

En particulier, le point 4) indique que seuls les contrôles spécifiques aux grandes entreprises sont susceptibles de générer des montants comptables moyens supérieurs à un modèle admissible. En se référant à la description que nous avons donnée des résultats du contrôle, le seuil de 75% peut sembler arbitraire puisque 10% des entreprises génèrent, de manière quasi-systématique, 90% des redressements. Ce seuil permet surtout d'assurer que les entreprises contrôlées sous la supervision d'un modèle pourront, dans la grande majorité des cas, se substituer aux contrôles habituels si la limitation des ressources ne peut être levée. Cela sans changer la nature de la politique de contrôle, du fait de l'absence de biais de sélection. Nous développons les deux premiers points dans la section qui suit. Puis, dans le dernier chapitre nous donnons les résultats et la mise en oeuvre, dans la pratique, du modèle.

2.3 Le processus de modélisation des données

La définition de quelques indicateurs permet de synthétiser les résultats et le processus opérationnel de la détection d'irrégularités. Toutefois, de nombreuses étapes sont nécessaires avant la validation d'un modèle. Dans le cas de l'apprentissage statistique, deux étapes sont toujours essentielles :
- le pré-traitement des données ;
- la recherche d'un modèle suffisamment générique pour répondre à un grand nombre de contraintes et suffisamment spécifique pour atteindre les performances annoncées.

Nous discutons de ces aspects et du passage d'une base de données à un modèle industriel. Dans le premier chapitre, nous avons montré comment caractériser n'importe quelle cotisation grâce à quatre paramètres : la masse salariale, le taux d'application de la cotisation, son assiette, et la proportion d'effectif concernée par la cotisation observée. Mais aucun de ces paramètres n'est certifié exact par le seul enregistrement d'une déclaration, sauf dans le cas où l'inspecteur du contrôle vérifie au sein de l'entreprise la situation de chaque salarié. Généralement, ces paramètres, hormis la masse salariale, ne sont pas enregistrés par l'URSSAF (du moins, nous n'y avions pas accès de manière simple à travers les bases de données). A la place, le montant de la cotisation est inscrit, ainsi que d'autres variables, liées, par exemple, au nombre de salariés ou au secteur d'activité. Nous les détaillons plus loin. Le système informatique, conformément à la masse salariale et aux éventuelles particularités déclarées par l'entreprise (contrats spécifiques, zone géographique favorisée, salaires pouvant bénéficier de mesures de réduction, autres mesures dérogatoires, ...), calcule les cotisations attendues et notifie, le cas échéant, les écarts avec les cotisations payées.

Notons que même si un taux d'application est défini par la législation, il comprend généralement une partie plafonnée, une autre, non plafonnée, et des conditions d'application qui peuvent entraîner une interprétation erronée de la part de l'entreprise, surtout si elle compte un grand nombre de salariés. L'assiette de cotisation, la part du salaire à laquelle s'applique la cotisation, est fixée également par la législation mais peut soutenir des exceptions. Le montant de la cotisation est, de manière générique, le produit de l'assiette et du taux de cotisation. Ce caractère générique est, en réalité, très extensible et nous l'illustrons, en détail, dans le quatrième chapitre consacré à la fraude dans le cadre du travail dissimulé.

De manière plus précise, la vérification des cotisations suppose de connaître chacun des taux des 900 catégories déclaratives de cotisation, leurs assiettes, le salaire brut de chaque salarié et chaque condition d'application d'une catégorie.

i) En supposant la disponibilité de toutes ces données, la problématique posée est alors autant liée à leur volume, plusieurs millions de salariés, qu'au choix d'un modèle capable de tirer parti d'autant d'informations.
ii) Une première simplification, massive, consiste alors à ne s'intéresser à l'information qu'au niveau de l'entreprise, puisque c'est elle qui est contrôlée. Pour chacune, on dispose alors uniquement des informations sur sa déclaration de cotisation, sur son historique de

relations avec l'URSSAF et sur une partie de ses caractéristiques structurelles.
iii) Une deuxième simplification permet d'éliminer toutes les informations permettant de caractériser l'entreprise. On ne souhaite voir apparaître aucun biais de sélection, par exemple un ciblage de l'entreprise basé sur son secteur d'activité, sa taille ou sa masse salariale. A la place, un processus de standardisation de toutes les données est mis en place, de sorte que la majorité des variables voient leurs valeurs comprises entres -1 et 1. Nous reprenons, là, le principe de l'invariance d'échelle : en observant des variables importantes, comme les cotisations, relativement à la masse salariale ou à l'effectif ou à d'autres "constantes" (relatives), l'information globale obtenue est plus pertinente. Par exemple, des changements dans la législation, très courants, ne perturbent pas la modélisation. Par la même occasion, on supprime ainsi une partie des biais éventuels qui apparaissent dans le ciblage des entreprises par l'URSSAF.
iv) Cependant, le résultat le plus important est la possibilité d'effectuer un *apprentissage incrémental*. Plus simplement, le modèle utilisé peut alors apprendre les données au fur et à mesure de leur arrivée, année après année, se mettre à jour de manière automatique, et être indépendant du volume des données.

De manière plus concise, simplifier et modéliser les données consiste à ne considérer que les entreprises (et non les salariés) et à éliminer toutes les dépendances à la législation, aux caractéristiques de l'entreprise, au volume des données ou à leur temporalité. Et, naturellement, sans aucune information nominative.

Dans la section précédente, nous avons résumé les deux types de variables disponibles :
- les codes-type de personnel
- Les variables liées à la déclaration de l'entreprise.

`Les codes-type de personnel`

Ce sont les catégories de cotisation à déclarer (et à verser) par chaque entreprise. Chaque catégorie correspond à une cotisation spécifiée par un intitulé (le nom), un taux, une assiette et des conditions d'applications. Par exemple, les heures supplémentaires sont une catégorie de cotisation, comme certains types de contrats ou encore les mesures de réduction, chacune défini par un code-type unique.

i) Les intitulés des codes-type sont des numéros, de 0 à 999, et nous disposons donc de 1000 variables (chacune associée à un code-type) qui peuvent être virtuellement déclarées par une entreprise. Comme la législation est susceptible de changer, nous retenons plutôt les numéros de -9 à 1023. Les codes-type virtuels que nous ajoutons sont destinés à accueillir des éventuelles modifications de la législation. Ils sont traités par un algorithme spécifique qui est chargé de détecter les nouveaux codes-type. Pour une meilleure lecture, nous détaillons rapidement la manière dont les codes-type de personnel sont fournis, en présentant un paradigme de la base de données constituée de chaque code-type :

Cotisation sociale	Exemple
SIRET	01234567
Numéro du Code-type	100
Taux plafonné	15.15
Taux déplafonné	20.95
Taux AT	01.60
Cotisation due	28 675.69
Assiette déplafonnée	77 572
Assiette plafonnée	74 307
Effectif exonéré moyen	NA

TABLE 2.5 – Représentation générique d'une cotisation sociale dans une base de données.

Le tableau ci-dessus est la représentation typique que nous avons de chaque catégorie de cotisation.
- Le SIRET est l'identifiant public de l'établissement. Une entreprise est constituée d'au moins un établissement. Il nous faut alors un identifiant unique, par entreprise, qui est le SIREN, lorsque l'entreprise a plusieurs établissements.
- Le numéro du code-type est le nom de la catégorie de cotisation et de la variable.
- Le taux plafonné est le taux de cotisation pour tous les salaires, avec une limite ne dépassant pas un plafond fixé par la Sécurité sociale. Par un exemple, pour un salaire dépassant le plafond, c'est ce dernier auquel sera appliqué le taux.
- Le taux déplafonné est le taux appliqué à tous les salaires, quels que soient leur montants.
- La cotisation due est calculée par le système d'information à partir des taux et assiettes, définis par la législation. Elle correspond au produit du taux et de l'assiette. La somme des cotisations plafonnée et déplafonnée constitue la cotisation due. Lorsqu'elle ne correspond pas à la cotisation effectivement payée par l'entreprise, un écart de cotisation, qui n'est pas nécessairement une irrégularité, est généré.
- Le taux AT est le taux associé aux Accidents du travail. Il varie selon le secteur d'activité de l'entreprise. Il est généralement associé à la cotisation obligatoire principale payée par l'entreprise. Et vaut donc 0 pour les autres.
- Les assiettes déplafonnées et plafonnées sont la somme, partielle ou totale, des salaires de tous les employés de l'entreprise concernée par la dudite cotisation. Le salaire maximal d'une assiette plafonnée est celui défini comme plafond par la Sécurité sociale.
- L'effectif exonéré moyen correspond au nombre de salariés ayant bénéficié de mesures de réduction sur leur déclaration de cotisations. La valeur NA (Non Attribué) est généralement assignée, lorsque l'effectif exonéré n'est pas connu ou que la catégorie de cotisation concernée ne supporte pas de mesures de réduction.

ii) Pour chaque code-type, la *cotisation due* est sa réalisation.
iii) Seul le *taux de la cotisation principale* (habituellement le *Cas général*, dont le code-type est 100) et le *taux AT* sont retenus comme variables liées aux taux. Les autres ne le sont pas car ils sont déjà intégrés dans le calcul de la cotisation due.
iv) Le *rapport entre assiette plafonnée et assiette déplafonnée* est également retenu.

v) L'effectif exonéré moyen, considéré relativement à l'ensemble des codes-type déclarés par l'entreprise, sert à évaluer le niveau de conformité de l'entreprise au regard de ses obligations déclaratives. Cette variable correspond à la "*compliance*".

Les codes-type de personnel permettent de construire 1037 variables, dont chacune peut être potentiellement déclarée par une entreprise. Lorsque l'entreprise ne déclare pas un code-type donné, sa valeur vaut simplement 0. La matrice constituée des n entreprises, en lignes, et des 1037 variables en colonnes, contient alors principalement la valeur 0. Cette particularité, ainsi que le grand nombre de variables, a un impact sur le modèle mathématique sous-jacent car il invalide de nombreuses méthodes inadaptées dans le traitement des matrices creuses (forte proportion de zéros).

`Les variables relatives aux caractéristiques de l'entreprise et à la politique de recouvrement`

Les autres variables ajoutées sont celles qui permettent de rendre plus homogène la matrice précédente. Nous mentionnons les variables disponibles non utilisées :
- le Type de personne, définissant l'entreprise comme une personne morale ou physique ;
- la domiciliation géographique ;
- le secteur d'activité ;
- la catégorie juridique de l'entreprise ;
- l'effectif ;
- la masse salariale ;

De notre point de vue, ces variables créent un biais de sélection principalement du fait des contrôles de plus en plus ciblés qui sont réalisés. *Dans le cadre de l'apprentissage statistique, la capacité qui prime est celle de généralisation. Dans le cas de la politique de contrôle, on souhaite plutôt une spécialisation. Celle-ci étant déjà réalisée à travers les contrôles effectués, il est inutile de la répliquer dans un modèle.* Par exemple, certains secteurs d'activité présentent plus de risques d'irrégularité que d'autres et les capacités d'apprentissage des modèles statistiques ont alors tendance à reproduire ce biais dans leur évaluation des entreprises, ce qui altère les capacités de généralisation. Dans le cas de la masse salariale et de l'effectif, il convient de traiter le problème différemment : il n'est pas souhaitable d'utiliser la masse salariale directement, car plus elle est importante, plus la probabilité est grande, pour une entreprise, d'être contrôlée par l'URSSAF. Nous souhaitons neutraliser l'influence de la masse salariale et l'ensemble des variables ayant une unité monétaire voient leurs valeurs être divisées par la masse salariale de l'entreprise. De même, on ne s'intéresse qu'aux *variations d'effectif*, et au *rapport entre le montant total des cotisations et l'effectif*. Puis, nous supprimons l'effectif et la masse salariale des variables.

Nous retenons alors les variables liées à la politique de recouvrement de l'URSSAF, et à diverses informations disponibles. Afin de faciliter la lecture, nous n'en donnons que les intitulés (la liste complète et les définitions étant détaillées dans l'annexe).
vi) Dans le cas de la politique de recouvrement de l'URSSAF, nous nous intéressons aux variables liées aux difficultés de l'entreprise : le *montant des écarts de cotisation*, les *pé-*

nalités, les *régularisations* suite à des *retards de paiement*, les *taxations d'office*, ainsi que le nombre de contrôles subis et la *date du dernier contrôle*.
vii) Les dernières variables disponibles sont celles liées à la *durée de vie* de l'entreprise, le *nombre de déclarations d'embauche* relativement à l'effectif rémunéré, le *salaire moyen relativement au SMIC*, la *fréquence de paiement* des cotisations, le *nombre de codes-type d'exonération* (mesures de réduction) relativement au nombre total de codes-type déclarés,...

Les caractéristiques des entreprises et celles liées au recouvrement permettent de disposer de 28 variables supplémentaires.

`Un algorithme de traitement générique`

La constitution d'une matrice ne suffit cependant pas à décrire les cotisations sociales et la relation avec les irrégularités. Une entreprise est contrôlée, au plus, sur ses trois dernières années de cotisation. Pour certaines cela peut être une année, ou deux, parmi les trois dernières. De plus, pour de nombreuses entreprises, des incohérences peuvent apparaître, comme une masse salariale ou un effectif nul. Dans d'autres cas, l'entreprise peut avoir cessé ses activités à un moment durant les trois dernières années. Il est alors nécessaire de définir un mécanisme qui puisse garantir la cohérence entre les informations des trois dernières années de cotisation, et entre ces informations et les contrôles effectués. Ce mécanisme doit, en particulier, permettre à l'algorithme de détection une actualisation de ses capacités.

i) Plus précisément, nous disposons d'un premier algorithme qui assure à l'algorithme de détection que les données fournies ne changeront pas de nature d'une année sur l'autre. Par exemple, si la législation change, par l'ajout ou la modification d'une catégorie de cotisation, il faut s'assurer que cela ne perturbe pas le modèle de détection.

ii) La deuxième propriété attendue du traitement des données est consécutive au point précédent. Les capacités d'apprentissage d'un modèle sont très étroitement liées au nombre d'exemples (ici les entreprises contrôlées) disponibles. Plus ce nombre est important, meilleures sont ces capacités. L'algorithme de traitement veille à ce que les données enregistrées de n'importe quelle année de cotisation soient assimilables par l'algorithme de détection à tout moment. Nous illustrons cette propriété dans les paragraphes qui suivent.

`L'apprentissage des données et l'évaluation`

Rappelons d'abord le processus canonique d'apprentissage et d'évaluation : les données et résultats disponibles forment l'échantillon sur lequel on souhaite effectuer l'apprentissage. Cet échantillon est subdivisé, aléatoirement, en trois sous-échantillons de taille différente ou de taille identique.
a) Un algorithme d'apprentissage utilise le sous-échantillon, dit d'entraînement ou d'apprentissage, dans lequel sont présents des exemples (données et résultats) de la relation à *apprendre*. Les données sont, ici, les entreprises et leurs déclarations de cotisation. Et les résultats sont ceux des contrôles.

b) Une fois la relation apprise, la validation de l'algorithme s'effectue en mesurant (et en recalibrant les paramètres du modèle le cas échéant) les erreurs (en particulier l'erreur de prédiction) sur l'échantillon de validation. Notons que l'étape de validation n'est pas nécessaire pour certains algorithmes.
c) Le troisième sous-échantillon, dit de test, sert à valider les erreurs. L'algorithme est testé sur ces données et l'erreur de prédiction mesurée lors de la validation doit être une borne supérieure de l'erreur de prédiction mesurée pour le test (appelée erreur de test).

Une fois ces trois étapes effectuées, l'algorithme doit être capable d'évaluer de nouvelles observations, sans avoir connaissance de leurs résultats, et doit pouvoir prédire l'erreur qui sera commise.
Ce point est fondamental : *la première étape (a) est la seule pour laquelle les résultats sont présentés à l'algorithme afin qu'il "apprenne" leur relation avec les données.*

Illustrons ce processus par un exemple sur les cotisations sociales. Supposons que le modèle doive évaluer le caractère régulier ou irrégulier des déclarations de toutes les entreprises d'Île-de-France afin de déterminer les contrôles à effectuer pour l'année 2013. Deux choix sont possibles :
- Le premier, et le plus commun à de nombreux algorithmes, consiste à fournir au modèle toutes les données relatives aux contrôles et aux déclarations des années 2010, 2011 et 2012 car, au plus, les trois dernières années de cotisation sont contrôlées. Pour des raisons opérationnelles, à la fin de l'année 2012, les déclarations de cette même année ne sont pas disponibles. Elles ne le sont que 6 mois plus tard. Pour évaluer les entreprises, le modèle a donc, à sa disposition, les données et contrôles des années 2010 et 2011 pour l'apprentissage. Une fois ce dernier effectué, il évalue toutes les déclarations de toutes les entreprises de ces deux années-là. La probabilité d'existence d'une irrégularité pour chaque entreprise détermine, avec le montant de redressement estimé et, éventuellement, d'autres critères, l'ordre de priorité des contrôles à effectuer.

Les données (des déclarations) des années précédant 2010 ne peuvent pas être utilisées, puisque les inspecteurs du contrôle ne pourront pas, par obligation légale, les vérifier. Rappelons que l'algorithme de détection évalue le caractère régulier ou irrégulier de toutes les déclarations ainsi que les montants de redressement estimés, relativement aux seules années qui peuvent être contrôlées. Chaque année, il faudrait donc effectuer l'apprentissage des données sans (pouvoir) tenir compte de l'historique. Par exemple, pour les entreprises à évaluer comme recommandations aux contrôles de l'année 2014, seules les données des années 2011, 2012 et 2013 pourraient être utilisées. Et ainsi de suite. Cette situation présente l'inconvénient de ne mettre à disposition qu'un nombre limité d'exemples à *apprendre*, les données associées aux contrôles des trois dernières années, et peut limiter les capacités de généralisation des modèles.

- A la place, nous faisons un second choix. L'apprentissage devient incrémental. De cette manière, l'algorithme acquiert une *mémoire* et *apprend* dans le même temps. Reprenons les exemples précédents et supposons que l'année 2006 soit la première année où des données et des résultats de contrôle sont disponibles pour le modèle.
i) Dans ce cas, nous pouvons "retourner en arrière" et effectuer l'apprentissage à partir

des informations des années 2006 à 2008. Le résultat est mémorisé par l'algorithme. Puis, un nouvel apprentissage est effectué pour les années 2007 à 2009, et *fusionne* avec le premier.
ii) Comme les résultats des contrôles effectués en 2010 sont connus, nous évaluons d'abord les entreprises pour cette année-là. Puis nous comparons les résultats à l'évaluation. Cependant, au moment de cette dernière, nous avons la possibilité de faire appel à la seule mémoire de l'algorithme, au dernier apprentissage effectué, ou à la fusion des deux. Pour le choix de la meilleure option, nous testons les trois et retenons la plus adaptée. Empiriquement, la fusion donne de meilleures performances. Continuons le processus. En 2011, un apprentissage est à nouveau effectué pour les années 2008 à 2010 et nous recommençons les étapes précédentes. Le modèle acquiert alors une mémoire de plus en plus importante et l'apprentissage devient de moins en moins nécessaire. Pour la recommandation aux contrôles de 2013, les exemples et résultats des années 2006 à 2009 sont mémorisés et sont testés contre l'apprentissage effectué par le même modèle, pour les années 2010 et 2011, sur l'échantillon de test des déclarations et résultats des contrôles de ces deux années. Notons qu'au moment de la disponibilité des données et résultats de l'année 2012, la recommandation est mise à jour en intégrant les nouvelles informations. En particulier, dans le modèle incrémental, on peut spécifier, modifier ou supprimer n'importe quelle partie de la mémoire.

L'approche par un modèle incrémental présente essentiellement des avantages face à des modèles plus conventionnels. Le risque principal est celui d'un changement dans la distribution des données. L'algorithme de traitement limite ce risque et justifie en grande partie la mise à l'écart (ou la transformation) de certaines variables. L'algorithme de détection doit également être conçu de façon à être peu sensible aux éventuels changements de paramètres de la distribution des données. Lorsque la distribution, elle-même, change, le problème devient beaucoup plus difficile. Dans ce cas cependant, tous les modèles d'apprentissage statistique en subissent l'effet.

`La matrice de travail et l'échantillon d'apprentissage`

Chaque nouvelle année, la matrice finale évaluée par le modèle est constituée des trois dernières années de cotisation de chaque entreprise. Le nombre d'entreprises est alors d'un peu plus de 300 000 et 1065 variables sont définies. Environ 5% des entreprises sont exclues de la matrice et sont considérés comme des cas non évaluables qui devraient être contrôlés. Environ 10% des entreprises auront été contrôlées sur, au moins, une de leurs trois dernières années de cotisation. Elles constituent la base de l'échantillon d'apprentissage, lequel a, en pratique, une taille définie par l'algorithme de traitement générique.

Nous rappelons la grande asymétrie des redressements effectués : 10% des entreprises contrôlées représentent 90% des redressements effectués. 40% des entreprises redressées, dans le cadre des contrôles comptables d'assiette, le sont pour un montant inférieur à 1000 euros. Ce montant est à rapporter à la masse salariale de l'entreprise, en particulier lorsque trois années de déclaration sont contrôlées.
- Lorsque les sommes redressées sont trop petites, les irrégularités attenantes peuvent être considérées comme un bruit qui perturbe l'apprentissage.

- Afin de filtrer ce bruit, nous définissons le *niveau d'irrégularité* comme le rapport entre le montant d'une irrégularité et la masse salariale des trois dernières années de cotisation. Puis, nous calculons un seuil en dessous duquel nous considérons que le niveau d'irrégularité est nul (si la masse salariale n'est pas trop importante). Ce seuil est déterminé grâce à l'utilisation de variables économiques et nous montrons sa mise en oeuvre dans le quatrième chapitre, consacré à la fraude dans le cadre du travail dissimulé. La disponibilité de variables économiques pour l'ensemble des entreprises n'est pas nécessaire à la détermination du seuil. *Ainsi, une irrégularité existe lorsque la différence entre cotisations vérifiées et cotisations reçues, par l'URSSAF, est positive mais n'est pas trop petite, relativement à la masse salariale de l'entreprise.*

Environ 90% des observations de la matrice ont pour valeur 0. Et 97% des variables sont les catégories de cotisation qui peuvent être virtuellement déclarées par une entreprise quelconque du Régime général de la Sécurité sociale. Comme la plupart des entreprises n'en remplissent qu'un petit nombre, de nombreux zéros (correspondant aux catégories ne nécessitant pas d'être déclarées) apparaissent.

Nous souhaitons également évaluer les montants de redressement lorsque la probabilité d'existence d'une irrégularité dans la déclaration de cotisations sociales de l'entreprise est plus grande que 0.5. Nous constituons alors un deuxième échantillon d'apprentissage identique au premier sauf pour leurs résultats : le caractère régulier ou irrégulier est remplacé par les montants redressés par les inspecteurs du contrôle.

`Aspects numériques`

L'algorithme de traitement générique prend en entrée les bases de données fournies par le service statistiques de l'URSSAF et produit en sortie la matrice de travail. Le processus requiert environ 18h pour l'ensemble des entreprises d'Île-de-France. Environ 50 Go de mémoire virtuelle sont nécessaires au traitement et nous utilisons une station de travail dotée de 30 Go de RAM (*Random Access Memory*) ou mémoire vive et d'un processeur à 3.8 Ghz (Nehalem) doté de 4 *coeurs* (*cores*) physiques et de l'*hyperthreading*. L'ensemble des tâches est réalisé avec le logiciel libre de calcul statistique *R* (www.r-project.org), dans sa version 64 bits.
Notons que la constitution et l'actualisation de la matrice de travail est faite, au plus, deux fois par an et le seul impératif est celui d'un mémoire vive en quantité suffisante. Cette contrainte peut cependant être partiellement levée (et le temps de traitement réduit) en divisant les tâches (et en optimisant l'algorithme de traitement).

`Schéma général`

L'apprentissage statistique a l'avantage de présenter un étroit lien entre la théorie, les algorithmes mis en place et leur exécution opérationnelle. En particulier, nous pouvons définir des modèles avec très peu d'hypothèses sur la nature de la relation entre les données et le phénomène analysé. Dans le cas des cotisations sociales, nous postulons simplement l'existence d'une relation entre les cotisations versées par les entreprises et la présence ou l'absence d'irrégularités lorsqu'une partie de ces entreprises est contrôlée pour cela.

Nous postulons également que la seule connaissance des résultats des contrôles et des déclarations de cotisation associées, chaque année, suffit à généraliser notre connaissance de la relation à toutes les entreprises, avec une grande probabilité de succès. Plusieurs difficultés apparaissent immédiatement :
- comment fournir des garanties de l'approche proposée ?
- Qu'apporte-t-elle de plus, alors que de nombreux efforts sont menés chaque année ?
- Peut-elle être mise en place sans déployer des moyens extraordinaires ?

La réponse à la première question est donnée par l'apprentissage statistique. Un de ses objets d'étude est, précisément, d'établir un cadre théorique et opérationnel dans lequel nous pouvons fournir des garanties avant que les modèles ne soient industrialisés. Nous consacrons le troisième chapitre au modèle et à ses propriétés ; dans le dernier chapitre, nous discutons des résultats.
La réponse à la deuxième question est la conséquence de la réalité opérationnelle. La tendance à l'augmentation des montants redressés est due à un ciblage plus important des entreprises pour lesquelles les enjeux financiers sont importants. Elles constituent, au plus, 10% de toutes les entreprises. Pour les autres, il existe une difficulté importante à identifier les irrégularités car l'exploitation des données requiert un niveau de traitement qu'il n'était pas possible de réaliser il y a encore quelques années. Plus précisément, l'augmentation de la puissance de calcul et les innovations algorithmiques permettent, aujourd'hui, d'effectuer des opérations de plus en plus complexes. Notons qu'ici, ces opérations ne sont que l'extension des efforts menés par les inspecteurs du contrôle.
Le troisième point résulte des deux premiers et constitue le point d'entrée du modèle proposé : *étendre la détection des irrégularités aux cotisations sociales à la totalité des entreprises a comme corollaire le développement d'un modèle théorique sous-jacent simple, souple, peu coûteux, capable de fournir des garanties et d'être évalué rapidement.*

Nous illustrons ci-dessous la manière dont est conçu le processus. Trois arguments le caractérisent :
- un modèle pour les données et un modèle pour l'inférence ;
- l'utilisation d'un grand nombre de modèles de base pour l'inférence (classifieurs), simples ;
- l'apprentissage incrémental des exemples.

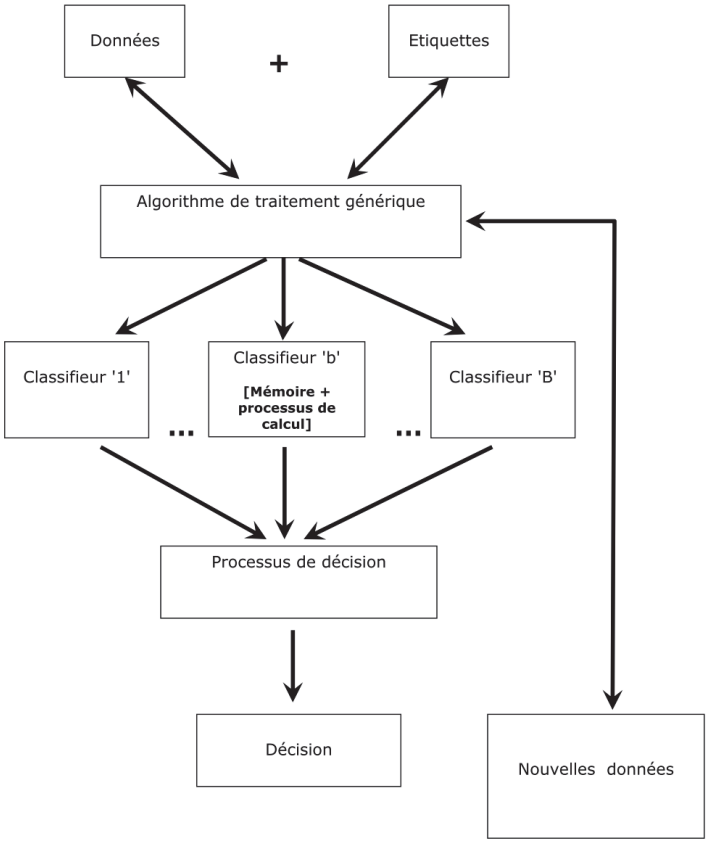

FIGURE 2.1 – Schéma général de la détection des irrégularités aux cotisations sociales.

Le graphique illustre la nature du processus que nous avons suivi. Les *données* sont les déclarations de cotisations, telles qu'elles figurent dans les bases de données de l'URSSAF, et les étiquettes sont les résultats des contrôles effectués. Ces deux éléments sont transformés par *l'algorithme de traitement générique* en une matrice de travail pour chacun des B *classifieurs* de base. Les classifieurs sont, chacun, constitués d'une *mémoire* et d'une règle de décision simples. Ils ont la particularité d'être très différents les uns des autres. Ils *apprennent* la relation entre les données et les étiquettes, puis transmettent, chacun, leur décision à un *processus* qui les organise pour produire la *décision complète*.
- Lorsque de *nouvelles données* arrivent sans étiquettes, l'algorithme de traitement générique les transforme à nouveau et les classifieurs de base prennent chacun une décision grâce à la mémoire des apprentissages précédents.
- Lorsque de *nouvelles données* arrivent avec des étiquettes, un nouvel apprentissage est effectué. Le processus de décision organise la manière dont la décision finale est prise ; elle est le résultat de la mémoire, de l'apprentissage courant ou d'une fusion des deux.

Dans le chapitre qui suit, nous définissons formellement les classifieurs de base et le processus de décision et montrons leurs propriétés théoriques.

Chapitre 3

Forêts uniformément aléatoires

Dans le cadre des outils d'apprentissage statistique fondés sur une agrégation de modèles de base, nous présentons les *forêts uniformément aléatoires*, une variante de l'algorithme de référence *Random Forests* (Breiman, 2001). Pour la classification, nous montrons sa convergence vers l'erreur de Bayes, en nous aidant des travaux de Devroye, Györfi et Lugosi (1996), puis de ceux de Biau, Devroye et Lugosi (2008). Les forêts uniformément aléatoires héritent des mêmes propriétés théoriques que les forêts aléatoires de Breiman. Elles s'en différencient par trois aspects : le tirage, avec remise, des variables candidates pour la construction de chaque région d'un arbre de décision (le modèle de base) ; le tirage des points de coupure selon la loi uniforme sur le support de chaque variable candidate dans la partition courante ; et le sous-échantillonnage des observations (à la place du *bootstrap*) dans le cas de la régression. Dans la pratique, le caractère incrémental, l'extrapolation ou encore la sélection locale de variables font partie des nouveaux outils intégrés dans le modèle et en étendent les applications. L'algorithme est disponible sous la forme d'un paquet R (*randomUniformForest*).

3.1 Introduction

Les forêts aléatoires font partie des modèles d'agrégation utilisées dans le cadre de l'apprentissage statistique. Leur idée fondatrice peut se résumer en deux points :
- la constitution d'un nombre important de modèles sous-jacents sur tout ou partie de l'échantillon d'apprentissage
- l'établissement d'une règle de décision basée sur le vote majoritaire (ou une version équivalente) des règles de décision de ces modèles.

Dans le cas des forêts aléatoires, les modèles sous-jacents sont des arbres de décision et la règle de décision est un vote à la majorité (classification), ou une moyenne (pondérée ou non) dans le cas de la régression, des résultats de toutes les règles de décision des modèles sous-jacents. Définir une forêt aléatoire revient à construire un arbre et sa règle de décision, puis à définir la règle de décision pour un nombre important d'arbres. De nombreuses variantes de forêts aléatoires existent dont, cependant, peu ont d'aussi bonnes performances opérationnelles que les forêts aléatoires de Breiman. Une motivation fondamentale de cet article est donc de proposer une variante dotée des mêmes propriétés, moins coûteuse en temps de calcul et plus simple à analyser du point de vue théorique.

Parmi les précurseurs immédiats des forêts aléatoires, figurent le *Bagging* (Breiman, 1996), pour l'aspect ensembliste, et la méthode *random subspaces* (Ho, 1998) qui introduit la sélection aléatoire de variables dans la construction de chaque arbre. Une particularité des forêts aléatoires, qu'il convient de souligner, est leur faible sensibilité à l'augmentation de la dimension du problème. Cet aspect les destine particulièrement à tous les phénomènes pour lesquels de nombreuses variables sont disponibles.

Nous commençons par introduire les arbres de décision d'une manière générale, à des fins de plus grande lisibilité. Nous nous intéressons uniquement aux arbres de décision binaires. Pour une revue claire de ces derniers, nous renvoyons à la thèse de Genuer (2010). Dans un second temps, nous définissons les *arbres de décision uniformément aléatoires*, pour lesquels nous montrons leur proximité, et leur consistance, avec le classifieur de Devroye, Györfi et Lugosi (1996, théorème 20.9) dont nous nous inspirons.
Puis, une troisième partie est consacrée à la consistance des *forêts uniformément aléatoires*, lesquelles étendent les propriétés des arbres de même nom, et nous y indiquons les différentes caractéristiques utiles à leur compréhension. Ces propriétés permettent de définir un algorithme dont nous proposons un paquet R (*randomUniformForest*) et que nous détaillons en partie. L'algorithme est introduit volontairement à la fin du document afin de simplifier la présentation. En effet, même s'il dérive de l'approche théorique, quelques optimisations sont nécessaires à la tenue de performances similaires à la version de référence. De plus, la multiplicité de possibilités permises par les forêts aléatoires, de la sélection de variables à l'apprentissage non supervisé, ainsi que les alternatives à la validation croisée ou le sous/sur échantillonnage, nécessite un traitement particulier. Nous décrivons un nouveau paradigme de ces possibilités en indiquant comment, avec peu d'efforts, construire des forêts uniformément aléatoires et les agréger. Notre point de vue traite essentiellement de la classification ; néanmoins dans la dernière partie, nous incluons plusieurs éléments importants dans le cadre de la régression. Tout au long du document, nous ferons référence aux forêts aléatoires de Breiman en essayant d'exposer clairement les changements apportés.

Nous définissons, ci-après, la notation utilisée.
La probabilité $\mathbb{P}$ est notée $\mathbf{P}$, l'espérance mathématique $\mathbb{E}$ est notée $\mathbf{E}$, la variance $\mathbb{V}$ est notée $\mathbf{Var}$, la covariance est notée $\mathbf{Cov}$. Dans le cadre de la classification, nous nous plaçons dans un problème à deux classes.
Une observation est définie comme un vecteur x de dimension d, où d est le nombre de variables du problème. A chaque observation est associée y, à valeurs discrètes, dans le cadre de la classification, et continues dans le cadre de la régression. Nous supposons qu'un classifieur possède toujours une règle de décision et peut se réduire à cette dernière. Une règle de décision n'est donc pas nécessairement un classifieur.
Un classifieur est défini comme une fonction $g(x) : \mathbb{R}^d \to \{0, 1\}$.
A chaque observation x, le classifieur associe la valeur $g(x)$ et le risque d'erreur est donné par $\mathbf{I}_{\{g(x)\neq y\}}$, où $\mathbf{I}$ désigne la *fonction indicatrice* qui vaut 1 si $g(x) \neq y$, 0 sinon.
Nous considérons donc un couple de variables aléatoires (X, Y) à valeurs dans $\mathbb{R}^d \times \{0, 1\}$, auquel on associe le classifieur $g(X)$ et la probabilité d'erreur $L(g)$, donnée par :

$$L = L(g) = \mathbf{E}\left\{\mathbf{I}_{\{g(X)\neq Y\}}\right\} = \mathbf{P}\left\{g(X) \neq Y\right\}.$$

Nous définissons également le meilleur classifieur parmi tous ceux possibles :

$$g^* = \underset{g:\mathcal{R}^d \to \{0,1\}}{\texttt{arg min}} \mathbf{P}\left\{g(X) \neq Y\right\}.$$

g^* est appelé le classifieur de Bayes et la probabilité d'erreur minimale (l'erreur de Bayes) est notée $L^* = L(g^*)$.
Si la distribution de (X, Y) est connue, alors g peut être calculée explicitement.
Généralement ce n'est pas le cas et, à la place, nous disposons d'un échantillon D_n avec :

$$D_n = \{(X_i, Y_i), 1 \leq i \leq n\},$$

où

$$X_i = (X_i^{(1)}, ..., X_i^{(j)}, ..., X_i^{(d)}),$$

est le vecteur des variables du problème pour l'observation i.

g est estimé par g_n défini par :

$$g_n = g_n(X, D_n).$$

La probabilité d'erreur conditionnelle est donnée par :

$$L_n = L(g_n) = \mathbf{P}\left\{g_n(X, D_n) \neq Y | D_n\right\}.$$

Un classifieur a la propriété de consistance si :

$$\lim_{n \to \infty} \mathbf{E}(L_n) = L^*.$$

La consistance est dite forte si :

$$L_n \overset{p.s.}{\to} L^*, \ quand \ n \to \infty.$$

Si un classifieur est consistant pour toutes les distributions de (X, Y) alors il est dit universellement consistant. La recherche de classifieurs consistants revient à déterminer des mesures du type :

$$\mathbf{P}\left\{L_n \geq L^* + \epsilon\right\},$$

que l'on cherche à majorer, avec $\epsilon \to 0$, quand $n \to \infty$.

En contrôlant ϵ, on peut alors déterminer les conditions dans lesquelles la consistance peut être obtenue. Cependant, la connaissance que nous avons de (X, Y) est limitée à l'échantillon D_n. Un estimateur de L_n est alors donné par sa contrepartie empirique, $\widehat{L}_n$, définie par :

$$\widehat{L}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{\{g_n(X_i, D_n) \neq Y_i\}}.$$

La consistance est la condition suffisante à la convergence de l'erreur du classifieur vers l'erreur la plus petite possible. Elle ne donne pas la vitesse de convergence de cette erreur, généralement déterminée par une borne de risque. Plus la vitesse de convergence est importante, plus vite un classifieur consistant atteint l'erreur de Bayes. En pratique, la vitesse de convergence a un impact immédiat sur le nombre de données nécessaire à la convergence vers l'erreur de Bayes.

3.2 Arbre de décision

3.2.1 Définition

Un arbre de décision est une structure algorithmique qui partitionne de manière récursive l'espace formé par les variables X. A chaque étape (noeud) du partitionnement, l'espace est divisé en plusieurs régions. On procède ainsi pour chaque nouvelle région créée, jusqu'à ce qu'un ou plusieurs critères d'arrêt soient atteints. Une région est définie comme une sous-partition dont la construction repose sur des règles spécifiques à l'arbre de décision. Une région qui ne peut plus être partitionnée est une feuille de l'arbre. La règle de décision de l'arbre attribue une unique classe à chaque feuille.

Chaque méthode de construction d'arbres propose plusieurs arguments d'optimisation, à chaque noeud, qui la différencient d'une autre. Le processus général est cependant très similaire à celui décrit ci-dessus. Nous nous intéressons plus particulièrement aux arbres de type CART, *Classification And Regression Trees* (Breiman, Friedman, Olshen et Stone, 1984) pour lesquels nous introduisons une version spécifique. La construction dyadique des régions est son point commun avec CART. Pour une revue de ces derniers, nous renvoyons également au chapitre 2 de la thèse de Gey (2002). Indépendamment de la règle de décision ou de l'agrégation des arbres, le principal mécanisme mis en oeuvre dans la construction d'un arbre est celui qui définit le partitionnement. Dans le cas du *Bagging* (Breiman, 1996), l'arbre de décision explore, à chaque étape du partitionnement, chaque variable et chaque observation afin de définir la région optimale. Le critère d'optimisation est local : on cherche d'abord pour chaque variable, l'observation optimale (le point de coupure) relativement au critère. Puis, parmi les variables candidates, on choisit celle dont la valeur pour le critère d'optimisation est la meilleure. Cette variable définit (avec l'observation correspondante) la région optimale parmi toutes celles évaluées. Ce mécanisme a lieu en tirant, avec remise, n points de D_n au début de la construction de l'arbre. Pour un résumé précis des différentes approches de construction de l'arbre de décision dans le cadre des méthodes ensemblistes, nous renvoyons à Lin et Jeon (2002). D'un point de vue théorique, la structure du classifieur et la définition de sa règle de décision sont les éléments essentiels de la recherche de consistance. On peut formaliser en une même approche cette recherche et la caractérisation du classifieur :
1- la construction et la modification d'une région ;
2- la définition des conditions d'arrêt ;
3- la définition d'une règle de décision.

3.2.2 Arbre de décision uniformément aléatoire

Nous définissons, comme dans CART, une structure qui partitionne récursivement les données en deux régions. A la différence de CART, le critère de sélection d'une région est appliqué sur des régions totalement aléatoires. Le partitionnement est poursuivi jusqu'à ce qu'un critère d'arrêt soit atteint. A chaque étape, seuls le nombre de régions et le choix de la région aléatoire optimale pour l'étape suivante sont importants. La règle de décision n'intervient que dans les régions terminales (celles pour lesquelles une condition

d'arrêt est atteinte). Elle se réduit à la classe majoritaire parmi les observations de la région évaluée.

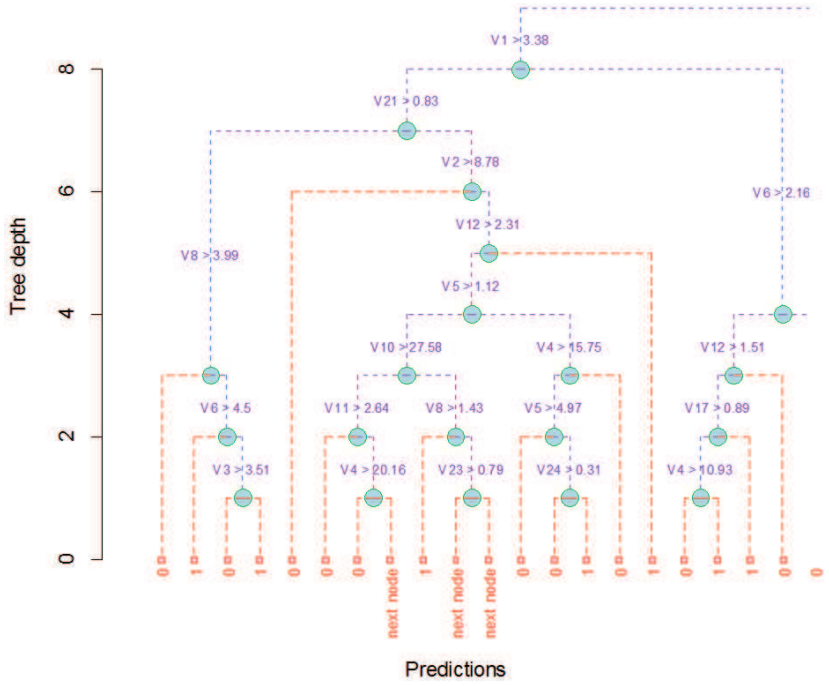

FIGURE 3.1 – Une partie d'un arbre de décision uniformément aléatoire.

Le graphique ci-dessus représente (en partie) un arbre de décision uniformément aléatoire. A chaque étape (les noeuds en bleu), on choisit la meilleure variable (V1, $V2$,...) qui sépare la partition courante en deux sous-partitions disjointes aléatoires. Par exemple, si $'V1 > 3.38'$ (premier noeud) alors toutes les observations au-dessus de la valeur 3.38, pour la variable $V1$, vont à droite. Sinon elles vont à gauche. L'indice de chaque observation de la variable candidate (ici $V1$) détermine le positionnement (à droite ou à gauche) de toute la partition. Les segments en rouge correspondent à des noeuds terminaux, ceux où la règle de décision s'applique. Ici, chaque région terminale a pour étiquette 0 ou 1. Les noeuds indiqués par "next node", mènent à d'autres parties de l'arbre.

Une définition possible d'un arbre de décision uniformément aléatoire est celle-ci :
un arbre de décision uniformément aléatoire est un arbre de décision binaire dans lequel les régions sont construites en choisissant les points de coupure de manière aléatoire. A chaque étape du partitionnement $\lceil \beta d \rceil$ variables, $\beta \geq 1/d$, sont tirées avec remise. Pour chaque variable candidate, un point de coupure, α, est tiré selon la loi Uniforme sur le support de la variable. La région optimale est celle qui maximise un critère d'entropie calculé pour chacune des $2\lceil \beta d \rceil$ régions aléatoires candidates. La poursuite du partitionnement dépend du critère, lors de la sélection d'une région et de sa région complémentaire, et de l'atteinte de conditions d'arrêt. La règle de décision de l'arbre n'est appliquée que pour les régions terminales.

Notons que, du point de vue théorique, la règle de décision doit être utilisée dans la construction de chaque région. Nous précisons cet élément dans la section qui suit. Nous formalisons les étapes de la construction ci-après. L'espace de départ est $\mathbb{R}^d$, où d est la dimension du problème, et nous définissons une région A ainsi :
A est une région de la partition $\mathcal{P}$ si, pour tout $B \in \mathcal{P}$, $A \cap B = \emptyset$ ou $A \subseteq B$.

Une région contient au plus n observations et au minimum une seule. Chaque partition possède exactement deux régions, la région A et sa région complémentaire A^C. L'arbre de décision construit $(k+1)$ régions (dont la première est la partition de départ et existe toujours) auxquelles appartiennent les observations. Le choix du nombre minimal d'observations a une conséquence théorique importante car la plupart des théorèmes de consistance sur les arbres de décision lient la convergence vers l'erreur de Bayes au rapport $\frac{k}{n}$ qui doit tendre vers 0, lorsque $k \to \infty$, quand $n \to \infty$. En quelque sorte, pour établir la consistance d'un arbre de décision, une condition nécessaire est, généralement, que le nombre d'observations dans les régions terminales ne soit pas trop petit. Par exemple, on peut spécifier que chaque région terminale contienne au moins $\left\lfloor \sqrt{n\mathtt{log}(n)} \right\rfloor$ observations. Dans la pratique, le nombre de régions dépend essentiellement des critères d'arrêt dans la construction de l'arbre et de la distribution du couple (X, Y). Pour des performances opérationnelles optimales, un arbre de décision uniformément aléatoire n'impose pas de limite au nombre de régions. Nous dérivons, à ce stade, deux variantes importantes qui distinguent un arbre uniformément aléatoire des arbres de type CART.

a) Dans CART, chaque région A est caractérisée par sa frontière : la variable $X^{(j)}$, $1 \leq j \leq d$, et le point de coupure x de sorte que pour tout $A \in \mathcal{P}$, $\left\{X^{(j)} \leq x_i\right\}$ avec i et j fixés. De même, pour tout $A^C \in \mathcal{P}$, $\left\{X^{(j)} > x_i\right\}$, où A^C est la région complémentaire de A. Supposons $A = \{A_1, ..., A_l, ..., A_k\}$, la suite de k régions construites par l'arbre de décision. Les régions $\{A_l, 1 \leq l \leq k\}$ sont alors des hyper-rectangles dont chaque frontière est définie par le choix des indices i et j les plus adaptés parmi les n^d possibilités de départ, ce nombre se réduisant chaque fois qu'une région terminale est atteinte. Dans le cas des forêts aléatoires de Breiman, pour chacune des régions, v variables, parmi d, sont tirées sans remise puis, pour chacune d'elles, on recherche l'observation x_i qui minimise un certain critère $\{\mathtt{G}(i, u, D_{n_l}), 1 \leq i \leq n_l, 1 \leq l \leq k, j_1 \leq u \leq j_v\}$, où n_l est le nombre d'observations de la l-ème région. La fonction $\mathtt{G}$ est le *critère d'impureté de Gini*.
Afin de simplifier, on pose $n \overset{def}{=} n_l$, le nombre d'observations de la partition courante.

Soit $n^{'} = \sum_{q=1}^{n} \mathbf{I}_{\left\{X_q^{(u)} \leq x_i\right\}}$,
$\mathtt{G}$ est définie, conditionnellement à D_n et pour i et u fixés, par :

$$\mathtt{G}(i, u, D_n) = \frac{n^{'}}{n} \sum_{c=0}^{1} \left\{ \frac{1}{n^{'}} \sum_{q=1}^{n} \mathbf{I}_{\{Y_q = c\}} \mathbf{I}_{\left\{X_q^{(u)} \leq x_i\right\}} \left(1 - \frac{1}{n^{'}} \sum_{q=1}^{n} \mathbf{I}_{\{Y_q = c\}} \mathbf{I}_{\left\{X_q^{(u)} \leq x_i\right\}}\right) \right\}$$
$$+ \frac{n - n^{'}}{n} \sum_{c=0}^{1} \left\{ \frac{1}{n - n^{'}} \sum_{q=1}^{n} \mathbf{I}_{\{Y_q = c\}} \mathbf{I}_{\left\{X_q^{(u)} > x_i\right\}} \left(1 - \frac{1}{n - n^{'}} \sum_{q=1}^{n} \mathbf{I}_{\{Y_q = c\}} \mathbf{I}_{\left\{X_q^{(u)} > x_i\right\}}\right) \right\}.$$

A est une région optimale si :

$$\textit{pour tout } A \in \mathcal{P}, \left\{ X_i^{(u^*)} \leq x_{i^*} | D_n \right\},\ j_1 \leq u \leq j_v, 1 \leq i \leq n,$$
$$\textit{pour tout } A^C \in \mathcal{P}, \left\{ X_i^{(u^*)} > x_{i^*} | D_n \right\},\ j_1 \leq u \leq j_v, 1 \leq i \leq n,$$

où (i^*, u^*) est obtenu en deux temps. Notons i_u, l'indice de la i-ème observation associé à la variable $X^{(u)}$:
- on cherche d'abord l'ensemble des indices optimaux $\{i_u^*, j_1 \leq u \leq j_v\}$ pour tous les u, donné par $\underset{i_u}{\texttt{arg min}}\ \texttt{G}(i_u, u, D_n)$;
- puis, on détermine u^*, l'indice de la variable optimale parmi les v variables candidates, donné par $\underset{u}{\texttt{arg min}}\ \texttt{G}(i_u^*, u, D_n)$.

Dans la version de référence des forêts aléatoires, le classifieur et sa règle de décision dépendent fortement des observations puisqu'elles sont toutes évaluées (pour les v variables tirées à chaque étape) afin de définir au mieux la région optimale. Le caractère aléatoire de l'arbre est exprimé en deux phases :
- d'abord au moment du tirage, avec remise, de n observations de D_n à chaque fois qu'un arbre est construit ;
- puis, au moment du tirage des v variables dans la construction de chaque région.
Par défaut dans l'algorithme *randomForest*, $v = \left\lfloor \sqrt{d} \right\rfloor$ pour la classification. *Le choix qui est fait est celui d'une région optimale parmi plusieurs régions aléatoires optimales.*

Dans le cas d'un arbre de décision uniformément aléatoire, le point de coupure est aléatoire et on tire, avec remise, $\lceil \beta d \rceil$ variables, $\beta \geq 1/d$. *Le choix qui est fait est celui d'une région aléatoire optimale parmi plusieurs régions aléatoires.* Si l'arbre est trop, ou totalement, aléatoire, ses performances seront limitées en pratique. S'il est trop optimisé, il n'a que peu d'intérêt dans une forêt aléatoire.

b) Supposons par la suite que $\beta > 1$. Une région A, d'un arbre de décision uniformément aléatoire, est optimale si :

$$\textit{pour tout } A \in \mathcal{P}, \left\{ X_i^{(j^*)} \leq \alpha_{j^*} | D_n \right\},\ 1 \leq j \leq d, 1 \leq i \leq n,$$
$$\textit{pour tout } A^C \in \mathcal{P}, \left\{ X_i^{(j^*)} > \alpha_{j^*} | D_n \right\},\ 1 \leq j \leq d, 1 \leq i \leq n,$$

où $\alpha_j \sim \mathcal{U}\left(\texttt{min}(X^{(j)}|D_n), \texttt{max}(X^{(j)}|D_n)\right)$ et $j^* = \underset{j \in \{1,\ldots,d\}}{\texttt{arg max}}\ \texttt{IG}(j, D_n)$.

Le critère à maximiser, `IG`, défini plus bas, est différent de `G` et procède d'une optimisation globale, au niveau de l'ensemble des variables candidates. Le point de coupure suit la loi Uniforme ; il doit simplement être issu d'une fonction aléatoire de X et être indépendant de Y. Quatre conditions mettent fin au partitionnement :
1 - le nombre minimal d'observations est atteint ;
2 - les observations de la région A_l ont toutes la même étiquette ;
3 - les observations de la région A_l sont toutes identiques pour X ;
4 - la fonction `IG` a pour maximum un seuil, supérieur ou égal à 0.

Les régions éligibles à un de ces quatre critères sont donc terminales et la règle de décision de l'arbre leur est appliquée.

Il nous faut définir la fonction `IG`, qui joue un rôle équivalent à $\widehat{L}_n$, l'erreur empirique. Pour cela, nous utilisons des outils de la théorie de l'information, en particulier l'entropie de Shannon (1948). `IG` est le gain d'information apporté par la connaissance de Y et de $Y|X$, et est défini comme suit :

$$\texttt{IG}(Y, X) = \texttt{H}(Y) - [\texttt{H}(Y|X \leq \alpha) + \texttt{H}(Y|X > \alpha)].$$

A chaque noeud de l'arbre, on écrit `IG` conditionnellement à D_n, ce qui donne :

$$\texttt{IG}(j, D_n) = \texttt{H}(Y|D_n) - \left[\texttt{H}\left(\left(Y|X^{(j)} \leq \alpha_j\right)|D_n\right) + \texttt{H}\left(\left(Y|X^{(j)} > \alpha_j\right)|D_n\right)\right].$$

$\texttt{H}(Y)$ est l'entropie de Shannon c'est-à-dire la quantité d'information introduite par la connaissance de Y. Elle est définie avec le logarithme de base 2, mais nous lui préférons le logarithme népérien :

$$\texttt{H}(Y) = \mathbf{E}\left[-\texttt{log}(\mathbf{P}(Y = y))\right].$$

Conditionnellement à D_n on a :

$$\texttt{H}(Y|D_n) = -\sum_{c=0}^{1}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{\{Y_i=c\}}\ \texttt{log}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{\{Y_i=c\}}\right)\right\},$$

et on impose $0\ \texttt{log}\ 0 = 0$, de sorte que $\texttt{H}(Y) \geq 0$.

Soit $n' = \sum_{i=1}^{n}\mathbf{I}_{\left\{X_i^{(j)} \leq \alpha_j\right\}}$,

$$\texttt{H}\left(\left(Y|X^{(j)} \leq \alpha_j\right)|D_n\right) = -\frac{n'}{n}\sum_{c=0}^{1}\left\{\frac{1}{n'}\sum_{i=1}^{n}\mathbf{I}_{\{Y_i=c\}}\mathbf{I}_{\left\{X_i^{(j)} \leq \alpha_j\right\}}\ \texttt{log}\left(\frac{1}{n'}\sum_{i=1}^{n}\mathbf{I}_{\{Y_i=c\}}\mathbf{I}_{\left\{X_i^{(j)} \leq \alpha_j\right\}}\right)\right\},$$

$$\texttt{H}\left(\left(Y|X^{(j)} > \alpha_j\right)|D_n\right) =$$
$$-\frac{n-n'}{n}\sum_{c=0}^{1}\left\{\frac{1}{n-n'}\sum_{i=1}^{n}\mathbf{I}_{\{Y_i=c\}}\mathbf{I}_{\left\{X_i^{(j)} > \alpha_j\right\}}\ \texttt{log}\left(\frac{1}{n-n'}\sum_{i=1}^{n}\mathbf{I}_{\{Y_i=c\}}\mathbf{I}_{\left\{X_i^{(j)} > \alpha_j\right\}}\right)\right\}.$$

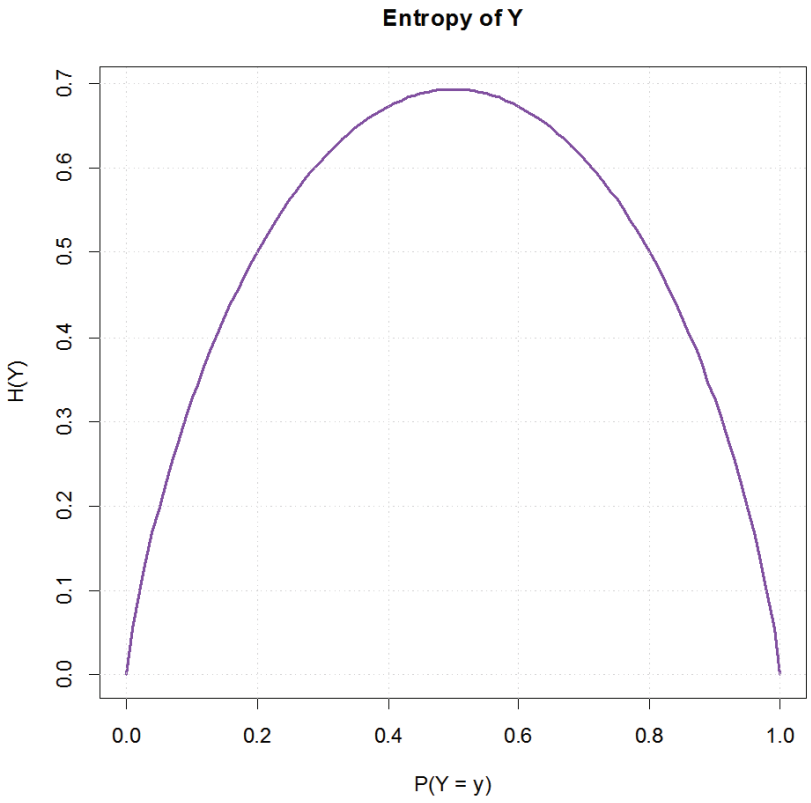

FIGURE 3.2 – La fonction d'entropie de Y pour le logarithme népérien.

La fonction d'entropie ci-dessus illustre le niveau d'incertitude que l'on a de Y. Pour l'ensemble des évènements de Y, l'incertitude est maximale lorsque leurs probabilités sont équi-réparties. Un point de vue équivalent consiste à dire que l'entropie est la valeur de l'incertitude qui résulte de la connaissance de Y. Si l'entropie est nulle, alors il n'y a plus d'incertitude. La fonction `IG` est, ainsi, la différence entre deux niveaux d'incertitude. Plus elle est grande, moins on a d'informations sur la variable observée et plus il est nécessaire de la conserver pour acquérir de l'information supplémentaire. Lorsque $\texttt{IG}(j, D_n), 1 \leq j \leq d$, est maximale, la variable $X^{(j^*)}$ est celle qui apporte le plus de gain d'information dans la définition de A.

Parmi les variantes de forêts utilisées en pratique, citons les *Extremely Randomized Trees* (Geurts et al., 2006), nommées également *Extra-Trees*, dont le choix du point de coupure repose sur un tirage uniforme d'une observation parmi celles considérées, et le choix de la variable à couper, sur un tirage, sans remise, de v variables parmi d, comme dans les forêts aléatoires de Breiman. Dans les Extra-Trees, il n'y a pas de modification de D_n pour la construction d'un arbre et le critère d'optimisation (*gain information ratio*) est plus complexe que le gain d'information `IG`.

3.2.3 Consistance

Nous avons défini la consistance comme la capacité du classifieur à converger vers l'erreur de Bayes, soit l'erreur qui rend le classifieur optimal. Un classifieur consistant ne garantit pas une erreur de prédiction faible. Cependant, s'il est optimal au sens de l'erreur de Bayes, alors on ne peut faire mieux que ce classifieur. Généralement, pour deux classifieurs consistants, la mesure de leurs performances s'effectue en comparant leurs vitesses de convergence, c'est-à-dire le nombre d'observations à partir duquel le classifieur tend à devenir optimal. Du point de vue théorique, la consistance est établie lorsque n tend vers l'infini, et dans la pratique, lorsque le nombre d'observations est assez important.

Notons R, la région associée à la partition $\mathcal{P}$. Pour un arbre uniformément aléatoire et pour une région $R \in \mathcal{P}$, de $\mathbb{R}^d$, la règle de décision est définie par :

$$g_{\mathcal{P}}(x, R) = g_{\mathcal{P}}(x) = \begin{cases} 1, \ si \ \sum_{i:X_i \in R} Y_i > \sum_{i:X_i \in R} (1 - Y_i), \ x \in R \\ \\ 0, \ sinon. \end{cases}$$

La règle de décision établit que la classe attribuée à une observation est celle qui correspond au vote majoritaire parmi les instances de Y dans la région R, laquelle appartient à la partition courante $\mathcal{P}$.
On note $L(R)$ l'espérance du risque d'erreur, définie par :

$$\begin{aligned} L(R) &= \mathbf{E}\left(\mathbf{I}_{\{X \in R, Y \neq y\}}\right) \\ &= \mathbf{P}(X \in R, Y \neq y) \\ &= \mathbf{I}_{\{Y=0\}}\mathbf{P}(X \in R, Y = 1) + \mathbf{I}_{\{Y=1\}}\mathbf{P}(X \in R, Y = 0) \\ &= \mathtt{min}(\nu_0(R), \nu_1(R)), \end{aligned}$$

où

$$\nu_j(R) = \mathbf{P}(X \in R, Y = j), j \in \{0, 1\}.$$

$L_n(R)$, l'espérance de l'erreur commise par le classifieur, est définie par :

$$\begin{aligned} L_n(R) &= \mathbf{E}\left(\mathbf{I}_{\left\{X \in R, \ g_{\mathcal{P}}(X) \neq Y\right\}}\right) \\ &= \mathbf{P}(Y = 1, g_{\mathcal{P}}(X) = 0, X \in R) + \mathbf{P}(Y = 0, g_{\mathcal{P}}(X) = 1, X \in R) \\ &= \mathbf{I}_{\left\{g_{\mathcal{P}}(X)=0\right\}}\mathbf{P}(X \in R, Y = 1) + \mathbf{I}_{\left\{g_{\mathcal{P}}(X)=1\right\}}\mathbf{P}(X \in R, Y = 0) \\ &= \mathtt{min}(\nu_0(R), \nu_1(R)), \end{aligned}$$

et la contrepartie empirique de $L_n(R)$, notée $\widehat{L}_n(R)$, est donnée par :

$$\begin{aligned} \widehat{L}_n(R) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\left\{X_i \in R, \ g_{\mathcal{P}}(X_i) \neq Y_i\right\}} \\ &= \mathtt{min}(\nu_{0,n}(R), \nu_{1,n}(R)), \end{aligned}$$

où

$$\nu_{j,n}(R) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{X_i \in R, Y_i = j\}}, j \in \{0, 1\}.$$

De même, nous notons $L(\mathcal{P})$, l'erreur sur la partition $\mathcal{P}$, donnée par :

$$L(\mathcal{P}) \stackrel{def}{=} \sum_{R \in \mathcal{P}} L(R),$$

l'erreur commise par le classifieur sur la partition, $L_n(\mathcal{P})$, définie par :

$$L_n(\mathcal{P}) \stackrel{def}{=} \sum_{R \in \mathcal{P}} L_n(R),$$

et sa contrepartie empirique $\widehat{L}_n(\mathcal{P})$, l'erreur commise en considérant $g_{\mathcal{P}}$ sur toutes les régions de $\mathcal{P}$. Naturellement, on souhaite minimiser $\widehat{L}_n(\mathcal{P})$ en minimisant $\sum_{R \in \mathcal{P}} \widehat{L}_n(R)$. Cependant, seules deux régions sont admissibles, à tout instant, lors de l'établissement de la règle de décision. Soit $T \subset R$, une région telle que pour tout $T \in \mathcal{P}$, $\left\{X^{(j)} \leq \alpha_j\right\}$, et $R - T$, la région complémentaire de T. Choisir la région T optimale, telle que $R - T$ soit sa région complémentaire, revient alors à minimiser la variation entre la somme des erreurs empiriques moyennes, $\widehat{L}_n(T) + \widehat{L}_n(R - T)$, issues du partitionnement et l'erreur empirique moyenne de la région courante, $\widehat{L}_n(R)$. Ainsi, T est définie par :

$$\underset{T}{\texttt{arg min}}\ \widehat{L}_n(T) + \widehat{L}_n(R - T) - \widehat{L}_n(R). \tag{3.1}$$

Dans ce cas, il suffit simplement de s'assurer que $\widehat{L}_n(R)$ est optimale pour assurer l'équivalence avec la minimisation de $\widehat{L}_n(\mathcal{P})$, ce qui est le cas dès que la règle de décision est appliquée à la partition de départ, $\mathbb{R}^d$.

`Le classifieur de Devroye, Györfi et Lugosi.`

Devroye et al. (1996) définissent un tel classifieur, associé à $g_{\mathcal{P}}$, et à L, et établissent sa consistance. Ils le nomment *greedy tree classifier* et nous le désignons par la suite *arbre de décision heuristique*. Un tel classifieur considère toutes les observations et toutes les variables à chaque étape de la construction de l'arbre, puis choisit la région qui minimise (3.1). C'est la seule hypothèse nécessaire à l'arbre de décision heuristique. En contrepartie, la complexité est en $O(n^d)$ puisqu'il faut explorer toutes les observations de toutes les variables pour construire une région.

Un arbre de décision uniformément aléatoire diffère d'un arbre de décision heuristique par une complexité moins importante et par l'utilisation de la fonction `IG` à la place de L dans le choix de T. La complexité d'un arbre de décision uniformément aléatoire est en $O(dn\texttt{log}(n))$ car, pour chaque région T, on tire le point de coupure, α, selon la loi Uniforme sur le support de X, pour chacune de ses composantes. On dispose alors, à chaque étape de la construction de l'arbre, de d points pour lesquels il faut rechercher les supports sur X, cette recherche nécessitant $(n-1)$ comparaisons. En assumant que l'arbre est équilibré, sa profondeur est d'ordre $\texttt{log}(n)$, ce qui entraîne une complexité totale, dans la construction de l'arbre, en $O(dn\texttt{log}(n))$. Notons que cette version implique que $\beta = 1$. Afin de simplifier l'analyse de l'arbre de décision uniformément aléatoire, nous considérons que le tirage des variables, pour la construction de chaque région, est effectué sans remise.

Dans la pratique, il est difficile d'utiliser le classifieur de Devroye, du fait du grand nombre de combinaisons à évaluer lors du choix d'une région optimale. Toutefois, son intérêt principal est de ne pas imposer de conditions sur le nombre d'observations dans chaque région terminale, mais uniquement sur le nombre de régions de l'arbre.

La règle de décision des deux classifieurs demeure la même. Pour l'échantillon d'apprentissage D_n, on récrit $g_{\mathcal{P}}$, tel que :

$$g_{\mathcal{P}}(x, R, D_n) = g_{\mathcal{P}}(x) = \begin{cases} 1, \ si \ \sum_{i=1}^{n} \mathbf{I}_{\{X_i \in R, Y_i = 1\}} > \sum_{i=1}^{n} \mathbf{I}_{\{X_i \in R, Y_i = 0\}}, \ x \in R \\ \\ 0, \ sinon. \end{cases}$$

Dans le cas d'un arbre de décision uniformément aléatoire, R est une région terminale (et la partition $\mathcal{P}$ possède, alors, au minimum 2 points) car la règle de décision n'intervient pas dans la construction des régions. Dans le cas de l'arbre de décision heuristique, $g_{\mathcal{P}}$ s'applique à chaque région candidate pour le choix de la région optimale. Pour établir la consistance de l'arbre de décision uniformément aléatoire, nous utilisons plusieurs arguments et adaptons le théorème de consistance obtenu par Devroye et al. Nous rappelons une partie de ces arguments ci-après, puis dans la preuve du théorème proposé.

Nous précisons d'abord comment est établie la consistance pour un arbre de décision : pour cela, nous cherchons une borne à $L(\mathcal{P})$ qui dépende de L^* et ϵ, pour tout $\epsilon > 0$. Comme L est inaccessible à cause des données, limitées à D_n, on souhaite que

$$\mathbf{P}(L_n(\mathcal{P}) > L^* + \epsilon) \to 0, \ quand \ n \to \infty.$$

Sous cette condition, $L_n(\mathcal{P})$, l'erreur commise par le classifieur sur la partition $\mathcal{P}$, est alors nécessairement mesurée sur des régions terminales de l'arbre, soit des régions telles que, pour toute région $R \in \mathcal{P}$, $T \subseteq R$, et $R - T$, sa région complémentaire :

$$\texttt{min}(\widehat{L}_n(T) + \widehat{L}_n(R - T) - \widehat{L}_n(R)) = 0. \tag{3.2}$$

Si de telles régions ne sont pas terminales, alors la relation (3.2) n'est pas vérifiée et on peut toujours trouver une sous-partition de $\mathcal{P}$ dont l'erreur est plus petite, si X admet une densité marginale.

L et $\texttt{IG}$ doivent également cohabiter au sein de l'arbre de décision uniformément aléatoire. Nous commençons par récrire $\texttt{H}$ en fonction de R :

$$\texttt{H}(Y) \stackrel{def}{=} \texttt{H}(Y, R) = \mathbf{E}\left[-\texttt{log}(\mathbf{P}(Y = j, X \in R))\right], j \in \{0, 1\}.$$

`Proposition 1.`
Soit R, la région associée à la partition $\mathcal{P}$, $T \subset R$, et $R - T$, sa région complémentaire. Si T et $R - T$ sont des régions terminales, alors

$$\underset{T}{\texttt{arg min}} \ L(T) + L(R - T) - L(R) = \underset{T}{\texttt{arg max}} \ \texttt{H}(Y, R) - (\texttt{H}(Y, R - T) + \texttt{H}(Y, T)).$$

Preuve.
Pour montrer l'égalité entre les deux termes de l'équation, il suffit de montrer que les deux fonctions atteignent les extremas pour la même probabilité ν.
$\mathtt{H}(Y,R) - (\mathtt{H}(Y,R-T) + \mathtt{H}(Y,T))$ atteint son maximum pour $\nu_0(R) = \nu_1(R) = 1/2$, $\nu_0(R-T) = 1 - \nu_0(T) = 1$ et $\nu_1(T) = 1 - \nu_0(T) = 0$. En effet, pour tout $R \in \mathcal{P}$,

$$\begin{aligned}\mathtt{max}(\ \mathtt{H}(Y,R)) &= \mathtt{max}(-\nu_0(R)\mathtt{log}(\nu_0(R)) - \nu_1(R)\mathtt{log}(\nu_1(R))) \\ &= \mathtt{max}(-\nu_0(R)\mathtt{log}(\nu_0(R)) - (1-\nu_0(R))\mathtt{log}(1-\nu_0(R))) \\ &= \mathtt{log}(2),\end{aligned}$$

qui est atteint pour $\nu_0(R) = \nu_1(R) = 1/2$.

$\mathtt{min}\,(\mathtt{H}(Y,R-T) + \mathtt{H}(Y,T)) = 0$ dès que :
$\nu_0(R-T) = 1 - \nu_1(R-T) = 1$ et $\nu_1(T) = 1 - \nu_0(T) = 0$.
Pour le premier terme de la proposition, on a alors :

$$\begin{aligned}\mathtt{min}(L(T) + L(R-T) - L(R)) =& \mathtt{min}\{\ \mathbf{P}(X \in T)\mathtt{min}(\nu_0(T), \nu_1(T)) \\ &+ \mathbf{P}(X \in (R-T))\mathtt{min}(\nu_0(R-T), \nu_1(R-T)) \\ &- \mathtt{min}(\nu_0(R), \nu_1(R))\} \\ =& \mathtt{min}\{\ \mathbf{P}(X \in T)\mathtt{min}(\nu_0(T), \nu_1(T)) \\ &+\ (1 - \mathbf{P}(X \in T))\mathtt{min}(\nu_0(R-T), \nu_1(R-T)) \\ &-\ \mathtt{min}(\nu_0(R), \nu_1(R))\ \} \\ =& 0.\square\end{aligned}$$

La *proposition 1* établit que IG et L, et donc $\widehat{L}_n$, sont équivalentes dans le choix de T pour la mesure de $L_n(\mathcal{P})$, lors de la construction des régions terminales de l'arbre. Il nous faut également considérer que $g_{\mathcal{P}}$ peut s'appliquer à n'importe quelle région comme dans le classifieur de Devroye et al. Cet argument s'établit ainsi : lors de la construction de l'arbre de décision uniformément aléatoire, nous utilisons la fonction IG pour déterminer la région optimale et nous calculons alors $g_{\mathcal{P}}$ et estimons $L_n(\mathcal{P})$. Les régions de l'arbre ne sont alors pas nécessairement terminales et l'interchangeabilité entre IG et L, en dehors des régions terminales, nécessite :
i) de ne pas avoir l'obligation d'explorer toutes les régions possibles à chaque étape de la construction de l'arbre,
ii) que la probabilité ν qui vérifie que le maximum de $(\widehat{L}_n(T) + \widehat{L}_n(R-T) - \widehat{L}_n(R))$ est atteint soit celle qui vérifie le minimum de la fonction IG. Ainsi, utiliser IG ne peut pas détériorer la détermination de la région T.
Dans le premier cas, nous montrons, dans la preuve du *théorème 1*, que la condition est vérifiée. Dans le second, nous faisons appel à la proposition qui suit :

Proposition 2.
Soit R, la région associée à la partition $\mathcal{P}$, $T \subset R$, et $R-T$, sa région complémentaire. Alors, pour la mesure de probabilité, ν, associée à R, T et $R-T$

$$\underset{T}{\mathtt{arg\ max}}\ (L(T) + L(R-T) - L(R)) = \underset{T}{\mathtt{arg\ min}}\ (\mathtt{H}(Y,R) - (\mathtt{H}(Y,R-T) + \mathtt{H}(Y,T))).$$

Preuve.
Calculons la probabilité ν pour le premier terme de l'égalité. On a :
$\texttt{min}(L(R)) = 0$, pour $\nu_0(R) = 1 - \nu_1(R) = 0$,
et $\texttt{max}\ (L(T)+L(R-T)) = 2$, pour $\nu_0(T) = 1-\nu_1(T) = 0$ et $\nu_0(R-T) = 1-\nu_1(R-T) = 0$.
En remplaçant dans H les probabilités ν_0 et ν_1, on obtient alors :
$\texttt{H}(Y, R) = -\nu_0(R)\texttt{log}(\nu_0(R)) - \nu_1(R)\texttt{log}(\nu_1(R)) = 0$,
$\texttt{H}(Y, R-T) = -\nu_0(R-T)\texttt{log}(\nu_0(R-T)) - \nu_1(R-T)\texttt{log}(\nu_1(R-T)) = 0$,
$\texttt{H}(Y, T) = -\nu_0(T)\texttt{log}(\nu_0(T)) - \nu_1(T)\texttt{log}(\nu_1(T)) = 0$.
D'où $\texttt{min}\ (\texttt{H}(Y, R) - (\texttt{H}(Y, R-T) + \texttt{H}(Y, T)) = 0$, ce qui vérifie l'égalité. □

On peut alors choisir l'un ou l'autre des critères d'optimalité, IG ou L, dans la construction de l'arbre de décision uniformément aléatoire. En pratique, l'utilisation de l'entropie produit de meilleures performances, du fait de ses propriétés de dérivabilité. De plus, il n'y a pas d'assimilation de $g_{\mathcal{P}}$ à la fois comme règle de décision et comme critère d'optimalité. La consistance de de l'arbre de décision uniformément aléatoire découle alors du théorème suivant :

Théorème 1.
Pour un classifieur associé à un arbre de décision uniformément aléatoire constitué de k_n régions, avec $k_n \to \infty$ et $k_n = o(\sqrt{n/\texttt{log}\ n})$, si X admet une densité marginale, alors $L_n \to L^*$, presque sûrement.

Preuve.
Le théorème reprend les travaux de Devroye, Györfi et Lugosi (1996, théorème 20.9) sur la consistance des arbres de décision heuristiques, en apportant des modifications pour son adaptation aux particularités des arbres de décision uniformément aléatoires.

Nous rappelons les mesures d'intérêt du problème :
$L(R) = \texttt{min}(\nu_0(R), \nu_1(R))$ est l'espérance de l'erreur commise par le classifieur sur la région R, avec $\nu_j(R) = \mathbf{P}(X \in R, Y = j), j \in \{0, 1\}$.
Un estimateur de $L(R)$ est $L_n(R)$ et, comme nous n'avons à disposition que l'échantillon d'apprentissage D_n, la contrepartie empirique de $L_n(R)$ est $\widehat{L}_n(R)$. La propriété de consistance est liée à la mesure L mais nécessite d'y inclure L_n et $\widehat{L}_n$ qui sont les estimateurs que l'on peut manipuler.

On a également $L(\mathcal{P}) \stackrel{def}{=} \sum_{R\in\mathcal{P}} L(R)$, l'erreur de classification sur la partition $\mathcal{P}$. Une partition $\mathcal{T}$ étend $\mathcal{P}$ en lui ajoutant la région T. La règle de décision de $g_{\mathcal{T}}$ est la même que $g_{\mathcal{P}}$, sauf sur $B \in \mathcal{P}$ et $B \supseteq T$.
Soit $\mathcal{G}_l, l \times l \times ... \times l$, une grille de $\mathbb{R}^d$. $\mathcal{G}_l$ est une sous-partition de $\mathbb{R}^d$ dont tous les éléments peuvent être évalués par la règle de décision. On a :

$$\forall\ \epsilon > 0,\ \exists\ l = l(\epsilon)\ /\ L(\mathcal{G}_l) \leq L^* + \epsilon.$$

En d'autres termes, on peut toujours trouver un certain nombre d'observations caractérisant $\mathcal{G}_l$, pour lesquelles l'erreur calculée est proche de l'erreur de Bayes.
Soit $\mathcal{Q}$, une partition telle que l'intersection de chaque $Q \in \mathcal{Q}$ et $G \in \mathcal{G}_l$ existe, au plus,

une fois. Alors,

$$L(\mathcal{Q}) \leq L(\mathcal{G}_l) \leq L^* + \epsilon.$$

Le point de vue est, ici, l'existence d'un lien entre l'erreur commise sur une partition et le nombre de points de celle-ci. Si l'erreur de Bayes et celle d'une partition bien définie de l'arbre ne sont liées que par le nombre de points de la partition, alors on peut établir une borne à l'erreur commise sur cette dernière. La difficulté consiste à trouver une condition générale sur l'influence du nombre de points sur la borne calculée, après s'être assuré que l'erreur commise sur la partition est optimale.
$\mathcal{Q}$ est un *affinement* de $\mathcal{G}_l$, soit une sous-partition de $\mathcal{G}_l$ dont l'erreur est toujours plus petite ou égale à celle que l'on fait sur $\mathcal{G}_l$.
$\mathcal{T}$ est une extension de $\mathcal{P}$ par Q, $Q \subseteq R \in \mathcal{P}$, s'il existe $\mathcal{T}$, telle que $\mathcal{T}$ contienne toutes les régions de $\mathcal{P}$ sauf R, Q et $R - Q$, la région complémentaire de Q.

`Lemme 1. Devroye, Györfi, Lugosi (1996).`
1- Soit $\mathcal{G}_l$, une partition finie telle que $L(\mathcal{G}_l) \leq L^* + \epsilon$.
2- Soit $\mathcal{P}$, une partition finie de $\mathbb{R}^d$ et $\mathcal{Q}$ un affinement *de $\mathcal{P}$ et $\mathcal{G}_l$.*

$\exists\ Q \in \mathcal{Q}$, et une extension de $\mathcal{P}$ par Q à $\mathcal{T}_Q$ /
si $L(\mathcal{P}) \geq L^* + \epsilon$, alors

$$L(\mathcal{T}_Q) - (L^* + \epsilon) \leq \left(1 - \frac{1}{|\mathcal{Q}|}\right)(L(\mathcal{P}) - (L^* + \epsilon)),$$

où $|\mathcal{Q}|$ est le nombre de régions de $\mathcal{Q}$.

Le lemme 1 indique que pour une sous-partition bien choisie de $\mathcal{P}$ et pour $\mathcal{G}_l$, il existe une région de la sous-partition pour laquelle la différence, entre l'erreur commise et l'erreur de Bayes, est pus petite ou égale à une fraction de celle sur la partition, à (la même fraction de) ϵ près. La conséquence est que plus on partitionne l'arbre, plus on se rapproche de l'erreur de Bayes. En particulier, pour chaque nouvelle sous-partition de l'arbre on dispose d'une relation entre l'erreur de Bayes, l'erreur de la région explorée par l'arbre et celle de la sous-partition. En contrepartie, le nombre de points de la partition courante diminue à mesure que l'on construit l'arbre, de sorte qu'au bout d'un certain temps, poursuivre le partitionnement ne présente plus d'intérêt pour la convergence. On souhaite alors que cette dernière soit plus (ou aussi) rapide que la perte d'observations consécutive à la construction de chaque nouvelle région. Ce lemme est applicable au classifieur de Devroye et al. (1996) mais également au classifieur de l'arbre de décision uniformément aléatoire, en exploitant un résultat donné dans sa preuve.

`Preuve du lemme 1.`
Soit $R \in \mathcal{P}$ et $Q_1, ..., Q_N$, les régions de $\mathcal{Q}$ incluses dans R.
On pose $p_i = \nu_0(Q_i)$, $q_i = \nu_1(Q_i)$, $p = \sum_{i=1}^{N} p_i$.
$L(Q_i) = \mathtt{min}(p_i, q_i)$, $L(R) = \mathtt{min}(p, q)$ et $L(R - Q_i) = \mathtt{min}(p - q_i, q - q_i)$.
On suppose que $p \leq q$. Si $\forall i,\ p_i \leq q_i$, alors,

$$\mathtt{min}(p, q) - \sum_{i=1}^{N} \mathtt{min}(p_i, q_i) = p - \sum_{i=1}^{N} p_i = 0.$$

Si $p_i > q_i$ et $i \in A$, tel que $|A| \geq 1$, un ensemble d'indices et $\Delta_i = p - q_i - (p - p_i) = p_i - q_i$, alors

$$\sum_{i \in A} \Delta_i = \sum_{i \in A} (p_i - q_i) = p - \sum_{i \notin A} p_i - \sum_{i \in A} q_i = p - \sum_{i \in A} \mathtt{min}(p_i, q_i).$$

La relation implique que

$$\max_{1 \leq i \leq N} \Delta_i \geq \frac{1}{|A|} \sum_{i \in A} \Delta_i \geq \frac{p - \sum_{i \in A} \mathtt{min}(p_i, q_i)}{|A|} \geq \frac{\sum_{i \in A} \mathtt{min}(p_i, q_i)}{N}$$

et

$$\max_{1 \leq i \leq N} (L(R) - L(Q_i) - L(R - Q_i)) \geq \frac{L(R) - \sum_{i=1}^{N} L(Q_i)}{N}.$$

Comme $L(Q) \leq (L^* + \epsilon)$, on obtient

$$\begin{aligned} \max_{Q \in \mathcal{Q}} (L(R_Q) - L(Q) - L(R_Q - Q)) &\geq \max_{R \in \mathcal{P}, Q \subseteq R} (L(R) - L(Q) - L(R - Q)) \\ &\geq \max_{R \in \mathcal{P}} \frac{L(R) - \sum_{Q \subseteq R} L(Q)}{|R|}, \end{aligned}$$

où $|R|$ est le nombre de régions de $\mathcal{Q}$ dans R. On en déduit

$$\begin{aligned} \max_{Q \in \mathcal{Q}} (L(R_Q) - L(Q) - L(R_Q - Q)) &\geq \frac{\sum_{R \in \mathcal{P}} [L(R) - \sum_{Q \subseteq R} L(Q)]}{\sum_{R \in \mathcal{P}} |R|} \\ &= \frac{\sum_{R \in \mathcal{P}} L(R) - \sum_{Q \in \mathcal{Q}} L(Q)}{\sum_{R \in \mathcal{P}} |R|} \\ &= \frac{L(\mathcal{P}) - L(\mathcal{Q})}{|\mathcal{Q}|}, \end{aligned}$$

$$\Rightarrow \max_{Q \in \mathcal{Q}} (L(R_Q) - L(Q) - L(R_Q - Q)) \geq \frac{L(\mathcal{P}) - L(\mathcal{Q})}{|\mathcal{Q}|}. \tag{3.3}$$

Comme $\mathcal{T}_Q$ est une extension de $\mathcal{P}$ par Q, $\mathcal{T}_Q$ contient toutes les parties de $\mathcal{P}$ sauf R, Q et $R - Q$ et

$$\max_{Q \in \mathcal{Q}} (L(R_Q) - L(Q) - L(R_Q - Q)) = \max_{Q \in \mathcal{Q}} (L(\mathcal{P}) - L(\mathcal{T}_Q)).$$

On a alors :

$$\begin{aligned} &\max_{Q \in \mathcal{Q}} (L(\mathcal{P}) - L(\mathcal{T}_Q) + (L^* + \epsilon)) \geq \frac{L(\mathcal{P}) - L(\mathcal{Q})}{|\mathcal{Q}|} + (L^* + \epsilon) \\ &\Rightarrow \max_{Q \in \mathcal{Q}} (-(L(\mathcal{T}_Q) - (L^* + \epsilon))) \geq \frac{L(\mathcal{P}) - (L^* + \epsilon)}{|\mathcal{Q}|} - (L(\mathcal{P}) - (L^* + \epsilon)). \end{aligned}$$

On en déduit

$$\begin{aligned} &\max_{Q \in \mathcal{Q}} (L(\mathcal{T}_Q) - (L^* + \epsilon)) \leq L(\mathcal{P}) - (L^* + \epsilon) - \frac{L(\mathcal{P}) - (L^* + \epsilon)}{|\mathcal{Q}|}, \\ &\Rightarrow L(\mathcal{T}_Q) - (L^* + \epsilon) \leq \left(1 - \frac{1}{|\mathcal{Q}|}\right) (L(\mathcal{P}) - (L^* + \epsilon)) . \square \end{aligned}$$

On remarque, de plus, que dans (3.3), le premier terme ne nécessite que la connaissance d'une seule région Q, celle qui maximise l'erreur, pour que le lemme soit vérifié. Comme

$$(3.3) \Rightarrow \max_{Q \in \mathcal{Q}} (L(\mathcal{T}_Q) - (L^* + \epsilon)) \leq \left(1 - \frac{1}{|\mathcal{Q}|}\right) (L(\mathcal{P}) - (L^* + \epsilon)),$$

n'importe quelle région $Q, Q \in \mathcal{Q}$, vérifie cette dernière inégalité. En particulier, si Q est choisie aléatoirement, l'inégalité n'est pas modifiée. La seule condition à satisfaire est la connaissance de $L(R_Q) - L(Q) - L(R_Q - Q)$ sur les d dimensions de la partition. Elle est satisfaite grâce à la *proposition 1* pour les régions terminales de l'arbre. Pour les régions non terminales la *proposition 2* assure que la fonction `IG`, pour le choix de de Q et $R_Q - Q$, ne peut pas détériorer la relation (3.3). Précisons que la relation (3.3) suppose soit, que toutes les régions possibles, Q, doivent être choisies, soit qu'on peut en sélectionner d aléatoirement, associées aux d dimensions. Les arbres de décision de type CART optimisent généralement la construction des régions sur chaque dimension, puis choisissent la région optimale parmi chacune des dimensions explorées.

L'application du lemme 1 à une partition $\mathcal{P}_i$ de l'arbre de décision entraîne l'existence d'une borne à l'erreur de prédiction de n'importe quelle sous-partition, bien choisie, $\mathcal{P}'_i$ de $\mathcal{P}_i$. Le nombre maximal de régions de $\mathcal{Q}$ détermine la forme de cette borne. Dans cette première étape, le point crucial est la nature de la fonction de minimisation de l'erreur empirique pour le choix des régions. On suppose $l \geq 2$. A la i-ème partition de l'arbre, $ni + dni(l-1) + l(l-1)(d-1)$ régions auront été examinées tout au plus pour l'arbre de décision heuristique (pour l'arbre de décision uniformément aléatoire, $n = 1$), le nombre de régions s'additionnant à mesure que l'arbre est construit, jusqu'à la partition $\mathcal{G}_l$. Chaque région potentielle peut contenir chacun des n points de l'espace, dans chacune de ses dimensions. On en déduit

$$|\mathcal{Q}| \leq l^2(d-1) + dnli.$$

Soit $\mathcal{P}'_i$, l'extension de $\mathcal{P}_i$ par $\mathcal{Q}$,alors, par application du lemme 1,

$$L(\mathcal{P}'_i) - (L^* + \epsilon) \leq \left(1 - \frac{1}{l^2(d-1) + dnli}\right) (L(\mathcal{P}_i) - (L^* + \epsilon)).$$

Soit $\mathcal{P}_{i+1}$, la partition qui suit immédiatement $\mathcal{P}_i$ et R_{i+1}, la région sélectionnée par minimisation de l'erreur empirique (ou par maximisation du gain d'information). On a :

$$\begin{aligned} L(\mathcal{P}_{i+1}) - L(\mathcal{P}'_i) &= (L(\mathcal{P}_{i+1}) - \widehat{L}_n(\mathcal{P}_{i+1})) + (\widehat{L}_n(\mathcal{P}_{i+1}) - \widehat{L}_n(\mathcal{P}'_i)) + (\widehat{L}_n(\mathcal{P}'_i) - L(\mathcal{P}'_i)) \\ &\leq (L(\mathcal{P}_{i+1}) - \widehat{L}_n(\mathcal{P}_{i+1})) + (\widehat{L}_n(\mathcal{P}'_i) - L(\mathcal{P}'_i)). \end{aligned}$$

$\mathcal{P}_i$ est donc la partition courante,
$\mathcal{P}'_i$ est une extension de $\mathcal{P}_i$ par $\mathcal{Q}$,
$\mathcal{P}_{i+1}$ est la sous-partition immédiate de $\mathcal{P}_i$.

La décomposition du terme de droite de la précédente inégalité donne :
$(L(\mathcal{P}_{i+1}) - \widehat{L}_n(\mathcal{P}_{i+1})) + (\widehat{L}_n(\mathcal{P'}_i) - L(\mathcal{P'}_i)) =$
$L(R_{i+1}) + L(R - Q - R_{i+1}) - \widehat{L}_n(R_{i+1}) - \widehat{L}_n(R - Q - R_{i+1}) + \widehat{L}_n(Q) - L(Q).$

Comme $\widehat{L}_n(Q) - \widehat{L}_n(R_{i+1}) - \widehat{L}_n(R - Q - R_{i+1}) \leq \widehat{L}_n(R)$, on obtient :

$$(L(\mathcal{P}_{i+1}) - \widehat{L}_n(\mathcal{P}_{i+1})) + (\widehat{L}_n(\mathcal{P'}_i) - L(\mathcal{P'}_i)) \leq \widehat{L}_n(R) - \inf_{R \in \mathcal{P}_i} L(R).$$

Grâce au lemme de Vapnik et Chervonenkis (1974c), on a :

$$\widehat{L}_n(R) - \inf_{R \in \mathcal{P}_i} L(R) \leq 2 \sup_{R \in \mathcal{P}_i} |\widehat{L}_n(R) - L(R)|,$$

et on en déduit

$$L(\mathcal{P}_{i+1}) - L(\mathcal{P'}_i) \leq 2 \sup_{R \in \mathcal{P}_i} |\widehat{L}_n(R) - L(R)|.$$

De façon plus générale, on considère $L(\mathcal{Z}_k)$, où k est le nombre de régions de l'arbre, la classe de toutes les régions de la forme $Z_0 - Z_1 - ... - Z_k$ tels que $Z_i \subseteq Z_0, 1 \leq i < k$, et tels que les $Z_0, Z_1, ..., Z_k$ soient mutuellement disjoints. Pour $i < k$,

$$L(\mathcal{P}_{i+1}) - L(\mathcal{P'}_i) \leq 2 \sup_{Z \in \mathcal{Z}_k} |\widehat{L}_n(Z) - L(Z)|.$$

Pour $Z \in \mathcal{Z}_k$ fixé, on a :

$$\begin{aligned} |\widehat{L}_n(Z) - L(Z)| &= |\mathtt{min}(\nu_{0,n}(Z), \nu_{1,n}(Z)) - \mathtt{min}(\nu_0(Z), \nu_1(Z))| \\ &\leq |\nu_{0,n}(Z) - \nu_0(Z)| + |\nu_{1,n}(Z) - \nu_1(Z)|, \end{aligned}$$

d'où

$$L(\mathcal{P}_{i+1}) - L(\mathcal{P'}_i) \leq V_n = 2 \sup_{Z \in \mathcal{Z}_k} (|\nu_{0,n}(Z) - \nu_0(Z)| + |\nu_{1,n}(Z) - \nu_1(Z)|).$$

On pose $G = \{V_n \leq \delta\}$, et on applique le lemme 1 à $\mathcal{P}_{i+1}$, la sous-partition immédiate de $\mathcal{P}_i$. On a :

$$L(\mathcal{P}_{i+1}) - L(\mathcal{P'}_i) \leq \delta,$$

et

$$\begin{aligned} L(\mathcal{P}_{i+1}) - (L^* + \epsilon) &\leq L(\mathcal{P'}_i) - (L^* + \epsilon) + \delta \\ &\leq L(\mathcal{P}_i) - (L^* + \epsilon) \left(1 - \frac{1}{l^2(d-1) + dnli}\right) + \delta. \end{aligned}$$

`Lemme 2.`
Soit a_n et b_n, deux suites de nombres positifs telles que $b_n \downarrow 0$, $b_0 < 1$ et δ fixé.
Si $a_{n+1} \geq a_n(1 - b_n) + \delta$, alors

$$a_{n+1} \leq a_0 e^{-\sum_{j=0}^{n} b_j} + (n+1)\delta.$$

`Preuve.`
La preuve du lemme 2 procède d'un raisonnement par récurrence :

$$\begin{aligned} a_1 &\leq a_0(1-b_0)+\delta, \\ et\ a_2 &\leq a_1(1-b_0)+\delta \\ &\leq a_0(1-b_0)(1-b_1)+\delta(1-b_1)+\delta. \end{aligned}$$

Par un raisonnement similaire,

$$\begin{aligned} a_3 &\leq a_2(1-b_2)+\delta \\ &\leq a_0(1-b_0)(1-b_1)(1-b_2)+\delta(1-b_1)(1-b_2)+\delta(1-b_2)+\delta. \end{aligned}$$

On en déduit

$$a_{n+1} \leq a_0 \prod_{i=0}^{n}(1-b_j)+\delta\sum_{m=1}^{n-1}\prod_{j=m}^{n}(1-b_j)+\delta.$$

Par ailleurs, comme

$$0<x<1 \Rightarrow \mathtt{log}(1-x) \leq -x,$$

il s'ensuit

$$\begin{aligned} a_0 \prod_{i=0}^{n}(1-b_j) &\leq a_0 e^{\sum_{j=0}^{n}\mathtt{log}(1-b_j)} \\ &\leq a_0 e^{-\sum_{j=0}^{n} b_j}, \end{aligned}$$

et

$$\delta\sum_{m=1}^{n-1}\prod_{j=m}^{n}(1-b_j)+\delta \leq \delta(n+1).\square$$

L'application du lemme 2 à $L(\mathcal{P}_k)-(L^*+\epsilon)$ entraîne

$$\begin{aligned} L(\mathcal{P}_k)-(L^*+\epsilon) &\leq (L(\mathcal{P}_0)-(L^*+\epsilon))e^{-\sum_{j=0}^{k-1}\frac{1}{l^2(d-1)+dnlj}}+\delta k \\ &\leq e^{-\int_0^{k-1}\left(\frac{1}{l^2(d-1)+dnlu}\right)\mathtt{d}u}+\delta k \\ &\leq e^{-\frac{1}{dnl}\mathtt{log}\left(\frac{l^2(d-1)+dnl(k-1)}{l^2(d-1)}\right)}+\delta k \\ &= \frac{1}{\left(1+\frac{d}{d-1}\frac{n(k-1)}{l}\right)^{\frac{1}{dnl}}}+\delta k. \end{aligned}$$

Posons

$$\delta k<\epsilon/2 \ \ et \ \ \phi=\frac{1}{\left(1+\frac{d}{d-1}\frac{n(k-1)}{l}\right)^{\frac{1}{dnl}}}<\frac{\epsilon}{2},$$

on a :

$$L(\mathcal{P}_k) - (L^* + \epsilon) \leq \phi + \delta k.$$

On rappelle que

$$L(\mathcal{P}_{i+1}) - L(\mathcal{P'}_i) \leq V_n = 2 \sup_{Z \in \mathcal{Z}_k} (|\nu_{0,n}(Z) - \nu_0(Z)| + |\nu_{1,n}(Z) - \nu_1(Z)|).$$

On a posé également $G = \{V_n \leq \delta\}$, et le lemme 1, appliqué à $\mathcal{P}_{i+1}$, la sous-partition immédiate de $\mathcal{P}_i$, entraîne la relation

$$L(\mathcal{P}_{i+1}) - L(\mathcal{P'}_i) \leq \delta,$$

équivalente à

$$\begin{aligned} L(\mathcal{P}_{i+1}) - (L^* + \epsilon) &\leq L(\mathcal{P'}_i) - (L^* + \epsilon) + \delta \\ &\leq L(\mathcal{P}_i) - (L^* + \epsilon) \left(1 - \frac{1}{l^2(d-1) + dnli}\right) + \delta, \end{aligned}$$

relation à laquelle nous venons d'appliquer le lemme 2. Finalement, pour $i + 1 = k$,

$$L(\mathcal{P}_{i+1}) - L(\mathcal{P'}_i) \leq \delta \Rightarrow L(\mathcal{P}_k) \leq L^* + 2\epsilon.$$

En récrivant la première inégalité, on obtient

$$L(\mathcal{P}_k) - (L^* + \epsilon) \leq L(\mathcal{P'}_{k-1}) - (L^* + \epsilon) + \delta,$$

$$\begin{aligned} \Rightarrow \mathbf{P}(L(\mathcal{P}_k) - L(\mathcal{P'}_{k-1}) \leq \delta) &\geq \mathbf{P}(V_n \leq \delta) \\ et\ \mathbf{P}(L(\mathcal{P}_k) \leq L^* + 2\epsilon) &\geq \mathbf{P}(V_n \leq \delta). \end{aligned}$$

On en déduit alors

$$\mathbf{P}(L(\mathcal{P}_k) > L^* + 2\epsilon) < \mathbf{P}(G^c) = \mathbf{P}(V_n > \delta).$$

Rappelons que

$$L_n(R) = \mathbf{P}\left\{X \in R, Y \neq g_{\mathcal{P}}(X)\right\},$$

$$L_n(\mathcal{P}) = \sum_{R \in \mathcal{P}} L_n(R).$$

Rappelons également que nous n'avons pas accès à $L(\mathcal{P}_k)$ dans la pratique, mais seulement aux versions empiriques de $L_n(\mathcal{P}_k)$, où k correspond au nombre de régions courantes. On introduit alors $\widehat{L}_n$ de sorte que :

$$L_n(\mathcal{P}_k) = (L_n(\mathcal{P}_k) - \widehat{L}_n(\mathcal{P}_k)) + (\widehat{L}_n(\mathcal{P}_k) - L(\mathcal{P}_k)) + L(\mathcal{P}_k).$$

Puisque

$$|L_n(\mathcal{P}_k) - \widehat{L}_n(\mathcal{P}_k)| \leq |(\widehat{L}_n(\mathcal{P}_k) - L(\mathcal{P}_k)|,$$

alors

$$\begin{aligned} L_n(\mathcal{P}_k) &\leq 2|\widehat{L}_n(\mathcal{P}_k) - L(\mathcal{P}_k)| + L(\mathcal{P}_k) \\ &\leq 2 \sum_{R \in \mathcal{P}_k} |\nu_0(R) - \nu_{0,n}(R)| + |\nu_1(R) - \nu_{1,n}(R)| + L(\mathcal{P}_k), \end{aligned}$$

et

$$L_n(\mathcal{P}_k) \leq 2W_n + L(\mathcal{P}_k),$$

avec $W_n = 2 \sum_{R \in \mathcal{P}_k} |\nu_0(R) - \nu_{0,n}(R)| + |\nu_1(R) - \nu_{1,n}(R)|$.

Pour établir la consistance de l'arbre de décision, on cherche alors une borne de $L_n(\mathcal{P}_k)$ qui dépende de L^*. On a montré qu'il existe $\epsilon > 0$, tel que $L(\mathcal{P}_k) > L^* + 2\epsilon$. On en déduit :

$$\mathbf{P}\{2W_n + L(\mathcal{P}_k) > L^* + 2\epsilon + 2\epsilon\} \leq \mathbf{P}\{2W_n > 2\epsilon\} + \mathbf{P}\{L(\mathcal{P}_k) > L^* + 2\epsilon\},$$

soit encore que

$$\mathbf{P}\{L_n(\mathcal{P}_k) > L^* + 4\epsilon\} \leq \mathbf{P}\{W_n > \epsilon\} + \mathbf{P}\{V_n > \delta\},$$

avec $\delta k < \epsilon/2$.

Posons $U_n(Z) = |\nu_0(Z) - \nu_{0,n}(Z)| + |\nu_1(Z) - \nu_{1,n}(Z)|$, pour une région Z, quelconque, de l'arbre.
Comme $V_n = 2 \sup_R(|\nu_0(R) - \nu_{0,n}(R)| + |\nu_1(R) - \nu_{1,n}(R)|)$, alors

$$V_n \leq \sup_Z(U_n(Z)) + \sup_{Z_1,\ldots,Z_k} \left(\sum_{i=1}^{k} U_n(Z_i) \right),$$

où les Z_i sont des régions disjointes de l'arbre. Il s'en suit que

$$V_n \leq (k+1) \sup_Z(U_n(Z))$$

et

$$W_n \leq (k+1) \sup_Z(U_n(Z)).$$

A ce stade, il faut alors faire appel à la théorie de Vapnik-Chervonenkis afin d'établir un résultat sur W_n et V_n. On note $s(\mathcal{A}, n)$ le nombre maximal de sous-ensembles de n points, qui peuvent être *pulvérisés* (correctement classés) par $\mathcal{A}$. On définit également $V_\mathcal{A}$, la dimension de Vapnik-Chervonenkis (ou dimension V-C) de $\mathcal{A}$, comme la plus grande valeur $k, k \geq 1$, telle que $s(\mathcal{A}, k) = 2^k, \ \forall\ k < n$.

`Exemples.`
(i) Si $\mathcal{A} = \{(-\infty, x_1] \times ... \times (-\infty, x_d]\}$*, alors* $V_\mathcal{A} = d$.
(ii) Si $\mathcal{A}$ *est la classe de tous les rectangles dans* $\mathbb{R}^d$*, alors* $V_\mathcal{A} = 2d$.
(iii) Si $\mathcal{A}$ *est la classes des convexes dans* $\mathbb{R}^d$*, alors* $V_\mathcal{A} = +\infty$.

Nous rappelons l'inégalité de Vapnik :

`Définition.`
Pour toute mesure de probabilité ν et pour une classe d'ensembles $\mathcal{A}$, et pour tout n et $\epsilon > 0$,

$$\mathbf{P}\{\sup_{A\in\mathcal{A}}|\nu_n(A) - \nu(A)| > \epsilon\} \leq 8s(\mathcal{A}, n)\mathsf{e}^{-n\epsilon^2/32}.$$

On en déduit :

$$\mathbf{P}\{\sup_{Z\in\mathcal{P}}|\nu_{0,n}(Z) - \nu_0(Z)| > \epsilon\} \leq 8s(\mathcal{P}, n)\mathsf{e}^{-n\epsilon^2/32}$$

et

$$\mathbf{P}\{\sup_{Z\in\mathcal{P}}|\nu_{1,n}(Z) - \nu_1(Z)| > \epsilon\} \leq 8s(\mathcal{P}, n)\mathsf{e}^{-n\epsilon^2/32},$$

d'où

$$\mathbf{P}\{\sup_{Z\in\mathcal{P}}(U_n(Z)) > \epsilon\} \leq 16s(\mathcal{P}, n)\mathsf{e}^{-n\epsilon^2/32}.$$

Comme $s(\mathcal{P}, n) = n^{2d}$ on obtient, pour W_n et V_n :

$$\mathbf{P}\{W_n > \epsilon\} \leq 16n^{2d}\mathsf{e}^{-n\left(\frac{\epsilon}{2(k+1)}\right)^2/32}$$

et

$$\mathbf{P}\{V_n > \delta\} \leq 16n^{2d}\mathsf{e}^{-n\left(\frac{\delta}{2(k+1)}\right)^2/32}.$$

La consistance est établie si le terme de droite de chacune des inégalités tend vers 0 lorsque n tend vers l'infini. Puisque $n\mathsf{e}^{-n} \to 0$ quand $n \to \infty$, il faut que l'argument de l'exponentielle, respectivement les termes en $\frac{n\epsilon^2}{k^2}$ et $\frac{n\delta^2}{k^2}$ dans les deux inégalités, tende vers l'infini lorsque n et k tendent vers l'infini. Pour la première inégalité,

$$\textit{si } k \to \infty \textit{ et } \frac{n}{k^2\log n} \to \infty, \textit{ quand } n \to \infty, \textit{ alors, pour tout } \epsilon > 0, \ \mathbf{P}\{W_n > \epsilon\} \to 0.$$

De même, puisque $\delta k < \epsilon/2$,

$$\textit{si } k \to \infty \textit{ et } \frac{n\delta^2}{k^2\log n} \to \infty, \textit{ quand } n \to \infty, \textit{ alors, pour tout } \delta > 0, \ \mathbf{P}\{V_n > \delta\} \to 0.$$

De la condition $\delta k < \epsilon/2$, on pose $\delta = \epsilon/(3k)$. On obtient :

$$\mathbf{P}\{V_n > \delta\} = \mathbf{P}\{V_n > \frac{\epsilon}{3k}\} \leq 16n^{2d}\mathsf{e}^{-n\left(\frac{\epsilon}{6(k^2+3k)}\right)^2/32}.$$

La condition sur k est alors modifiée et

$$\textit{si } k \to \infty \textit{ et } \frac{n}{k^3\log n} \to \infty, \textit{ quand } n \to \infty, \textit{ alors, pour tout } \epsilon > 0, \ \mathbf{P}\{V_n > \epsilon\} \to 0.$$

Sous cette dernière condition, on a :

$$\mathbf{P}\{L_n(\mathcal{P}_k) > L^* + 4\epsilon\} \to 0.$$

Le lemme de Borel-Cantelli permet de conclure sur la convergence vers 0 de $L_n(\mathcal{P}_k) - L^*$. La convergence peut être améliorée grâce au théorème suivant qui offre une borne de risque plus précise pour W_n et V_n.

`Théorème. Lugosi, Nobel (1993).`
Soit $X_1, ..., X_n$ des vecteurs aléatoires i.i.d. dans $\mathbb{R}^d$ de mesure μ et de mesure empirique μ_n. Soit $\mathcal{A}$ une collection de partitions de $\mathbb{R}^d$. Alors, pour tout $M < \infty$ et $\epsilon > 0$,

$$\mathbf{P}\left\{ \sup_{\mathcal{P}^{(M)} \in \mathcal{A}} \left(\sum_{A \in \mathcal{P}^{(M)}} |\mu_n(A) - \mu(A)| \right) > \epsilon \right\} \leq 8 s(\mathcal{A}^M, n) e^{-n\epsilon^2/512} + e^{-n\epsilon^2/2}.$$

On en déduit :

$$\mathbf{P}\left\{ \sup_{\mathcal{P}^{(k)} \in \mathcal{P}} \left(\sum_{Z \in \mathcal{P}^{(k)}} |\nu_{0,n}(Z) - \nu_0(Z)| \right) > \epsilon \right\} \leq 8 s(\mathcal{P}^{(k)}, n) \mathrm{e}^{-n\epsilon^2/512} + \mathrm{e}^{-n\epsilon^2/2},$$

avec $\mathcal{P}^{(k)} \stackrel{def}{=} \mathcal{P}_k$.

L'application du théorème à $U_n(Z)$ entraîne

$$\mathbf{P}\{\sup_{Z \in \mathcal{P}} \sum U_n(Z) > \epsilon\} \leq 16 s(\mathcal{P}, n) \mathrm{e}^{-n\epsilon^2/512} + 2\mathrm{e}^{-n\epsilon^2/2}.$$

En rappelant l'inégalité de Vapnik appliquée à $U_n(Z)$:

$$\mathbf{P}\{\sup_{Z \in \mathcal{P}}(U_n(Z)) > \epsilon\} \leq 16 s(\mathcal{P}, n) \mathrm{e}^{-n\epsilon^2/32},$$

et en remarquant que :

$$\begin{aligned} V_n &= 2 \sup_{R \in \mathcal{P}} (|\nu_0(R) - \nu_{0,n}(R)| + |\nu_1(R) - \nu_{1,n}(R)|) \\ &\leq \sup_Z (U_n(Z)) + \sup_{Z_1, ..., Z_k} \left(\sum_{i=1}^k U_n(Z_i) \right), \end{aligned}$$

on en déduit pour V_n :

$$\mathbf{P}\left\{ V_n > \frac{\epsilon}{3k} \right\} \leq 16 n^{2d} \left(\mathrm{e}^{-n\left(\frac{\epsilon}{6k}\right)^2/32} + \mathrm{e}^{-n\left(\frac{\epsilon}{6k}\right)^2/512} \right) + 2\mathrm{e}^{-n\left(\frac{\epsilon}{6k}\right)^2/2}.$$

Il s'en suit que :

$$L_n(\mathcal{P}_k) \stackrel{p.s.}{\to} L^*, \ si \ \frac{n}{k^2 \log n} \to \infty. \square$$

3.3 Forêt uniformément aléatoire

Une forêt uniformément aléatoire est une collection d'arbres de décision uniformément aléatoires et sa règle de décision est définie par le vote majoritaire des règles de chaque arbre (ou leur moyenne dans le cas de la régression).
Soit $\bar{g}_{\mathcal{P}}^{(B)}$, $B > 1$, la règle de décision associée à la forêt uniformément aléatoire.
Dans le cas de la classification, elle est définie par :

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = \begin{cases} 1, \ si \ \sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x)=1\}} > \sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x)=0\}} \\ \\ 0, \ sinon. \end{cases}$$

Comme $g_{\mathcal{P}}$ est un classifieur consistant, nous pouvons faire appel aux résultats de Biau, Devroye et Lugosi (2008) établissant la consistance de la forêt aléatoire, lorsque l'arbre de décision est consistant.

`Proposition. Biau, Devroye, Lugosi (2008).`
Soit $\{g_{\mathcal{P}}\}$ une suite de classifieurs consistants. Alors, $\bar{g}_{\mathcal{P}}^{(B)}$, le classifieur de la forêt aléatoire résultante, est aussi consistant.

`Preuve.`
Nous reprenons la preuve de Biau, Devroye et Lugosi en y apportant des éléments complémentaires.

Soit

$$\eta(x) = \mathbf{P}(Y = 1|X = x),$$

et g^*, le classifieur de Bayes, tel que :

$$g^*(x) = \begin{cases} 1, \ si \ \eta(x) > 1/2 \\ 0, \ sinon. \end{cases}$$

On rappelle $L_n(R)$, la probabilité d'erreur de $g_{\mathcal{P}}$ dans la région R, définie par :

$$L_n(R) = \mathbf{P}(g_{\mathcal{P}}(X) \neq Y, X \in R),$$

et

$$L_n(\mathcal{P}) = \mathbf{P}(g_{\mathcal{P}}(X) \neq Y).$$

On rappelle également l'erreur de Bayes,

$$L^* = \mathbf{P}\{g^*(X) \neq Y\} = \texttt{min}(\eta(x), 1 - \eta(x)).$$

La consistance de $g_{\mathcal{P}}$ équivaut à écrire que $L_n(\mathcal{P}) \to L^*$, presque sûrement. On a, en suivant les arguments de Devroye, Györfi, Lugosi (1996, théorème 2.1) :

$$\begin{aligned} \mathbf{P}\{g_{\mathcal{P}}(X) \neq Y|X \in R\} &= 1 - \mathbf{P}\{g_{\mathcal{P}}(X) = Y|X \in R\} \\ &= 1 - (\mathbf{P}\{Y = 1, g_{\mathcal{P}}(X) = 1|X \in R\} + \mathbf{P}\{Y = 0, g_{\mathcal{P}}(X) = 0|X \in R\}) \\ &= 1 - \left(\mathbf{I}_{\{g_{\mathcal{P}}(x)=1\}}\mathbf{P}\{Y = 1|X \in R\} + \mathbf{I}_{\{g_{\mathcal{P}}(x)=0\}}\mathbf{P}\{Y = 0|X \in R\}\right) \\ &= 1 - \left(\mathbf{I}_{\{g_{\mathcal{P}}(x)=1\}}\eta(x) + \mathbf{I}_{\{g_{\mathcal{P}}(x)=0\}}(1 - \eta(x))\right). \end{aligned}$$

On obtient alors :

$$\begin{aligned}
&\mathbf{P}\{g_{\mathcal{P}}(X) \neq Y | X \in R\} - \mathbf{P}\{g^*_{\mathcal{P}}(X) \neq Y | X \in R\} \\
&\quad = \eta(x)\left(\mathbf{I}_{\{g^*_{\mathcal{P}}(x)=1\}} - \mathbf{I}_{\{g_{\mathcal{P}}(x)=1\}}\right) + (1-\eta(x))\left(\mathbf{I}_{\{g^*_{\mathcal{P}}(x)=0\}} - \mathbf{I}_{\{g_{\mathcal{P}}(x)=0\}}\right) \\
&\quad = \eta(x)\left(\mathbf{I}_{\{g^*_{\mathcal{P}}(x)=1\}} - \mathbf{I}_{\{g_{\mathcal{P}}(x)=1\}}\right) + (1-\eta(x))\left(1 - \mathbf{I}_{\{g^*_{\mathcal{P}}(x)=1\}} - 1 + \mathbf{I}_{\{g_{\mathcal{P}}(x)=1\}}\right) \\
&\quad = (2\eta(x)-1)\left(\mathbf{I}_{\{g^*_{\mathcal{P}}(x)=1\}} - \mathbf{I}_{\{g_{\mathcal{P}}(x)=1\}}\right).
\end{aligned}$$

Supposons que $\eta(x) > 1/2$. On en déduit :

$$\begin{aligned}
\mathbf{P}\{g_{\mathcal{P}}(X) \neq Y | X \in R\} &= (2\eta(x)-1)\left(\mathbf{I}_{\{g^*_{\mathcal{P}}(x)=1\}} - \mathbf{I}_{\{g_{\mathcal{P}}(x)=1\}}\right) + \mathbf{P}\{g^*_{\mathcal{P}}(X) \neq Y | X \in R\} \\
&= (2\eta(x)-1)\left(\mathbf{I}_{\{g^*_{\mathcal{P}}(x)=1\}} - \mathbf{I}_{\{g_{\mathcal{P}}(x)=1\}}\right) + 1 - \eta(x) \\
&= (2\eta(x)-1)\left(\mathbf{I}_{\{g_{\mathcal{P}}(x)=0\}} - \mathbf{I}_{\{g^*_{\mathcal{P}}(x)=0\}}\right) + 1 - \eta(x).
\end{aligned}$$

Comme $g_{\mathcal{P}}$ est un classifieur consistant, on a :

$$\mathbf{E}(L_n(R)) \to L^*,$$

qui équivaut à écrire :

$$\begin{aligned}
&\mathbf{P}\{g_{\mathcal{P}}(x) \neq Y\} \to \mathbf{P}\{g^*_{\mathcal{P}}(x) \neq Y\} \\
\Rightarrow\ &\mathbf{P}\{g_{\mathcal{P}}(x) = 0\} \to \mathbf{P}\{g^*_{\mathcal{P}}(x) = 0\}
\end{aligned}$$

et

$$\mathbf{P}\{g_{\mathcal{P}}(x) = 0\} \to 0,$$

puisque $L^* = \mathtt{min}(\eta(x), 1-\eta(x)) = 1-\eta(x)$.
Pour le classifieur $\bar{g}^{(B)}_{\mathcal{P}}$,

$$\mathbf{P}\left\{\bar{g}^{(B)}_{\mathcal{P}}(x) = 0\right\} = \mathbf{P}\left\{\frac{1}{B}\sum_{b=1}^{B} \mathbf{I}_{\{g^{(b)}_{\mathcal{P}}(x)=0\}} > 1/2\right\}.$$

Grâce à l'inégalité de Markov, on a :

$$\begin{aligned}
\mathbf{P}\left\{\frac{1}{B}\sum_{b=1}^{B} \mathbf{I}_{\{g^{(b)}_{\mathcal{P}}(x)=0\}} > 1/2\right\} &\leq 2\mathbf{E}\left\{\frac{1}{B}\sum_{b=1}^{B} \mathbf{I}_{\{g^{(b)}_{\mathcal{P}}(x)=0\}}\right\} \\
&= 2\mathbf{P}\left\{g_{\mathcal{P}}(x) = 0\right\} \to 0.\square
\end{aligned}$$

On en déduit la consistance de la forêt uniformément aléatoire :

`Proposition 3.`
La forêt uniformément aléatoire est consistante, dès que X admet une densité marginale dans $\mathbb{R}^d$, si $k_n \to \infty$ et $k_n = o\left(\sqrt{n/\mathtt{log}\ n}\right)$, quand $n \to \infty$.

`Preuve.`
Comme l'arbre uniformément aléatoire est consistant (*théorème 1*), il suffit d'appliquer

la proposition de Biau, Devroye et Lugosi pour obtenir la consistance de la forêt uniformément aléatoire. □

Notons que dans le cas de la régression, la règle de décision s'écrit :

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = \frac{1}{B}\sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(x),$$

avec

$$g_{\mathcal{P}}(x,R) = g_{\mathcal{P}}(x) = \frac{1}{\sum_{i=1}^{n}\mathbf{I}_{\{X_i\in R\}}}\sum_{i=1}^{n} Y_i\mathbf{I}_{\{X_i\in R\}},\ x\in R.$$

3.4 Forêt uniformément aléatoire *incrémentale*

La forêt uniformément aléatoire *incrémentale* exploite le fait que les données ne sont pas toujours disponibles en même temps. Nous introduisons, ici, le néologisme *incrémental* (à la place d'incrémentiel) et le conservons tout au long du document. Sa règle de décision est une *règle de décision agrégée*, soit, une règle de décision de S, avec $S > 1$, forêts uniformément aléatoires issues de S sous-échantillons aléatoires de D_n, chacun de taille $m_s, 1 \leq s \leq S$, avec $m_s << n$ et n grand. Elle est notée $\bar{g}_{\mathcal{P},big}^{(B)}$ et définie par :

$$\bar{g}_{\mathcal{P},big}^{(B)} = \begin{cases} 1,\ si\ \sum_{s=1}^{S}\sum_{b=1}^{B}\mathbf{I}_{\left\{g_{\mathcal{P}_s}^{(b)}(x)=1\right\}} > \sum_{s=1}^{S}\sum_{b=1}^{B}\mathbf{I}_{\left\{g_{\mathcal{P}_s}^{(b)}(x)=0\right\}} \\ \\ 0,\ sinon. \end{cases}$$

où $\mathcal{P}_s$ est la partition d'origine du sous-échantillon s, de taille m_s.
Dans le cas de la régression :

$$\bar{g}_{\mathcal{P},big}^{(B)}(x) = \frac{1}{BS}\sum_{s=1}^{S}\sum_{b=1}^{B} g_{\mathcal{P}_s}^{(b)}(x).$$

`Proposition 4.`
La forêt uniformément aléatoire incrémentale est consistante, dès que X admet une densité marginale dans $\mathbb{R}^d$, si $\frac{n}{S}\to\infty, k_n\to\infty$ et $k_n = o\Big(\sqrt{(n/S)\texttt{log}(n/S)}\Big)$, quand $n\to\infty$.

`Preuve.`
il suffit que le rapport n/S tende vers l'infini. La forêt uniformément aléatoire incrémentale se comporte alors comme une forêt uniformément aléatoire et l'application du théorème de Biau, Devroye et Lugosi (2008, théorème 6) entraîne la consistance. □

Notons qu'ici, il n'y a toujours qu'une seule règle de décision pour la forêt aléatoire et non une moyenne des règles de décision de chaque sous-échantillon évalué. Si les paramètres de la distribution de X ne changent pas d'un sous-échantillon à l'autre, la construction de la forêt aléatoire est équivalente au *Subagging* ("SUBsample AGGregatING"), énoncé par Friedman et Hall (1999, 2007) et formalisé par Bühlmann et Yu (2000), dont le principe général est marqué par l'utilisation d'un sous-échantillon aléatoire de D_n pour chaque

classifieur de base du modèle ensembliste. Si les paramètres de la distribution changent ou si la loi n'est plus la même, le problème est beaucoup plus difficile à résoudre puisque l'information fournie par D_n peut devenir inadaptée sur l'échantillon de test. Lorsque la loi de probabilité change, la règle de décision est revue. Ce cas est typique des données qui arrivent périodiquement, ou dépendent intrinsèquement du temps, et un changement de loi implique de supprimer les arbres les moins adaptés, par exemple ceux les moins récents, et de faire dépendre du temps le nombre d'arbres à définir. Pour détecter un changement de loi, un test de Kolmogorov-Smirnov peut être appliqué entre deux sous-échantillons consécutifs pour la variable la plus importante du premier (échantillon) d'entre eux. Son résultat détermine le nombre d'arbres associé au dernier sous-échantillon. La règle de décision de la forêt uniformément aléatoire incrémentale est alors notée $\bar{g}^{(T)}_{\mathcal{P},inc}$ et définie par :

$$\bar{g}^{(T)}_{\mathcal{P},inc} = \begin{cases} 1, \ si \ \sum_{t=1}^{T}\sum_{b=1}^{B_t} \mathbf{I}_{\left\{g^{(b)}_{\mathcal{P}_t}(x)=1\right\}} > \sum_{t=1}^{T}\sum_{b=1}^{B_t} \mathbf{I}_{\left\{g^{(b)}_{\mathcal{P}_t}(x)=0\right\}} \\ \\ 0, \ sinon. \end{cases}$$

et pour la régression,

$$\bar{g}^{(T)}_{\mathcal{P},inc}(x) = \frac{1}{\sum_{t=1}^{T} B_t} \sum_{t=1}^{T}\sum_{b=1}^{B_t} g^{(b)}_{\mathcal{P}_t}(x),$$

où $\mathcal{P}_t$ est la partition des données pour la période t,
B_t, le nombre d'arbres de décision construits pour la période t,
T, le nombre de périodes.

Dans cette version, la dépendance au temps est plus explicite et la dernière partition, $\mathcal{P}_T$, est généralement celle qui contient le plus d'arbres. Dans la pratique et dans certaines situations, une forêt uniformément aléatoire incrémentale peut former un phénomène de *mémoire* avec l'augmentation du nombre d'arbres et de données. Cette mémoire se traduit par un recours facultatif à certains échantillons d'apprentissage, le nombre d'arbres étant suffisamment important. Les forêts uniformément aléatoires incrémentales profitent du fait que la corrélation entre les arbres tend à être plus faible que dans le cadre classique, du fait de l'absence d'optimisation des points de coupure, ce qui compense, en partie, la perte de précision induite par un plus faible nombre d'observations pour la construction de chaque arbre.

3.5 Propriétés et extensions

Une forêt uniformément aléatoire dérive ses propriétés de celles des forêts aléatoires de Breiman. Ces propriétés concernent essentiellement la convergence de l'erreur de prédiction et les bornes de risque. La première est une application de la loi des grands nombres aux forêts aléatoires et assure que l'erreur empirique tend bien vers la vraie erreur de prédiction du modèle. La deuxième assure que cette erreur de prédiction admet une borne supérieure explicite. Nous reprenons la notation de Breiman afin d'assurer une même lisibilité avec l'article original présentant les forêts aléatoires.

3.5.1 Erreur de prédiction

Nous analysons, ici, l'erreur de prédiction de la forêt uniformément aléatoire dans le cas de la classification et de la régression.

`Erreur de prédiction dans le cas de la classification`

Dans le cas de la classification, l'erreur de prédiction, ou erreur de généralisation, est l'erreur moyenne commise par un classifieur sur l'appartenance d'une observation à une classe du problème. Lorsqu'elle est calculée pour toutes les observations possibles, l'erreur de prédiction est l'espérance de l'erreur commise par le classifieur. Supposons que l'on s'intéresse à la différence (ou la marge) entre la proportion, sur tous les arbres, du nombre de cas correctement classés et du nombre de cas incorrectement classés. La probabilité pour que cette différence soit négative est l'erreur de prédiction.
Plus formellement, notons mg cette marge. Alors,

$$mg(X,Y) = \left(\frac{1}{B}\sum_{b=1}^{B}\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X)=Y\}}\right) - \left(\frac{1}{B}\sum_{b=1}^{B}\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X)\neq Y\}}\right)$$

et l'erreur de prédiction, PE^*, s'écrit :

$$PE^* = \mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(mg(X,Y) < 0\right).$$

Soit $\widehat{PE}^*$, la contrepartie empirique de PE^*. On a :

$$\begin{aligned}
\widehat{PE}^* &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{\{mg(X_i,Y_i)<0\}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{\left\{\frac{1}{B}\left(\sum_{b=1}^{B}\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X_i)=Y_i\}}\right)-\frac{1}{B}\left(\sum_{b=1}^{B}\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X_i)\neq Y_i\}}\right)<0\right\}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{\left\{\left(\sum_{b=1}^{B}\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X_i)=Y_i\}}\right)<\left(\sum_{b=1}^{B}\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X_i)\neq Y_i\}}\right)\right\}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{\left\{g_{\mathcal{P}}^{(B)}(X_i)\neq Y_i\right\}} \\
&= \overline{PE}^*.
\end{aligned}$$

Supposons que chacun des arbres de décision dépende d'un paramètre aléatoire θ,
soit que $g_{\mathcal{P}}(X) \overset{def}{=} g_{\mathcal{P}}(X,\theta)$ et $g_{\mathcal{P}}^{(b)}(X) \overset{def}{=} g_{\mathcal{P}}(X,\theta_b)$, avec θ, le paramètre qui caractérise la structure aléatoire d'un arbre et différencie donc chaque arbre des autres.

`Théorème. Breiman (2001, théorème 1.2).`
Lorsque $B \to \infty$, et pour presque sûrement tous les arbres de décision $g_{\mathcal{P}}(X,\theta_1), ..., g_{\mathcal{P}}(X,\theta_B)$, d'une forêt aléatoire,

$$PE^* \overset{p.s.}{\to} PE = \mathbf{P}_{\mathbf{X},\mathbf{Y}}\{\mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta)=Y) - \mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta)\neq Y) < 0\}.$$

Ce résultat est une application de la loi forte des grands nombres. Il a pour conséquence pratique de limiter le nombre d'arbres nécessaires à la convergence de l'erreur de prédiction de la forêt aléatoire. Comme l'estimateur tend vers la vraie erreur, au delà d'un certain nombre d'arbres, il n'y a plus d'amélioration possible. Les deux principales conséquences sont, cependant, de deux types :

i) La forêt aléatoire ne peut pas *sur-apprendre* les données, bien que l'erreur de prédiction sur D_n, l'échantillon d'entraînement ou d'apprentissage, puisse être nulle et le soit souvent en pratique. L'erreur d'entraînement n'est cependant jamais utilisée pour estimer l'erreur de prédiction de la forêt aléatoire. A la place, *l'erreur OOB*, que nous décrivons plus loin, est généralement plus réaliste et plus proche de la vraie erreur de prédiction, en particulier quand n est suffisamment grand et que l'échantillon de test est également plus grand que l'échantillon d'entraînement. Il convient de noter que le *sur-apprentissage* n'est évité que sous les hypothèses de non dépendance pour presque tous les arbres. La non dépendance se traduit par une corrélation moyenne faible entre les arbres et peut être explicitement mesurée grâce à D_n. Du point de vue empirique, nous avons noté, dans le cas de la régression, que la corrélation moyenne était généralement beaucoup plus élevée que dans le cas de la classification. Toutefois, la corrélation est, alors, celle entre les résidus de l'estimation.

ii) La deuxième conséquence, et peut-être une façon simple avec la borne de risque de Breiman, de comprendre les bonnes performances des forêts aléatoires, est la loi forte des grands nombres elle-même. Une forêt aléatoire peut être vue comme une méthode de Monte Carlo particulière destinée à estimer une espérance.

Quand $n \to \infty$, $g_{\mathcal{P}}(X, \theta)$ est une approximation aléatoire (par construction) de $\mathbf{E}(Y|X, \theta)$. En générant B arbres indépendants et en appliquant la loi des grands nombres, on a :

$$\frac{g_{\mathcal{P}}(X, \theta_1) + g_{\mathcal{P}}(X, \theta_2) + ... + g_{\mathcal{P}}(X, \theta_B)}{B} \overset{p.s.}{\to} \mathbf{E}_\theta(\mathbf{E}(Y|X, \theta)), \ quand\ B \to \infty.$$

Le caractère particulier est à lier au fait que la variance de la forêt tend à être faible alors qu'elle est importante pour un arbre et doit le demeurer ; les approches théoriques mettent en évidence la réduction de variance face à un arbre de décision mais ne caractérisent pas la manière dont la variance peut être contrôlée tout en réduisant la corrélation.

Pour préciser les points i) et ii), nous procédons, à ce stade, à une analyse séparée dans le cas de la régression. L'absence de cadre unifié pour l'analyse de l'erreur de prédiction est aussi un argument à cette séparation. Toutefois, les travaux de Domingos (2000), puis de James (2003), fournissent un cadre favorable à une même analyse pour la régression et la classification. Ils font suite à ceux de Geman, Bienenstock et Doursat (1992), lesquels formalisent initialement le concept de la décomposition biais-variance de l'erreur de prédiction.

`Erreur de prédiction dans le cas de la régression`

Dans le cas de la régression, l'erreur de prédiction d'un arbre de décision correspond à l'espérance de l'erreur (risque) quadratique, entre Y et la valeur de la règle de décision $g_{\mathcal{P}}$, pour tous les arbres de paramètre θ et pour toutes les observations possibles. On la note $PE(g_{\mathcal{P}}(X,\theta))$ et elle est définie par :

$$PE(arbre) \overset{def}{=} PE(g_{\mathcal{P}}(X,\theta)) = \mathbf{E}_{\theta}\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}(X,\theta))^2.$$

Un estimateur de l'erreur de prédiction est l'erreur quadratique moyenne d'un arbre, définie par :

$$PE^*(arbre) \overset{def}{=} PE^*(g_{\mathcal{P}}(X,\theta)) = \frac{1}{B}\sum_{b=1}^{B}\left(\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}(X,\theta_b))^2\right).$$

Précisons que les notations PE et PE^* sont définies afin de préserver la cohérence avec la classification. Dans le cas de la régression, Breiman définit l'erreur de prédiction sous la notation PE^* et n'affecte pas de notation à l'erreur quadratique moyenne. L'espérance de l'erreur quadratique de la forêt uniformément aléatoire est définie par :

$$PE(forêt) \overset{def}{=} PE(\bar{g}_{\mathcal{P}}^{(B)}(X)) = \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \mathbf{E}_{\theta}g_{\mathcal{P}}(X,\theta)\right)^2.$$

Et l'espérance de l'erreur quadratique de l'estimateur de la forêt est donnée par :

$$PE^*(forêt) \overset{def}{=} PE^*(\bar{g}_{\mathcal{P}}^{(B)}(X)) = \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \frac{1}{B}\sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(X,\theta_b)\right)^2.$$

`Théorème. Breiman (2001, théorème 11.1).`
Quand $B \to \infty$*,*

$$\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \frac{1}{B}\sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(X,\theta_b)\right)^2 \overset{p.s.}{\to} \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \mathbf{E}_{\theta}g_{\mathcal{P}}(X,\theta)\right)^2.$$

Ce théorème est également une application de la loi forte des grands nombres et permet de conclure sur la convergence de l'estimateur de la forêt aléatoire vers l'espérance de l'arbre de décision. Il nous permet également d'aborder simplement la décomposition biais-variance de l'erreur de prédiction et d'explorer la manière dont la forêt aléatoire améliore l'arbre de décision. En particulier, l'éclairage important est le paradoxe apparent entre la variance des arbres de décision que l'on espère importante et celle de la forêt aléatoire qui doit demeurer faible, sans quoi l'erreur de prédiction n'est pas suffisamment réduite relativement à celle sur l'arbre.

`Erreur de prédiction et décomposition biais-variance`

Dans le cas de la régression, l'estimateur de la forêt aléatoire est simplement la moyenne des résultats des règles de décision de chaque arbre. On a :

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = \frac{1}{B}\sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(x).$$

Soit Y, une variable aléatoire à valeurs dans $\mathbb{R}$ et $\hat{Y}$, un estimateur de Y. Plus précisément, on suppose que toute réalisation de Y peut être estimée par une réalisation de $\hat{Y}$. On suppose également que la loi des grands nombres s'applique. L'erreur quadratique moyenne entre Y et son estimateur $\hat{Y}$ s'écrit :

$$\begin{aligned}
\mathbf{E}(Y-\hat{Y})^2 &= \mathbf{E}(Y-\mathbf{E}(Y)+\mathbf{E}(Y)-\hat{Y})^2 \\
&= \mathbf{E}\left\{Y-\mathbf{E}(Y)\right\}^2 + \mathbf{E}\left\{\mathbf{E}(Y)-\hat{Y}\right\}^2 + 2\mathbf{E}\left\{(Y-\mathbf{E}(Y))(\mathbf{E}(Y)-\hat{Y})\right\} \\
&= \mathbf{E}\left\{Y-\mathbf{E}(Y)\right\}^2 + \mathbf{E}\left\{\mathbf{E}(Y)-\hat{Y}\right\}^2 + 2\left(\mathbf{E}(Y)\mathbf{E}(\hat{Y}) - \mathbf{E}(Y\hat{Y})\right) \\
&= \mathbf{E}\left\{Y-\mathbf{E}(Y)\right\}^2 + \mathbf{E}\left\{\mathbf{E}(Y)-\mathbf{E}(\hat{Y})+\mathbf{E}(\hat{Y})-\hat{Y}\right\}^2 + 2\left(\mathbf{E}(Y)\mathbf{E}(\hat{Y}) - \mathbf{E}(Y\hat{Y})\right) \\
&= \mathbf{E}\left\{Y-\mathbf{E}(Y)\right\}^2 + \left(\mathbf{E}(Y)-\mathbf{E}(\hat{Y})\right)^2 + \mathbf{E}\left\{\mathbf{E}(\hat{Y})-\hat{Y}\right\}^2 + 2\left(\mathbf{E}(Y)\mathbf{E}(\hat{Y}) - \mathbf{E}(Y\hat{Y})\right).
\end{aligned}$$

Considérons maintenant que Y est associé à l'échantillon $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$ et que :

$$Y = f(X) + \epsilon, \tag{3.4}$$

où f est une fonction déterministe des observations x,
et ϵ est une variable aléatoire indépendante de X, avec $\mathbf{E}(\epsilon) = 0$.
On suppose que $\hat{Y}$ est un estimateur de $f(X)$. On a :

$$\begin{aligned}
\mathbf{E}(\hat{Y})\mathbf{E}(Y) - \mathbf{E}(Y\hat{Y}) &= \mathbf{E}\left\{\hat{Y}f(X)\right\} - \mathbf{E}\left\{(f(X)+\epsilon)\hat{Y}\right\} \\
&= 0.
\end{aligned}$$

Finalement,

$$\begin{aligned}
\mathbf{E}(Y-\hat{Y})^2 &= \mathbf{E}\left\{Y-\mathbf{E}(Y)\right\}^2 + \left(\mathbf{E}(Y-\hat{Y})\right)^2 + \mathbf{Var}(\hat{Y}) \\
&= \mathbf{E}(\epsilon^2) + \left(f(X)-\mathbf{E}(\hat{Y})\right)^2 + \mathbf{Var}(\hat{Y})
\end{aligned}$$

$\Leftrightarrow$ *Espérance de l'erreur de prédiction* $= Bruit + Biais^2 + Variance$.

Si les erreurs résiduelles sont centrées et si f est déterministe, l'erreur de prédiction dépend d'un bruit résiduel qui ne peut être réduit car il n'est pas contrôlé par l'estimateur. Ce bruit peut être assimilé à une sorte d'erreur de Bayes dans le cas de la régression. L'erreur de prédiction dépend également de l'espérance (au carré) du biais de l'estimateur et de sa variance.

En s'abstrayant de la relation (3.4), et en ne faisant plus d'hypothèses sur les erreurs, hormis l'existence des moments d'ordre 1 et 2, ou sur f, on a donc un terme de covariance supplémentaire dans la décomposition de l'erreur de prédiction. Ecrivons la décomposition biais-variance, pour tous les paramètres θ d'un arbre de décision uniformément aléatoire. On pose :

$g_{\mathcal{P}}(X,\theta) \overset{def}{=} \hat{Y}$, $\mathbf{E}(Y-\hat{Y})^2 \overset{def}{=} \mathbf{E}_\theta\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}(X,\theta))^2$ et $\epsilon \overset{def}{=} Y - \mathbf{E}(Y)$.

L'espérance de l'erreur de prédiction de l'arbre est alors donnée par :

$$\mathbf{E}_\theta \mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}(X,\theta))^2 = \\ \mathbf{E}(\epsilon^2) + \mathbf{E}_\theta \left\{ \mathbf{E}_{\mathbf{X},\mathbf{Y}} \left[Y - g_{\mathcal{P}}(X,\theta) \right] \right\}^2 + \mathbf{E}_\theta \mathbf{Var}_{\mathbf{X}}(g_{\mathcal{P}}(X,\theta)) - 2\mathbf{E}_\theta \mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(g_{\mathcal{P}}(X,\theta), Y),$$

avec

$$g_{\mathcal{P}}(x,R) = g_{\mathcal{P}}(x) = \frac{1}{\sum_{i=1}^n \mathbf{I}_{\{X_i \in R\}}} \sum_{i=1}^n Y_i \mathbf{I}_{\{X_i \in R\}}, \ x \in R,$$

où R est la région à laquelle appartient l'observation x.

De même, pour la forêt uniformément aléatoire, on obtient :

$$\mathbf{E}_{\mathbf{X},\mathbf{Y}} \left(Y - \mathbf{E}_\theta g_{\mathcal{P}}(X,\theta) \right)^2 = \\ \mathbf{E}(\epsilon^2) + \left\{ \mathbf{E}_{\mathbf{X},\mathbf{Y}} \left[Y - \mathbf{E}_\theta g_{\mathcal{P}}(X,\theta) \right] \right\}^2 + \mathbf{Var}_{\mathbf{X}} \left(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta) \right) - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta), Y).$$

La difficulté est, ici, l'écriture non triviale de la règle de décision, laquelle dépend de la profondeur de l'arbre et des réalisations de Y dans la région correspondante. Pour caractériser la réduction de l'erreur de prédiction de la forêt face à celle d'un arbre, Breiman s'intéresse plutôt à l'estimateur de la forêt aléatoire et montre que sa variance est plus faible que celle de n'importe quel arbre de décision sous-jacent. De ce fait, l'erreur de prédiction l'est aussi, à condition que l'espérance du biais de l'arbre de décision soit nulle, soit que :

$$\mathbf{E}(Y) = \mathbf{E}_{\mathbf{X}}(g_{\mathcal{P}}(X,\theta)).$$

Sous cette dernière hypothèse, on obtient alors trois types de résultats.
i) Il est plus adapté de diminuer le biais d'un arbre de décision aléatoire que sa variance. Le biais n'est pas modifié par la forêt aléatoire alors que c'est le cas pour la variance.
ii) La variance de l'arbre de décision doit être importante, mais pas trop, afin que la loi des grands nombres continue de s'appliquer.
iii) Les arbres de décision doivent être aussi peu corrélés que possible.

Dans le cas de la classification, on obtient également une décomposition biais-variance de l'erreur de test. On dispose du même échantillon D_n, mais on a maintenant $Y \in \{0,1\}$ et il existe une relation, inconnue, entre Y et X. On a :

$$\begin{aligned} &\mathbf{P}_{\mathbf{X},\mathbf{Y}}(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y) \\ &\quad = \mathbf{Var}(Y) + \left\{ \mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - \bar{g}_{\mathcal{P}}^{(B)}(X)) \right\}^2 + \mathbf{Var}_{\mathbf{X}}(\bar{g}_{\mathcal{P}}^{(B)}(X)) - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\bar{g}_{\mathcal{P}}^{(B)}(X), Y) \\ &\quad = \mathbf{P}(Y=y)\mathbf{P}(Y \neq y) + \left\{ \mathbf{P}(Y=y) - \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=y) \right\}^2 \\ &\qquad + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=y)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq y) - 2\mathbf{E}\left\{ \left[\bar{g}_{\mathcal{P}}^{(B)}(X) - \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=y) \right] \left[Y - \mathbf{P}(Y=y) \right] \right\} \\ &\quad = \mathbf{P}(Y=1) + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - 2\mathbf{E}\left\{ Y \bar{g}_{\mathcal{P}}^{(B)}(X) \right\}. \end{aligned}$$

Du point de vue théorique, la construction de la forêt aléatoire nécessite alors d'agir fortement sur le biais des arbres de décision et de leur permettre d'atteindre une variance suffisamment grande. Pour ces deux raisons, les forêts totalement aléatoires ne peuvent atteindre des performances opérationnelles optimales, même si leur caractère consistant

est démontré. Agir sur le biais consiste à construire des arbres larges et profonds et aboutit à une erreur d'entraînement nulle sur D_n. Agir sur la variance nécessite de définir une *randomisation* non triviale.

Poursuivons la décomposition en examinant de plus près la variance de la forêt aléatoire. On a :

$$\begin{aligned}\mathbf{Var}\left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right) &= \frac{1}{B^2}\sum_{b=1}^{B}\sum_{c=1}^{B}\mathbf{Cov}\left(g_{\mathcal{P}}^{(b)}(X), g_{\mathcal{P}}^{(c)}(X)\right) \\ &= \frac{1}{B^2}\left[\sum_{b=1}^{B}\mathbf{Var}\left(g_{\mathcal{P}}^{(b)}(X)\right) + 2\sum_{1\leq b<c\leq B}\mathbf{Cov}\left(g_{\mathcal{P}}^{(b)}(X), g_{\mathcal{P}}^{(c)}(X)\right)\right].\end{aligned}$$

Supposons que les variances des arbres soient identiques. On obtient :

$$\begin{aligned}\mathbf{Var}\left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right) &= \frac{1}{B^2}\left[B\mathbf{Var}\left(g_{\mathcal{P}}(X)\right) + B(B-1)\rho\mathbf{Var}\left(g_{\mathcal{P}}(X)\right)\right] \\ &= \frac{\mathbf{Var}\left(g_{\mathcal{P}}(X)\right)}{B} + \frac{(B-1)\rho\mathbf{Var}\left(g_{\mathcal{P}}(X)\right)}{B} \\ &= \rho\mathbf{Var}\left(g_{\mathcal{P}}(X)\right) + \frac{(1-\rho)}{B}\mathbf{Var}\left(g_{\mathcal{P}}(X)\right).\end{aligned}$$

Quand $B \to \infty$, la variance de la forêt aléatoire dépend uniquement de la variance de l'arbre et de la corrélation entre les arbres (pris deux à deux). Si la variance de l'arbre est trop grande, la diminution consécutive de la corrélation ne suffit pas à réduire suffisamment l'erreur de prédiction. Si la variance est trop petite, la corrélation est trop importante et l'erreur de prédiction ne baisse pas non plus. Ce cas est aussi le moins optimal, car il invalide l'application de la loi forte des grands nombres. On obtient alors, *si l'arbre de décision est sans biais et si sa variance est constante* :

$$\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - \mathbf{E}_\theta g_{\mathcal{P}}(X,\theta))^2 = \mathbf{E}(\epsilon^2) + \rho\mathbf{Var}_{\mathbf{X}}(g_{\mathcal{P}}(X,\theta)) - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta), Y). \quad (3.5)$$

En pratique, la variance des arbres n'est cependant pas homoscédastique et l'hypothèse d'une absence de biais de l'arbre de décision n'est pas garantie. Breiman apporte une réponse à ce problème en fournissant directement une borne supérieure à l'erreur de prédiction de la forêt aléatoire. Elle a l'avantage de faire explicitement référence à l'erreur de prédiction de l'arbre de décision, ce qui évite d'avoir à analyser la variance d'un arbre. Nous l'examinons d'abord dans le cas de la régression, puis dans le cas de la classification. Puis nous montrons comment les mettre en oeuvre en pratique.

3.5.2 Bornes de risque

Les forêts aléatoires forment d'abord une structure algorithmique qui peut prendre de nombreuses formes. Leur point le plus commun avec les probabilités est la loi des grands nombres, dès lors que l'on considère un arbre de décision comme une sorte de variable aléatoire. Leur autre caractéristique consiste à mêler l'aspect aléatoire à la recherche d'optimalité, c'est-à-dire la capacité de la forêt, en utilisant tous les arbres, à générer une marge (la différence entre les observations bien classées et celles mal classées) élevée.

Plus simplement, la forêt aléatoire transforme un ensemble de *variables* aléatoires en un classifieur *optimal*, dans le sens où la forêt est plus performante (presque sûrement) que chacun des arbres. Pour formaliser ce concept, Breiman introduit la notion de *force* qui caractérise l'espérance de la marge, et la *corrélation* qui caractérise la dépendance (entre les arbres) et le caractère aléatoire de la forêt.

Comprendre et améliorer les forêts aléatoires nécessite de bien appréhender ces deux aspects. Notons que leur implémentation numérique a de nombreux avantages et permet, par exemple, de déterminer assez précisément les modifications qui bénéficient ou non à la forêt construite. Nous explicitons donc la force et la corrélation et montrons comment elles permettent d'établir une borne de risque pour la forêt aléatoire.

`Borne de risque dans le cas de la régression`

Pour la régression, le théorème suivant transforme l'analyse des forêts aléatoires en deux problématiques principales :
- l'existence d'une borne de risque qui permet d'aborder l'essentiel des propriétés théoriques numériquement ;
- une fois la borne de risque connue, la recherche d'une vitesse de convergence universelle.

`Théorème. Breiman (2001, théorème 11.2).`
On suppose que pour tout θ, $\mathbf{E}(Y) = \mathbf{E}_{\mathbf{X}}(g_{\mathcal{P}}(X, \theta))$.
Alors,

$$PE(forêt) \leq \bar{\rho} PE(arbre).$$

De plus, l'erreur de prédiction théorique de la forêt (uniformément) aléatoire admet une formulation explicite donnée par :

$$PE(forêt) = PE(\mathbf{E}_\theta g_{\mathcal{P}}(X, \theta)) = \bar{\rho} \left(\mathbf{E}_\theta \sqrt{\mathbf{E}_{\mathbf{X},\mathbf{Y}} \left(Y - g_{\mathcal{P}}(X, \theta)\right)^2} \right)^2,$$

où $\bar{\rho}$ est la corrélation moyenne entre les erreurs résiduelles de tous les arbres de la forêt. On rappelle que l'erreur de prédiction de l'arbre est l'espérance de son erreur quadratique, soit :

$$PE(arbre) \stackrel{def}{=} PE(g_{\mathcal{P}}(X, \theta)) = \mathbf{E}_\theta \mathbf{E}_{\mathbf{X},\mathbf{Y}} (Y - g_{\mathcal{P}}(X, \theta))^2.$$

A La différence de la relation (3.5), il n'y a pas d'hypothèse sur la variance des arbres et l'espérance du biais de l'arbre est supposée nulle. La corrélation entre les erreurs résiduelles des arbres est le principal facteur de réduction de la variance et, donc, de l'erreur de prédiction. La minimisation de l'erreur quadratique moyenne passe principalement par une réduction de la corrélation, en augmentant le nombre et la profondeur des arbres ainsi que le nombre de variables candidates pour la construction de chaque région. Notons que la réduction de la corrélation se traduit par une réduction du biais de la forêt (qui n'est pas nul, dans la pratique) et une augmentation de la variance des arbres. Une manière (radicale) de réduire la corrélation dans les forêts uniformément aléatoires consiste à ne plus utiliser les Y_i, en les substituant, en partie ou en totalité, par un échantillon aléatoire,

bien choisi, de distribution gaussienne, de moyenne $\overline{Y}$ et de variance $c\widehat{\mathbf{Var}}(Y), c > 1$. Cette approche, bien qu'iconoclaste, produit des résultats intéressants et a l'avantage de protéger du sur-apprentissage.

On note $\overline{PE^*}(g_{\mathcal{P}})$, l'estimateur de l'erreur quadratique moyenne de B arbres uniformément aléatoires :

$$\overline{PE^*}(arbre) = \overline{PE^*}(g_{\mathcal{P}}(X,\theta)) = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g_{\mathcal{P}}^{(b)}(X_i,\theta)\right)^2\right).$$

L'estimateur de l'erreur de prédiction de la forêt est donné par :

$$\widehat{PE^*}(\bar{g}_{\mathcal{P}}^{(B)}(X)) = \hat{\bar{\rho}}\left(\frac{1}{B}\sum_{b=1}^{B}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g_{\mathcal{P}}^{(b)}(X_i)\right)^2}\right)^2,$$

et $\hat{\bar{\rho}}$, l'estimateur de la corrélation moyenne entre les erreurs résiduelles des B arbres, définie par :

$$\begin{aligned}\hat{\bar{\rho}} &= \frac{\frac{1}{B}\sum_{1\leq b<c\leq B}\widehat{\mathbf{Cov}}_{\theta,\theta'}(Y - g_{\mathcal{P}}^{(b)}(X,\theta), Y - g_{\mathcal{P}}^{(c)}(X,\theta'))}{\frac{1}{B}\sum_{1\leq b<c\leq B}\sqrt{\widehat{\mathbf{Var}}_{\theta}(Y - g_{\mathcal{P}}^{(b)}(X,\theta))}\sqrt{\widehat{\mathbf{Var}}_{\theta'}(Y - g_{\mathcal{P}}^{(c)}(X,\theta'))}} \\ &= \frac{\sum_{1\leq b<c\leq B}\hat{\rho}_{b,c}(\theta,\theta')\sqrt{\widehat{\mathbf{Var}}_{\theta}(Y - g_{\mathcal{P}}^{(b)}(X,\theta))}\sqrt{\widehat{\mathbf{Var}}_{\theta'}(Y - g_{\mathcal{P}}^{(c)}(X,\theta'))}}{\sum_{1\leq b<c\leq B}\sqrt{\widehat{\mathbf{Var}}_{\theta}(Y - g_{\mathcal{P}}^{(b)}(X,\theta))}\sqrt{\widehat{\mathbf{Var}}_{\theta'}(Y - g_{\mathcal{P}}^{(c)}(X,\theta'))}}.\end{aligned}$$

Dans le cas de la régression, on dispose de trois mesures permettant d'avoir des garanties sur l'erreur de test (l'erreur quadratique moyenne) :
i) l'erreur de prédiction théorique de la forêt uniformément aléatoire, estimée à partir des données OOB et qui agit, alors, comme un critère de sélection de modèle ;
ii) la borne de risque de Breiman (estimée également à partir des données OOB) qui constitue une borne supérieure de l'erreur de prédiction théorique ;
iii) l'erreur quadratique moyenne OOB qui est un estimateur de l'erreur de test.
Ces trois mesures présentent l'intérêt de faire cohabiter l'analyse numérique et l'analyse théorique comme outils d'exploration. En particulier, la borne de risque de Breiman donne également un sens à la dépendance entre les arbres et permet d'observer les changements de structure. Dans le cadre des théorèmes de consistance sur les arbres de décision, un intérêt peut être l'observation de l'effet d'une limitation du nombre de régions sur l'erreur de prédiction théorique.

`Borne de risque dans le cas de la classification`

Posons :

$$mr(X,Y) = \mathbf{P}_{\theta}(g_{\mathcal{P}}(X,\theta) = Y) - \mathbf{P}_{\theta}(g_{\mathcal{P}}(X,\theta) \neq Y),$$

la limite de $mg(X,Y)$.
On note $s, s > 0$, la force du classifieur, définie par :

$$s = \mathbf{E}_{\mathbf{X},\mathbf{Y}}\{mr(X,Y)\}.$$

Comme $PE = \mathbf{P}_{\mathbf{X},\mathbf{Y}}(mr(X,Y) < 0)$, on a :

$$PE = \mathbf{P}_{\mathbf{X},\mathbf{Y}}\{\, \mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) \neq Y) - \mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) = Y) - \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left[mr(X,Y)\right] \geq -\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left[mr(X,Y)\right]\},$$

et l'inégalité de Bienaymé-Tchebychev permet de conclure :

$$PE^* \leq \frac{\mathbf{Var}(mr)}{s^2}.$$

On peut décrire plus explicitement la borne de risque. mr s'écrit encore :

$$mr(X,Y) = \mathbf{E}_\theta\left[\mathbf{I}_{\{g_{\mathcal{P}}(X,\theta)=Y\}} - \mathbf{I}_{\{g_{\mathcal{P}}(X,\theta)\neq Y\}}\right].$$

Soit,

$$rmg(\theta,X,Y) = \mathbf{I}_{\{g_{\mathcal{P}}(X,\theta)=Y\}} - \mathbf{I}_{\{g_{\mathcal{P}}(X,\theta)\neq Y\}},$$

la marge entre une observation classée correctement et une autre, incorrectement classée. On a :

$$mr(X,Y) = \mathbf{E}_\theta\left[rmg(\theta,X,Y)\right],$$

et pour tous les θ, θ' tels que θ et θ' soient indépendants,

$$\begin{aligned} mr(X,Y)^2 &= \{\mathbf{E}_{\theta,\theta'}\left[rmg(\theta,X,Y)\right]\}^2 \\ &= \mathbf{E}_{\theta,\theta'}\left[rmg(\theta,X,Y)rmg(\theta',X,Y)\right], \end{aligned}$$

et

$$\mathbf{Var}(mr) = \mathbf{E}\left[(mr - \mathbf{E}(mr))^2\right],$$

$$\begin{aligned} &\mathbf{Var}(mr(X,Y)) \\ &\quad= \mathbf{E}_{\theta,\theta'}\{[\mathbf{E}_{\theta,\theta'}(rmg(\theta,X,Y)rmg(\theta',X,Y)) - \mathbf{E}_{\theta,\theta'}[\mathbf{E}_{\theta,\theta'}(rmg(\theta,X,Y)rmg(\theta',X,Y))]]^2\} \\ &\quad= \mathbf{E}_{\theta,\theta'}\left\{\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(rmg(\theta,X,Y)rmg(\theta',X,Y))\right\} \\ &\quad= \mathbf{E}_{\theta,\theta'}\left\{\rho(\theta,\theta')\sqrt{\mathbf{Var}(\theta)}\sqrt{\mathbf{Var}(\theta')}\right\}, \end{aligned}$$

où $\rho(\theta,\theta')$ et $\mathbf{Var}(\theta)$ sont, respectivement, la corrélation entre deux arbres (caractérisés grâce à θ et θ') et la variance d'un arbre, cette dernière s'interprétant ici comme l'erreur de prédiction d'un arbre à un terme près. On obtient alors :

$$\begin{aligned} \mathbf{Var}(mr(X,Y)) &= \bar{\rho}\left(\mathbf{E}_\theta\left(\sqrt{\mathbf{Var}(\theta)}\right)^2\right) \\ &\leq \bar{\rho}\left(\mathbf{E}_\theta\left(\mathbf{Var}\left(\theta\right)\right)\right). \end{aligned}$$

$\bar{\rho}$ est la corrélation moyenne entre tous les arbres, pris deux à deux, de la forêt aléatoire. Plus elle est faible, moins il y a de dépendance entre les arbres et plus l'erreur de prédiction tend à diminuer. $\mathbf{Var}(\theta)$ joue le même rôle que $\mathbf{Var}(mr(X,Y))$, mais pour un arbre. Plus la variance d'un arbre est élevée, plus le nombre de grandes variations de marge est

important. Un arbre avec une variance élevée aura tendance a produire des résultats (sur un même échantillon de test) très différents pour plusieurs apprentissages sur un même échantillon d'entraînement. Si la corrélation moyenne entre tous les arbres est faible, une variance moyenne élevée n'aura que peu d'impact sur la forêt aléatoire. De façon contre-intuitive, il est même préférable que les arbres possèdent des variances importantes. Nous montrons plus explicitement cette propriété en récrivant la corrélation moyenne :

$$\bar{\rho} = \frac{\mathbf{E}_{\theta,\theta'}[\rho(\theta,\theta')\sqrt{\mathbf{Var}(\theta)}\sqrt{\mathbf{Var}(\theta')}]}{\mathbf{E}_{\theta,\theta'}[\sqrt{\mathbf{Var}(\theta)}\sqrt{\mathbf{Var}(\theta')}]},$$

et

$$\begin{aligned}\mathbf{E}_\theta(\mathbf{Var}(\theta)) &\leq \mathbf{E}_\theta[\mathbf{E}_{\mathbf{X},\mathbf{Y}}(rmg(\theta,X,Y))]^2 - s^2 \\ &\leq 1 - s^2.\end{aligned}$$

On en déduit :

`Théorème. Breiman (2001, théorème 2.3).`
Une borne supérieure pour l'erreur de généralisation de la forêt aléatoire est donnée par

$$PE^* \leq \frac{\bar{\rho}(1-s^2)}{s^2}.$$

Pour la classification également, l'erreur de généralisation est plus faible pour la forêt que pour un arbre quelconque. On peut différencier, ici, le caractère aléatoire, à travers la corrélation moyenne, $\bar{\rho}$, qui doit demeurer basse et l'optimalité, la marge s^2, qu'il faut maintenir haute. En pratique, chacun des paramètres a un effet opposé sur l'autre et une des approches pour faire décroître l'erreur de prédiction consiste à observer le comportement de ces deux mesures en modifiant D_n ou en changeant la structure des arbres.

Pour $Y \in \{0,1\}$, $mr(X,Y)$ s'écrit encore

$$mr(X,Y) = 2\mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) = Y) - 1,$$

et une condition suffisante pour que $s > 0$ est donnée par :

$$\mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) = Y) > 1/2.$$

Les arbres de décision peuvent donc être des classifieurs *faibles*, soit des classifieurs avec une erreur de prédiction élevée mais inférieure à 50%, sans que cela ne nuise à la diminution de l'erreur de prédiction de la forêt. Ce dernier point justifie que certaines versions de forêts aléatoires ne nécessitent que de peu optimiser les arbres de décision.

3.5.3 L'erreur *Out-of-bag* (OOB)

Lors de la construction de la forêt aléatoire, D_n est re-échantillonné avec remise (ou sous-échantillonné) pour chacun des arbres. D_n a toujours la même taille (pour le tirage avec remise) mais seule une partie des observations est utilisée pour la construction d'un

arbre. L'autre partie sert à estimer l'erreur de prédiction de la forêt aléatoire. Cette erreur de prédiction est appelée erreur OOB. Le tirage avec remise (*bootstrap*) permet de libérer environ 36.8% des données pour l'estimation OOB. On arrive à disposer, néanmoins, des n observations pour le calcul de l'erreur OOB, mais de seulement b arbres parmi les B. En effet, chaque arbre ne rencontre qu'une partie des données tandis que la forêt, elle, rencontre l'ensemble des données sur une partie des arbres. Pour rendre plus clair le processus, nous l'explicitons ci-dessous :
i) pour les B arbres de la forêt, nous notons les indices des données OOB puis les filtrons, de façon à éliminer les redondances dues au tirage avec remise de D_n et celles induites par le nombre d'arbres. Chacun des indices est ainsi unique ;
ii) comme aucun arbre ne travaille sur toutes les données, pour chaque observation de D_n, il existe au moins un arbre qui ne l'aura pas classée ;
iii) chacune de ces observations est une donnée OOB (donc jamais rencontrée) pour b arbres de décision parmi les B ;
iv) chaque observation OOB est alors classée par b arbres de décision, et la valeur de la règle de décision reste le vote majoritaire (des b arbres) ou la moyenne dans le cas de la régression ;
v) l'erreur de prédiction OOB est le rapport entre le nombre de cas incorrectement classés par la règle de décision de la forêt et n, le nombre d'observations.
On a, dans le cas de la classification, pour une observation x et pour D_n uniquement :

$$\bar{g}^{(B)}_{\mathcal{P},oob}(x) = \begin{cases} 1, \ si \ \sum_{b=1}^{B} \mathbf{I}_{\left\{g^{(b)}_{\mathcal{P}}(x)=1\right\}} \mathbf{I}_{\left\{b\in G^{-}(x,B)\right\}} > \sum_{b=1}^{B} \mathbf{I}_{\left\{g^{(b)}_{\mathcal{P}}(x)=0\right\}} \mathbf{I}_{\left\{b\in G^{-}(x,B)\right\}} \\ \\ 0, \ sinon. \end{cases}$$

où $\bar{g}^{(B)}_{\mathcal{P},oob}$ est la règle de décision de la forêt aléatoire réduite à B' arbres, $B' \simeq \lceil \exp(-1)B \rceil$, $G^{-}(x,B)$ est l'ensemble des arbres, parmi les B possibles, n'ayant jamais classé l'observation x.
Plus précisément, la règle de décision, dans le cas OOB, n'est valable que sur D_n et a lieu en deux temps :
- dans le premier, on construit la règle de chaque arbre pour D_n ;
- Dans un second temps, la règle de la forêt, du point de vue OOB, est une règle de décision qui omet tous les arbres ayant déjà rencontré l'observation x.

L'erreur de prédiction OOB (comme estimateur de l'erreur de prédiction de la forêt), pour la classification, s'écrit alors :

$$\overline{PE}^{*(B)}_{oob} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{\left\{\bar{g}^{(B)}_{\mathcal{P},oob}(X_i)\neq Y_i\right\}}.$$

Dans le cas de la régression, l'erreur quadratique moyenne OOB estime l'erreur quadratique moyenne sur les données de test. Elle est donnée par :

$$\overline{PE}^{*}(\bar{g}^{(B)}_{\mathcal{P},oob}(X)) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{g}^{(B)}_{\mathcal{P},oob}(X_i))^2.$$

Mais on dispose également d'un (second) estimateur explicite de l'erreur de prédiction théorique de la forêt défini par :

$$\widehat{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X)) = \bar{\hat{\rho}}_{oob}\left(\frac{1}{\sum_{b=1}^{B}\mathbf{I}_{\left\{b\in G^-(x,B)\right\}}}\sum_{b=1}^{B}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(Y_i - g^{(b)}_{\mathcal{P}}(X_i)\right)\mathbf{I}_{\left\{b\in G^-(x,B)\right\}}\right)^2}\right)^2,$$

avec

$$\bar{g}^{(B)}_{\mathcal{P},oob}(x) = \frac{1}{\sum_{b=1}^{B}\mathbf{I}_{\left\{b\in G^-(x,B)\right\}}}\sum_{b=1}^{B} g^{(b)}_{\mathcal{P}}(x)\mathbf{I}_{\left\{b\in G^-(x,B)\right\}}.$$

Dans la pratique, l'erreur de prédiction OOB permet de se passer d'un échantillon de validation, tout comme de la validation croisée, pour estimer l'erreur de prédiction. Les résultats fournissent généralement une borne supérieure proche de l'erreur de test.
Le classifieur OOB procure également d'autres avantages comme la possibilité d'estimer le biais (dans le cas de la régression) de la forêt. L'estimateur permet, si le biais n'est pas trop petit, de réduire l'erreur de prédiction dans certains cas, en traitant le biais hors de la forêt aléatoire, par *post-processing*.

`Bornes de risque OOB`

si n et B sont assez grands, les estimateurs OOB agissent comme des bornes supérieures de l'erreur de test.
a) Dans le cas de la régression, on a :

$$\widehat{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X)) \leq \hat{\bar{\rho}}_{oob}\widehat{PE}(g_{\mathcal{P},oob}(X,\theta)),$$

avec

$$\widehat{PE}(g_{\mathcal{P},oob}(X,\theta)) = \frac{1}{\sum_{b=1}^{B}\mathbf{I}_{\left\{b\in G^-(x,B)\right\}}}\sum_{b=1}^{B}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left(Y_i - g^{(b)}_{\mathcal{P}}(X_i)\right)\mathbf{I}_{\left\{b\in G^-(x,B)\right\}}\right)^2\right).$$

$\widehat{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X))$ *est l'estimateur de l'erreur de prédiction théorique de la forêt (uniformément) aléatoire et constitue l'erreur de prédiction que l'on peut espérer atteindre sur les données de test.*

Lorsque l'erreur quadratique moyenne, $\overline{PE}^*(\bar{g}^{(B)}_{\mathcal{P}}(X))$, lui est supérieure sur les données de test, il est généralement possible d'améliorer cette dernière en agissant sur les arbres de décision. Idéalement, $\overline{PE}^*(\bar{g}^{(B)}_{\mathcal{P}}(X)) \leq \hat{\bar{\rho}}_{oob}\widehat{PE}(g_{\mathcal{P},oob}(X,\theta))$ mais la convergence de l'erreur quadratique moyenne vers l'erreur de prédiction théorique de la forêt uniformément aléatoire n'est pas établie. On peut cependant obtenir une borne de risque non asymptotique de l'erreur quadratique moyenne grâce à l'évaluation OOB.

Proposition 5.
On considère un échantillon d'entraînement $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$, $Y \in \mathbb{R}$, *associé au classifieur* $\bar{g}^{(B)}_{\mathcal{P},oob}$. *On suppose un échantillon de test* D_{N-n}, *de taille* $N-n$ *et associé au classifieur* $\bar{g}^{(B)}_{\mathcal{P}}$, *de même distribution que* D_n *et dont les valeurs prises par* Y *sont inconnues. Pour chacun des échantillons, on suppose que :*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \bar{g}^{(B)}_{\mathcal{P},oob}(X_i)\mathbf{I}_{\{(X_i,Y_i)\in D_n\}} \approx \frac{1}{N-n}\sum_{i=n+1}^{N} Y_i \bar{g}^{(B)}_{\mathcal{P}}(X_i)\mathbf{I}_{\{(X_i,Y_i)\in D_{N-n}\}}, \tag{3.6}$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 \approx \frac{1}{N-n}\sum_{i=n+1}^{N} Y_i^2. \tag{3.7}$$

Si n *et* B *sont assez grands et si*

$$\left|\frac{1}{N-n}\sum_{i=n+1}^{N} \bar{g}^{(B)}_{\mathcal{P}}(X_i)\right| < \left|\frac{1}{n}\sum_{i=1}^{n} \bar{g}^{(B)}_{\mathcal{P},oob}(X_i)\right| \quad et \quad \widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}^{(B)}_{\mathcal{P}}(X)) < \widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}^{(B)}_{\mathcal{P},oob}(X)), \tag{3.8}$$

alors, pour n'importe quel échantillon de test assez grand,

$$\overline{PE}^{*}(\bar{g}^{(B)}_{\mathcal{P}}(X)) \leq \overline{PE}^{*}(\bar{g}^{(B)}_{\mathcal{P},oob}(X)).$$

Preuve.
On rappelle que $\overline{PE}^{*}(\bar{g}^{(B)}_{\mathcal{P},oob}(X))$ est l'erreur de prédiction OOB calculée sur l'échantillon d'entraînement et correspondant à une estimation de l'erreur de prédiction sur les données de test, $\overline{PE}^{*}(\bar{g}^{(B)}_{\mathcal{P}}(X))$. L'erreur de prédiction s'écrit :

$$PE(\bar{g}^{(B)}_{\mathcal{P}}(X)) = \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right)^2,$$

et sa décomposition donne :

$$\begin{aligned}
&PE(\bar{g}^{(B)}_{\mathcal{P}}(X)) \\
&\quad = \mathbf{Var}(Y) + \left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left[Y - \mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right]\right\}^2 + \mathbf{Var}_{\mathbf{X}}\left(\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right) - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta), Y) \\
&\quad = \mathbf{E}(Y^2) + \left(\mathbf{E}_{\mathbf{X},\mathbf{Y}}\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right)^2 - 2\mathbf{E}_{\mathbf{X},\mathbf{Y}}\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\mathbf{E}(Y) + \mathbf{Var}_{\mathbf{X}}\left(\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right) \\
&\qquad - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta), Y) \\
&\quad = \mathbf{E}(Y^2) + \left(\mathbf{E}_{\mathbf{X},\mathbf{Y}}\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right)^2 + \mathbf{Var}_{\mathbf{X}}\left(\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right) - 2\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta) Y\right).
\end{aligned}$$

On obtient :

$$PE^{*}(\bar{g}^{(B)}_{\mathcal{P}}(X)) = \mathbf{E}(Y^2) - 2\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y \bar{g}^{(B)}_{\mathcal{P}}(X)\right\} + \left\{\mathbf{E}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right)\right\}^2 + \mathbf{Var}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right).$$

Sous les relations (3.6), (3.7) et (3.8), on en déduit, pour les versions empiriques, $\overline{PE}^{*}(\bar{g}^{(B)}_{\mathcal{P}}(X)) \leq \overline{PE}^{*}(\bar{g}^{(B)}_{\mathcal{P},oob}(X))$. □

b) Dans le cas de la classification, nous fournissons également le même type de borne pour l'erreur de test. A la différence de la régression, elle est ici reliée à la borne de risque de Breiman, donnée, dans la version OOB, par :

$$\overline{PE}^{*(B)}_{oob} \leq \frac{\hat{\bar{\rho}}_{oob}(1 - \hat{s}^2_{oob})}{\hat{s}^2_{oob}}.$$

Proposition 6.
On considère un échantillon d'entraînement $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$, avec $Y \in \{0, 1\}$, associé au classifieur $\bar{g}^{(B)}_{\mathcal{P},oob}$. On suppose un échantillon de test D_{N-n}, de taille $N - n$, associé au classifieur $\bar{g}^{(B)}_{\mathcal{P}}$ et de même distribution que D_n, dont les classes correspondant à chaque observation sont inconnues. Pour chacun des échantillons, on suppose que :

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{Y_i=1\}} \approx \frac{1}{N-n}\sum_{i=n+1}^{N} \mathbf{I}_{\{Y_i=1\}}, \tag{3.9}$$

$$\hat{\rho}_{\bar{g}^{(B)}_{\mathcal{P}}(X),Y} \approx \hat{\rho}_{\bar{g}^{(B)}_{\mathcal{P},oob}(X),Y}. \tag{3.10}$$

Si n et B sont assez grands et si

$$\sqrt{\widehat{\mathbf{Var}_{\mathbf{X}}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right)} - \sqrt{\widehat{\mathbf{Var}_{\mathbf{X}}}\left(\bar{g}^{(B)}_{\mathcal{P},oob}(X)\right)} > \frac{\left(1 - \frac{2}{n}\sum_{i=1}^{n} \mathbf{I}_{\{Y_i=1\}}\right)\left(\frac{1}{N-n}\sum_{i=n+1}^{N} \mathbf{I}_{\left\{\bar{g}^{(B)}_{\mathcal{P}}(X_i)=1\right\}} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\left\{\bar{g}^{(B)}_{\mathcal{P},oob}(X_i)=1\right\}}\right)}{2\hat{\rho}\sqrt{\widehat{\mathbf{Var}}(Y|D_n)}} \tag{3.11}$$

alors, pour n'importe quel échantillon de test assez grand,

$$\overline{PE}^{*} \leq \overline{PE}^{*(B)}_{oob}.$$

Preuve.
La décomposition biais-variance-covariance dans la cas de la classification binaire, à valeurs dans $\{0, 1\}$, est associée à l'erreur de test et donnée par :

$$\begin{aligned}
&\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X) \neq Y\right) \\
&\quad = \mathbf{Var}(Y) + \left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \bar{g}^{(B)}_{\mathcal{P}}(X)\right)\right\}^2 + \mathbf{Var}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right) - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X), Y\right),
\end{aligned}$$

avec

$$\begin{aligned}
\mathbf{Var}(Y) &= \mathbf{P}(Y=1)\mathbf{P}(Y=0), \\
\left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - \bar{g}^{(B)}_{\mathcal{P}}(X))\right\}^2 &= \left\{\mathbf{P}(Y=1) - \mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1)\right\}^2, \\
\mathbf{Var}_{\mathbf{X}}(\bar{g}^{(B)}_{\mathcal{P}}(X)) &= \mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=0)\mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1), \\
\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\bar{g}^{(B)}_{\mathcal{P}}(X), Y) &= \mathbf{E}\left\{\left[\bar{g}^{(B)}_{\mathcal{P}}(X) - \mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1)\right][Y - \mathbf{P}(Y=1)]\right\}.
\end{aligned}$$

Il s'en suit que :

$$\begin{aligned}
&\mathbf{Var}(Y) + \left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - \bar{g}^{(B)}_{\mathcal{P}}(X))\right\}^2 \\
&\quad = \mathbf{P}(Y=1)\mathbf{P}(Y=0) + \{\mathbf{P}(Y=1)\}^2 - 2\mathbf{P}(Y=1)\mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1) + \left\{\mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1)\right\}^2 \\
&\quad = \mathbf{P}(Y=1)\left(\mathbf{P}(Y=0) + \mathbf{P}(Y=1) - 2\mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1)\right) + \left\{\mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1)\right\}^2 \\
&\quad = \mathbf{P}(Y=1)\left(1 - 2\mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1)\right) + \left\{\mathbf{P}(\bar{g}^{(B)}_{\mathcal{P}}(X)=1)\right\}^2,
\end{aligned}$$

et

$$\begin{aligned}
&\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\bar{g}_{\mathcal{P}}^{(B)}(X), Y) \\
&\quad = \mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{P}(Y=1)\mathbf{E}\left\{\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\mathbf{E}\left\{Y\right\} \\
&\quad\quad + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\mathbf{P}(Y=1) \\
&\quad = \mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{P}(Y=1)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1).
\end{aligned}$$

On en déduit :

$$\begin{aligned}
&\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) \\
&\quad = \mathbf{P}(Y=1)\left(1 - 2\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right) + \left\{\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right\}^2 + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=0)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) \\
&\quad\quad - 2\left(\mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{P}(Y=1)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right) \\
&\quad = \mathbf{P}(Y=1) + \left\{\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right\}^2 + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=0)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - 2\mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\}.
\end{aligned}$$

D'où

$$\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) = \mathbf{P}(Y=1) + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - 2\mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\},$$

avec

$$\begin{aligned}
\mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} &= \mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\bar{g}_{\mathcal{P}}^{(B)}(X), Y) + \mathbf{E}(Y)\mathbf{E}(\bar{g}_{\mathcal{P}}^{(B)}(X)) \\
&= \rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} + \mathbf{E}(Y)\mathbf{E}(\bar{g}_{\mathcal{P}}^{(B)}(X)) \\
&= \rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} + \mathbf{P}(Y=1)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1),
\end{aligned}$$

où ρ est le coefficient de corrélation entre $\bar{g}_{\mathcal{P}}^{(B)}(X)$ et Y. On obtient :

$$\begin{aligned}
&\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) \\
&\quad = \mathbf{P}(Y=1) + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)(1 - 2\mathbf{P}(Y=1)) - 2\rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))}.
\end{aligned}$$

$\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) - \mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P},oob}^{(B)}(X) \neq Y\right) \leq 0$ équivaut à écrire :

$$\begin{aligned}
&\left(\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - \mathbf{P}(\bar{g}_{\mathcal{P},oob}^{(B)}(X)=1)\right)(1 - 2\mathbf{P}\left(Y=1\right)) \\
&\qquad\qquad -2\sqrt{\mathbf{Var}(Y)}\left(\rho\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} - \rho_{oob}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))}\right) \leq 0.
\end{aligned}$$

Finalement,

$$\begin{aligned}
&\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) - \mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P},oob}^{(B)}(X) \neq Y\right) \leq 0 \ si \\
&\rho\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} - \rho_{oob}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))} > \frac{(1 - 2\mathbf{P}\left(Y=1\right))\left(\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - \mathbf{P}(\bar{g}_{\mathcal{P},oob}^{(B)}(X)=1)\right)}{2\sqrt{\mathbf{Var}(Y)}}.
\end{aligned}$$

Sous les relations (3.9) et (3.10), on en déduit la relation (3.11) et, pour les versions empiriques de l'erreur de prédiction, $\overline{PE}^{*} \leq \overline{PE}_{oob}^{*(B)}$.□

`Remarques:` les (estimateurs des) deux bornes de risque de Breiman ne sont pas toujours vérifié(e)s et peuvent conduire à une confusion avec les bornes de risque OOB présentées dans les propositions 5 et 6. Cela est le cas lorsque les classes sont déséquilibrées et dans un certain nombre d'autres situations.
Pour la classification, lorsque l'erreur OOB est supérieure à la borne de Breiman, la forêt (uniformément) aléatoire n'est pas paramétrée de manière optimale et il existe un risque de sur-apprentissage du modèle. Il convient alors de contrôler les hyper-paramètres du modèle. Ces derniers désignent les mécanismes qui agissent sur tous les arbres, du fait de leur caractère aléatoire, afin de stabiliser ou de minimiser l'erreur de prédiction. Citons, par exemple, le rééquilibrage de classes (Chen et al., 2004) ou leur perturbation. La méthode la plus simple est, généralement, la réduction du nombre de variables candidates à la construction de chaque région d'un arbre. Plus ce nombre diminue, plus le caractère aléatoire de la forêt s'accroît, ce qui réduit fortement le risque de sur-apprentissage (mais tend à augmenter l'erreur de prédiction).
Dans la littérature sur les forêts aléatoires, la borne de risque de Breiman n'est pas unanimement perçue comme optimale. Sur les différents jeux de données que nous avons évalués, elle s'est révélée proche de l'erreur de test lorsque le nombre d'observations dans l'échantillon d'entraînement était suffisant ($>$500), avec les paramètres par défaut du modèle. Lorsque le nombre d'observations est trop petit, que les classes d'un problème sont déséquilibrées ou que les variables du problème sont trop corrélées, une calibration des erreurs est nécessaire pour obtenir une valeur optimale de la borne de Breiman.
Dans le cas de la classification, le résultat de la calibration a lieu sur l'échantillon d'entraînement en comparant l'erreur OOB à la borne de risque de Breiman. Celle-ci doit être supérieure à la première et deux des façons les plus simples pour l'effectuer sont la réduction du nombre de variables candidates à la construction de chaque région d'un arbre, et, le cas échéant, le rééquilibrage des classes.

Dans le cas de la régression, la borne de risque de Breiman ne concerne que l'erreur de prédiction théorique de la forêt aléatoire. Plus précisément, la borne de risque n'est pas une borne supérieure de l'erreur quadratique moyenne. Elle détermine les améliorations possibles dans le modèle, en la comparant avec l'erreur quadratique moyenne OOB. Un échantillon de validation peut, ici, être utile pour la calibration des paramètres, conjointement à l'utilisation de l'erreur OOB comme borne supérieure de l'erreur quadratique moyenne (sur les données de test).

De manière générale, la borne de risque de Breiman nous paraît nécessaire pour optimiser ou garantir les performances sur des problématiques industrielles, sensibles ou de long terme, mettant en jeu des forêts aléatoires. En particulier, elle permet, dans le cas des forêts uniformément aléatoires, d'unifier les différentes mesures d'erreur. Notons également que, dans le cas de la classification, la borne de risque introduit le problème de la vitesse de convergence comme un problème numérique. Comme c'est une borne supérieure, elle implique que toute autre borne de risque dépendant de n doive lui être inférieure. De même, une telle borne ne pourrait, bien sûr, être plus petite que l'erreur de prédiction de la forêt (si B est assez grand) puisque celle-ci converge vers la vraie erreur du modèle.

3.5.4 Sélection de variables

Les forêts aléatoires sont fréquemment utilisées quand la dimension de l'espace d'entrée des variables est grande. Elles permettent, notamment, de définir les variables les plus importantes du phénomène observé. Biau (2012) présente un modèle légèrement différent de la version de Breiman et montre que la vitesse de convergence de la forêt ne dépend que des variables les plus importantes, quel que soit leur nombre total. Nous considérons ici la classification, pour laquelle le gain d'information (la fonction $\texttt{IG}$) sélectionne les variables les plus informatives et donc les plus décisives, et supposons qu'il n'y a pas de biais car les points de coupure sont tirés aléatoirement. Notons que l'absence de biais de sélection est conditionnée à la capacité prédictive. Strobl et al. (2008) montrent que les forêts aléatoires peuvent présenter un biais dans la sélection des variables les plus importantes. Plus précisément, lorsque des variables explicatives sont corrélées ou lorsqu'elles sont de nature différente (continues et catégorielles par exemple) la sélection nécessite d'être attentif au nombre de variables participant à la construction de chaque région d'un arbre. Notre point de vue consiste à indiquer que les variables les plus décisives dans une forêt uniformément aléatoire sont celles qui contribuent le plus aux capacités de prédiction de l'algorithme mais pas nécessairement celles qui permettent d'interpréter au mieux le phénomène observé. Pour chaque variable de chaque région (et de chaque arbre), le score cumulé des gains d'information fournit le degré d'importance. Plus la variable est importante, plus son score est élevé. Nous posons $\texttt{VI}(j)$, le score d'importance de la j-ème variable. Pour la forêt uniformément aléatoire, et dans le cas de la classification, on a :

$$\texttt{VI}(j^*) = \sum_{b=1}^{B} \sum_{l=1}^{k} \texttt{IG}_{b,l}(j^*, D_n),$$

et, dans le cas de la régression :

$$\texttt{VI}(j^*) = \sum_{b=1}^{B} \sum_{l=1}^{k} \texttt{L}_{2_{b,l}}(j^*, D_n),$$

où j^* est l'indice de la variable ayant eu les plus grands gains d'informations parmi les régions candidates, à chaque étape de la construction des arbres de décision.

Le rapport $\texttt{VI}(j)/\sum_{j=1}^{d} \texttt{VI}(j)$ mesure l'influence relative des variables et, lorsque d est grand, permet la réduction de dimension. Cette dernière est validée en mesurant l'erreur de prédiction sur un échantillon de validation une fois la réduction effectuée. Dans le cas de données médicales mettant en jeu des milliers de gènes, la sélection de variables par une forêt uniformément aléatoire puis la classification par une forêt aléatoire de Breiman fournit une alternative à la réduction de dimension par d'autres méthodes. Nombre d'articles pointent le manque "d'interprétabilité" des forêts aléatoires. Nous renvoyons le lecteur à la page personnelle de Breiman et Cutler
(http ://www.stat.berkeley.edu/∼breiman/RandomForests/cc_home.htm). De nombreux éléments y sont développés pour une meilleure interprétation des résultats fournis. Notons également que l'importance des variables est directement liée à la diversité des observations de chacune d'elles. Moins il y a de dispersion entre les observations, moins la variable correspondante sera choisie par l'algorithme. Une conséquence indirecte de

ce phénomène apparaît dans le cas des matrices creuses : Les données constantes trop nombreuses tendent à être éliminées (ainsi que les variables pour la partition associée) par l'algorithme très rapidement au profit de celles où la dispersion reste élevée. Cette propriété permet une adaptation naturelle aux problèmes de type *sparse* et peut être vue comme une réduction de dimension implicite.

`Sélection locale de variables`

Une forêt uniformément aléatoire permet également d'effectuer une sélection de variables localement, en liant chaque observation de l'échantillon aux variables les plus importantes de la prédiction associée et à leurs interactions. Cette méthode restreint les variables à celles qui ont un lien immédiat avec la règle de décision appliquée à l'observation x. Pour chacune d'elles, on retient la variable associée à la prédiction pour chaque arbre de la forêt aléatoire. Pour l'ensemble des observations et des arbres, la variable la plus importante est alors celle qui est observée le plus grand nombre de fois. Cette opération est effectuée, à nouveau, en masquant l'une après l'autre chaque variable. La sélection locale donne plus d'influence aux variables les plus prédictives et à celles dont les interactions avec d'autres sont les plus importantes, tout en permettant une interprétation plus précise du phénomène observé. Elle a également la particularité d'être applicable aussi bien à l'échantillon d'entraînement qu'à celui de test.

`Définition.` *Une variable est importante (localement) au premier ordre si, pour une même observation, et pour tous les arbres, elle a la plus grande fréquence d'apparition dans une région terminale.*

Posons $\texttt{LVI}^{(b)}(j,i)$, le score d'importance locale de la réalisation i et de la variable d'indice j de X, pour le b-ème arbre,
$\texttt{LVI}(j,i)$ le score pour tous les arbres de la réalisation i et de la variable d'indice j de X,
$\texttt{LVI}(j,.)$, le score de la variable d'indice j de X,
$R(j,\alpha_j) = \{X | X^{(j)} \leq \alpha_j\}$, une région candidate, et terminale, de l'arbre dont toutes les observations sont inférieures à α_j, pour la variable $X^{(j)}$, $R^C(j,\alpha_j) = \{X | X^{(j)} > \alpha_j\}$, la région complémentaire de R. On définit $\texttt{LVI}^{(b)}(j,i)$ par :

$$\texttt{LVI}^{(b)}(j,i)$$
$$= \mathbf{I}_{\{\texttt{IG}_b(j^*,D_n) \geq \texttt{IG}_b(j,D_n)\}} \left(\mathbf{I}_{\left\{g_{\mathcal{P}}^{(b)}(X_i,R(j^*,\alpha_{j^*}))=g_{\mathcal{P}}^{(b)}(X_i,R(j,\alpha_j))\right\}} + \mathbf{I}_{\left\{g_{\mathcal{P}}^{(b)}(X_i,R^C(j^*,\alpha_{j^*}))=g_{\mathcal{P}}^{(b)}(X_i,R^C(j,\alpha_j))\right\}} \right).$$

Comme la règle de décision n'est appliquée qu'aux régions terminales, l'indicatrice associée établit que l'on se trouve dans une telle région, pour laquelle on compare la règle de décision de la variable (d'indice) j avec celle définie comme optimale par l'arbre. On agit de même avec la fonction `IG` afin d'éviter les confusions lorsque la règle de décision assigne la même classe à plusieurs variables.
Pour la j-ème variable de X et une observation i, on a :

$$\texttt{LVI}(j,i) = \sum_{b=1}^{B} \texttt{LVI}^{(b)}(j,i).$$

Le score sur tous les arbres et toutes les observations de la variable $X^{(j)}$ est donné par :

$$\texttt{LVI}(j,.) = \sum_{i=1}^{n} \texttt{LVI}(j,i).$$

On obtient ainsi le score d'importance locale de chaque variable. Un avantage de la sélection locale de variables est son comportement en termes de dynamique. Par opposition à la sélection basée sur le gain d'information, qui peut être vue comme une sélection globale, l'importance locale s'exprime aussi relativement aux données de test ce qui permet de tenir compte d'interactions absentes de l'échantillon initial. Plus précisément, l'importance globale des variables est définie sur le modèle ; l'importance locale y ajoute une prise en compte plus explicite des données.

`Interactions, importance et dépendance partielles`

Pour une meilleure interprétation des résultats d'une forêt uniformément aléatoire, on peut s'intéresser aux interactions entre variables ainsi qu'à la dépendance de chacune d'elles avec la variable à expliquer (Y), en tenant compte de l'ensemble des variables du problème.
Notons $\texttt{LVI}_{(1)}(j,i), \texttt{LVI}_{(2)}(j,i), ..., \texttt{LVI}_{(d)}(j,i)$ les pseudo-statistiques d'ordre de $X^{(j)}$ pour tous les arbres et pour l'observation i. Nous indiquons le terme *pseudo* car ces statistiques d'ordre ne portent pas sur les observations de X mais sur une fonction de score (`LVI`) dont les réalisations sont définies à partir de X. On souhaite qu'en première position corresponde la plus grande valeur de $\texttt{LVI}(j,i)$. De manière plus générale, on définit alors $\texttt{LVI}_{(d-q+1)}(j,i), 1 \leq q \leq d$, le score d'importance d'ordre $d-q+1$, et de position q, de $X^{(j)}$ associé à sa i-ème réalisation. Par exemple, $\texttt{LVI}_{(d-1)}(j,i)$ est le score d'importance de $X^{(j)}$ et de sa i-ème réalisation en considérant la fréquence d'apparition de $X^{(j)}$ dans une région terminale à la deuxième position, $q = 2$, la première étant la plus grande. On souhaite ensuite, pour chaque variable, connaître le score d'importance de position q pour toutes les réalisations. Notons $\texttt{SLI}_q(j,.)$ ce score. On a :

$$\texttt{SLI}_q(j,.) = \sum_{i=1}^{n} \texttt{LVI}_{(d-q+1)}(j,i).$$

On peut également définir la notion d'`importance partielle`. *Une variable est (partiellement) importante si, pour une même observation et à tous les ordres, elle a la plus grande fréquence d'apparition dans une région terminale.* Le score d'importance partielle est donné par :

$$\texttt{SLI}(j,.) = \sum_{q=1}^{d} \texttt{SLI}_q(j,.) \tag{3.12}$$

L'importance partielle a un intérêt pratique. Par exemple, dans le cas de la classification, on peut isoler, pour chaque classe, les variables les plus décisives. Dans le cas de la régression, on peut s'intéresser aux variables qui contribuent le plus à de grandes valeurs de la variable à expliquer.

Dans le cas de la recherche d'interactions, pour chaque variable, on obtient plusieurs scores correspondant aux différentes valeurs de q. A la position 1, correspond le facteur d'importance le plus élevé. A chaque nouvelle position, ce facteur perd en importance. Pour chacun des facteurs, on peut classer les variables selon leur score et mesurer leurs interactions grâce à un tableau de contingence.

`Définition.` *Une variable interagit avec une autre si, pour une même observation, et pour tous les arbres, les deux ont respectivement la première et la seconde plus grande fréquence d'apparition dans une région terminale.*

On peut remarquer que les interactions ne sont ici envisagées que pour les variables qui apparaissent le plus souvent (en première et deuxième position) dans une région terminale. En particulier, pour deux variables j et j' de X, on note $\texttt{VII}_{(1,2)}(j, j')$, les interactions d'ordre 1 et 2, définie par :

$$\texttt{VII}_{(1,2)}(j, j') = \frac{\texttt{SLI}_1(j, .) + \texttt{SLI}_2(j', .)}{n}.$$

Les interactions, une fois standardisées, ont pour principal avantage d'évaluer à la fois la dépendance locale entre variables et leur influence sur la variable à expliquer. L'interprétation du phénomène observé est alors plus complète grâce au lien entre variables influentes et variables d'importance moindre.

Un des inconvénients des méthodes non paramétriques est la difficulté à établir un lien entre variable à expliquer et variables explicatives, comme il est possible de le faire dans une régression linéaire, par exemple. Les forêts aléatoires permettent d'obtenir ce type de relation en produisant un graphe de `dépendance partielle` qui permet de mieux visualiser la relation de chaque variable explicative avec celle à expliquer. La dépendance partielle de position q, notée $\texttt{pD}_{(q)}$, est définie par :

$$\texttt{pD}_{(q)}(Y, X) = \left\{ \left(\left(\bar{g}_{\mathcal{P}}^{(B)}(X_i, R(j^*, \alpha_{j^*})), X_i^{(j^*)} \right) | \texttt{SLI}_q(j^*, i) > 0 \right), 1 \leq i \leq n \right\},$$

où j^* est la variable d'indice j de X pour laquelle la règle de décision $\bar{g}_{\mathcal{P}}^{(B)}$ de la forêt aléatoire est définie. On note $\texttt{Imp}(d)$, l'ensemble des variables localement importantes parmi les d variables du problème ; j^* appartient alors nécessairement à $\texttt{Imp}(d)$.

La dépendance partielle est définie conditionnellement à l'information de position q contenue dans $\texttt{SLI}_q(j^*, i)$. Celle-ci doit être positive. Lorsque c'est le cas, $X^{(j^*)}$ s'exprime à travers la prédiction associée à l'observation i pour, au moins, un arbre (à la position q). Moins la variable $X^{(j^*)}$ est importante, moins il y a de prédictions qui lui sont associées. Dans ce cas, il faut alors s'intéresser à plusieurs ordres d'importance, afin de vérifier que cette variable ne détermine aucune tendance, et la dépendance partielle se récrit en tenant compte de cette possibilité. On définit la dépendance partielle à tous les ordres, pD, par :

$$\texttt{pD}(Y, X) = \left\{ \left(\left(\bar{g}_{\mathcal{P}}^{(B)}(X_i, R(j^*, \alpha_{j^*})), X_i^{(j^*)} \right) \middle| \sum_{q=1}^{d} \texttt{SLI}_q(j^*, i) > 0 \right), 1 \leq i \leq n \right\}.$$

Ainsi on dispose de, virtuellement, plus de points pour visualiser la relation entre variable à expliquer et variable explicative. La dépendance partielle mesure cette relation en tenant compte de l'ensemble des variables, du fait de la règle de décision de la forêt aléatoire, et s'interprète de la même manière que les coefficients d'une régression linéaire, et plus simplement encore par une visualisation. Notons que le *Stochastic Gradient Boosting* (Friedman, 2002) est un des premiers modèles à proposer un modèle de la dépendance partielle. Notre approche est différente de celle de Friedman mais produit des résultats similaires.

3.5.5 Prédictions, extrapolation et valeurs extrêmes

Contrairement aux modèles paramétriques, il est plus délicat d'extrapoler, dans le cas de la régression, avec une forêt aléatoire, lorsque les réalisations de X se trouvent hors du support observé jusque là. Si X, pour les données de test, est en dehors des valeurs prises dans D_n, la prédiction risque d'être éloignée de la réalité. Une solution est de perturber Y dans D_n mais on n'a moins de contrôle sur le modèle ; une alternative consiste à paramétrer la dépendance partielle pour les variables les plus importantes. Pour des réalisations de X non comprises dans le support observé jusqu'alors, ou bien extrêmes, la prédiction est alors donnée par la moyenne des prédictions de chaque modèle paramétrique et de la règle de décision de la forêt aléatoire. Notons que dans le cas de valeurs extrêmes, il est préférable de s'intéresser également aux bornes des intervalles de confiance et à leur voisinage. La prédiction prend alors une forme différente et peut, par exemple, être donnée avec une certaine probabilité dans un intervalle et avec une autre probabilité (beaucoup plus faible) dans un autre intervalle, mais avec intensité bien plus grande. Pour chaque variable importante, j^*, du problème, nous considérons le modèle linéaire simple, pour $X^{(j^*)}$, défini par :

$$\left(\bar{g}_{\mathcal{P}}^{(B)}(X, R(j^*, \alpha_{j^*}))\,\middle|\,\sum_{q=1}^{d}\texttt{SLI}_q(j^*,.)>0\right) = a^{(j^*)} + b^{(j^*)}\left(X^{(j^*)}\,\middle|\,\sum_{q=1}^{d}\texttt{SLI}_q(j^*,.)>0\right)+\epsilon,$$

où $\epsilon \sim \mathcal{N}(0,\sigma^2)$.
Nous en déduisons un estimateur partiel de Y pour chaque $X^{(j^*)}$, noté $\hat{Y}^{(j^*)}$, et défini par :

$$\hat{Y}^{(j^*)} = \hat{a}^{(j^*)} + \hat{b}^{(j^*)}\left(X^{(j^*)}\,\middle|\,\sum_{q=1}^{d}\texttt{SLI}_q(j^*,.)>0\right),$$

avec

$$\hat{a}^{(j^*)} = \frac{1}{\sum_{i=1}^{n}\mathbf{I}_{\left\{\sum_{q=1}^{d}\texttt{SLI}_q(j^*,i)>0\right\}}}\sum_{i=1}^{n}\bar{g}_{\mathcal{P}}^{(B)}(X_i, R(j^*,\alpha_{j^*}))\mathbf{I}_{\left\{\sum_{q=1}^{d}\texttt{SLI}_q(j^*,i)>0\right\}}$$
$$-\hat{b}^{(j^*)}\frac{1}{\sum_{i=1}^{n}\mathbf{I}_{\left\{\sum_{q=1}^{d}\texttt{SLI}_q(j^*,i)>0\right\}}}\sum_{i=1}^{n}X_i^{(j^*)}\mathbf{I}_{\left\{\sum_{q=1}^{d}\texttt{SLI}_q(j^*,i)>0\right\}}.$$

et

$$\hat{b}^{(j^*)} = \frac{\mathbf{Cov}\left(\left(X^{(j^*)}, \bar{g}_{\mathcal{P}}^{(B)}(X, R(j^*,\alpha_{j^*}))\right)\,|\,\sum_{q=1}^{d}\texttt{SLI}_q(j^*,.)>0\right)}{\mathbf{Var}\left(X^{(j^*)}|\,\sum_{q=1}^{d}\texttt{SLI}_q(j^*,.)>0\right)}.$$

On obtient alors l'estimateur de Y pour les valeurs extrêmes, ou en dehors du support de X pendant la phase d'apprentissage. Il est noté $\bar{g}^{(B)}_{\mathcal{P},ev}$ et donné par :

$$\bar{g}^{(B)}_{\mathcal{P},ev}(x) = \frac{1}{1+\sum_{j=1}^{d}\mathbf{I}_{\{j\in \texttt{Imp}(d)\}}\mathbf{I}_{\left\{\hat{b}^{(j)}\neq 0\right\}}}\left(\sum_{j=1}^{d}\left(\hat{y}^{(j)}\mathbf{I}_{\{j\in \texttt{Imp}(d)\}}\mathbf{I}_{\left\{\hat{b}^{(j)}\neq 0\right\}}\right)+\bar{g}^{(B)}_{\mathcal{P}}(x)\right).$$

La règle de décision pour les valeurs extrêmes, $\bar{g}^{(B)}_{\mathcal{P},ev}$, est identique à la règle de décision de la forêt uniformément aléatoire s'il n'existe aucune dépendance partielle linéaire, croissante ou décroissante, entre Y et X. Un test de Fischer permet de s'assurer de la non nullité de $\hat{b}^{(j)}$.
$\bar{g}^{(B)}_{\mathcal{P},ev}$ est utilisable dès qu'une observation de la variable j^*, de X, est en dehors du support observé dans D_n. Dans ce type de situation, l'extrapolation est accompagnée d'un intervalle de confiance. Celui-ci est défini, spécifiquement, ici à partir de $\underset{j\in \texttt{Imp(d)}}{\texttt{max}}\ \hat{Y}^{(j)}$ si l'observation x est la plus grande observée relativement à D_n, et $\underset{j\in \texttt{Imp(d)}}{\texttt{min}}\ \hat{Y}^{(j)}$ si elle est la plus petite.

3.5.6 Intervalles de prédiction et de confiance

Dans le cas de la régression, la structure des forêts aléatoires permet de définir simplement des intervalles de prédiction pour chaque réalisation de Y à partir desquels on peut, éventuellement, construire un intervalle de confiance pour un paramètre. On note $\hat{q}_\alpha(g_{\mathcal{P}}(X_i,\theta))$, le quantile empirique d'ordre α, $0<\alpha<1$, de la distribution de $g_{\mathcal{P}}$ pour la i-ème observation.

Avec une probabilité $1-\alpha$*, et pour tout* $i\in[1,n]$*, l'intervalle de prédiction bootstrap pour chaque* Y_i *est tel que* :

$$Y_i \in \left[\hat{q}_{\alpha/2}(g_{\mathcal{P}}(X_i,\theta)), \hat{q}_{1-\alpha/2}(g_{\mathcal{P}}(X_i,\theta))\right].$$

Toutefois, cet intervalle de prédiction est généralement trop large (en particulier quand une seule valeur, comme 0, est très souvent présente dans les réalisations de Y). Nous fournissons une procédure qui définit un intervalle de prédiction moins large que le précédent tout en proposant des propriétés empiriques intéressantes.

Pour chaque Y_i, nous procédons comme suit :
- nous tirons, sans remise, S estimations de $G(X_i,B)=\{g^{(1)}_{\mathcal{P}}(X_i), g^{(2)}_{\mathcal{P}}(X_i),...,g^{(B)}_{\mathcal{P}}(X_i)\}$, où S est le nombre de valeurs uniques de $G(X_i,B)$. Puis, nous calculons un premier estimateur de R_i défini par :

$$\bar{g}^{(S)}_{\mathcal{P}}(X_i)=\frac{1}{S}\sum_{s=1}^{S}g^{(s)}_{\mathcal{P}}(X_i),$$

où $g^{(s)}_{\mathcal{P}}(X_i)$ est la valeur du s-ème arbre de décision, choisi aléatoirement, parmi les S.
- Cette opération est répétée K fois afin de constituer un sous-échantillon de taille K, $\{\bar{g}^{(S_1)}_{\mathcal{P}}(X_i), \bar{g}^{(S_2)}_{\mathcal{P}}(X_i),...,\bar{g}^{(S_K)}_{\mathcal{P}}(X_i)\}$.

- Nous considérons ensuite $\hat{q}_\alpha(\bar{g}_{\mathcal{P}}^{(S)}(X_i,))$, le quantile empirique d'ordre α de la distribution de $\bar{g}_{\mathcal{P}}^{(S)}$ pour les valeurs des K règles de décision de la i-ème observation.
- Ce quantile n'est pas directement utilisé pour encadrer Y_i. Dans le cas d'une distribution proche de la loi exponentielle, la concentration des petites valeurs de la variable à estimer est le principal problème. Dans le voisinage d'un quantile, nous supposons que la distribution est approximativement gaussienne et définissons l'intervalle de prédiction suivant :
avec une probabilité approchée $1-\alpha$,

$$Y_i \in \left[\hat{q}_{\alpha/2}(\bar{g}_{\mathcal{P}}^{(S)}(X_i)) + z_{\alpha/2}\sqrt{\frac{\widehat{\mathbf{Var}}_{\theta_\mathbf{S}}(g_{\mathcal{P}}(X_i,\theta_S))}{S}}, \hat{q}_{1-\alpha/2}(\bar{g}_{\mathcal{P}}^{(S)}(X_i)) + z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathbf{Var}}_{\theta_\mathbf{S}}(g_{\mathcal{P}}(X_i,\theta_S))}{S}}\right],$$

où $z_{\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la loi $\mathcal{N}(0,1)$,
et $\widehat{\mathbf{Var}}_{\theta_\mathbf{S}}(g_{\mathcal{P}}(X_i,\theta_S))$ est la variance empirique de la règle de décision dont les valeurs sont uniques pour X_i.

L'intervalle de prédiction donne, ici, un point de vue globalement plus réaliste que le précédent en construisant une borne inférieure dont le principal argument est d'être, par exemple, différent de 0 beaucoup plus souvent qu'en la construisant à partir de la distribution de $g_{\mathcal{P}}$. On obtient, par exemple, simplement un intervalle de confiance de $\bar{Y}$. Posons :

$$\tilde{q}_{\alpha/2}(g_{\mathcal{P}}(X_i,\theta)) = \hat{q}_{\alpha/2}(\bar{g}_{\mathcal{P}}^{(S)}(X_i)) + z_{\alpha/2}\sqrt{\frac{\widehat{\mathbf{Var}}_{\theta_\mathbf{S}}(g_{\mathcal{P}}(X_i,\theta_S))}{S}},$$

avec une probabilité approchée $1-\alpha$,

$$\bar{Y} \in \left[\frac{1}{n}\sum_{i=1}^{n}\tilde{q}_{\alpha/2}(g_{\mathcal{P}}(X_i,\theta)), \frac{1}{n}\sum_{i=1}^{n}\tilde{q}_{1-\alpha/2}(g_{\mathcal{P}}(X_i,\theta))\right].$$

- L'intervalle de confiance est ensuite validé grâce aux informations OOB ou bien avec un échantillon de validation.

3.5.7 *Big data* et inférence

De nombreux articles illustrent le rôle de plus en plus important, voire décisif, des très grands volumes de données, appelés plus communément *big data*, dans l'inférence. Pour les méthodes d'apprentissage statistique, deux aspects majeurs de ce rôle sont le lien avec la capacité de mieux prédire et celle, plus immédiate, de simplement pouvoir traiter, décrire et résumer de tels volumes. La parallélisme est un premier pas dans ce domaine et les forêts aléatoires s'y prêtent bien, puisque les arbres sont construits indépendamment les uns des autres. Nous nous intéressons aux modèles prédictifs, lorsque les données sont en très grand nombre, à travers deux aspects :
- sous leur forme spatiale, dans le cas où il y a un besoin (éventuel) d'exhaustivité.
- Sous leur forme temporelle, lorsque, par exemple, de nouvelles données à fort potentiel ne sont pas immédiatement disponibles.
Quand n est très grand, par exemple $n > 10^9$, aux difficultés calculatoires sévères

s'ajoutent des problèmes de stabilité de la prédiction. Dans le cas des forêts aléatoires, les arbres tendent alors à être très profonds sans que l'augmentation de leur nombre ne suffise à diminuer la variance de l'estimateur. Les problèmes calculatoires sont, eux, d'abord liés aux temps d'accès. Par exemple, sous le logiciel R, largement répandu dans la communauté scientifique, le stockage d'une matrice, avec $n = 10^7$ et $d = 100$, prend près de 8 Go. En y ajoutant les temps de calcul, les objets du *big data* se révèlent hors de portée d'un ordinateur personnel ou d'une station de travail. Même pour des centres de calcul, les ressources peuvent se révéler coûteuses, notamment du fait de la location des capacités de calcul (solution qui tend à remplacer l'achat de gros ordinateurs). Pour des grandes valeurs de n, Le *bootstrap* et le sous-échantillonnage, efficaces dans le cas des forêts (uniformément) aléatoires, doivent être réadaptés. Notons que la réduction de la dimension d n'est pas abordée, car elle est traitée implicitement via la sélection de variables. Parallèlement, réduire la dimension en même temps que n pose des problèmes importants d'optimisation et invalide la consistance proposée pour les forêts uniformément aléatoires.

Sur le plan théorique et pratique, Jordan, Kleiner, Sarkar et Talwalkar (2012) traitent le problème de l'estimation, lorsque n est très grand.
- Les n observations sont subdivisées en S sous-échantillons de taille fixe m.
- Dans une seconde étape, une méthode de Monte Carlo est associée à un tirage avec remise de n observations, parmi m, pour chacun des S échantillons. Cette dernière opération est répétée r fois de façon à construire un estimateur intermédiaire.
Il n'y a pas de problème calculatoire ici, malgré le tirage de n observations, car seules les m observations sont stockées.
- Dans la troisième étape, L'estimateur final est alors calculé en faisant la moyenne des S estimateurs intermédiaires calculés grâce à la méthode de Monte Carlo.
Dans leur approche, un point central est le choix de m, lequel ne doit être ni trop petit, au risque d'empêcher la convergence de l'estimateur final, ni trop grand, au risque de ne finalement pas réduire les temps de calcul. La complexité du problème est en $O(rmS)$ si la complexité d'un seul estimateur est en $O(m)$. Si on dispose d'un nombre d'unités de calcul U pouvant traiter le problème en mémoire, la complexité se transforme en $O(rm\frac{S}{U})$ et les temps de calcul sont beaucoup moins importants, notamment du fait de l'élimination (ou de la réduction) d'une grande partie des temps d'accès aux données.

Lorsque n est très grand, nous proposons une méthode reprenant une partie des éléments précédents pour l'estimation en utilisant une forêt uniformément aléatoire incrémentale, pour laquelle la dépendance intrinsèque aux observations est faible. Puisque l'arbre de décision est construit en partitionnant l'espace, une forêt incrémentale peut également l'être de la même manière :
i) partitionner l'espace ;
ii) pour chaque partition, construire une forêt aléatoire sans définir sa règle de décision ;
iii) agréger les forêts pour n'en former qu'une, en utilisant les règles de décision de tous les arbres.

- Dans la pratique, Il faut tout d'abord choisir m de sorte qu'une forêt aléatoire calculée sur un échantillon de taille m n'ait pas une erreur de prédiction trop grande face à celle sur l'ensemble des données, tout en réduisant significativement les temps de calcul.

- Tout comme Jordan et al., nous divisons les données en S sous-échantillons aléatoires de taille m. Chacun d'eux constitue une sous-partition dans laquelle nous créons une forêt uniformément aléatoire de B arbres.
- Les S forêts uniformément aléatoires incluent, par construction, une méthode de Monte Carlo interne. Mais, dans ce cas précis, leur règle de décision est en suspens, afin de bénéficier d'une corrélation moyenne plus faible dans l'étape finale. L'idée est ici de compenser la perte d'optimalité (la marge) induite par la réduction de n.
- Dans l'étape d'agrégation, pour les $B \times S$ arbres ensemble, la règle de décision de la forêt uniformément aléatoire incrémentale est appliquée.

Le corollaire à cette approche est que la règle de décision de chaque arbre est issue de la distribution des n observations, laquelle est répliquée par le tirage uniforme dans la construction des S sous-échantillons. Il est également important de conserver la philosophie de Breiman (grande variance des arbres et corrélation faible) tout en réduisant les temps de calcul, puisque construire $B \times S$ arbres uniformément aléatoires avec des échantillons de taille $m = \lceil n/S \rceil$ est beaucoup moins coûteux que de construire B arbres avec un échantillon de taille n, lorsque n est grand, car la profondeur des arbres est alors plus importante et impacte en cascade l'algorithme. Le problème du volume à traiter disparaît, puisque n'importe quelle unité de calcul peut maintenant traiter une partie des données de manière indépendante. A la place, apparaît un problème de nombre de ces unités et de précision de l'estimateur. La complexité est en $O(B\beta dm(S/U)\texttt{log}(S/U))$, où $B = r$ et d est la dimension du problème. Idéalement, $S \leq U$ et on peut poser $m = \lceil (n/U)^\gamma \rceil, 0 < \gamma < 1$. Pour $U = 1$, Jordan et al. proposent $\gamma = 0.7$ comme un choix adapté. Notons que dans ce cas, sous R, la matrice proposée en exemple ne prend plus que 600 Mo pour le calcul, soit 12 fois moins de place.

De la même manière que précédemment, on peut voir le problème des gros volumes comme un problème temporel. A chaque nouvelle période de temps, s'ajoute un (grand) nombre de données qui peuvent améliorer la prédiction ou la compréhension du problème. Dans ce cas, l'apprentissage statistique est incrémental et s'assimile à un apprentissage par l'expérience. Pour une forêt uniformément aléatoire, le paradigme précédent ne varie pas : comme la dépendance aux observations est faible, les nouvelles données sont vues simplement comme un nouvel échantillon de même distribution. Lorsque cette dernière change, il suffit alors de ne faire *voter* que les arbres les plus récents, mais d'autres techniques sont possibles. Dans ce nouvel échantillon tout comme dans celui qui le précède, chaque observation a la même probabilité d'appartenir à une région particulière.

Pour illustrer notre propos, nous nous intéressons à la détection des irrégularités aux cotisations sociales en France, pour laquelle moins de 10% de l'ensemble des déclarations de cotisation des entreprises est contrôlé chaque année et moins de 5% de l'ensemble se révèle être véritablement assimilable à des cas d'irrégularités de cotisations au détriment de la Sécurité sociale. En observant plus précisément ces 5%, un peu plus de la moitié correspond à un montant d'irrégularités assez significatif, soit des sommes supérieures à 1000 euros pour 3 années de cotisations contrôlées et contrôlables. La difficulté est, dans ce cas précis, la complexité à utiliser l'historique des données ; par obligation légale, seul tout ou partie des trois dernières années de déclarations peut être contrôlé(e). De plus,

même si l'utilisation de l'historique demeure possible, la masse des données tend à devenir trop importante au bout d'un certain temps. Il faut donc, chaque année, recalculer le modèle sur de nouvelles données et omettre les précédentes, à moins que l'algorithme ne puisse tirer profit des anciennes données pour améliorer ses capacités de prédiction.

Notons qu'ici, le caractère big data est un problème à la fois d'inférence et d'assimilation de données. On a $d > 1000$ et $n > 120000 \times T$, où T est le nombre de périodes pour lesquelles on obtient un nouveau jeu de données. T augmente chaque année et il faut intégrer toutes les périodes de manière automatique sans altérer l'inférence réalisée jusque là. Un algorithme qui prend en compte uniquement la dernière période dispose d'au plus 10% des données pour l'apprentissage et la validation, et 90% des données à évaluer. Un algorithme qui prend en compte toutes les périodes à la fois dispose d'un échantillon d'apprentissage T fois plus grand. En contrepartie, il ne peut pas s'adapter à des changements dans les données, lesquels apparaissent, par exemple, lorsque la législation en matière de cotisations sociale est modifiée. Une autre difficulté est l'*évolutivité* intrinsèque. La complexité de l'inférence augmente avec T alors que les unités de calcul sont en nombre limité.

La forêt uniformément aléatoire incrémentale permet *d'apprendre et de mémoriser* les données au fur et à mesure de leur arrivée. Le processus est un peu plus spécifique et utilise la règle de décision $\bar{g}^{(T)}_{\mathcal{P},inc}$. Il est défini comme suit :

i) construire une forêt uniformément aléatoire pour la période T_i et l'échantillon $D_{n,i}$;
ii) pour la nouvelle période T_{i+1}, construire une nouvelle forêt uniformément aléatoire avec l'échantillon $D_{n,i+1}$ (et éventuellement un sous-échantillon de $D_{n,i}$) ;
iii) évaluer les capacités (la mémoire) de la première forêt avec $D_{n,i+1}$. Comparer avec l'apprentissage effectué par la nouvelle forêt ;
iv) à la fin de la période T_{i+1}, si la mémoire est moins performante, fusionner les deux forêts et utiliser la règle de décision agrégée pour la nouvelle forêt ainsi formée ;
v) Si la forêt uniformément aléatoire a une trop grande taille ou si la distribution change, ne faire, éventuellement, voter qu'une partie des arbres de manière aléatoire ou (respectivement) les arbres les plus récents. Poser $i = i + 1$ et recommencer les étapes ii) à v).

En comparaison de $\bar{g}^{(B)}_{\mathcal{P},big}$, la règle de décision évolue, ici, au fil du temps. Dans le cas des cotisations sociales, cette approche améliore fortement les résultats ($> +10\%$) car on dispose en quelque sorte de, virtuellement, plus d'exemples d'entraînement, tout en conservant les propriétés importantes de la forêt aléatoire. Le point central est que les données observables de manière périodique sont vues par le modèle comme des versions de D_n avec la même distribution. Dans ce type d'approche, il est préférable de prêter attention au *pre-processing* initial des données. Les capacités de mémorisation tendent à beaucoup mieux s'exprimer et certains nouveaux échantillons d'entraînement peuvent même devenir facultatifs, la forêt aléatoire ayant d'aussi bons résultats sans eux. Cependant, malgré les propriétés d'invariance d'échelle des arbres (Devroye et al., 1996), une version incrémentale peut être sensible à des changements dans la structure des données.

3.6 Implémentations et aspects numériques

Nous décrivons dans cette section une implémentation numérique d'une forêt uniformément aléatoire. Dans sa version initiale, elle ne possède que peu de différences avec le point de vue théorique. Puis, nous indiquons deux applications dont le processus diffère de celles des forêts aléatoires de Breiman : la sélection de variables et le traitement de grands volumes de données. Nous terminons cette section en proposant plusieurs outils d'interprétation ainsi que des exemples sur des données synthétiques et réelles.

3.6.1 Algorithme

Dans les algorithmes de forêts aléatoires, plusieurs paramètres cohabitent et peuvent modifier assez sensiblement les performances. Deux d'entre eux sont très influents :
- le nombre d'arbres, B, de la forêt ;
- le nombre de variables admissibles pour la construction des régions de chaque arbre.
Empiriquement, à partir de 200 arbres, l'erreur de prédiction ne décroît plus que lentement et pour $B = 500$, les améliorations sont marginales. Lorsque n est grand, il convient d'essayer de plusieurs valeurs de B, en particulier lorsqu'on souhaite tester l'effet de certains paramètres. Pour des raisons liées aux temps de calcul, nous proposons, par défaut, $B = 100$.
Le nombre de variables admissibles pour la construction de chaque région demeure le paramètre le plus sensible aux résultats. Dans le cas des forêts aléatoires, et pour la classification, on tire généralement $\lceil\sqrt{d}\rceil$ à chaque étape du partitionnement. Dans le cas des forêts uniformément aléatoires, le nombre de variables admissibles dépend du paramètre β. A chaque étape du partitionnement, on tire aléatoirement, avec remise, $\lceil\beta d\rceil$ variables. Ainsi, à chaque étape, une même variable peut être tirée plusieurs fois. Par défaut $\beta = 4/3$. Il est également possible de spécifier un β aléatoire, généré par l'algorithme, d'effectuer un *Bagging* (pour la détermination de chaque région, toutes les variables sont utilisées) d'arbres de décision uniformément aléatoires, ou de construire une forêt totalement aléatoire, $\beta = 1/d$ (à chaque étape on tire une seule variable aléatoirement et un point de coupure, selon la loi uniforme sur le support de la variable).

`Remarque:` afin d'exploiter pleinement les capacités des forêts aléatoires, Breiman préconise un tirage avec remise (*bootstrap*) des observations pour la construction de chaque arbre. Pour la classification, cela fonctionne également pour les forêts uniformément aléatoires. Mais pas dans le cas de la régression, pour laquelle nous utilisons, par défaut, le *subsampling* (sous-échantillonnage sans remise) dont les résultats sont généralement meilleurs. Par défaut, nous fixons sa valeur à 0.7. Cependant, lorsque la valeur prédictive est le seul critère, il peut être opportun de considérer tout l'échantillon.

- Sauf dans des cas spécifiques, la profondeur des arbres ou le nombre minimal d'observations dans les régions terminales sont optimaux pour leur valeurs par défaut (profondeur maximale et nombre minimal d'observations à 1).
- De très nombreux autres paramètres peuvent être modifiés dans l'algorithme. Citons le rééquilibrage de classes, la perturbation des sorties, la combinaison linéaire de variables. L'algorithme supporte également les variables à valeurs catégorielles.

Trois éléments sont caractéristiques d'une forêt uniformément aléatoire :
i) le tirage des points de coupure suit la loi Uniforme sur le support de X ;
ii) pour la construction de chaque région, on tire $\lceil \beta d \rceil$ variables, $\beta \geq 1/d$, avec remise ;
iii) le gain d'information `IG`, que l'on maximise pour la classification, et, dans le cas de la régression, la fonction $\mathtt{L}_2$ définie par :

$$\mathtt{L}_2(j, D_n) = \sum_{i=1}^{n} \left(Y_i \mathbf{I}_{\left\{X_i^{(j)} \leq \alpha_j\right\}} - \hat{Y}_A \mathbf{I}_{\left\{X_i^{(j)} \leq \alpha_j\right\}} \right)^2 + \sum_{i=1}^{n} \left(Y_i \mathbf{I}_{\left\{X_i^{(j)} > \alpha_j\right\}} - \hat{Y}_{A^C} \mathbf{I}_{\left\{X_i^{(j)} > \alpha_j\right\}} \right)^2,$$

que l'on cherche à minimiser, avec

$$\hat{Y}_A = \frac{1}{\sum_{i=1}^{n} \mathbf{I}_{\left\{X_i^{(j)} \leq \alpha_j\right\}}} \sum_{i=1}^{n} Y_i \mathbf{I}_{\left\{X_i^{(j)} \leq \alpha_j\right\}} \; et \; \hat{Y}_{A^C} = \frac{1}{\sum_{i=1}^{n} \mathbf{I}_{\left\{X_i^{(j)} > \alpha_j\right\}}} \sum_{i=1}^{n} Y_i \mathbf{I}_{\left\{X_i^{(j)} > \alpha_j\right\}}.$$

La fonction $\mathtt{L}_2$ peut être remplacée par n'importe quel autre critère car le point essentiel est la recherche d'un minimum global sur la partition. Nous rappelons également les conditions d'arrêt au partitionnement d'une région :
1 - le nombre minimal d'observations vaut 1,
2 - les observations d'une région ont toutes le même label,
3 - les observations d'une région sont toutes identiques pour l'ensemble des X,
4 - pour la classification, la fonction `IG` est inférieure à un seuil, supérieur ou égal à 0, pour chacune des $\lceil \beta d \rceil$ variables ; pour la régression, $\mathtt{L}_2$ est plus petite ou égale qu'un seuil, qui peut dépendre de k et de Y ou qui vaut 0.

`Arbre de décision uniformément aléatoire` :

1- tirer, avec remise pour la classification et sans remise (ou en sous-échantillonnant) pour la régression, un échantillon D_n, ou un sous-échantillon si n est grand,
a) tirer $\lceil \beta d \rceil$ variables avec remise, parmi les d,
b) pour chacune des $\lceil \beta d \rceil$ variables, tirer α selon la loi Uniforme sur le support de chacune des variables,
2- pour chacune des $\lceil \beta d \rceil$ variables, choisir le couple (j^*, α_{j^*}) qui maximise $\mathtt{IG}(j, D_n)$, pour la classification, et qui minimise $\mathtt{L}_2(j, D_n)$ pour la régression, $j \in [1, d]$,
3- le couple (j^*, α_{j^*}) définit les frontières des deux nouvelles régions A et A^C,
4- si un critère d'arrêt est atteint, arrêter le partitionnement et appliquer la règle de décision $g_{\mathcal{P}}$,
5- sinon, recommencer les étapes 1 à 5 pour A et pour A^C.

`Forêt uniformément aléatoire` :

1- Pour b allant de 1 à B, construire un arbre de décision uniformément aléatoire et sa règle de décision
2- Pour les B arbres de décision construits, appliquer la règle de décision $\bar{g}_{\mathcal{P}}^{(B)}$.

Dans une forêt uniformément aléatoire, il n'y a pas de parcours des observations pour déterminer le meilleur point de coupure. Elles ne servent uniquement qu'à positionner

les régions dans l'espace. Cette étape permet d'accélérer les calculs. A contrario, le tirage, avec remise, des $\lceil \beta d \rceil$ variables peut pénaliser l'algorithme, notamment en grande dimension. Pour réduire la complexité, en $O(B\beta dn \log n)$, la sélection de variables et le sous-échantillonnage se révèlent utiles et offrent, dans certains cas, des gains de performance.

L'ensemble des détails de l'implémentation des forêts uniformément aléatoires est disponible dans le manuel de référence, sur la page web :
http ://cran.r-project.org/web/packages/randomUniformForest/index.html

3.6.2 Aspects numériques

Nous présentons dans cette section plusieurs exemples de résultats et visualisations des forêts uniformément aléatoires sur des données synthétiques et réelles. Tous les calculs sont réalisés grâce au logiciel libre R.

`Protocole et logiciel`

- Nous utilisons 2 jeux de données synthétiques définis et 2 jeux de données réelles.
- Pour chaque jeu et sauf mention contraire, 50% des données sont tirées aléatoirement et utilisées comme échantillon d'entraînement.
- Plusieurs algorithmes, disponibles sous R sous la forme de *packages* (briques logicielles), sont utilisés conjointement aux forêts uniformément aléatoires afin de mesurer la pertinence des résultats face à l'état de l'art :

randomForest, la version R de l'algorithme de référence des forêts aléatoires de Breiman (2001).
extraTrees, la version R (qui n'est pas celle des auteurs) d'une variante de forêts aléatoires, *Extremely Randomized Trees* (Geurts, Ernst, Wehenkel (2006)).
gbm, une implémentation du Stochastic Gradient Boosting de Friedman (2002).
e1071, une implémentation des SVM (*Support Vector Machines*) de Vapnik (1995).
randomGLM (Song, Langfelder, Horvath (2013)), un algorithme qui reprend les idées du Bagging en les adaptant aux modèles linéaires généralisés.
glmnet (Friedman, Hastie, Tibshirani (2010)) qui exploite les idées du *LASSO* (Tibshirani (1996)) et de la régression *RIDGE* (Tikhonov (1963)) en les appliquant aux modèles linéaires.
rpart une implémentation R des arbres de décision CART (Breiman, Friedman, Olshen, Stone (1984)).
randomUniformForest est l'implémentation qui désigne l'algorithme de référence des forêts uniformément aléatoires.

- Tous les algorithmes sont utilisés avec leurs paramètres par défaut, sauf pour le deuxième jeu de données et pour le *package gbm* dont la version par défaut donne des résultats éloignés de ceux avec les paramètres que nous avons pu spécifier.
- Pour chaque algorithme, nous mesurons la proportion d'observations mal classées sur les données de test (erreur de prédiction ou de test), pour la classification, et l'erreur

quadratique moyenne dans le cas de la régression.
- Nous n'effectuons pas de validation croisée, car nous souhaitons nous placer dans un cadre *industriel*, pour lequel on ne dispose que d'un seul échantillon de test, le même pour tous les algorithmes, et d'une seule erreur de prédiction à fournir à l'opérateur.
De plus, certains algorithmes disposent de nombreuses options leur permettant d'améliorer significativement leurs performances, en particulier lorsque les données présentent certaines caractéristiques, et une validation croisée nécessite alors d'optimiser, pour chaque algorithme et chaque jeu de données, de nombreux paramètres.
Pour ne pas pénaliser les modèles aléatoires, nous prenons, pour ces derniers, le meilleur résultat parmi 10 essais. Toutefois, sur les deux premiers jeux de données, nous effectuons chaque calcul 10 fois en générant autant de fois un échantillon aléatoire d'apprentissage portant sur 50% des observations. La moyenne et l'écart-type de l'erreur, entre parenthèses, sont reportés.
- Afin d'illustrer certaines des capacités des forêts (uniformément) aléatoires, nous indiquons toutes les informations menant à une meilleure interprétation des résultats, en particulier sur les jeux de données réelles.
- Sous R, la plupart des algorithmes proposent une même procédure pour apprendre, puis prédire les données. Sa forme générique est la suivante :

$$\texttt{LearningModel} = \texttt{Algorithm}\,(\texttt{trainData}, \texttt{trainLabels})$$

où `LearningModel` est le modèle résultant de l'apprentissage par l'algorithme,
`Algorithm` est (le nom de) l'algorithme d'apprentissage,
`trainData` est la matrice des données d'apprentissage (sans les réalisations de la variable à expliquer),
`trainLabels` est le vecteur des réponses (classes dans le cas de la classification, valeurs continues dans le cas de la régression) de la variable à expliquer, pour les données d'apprentissage.

Dans quelques cas, il faut spécifier explicitement le format de `trainData` et `trainLabels` car l'algorithme ne fait pas automatiquement la distinction entre classification et régression ou bien, a besoin d'un format spécifique pour les données d'apprentissage. Pour forcer le format matriciel numérique sous R, on peut utiliser la commande 'as.matrix()'. Rendre explicite l'utilisation de la classification nécessite la commande 'as.factor()', appliquée à `trainLabels`. Une fois les données apprises, la prédiction prend la forme générique suivante :

$$\texttt{predictedLabels} = \textit{predict}\,(\texttt{LearningModel}, \texttt{testData})$$

où `predictedLabels` est le vecteur des valeurs prédites par le modèle,
`testData` est la matrice des données de test,
'*predict*()' est la fonction générique sous R (son appel est explicite et R fait le lien avec la vraie fonction de prédiction du modèle) qui sert à prédire des valeurs à partir d'un modèle et de données de test.

Notons que certains algorithmes acceptent cette dernière formulation sans fournir les résultats sous la forme d'un vecteur. La forêt uniformément aléatoire distingue nativement la classification de la régression, sous certaines conditions, et sa version par défaut,

une fois les données chargées, nécessite 3 lignes sous R comme la plupart des algorithmes :

```
# si le paquet n'est pas encore installé
install.packages("randomUniformForest")
# s'il est déjà installé
library(randomUniformForest)
rUFLearningModel = randomUniformForest(trainData, trainLabels)
rUFPredictedLabels = predict(rUFLearningModel, testData)
```

Il est aussi possible de spécifier de nombreuses options ou d'effectuer la prédiction dans le même temps que la modélisation. Par exemple :

```
rUFLearningModel = randomUniformForest(trainData, trainLabels,
xtest = testData, ytest = testLabels)
rUFLearningModel
```

...effectue la modélisation sur les données d'entraînement et la prédiction sur un échantillon de test, puis affiche un résumé des résultats.

Nous terminons cette description par quelques mots sur R, le logiciel d'analyse statistique, de calcul et de modélisation, très utilisé au niveau académique et de plus en plus au niveau industriel. Un problème souvent discuté est la capacité des algorithmes, sous licence libre et sous R, à résister aux contraintes de l'industrialisation.
- En matière de mémoire centrale (16 To supportés pour les versions 64 bits), de stabilité et d'exportation des résultats, R est une alternative à de nombreux logiciels. En particulier, les détails d'un problème peuvent être poussés à leurs extrêmes. Lorsqu'une fonction n'existe pas sous R, malgré ses nombreuses extensions (> 5000), il suffit de la programmer dans le langage R, proche du C, ou un autre (C, C++, Fortran, Java, ...).
- Pour le traitement de gros volumes de données, des packages tels que *bigmemory* (Kane, Emerson, 2011) proposent de les manipuler ou stocker. D'autres packages comme *plyr* (Wickham, 2011), associés à *bigmemory*, rendent transparente la philosophie MapReduce, paradigme de l'inférence en big data.
- En matière de vitesse, R est généralement plus lent que ses concurrents. Cependant, il existe des packages, en particulier *Rcpp* (Eddelbuettel, Francois, 2011), qui permettent le mélange de code R et C++. La vitesse est alors du même ordre que celle du langage le plus rapide. Cela nécessite cependant un profilage du code source sur les parties critiques, en contrepartie d'un prototypage sous R bien plus simple.
Dans ce type d'utilisation (R, C++, *bigmemory ou data.table* + *plyr*), les contraintes sont largement levées et le passage à l'étape industrielle des algorithmes devient naturel. En particulier, les contraintes liées au volume ne dépendent plus que des capacités de stockage physiques. Notons que lorsqu'un environnement distribué (Hadoop) existe déjà pour les données, des packages comme *rmr2* (Revolution Analytics, 2013) permettent de les interfacer rapidement avec un algorithme sous R.

i) `Données synthétiques : classification`

Nous utilisons, pour la classification, un jeu de données artificiel (Synth_DataSet_1000_100) défini ainsi : $n = 1000$ et $p = 100$, et on considère une variable aléatoire $Z^{(j)}, 1 \leq j \leq p$, telle que $Z^{(j)} \sim \mathcal{U}_{[-10,10]}$.On définit également deux sources de bruit ϵ_1 et ϵ_2 données par $\epsilon_1 \sim \mathcal{U}_{[-1,1]}$ et $\epsilon_2 \sim \mathcal{U}_{[-1,1]}$ et on pose ensuite, pour $j \in [1, d]$,

$$X^{(j)} \sim \mathcal{N}(z_j, z_j^2),$$

où z_j est une réalisation de $Z^{(j)}$.
Y est une variable aléatoire à valeurs dans $\{0, 1\}$, telle que pour tout $y \in Y$, et tout x de X,

$$y = \mathbf{I}_{\left\{\frac{1}{n}\sum_{i=1}^{n} R(X_i,\epsilon) \leq R(x,\epsilon)\right\}},$$

avec $R(X, \epsilon) = 2\left(X^{(1)}X^{(2)} + X^{(3)}X^{(4)}\right) + \epsilon_1 X^{(5)} + \epsilon_2 X^{(6)}$.

Le code R suivant permet de simuler X puis Y :

```
set.seed(2014) # pour la reproductibilité
n = 1000; p = 100;
# fonctions disponibles dans le paquet randomUniformForest
X = simulationData(n,p)
X = fillVariablesNames(X)
epsilon1 = runif(n,-1,1)
epsilon2 = runif(n,-1,1)
rule = 2*(X[,1]*X[,2] + X[,3]*X[,4]) + epsilon1*X[,5] + epsilon2*X[,6]
Y = ifelse(rule > mean(rule), 1,0)

# division de l'échantillon en entraînement et test.
train_test = init_values(X, Y, sample.size = 1/2)
X1 = train_test$xtrain
Y1 = as.factor(train_test$ytrain)
X2 = train_test$xtest
Y2 = as.factor(train_test$ytest)
```

	Random Forests	ExtRaTrees	GBM	GBM (optimized)
Test error	0.1528 (0.0164)	0.1394 (0.02)	0.2248 (0.0241)	0.132 (0.0216)

	SVM	GLMnet	CART	Random Uniform Forests
Test error	0.1624 (0.0216)	0.1526 (0.0194)	0.1738 (0.0153)	0.1286 (0.014)

TABLE 3.1 – Classification (Synth_DataSet_1000_100) : artificial data. n = 1000, p = 100, test sample = 50%. Cross-validation (2-folds, 10 times). No tuning.

Les résultats permettent, dans un premier temps, de mesurer l'écart qui peut séparer les paramètres par défaut d'autres, plus optimaux (Boosting (optimized)). Les paramètres 'optimisés' du Boosting sont liés au nombre et à la profondeur des arbres, au nombre minimal d'observations et au facteur d'apprentissage de l'algorithme. Dans le package gbm

ils correspondent aux arguments "*ntree = 500, interaction.depth = 24, n.minobsinnode = 1, shrinkage = 0.05*" et nous ne les changeons plus dans tout le reste du document. Un avantage des forêts aléatoires est l'absence de réglages nécessaires à des performances intéressantes.

ii) `Données synthétiques : régression`

Pour la régression, nous utilisons un des jeux de données, dits *Friedman data sets*, dont l'ensemble est disponible sur Weka (www.cs.waikato.ac.nzml/weka/datasets.html). Les données choisies correspondent au fichier 'fri_c3_1000_25' pour lequel $n = 1000$ et $p = 25$. La variable à expliquer (Y) correspond à la 26e colonne du fichier. La fonction de régression est définie par :

$$Y = 10sin(\pi X^{(1)} X^{(2)}) + 20(X^{(3)} - 0.5)^2 + 10X^{(4)} + 5X^{(5)} + \epsilon,$$

où $\epsilon \sim \mathcal{N}(0, 1)$.

```
dataset = "fri_c3_1000_25.txt" # nom du fichier
XY = read.table(dataset)      # lecture des données
Response = 26    # colonne de la variable à prédire

# extraction
Y = extractYFromData(XY, whichColForY = Response)$Y
X = extractYFromData(XY, whichColForY = Response)$X

# initialisation
set.seed(2014)
train_test = init_values(X, Y, sample.size = 1/2)
X1 = train_test$xtrain
Y1 = train_test$ytrain
X2 = train_test$xtest
Y2 = train_test$ytest
```

	Random Forests	ExtRaTrees	GBM
Mean squared error	0.1219 (0.01)	0.0954 (0.0075)	0.0801 (0.0073)

	CART	Random Uniform Forests
Mean squared error	0.3156 (0.035)	0.0907 (0.0085)

TABLE 3.2 – Regression (fri_c3_1000_25) : artificial data. n = 1000, p = 25, test sample = 50%. Cross-validation (2-folds, 10 times). Tuning all models.

Dans cette comparaison, nous avons essayé d'optimiser tous les modèles, à cause des écarts importants avec les paramètres par défaut. Malgré cela, des modèles comme GLMnet (0.72) et les SVM (0.67), fonction *tune.svm* du paquet *e1071*, ont continué à avoir des erreurs trop importantes. Nous les avons donc enlevés. Seul CART a été laissé avec ses réglages par défaut. Les optimisations portent principalement sur le paramètre *mtry*

des forêts aléatoires. Le mécanisme générique de réduction de l'erreur a été utilisé dans le cas des forêts uniformément aléatoires. Il est utile lorsque les observations ne sont pas trop nombreuses ou lorsque le biais est important.

iii) `Données réelles : régression`

Nous illustrons, ici, de manière plus importante les propriétés théoriques, la visualisation et l'interprétation que l'on peut effectuer avec les forêts uniformément aléatoires. Les données sont disponibles via le package 'openair' (CarsLaw, Ropkins (2012)) issu de l'*Environmental Research Group* (King's College, London) dont le but est de permettre une meilleure compréhension des déterminants de la pollution atmosphérique à Londres. Le site web du projet (http ://www.openair-project.org) fournit un descriptif complet des objectifs. Nous présentons rapidement, et pour une meilleure compréhension, la manière dont sont récupérées et résumées les données via R :

```
install.packages("openair")
require(openair)
data(mydata)
summary(mydata)
```

$n = 65533$ et $p = 10$. Les variables sont définies ainsi :

```
list("date")
Observation date/time stamp in year-month-day hour:minute:second format (POSIXct).
list("ws")
Wind speed, in m/s, as numeric vector.
list("wd")
Wind direction, in degrees from North, as a numeric vector.
list("nox")
Oxides of nitrogen concentration, in ppb, as a numeric vector.
list("no2")
Nitrogen dioxide concentration, in ppb, as a numeric vector.
list("o3")
Ozone concentration, in ppb, as a numeric vector.
list("pm10")
Particulate PM10 fraction measurement, in ug/m3 (raw TEOM), as a numeric vector.
list("so2")
Sulfur dioxide concentration, in ppb, as a numeric vector.
list("co")
Carbon monoxide concentration, in ppm, as a numeric vector.
list("pm25")
Particulate PM2.5 fraction measurement, in ug/m3, as a numeric vector.
```

Les observations vont de 1998 à 2005 et la première variable enregistre la date complète (année/mois/jour et heure) pour chacun des 7 polluants atmosphériques constituant les données. La vitesse et la direction du vent sont également des variables. Ce jeu de données a la particularité de ne pas explicitement présenter de variable à expliquer. Cependant, pour les mesures de qualité de l'air, le NO_2 (dioxyde d'azote) et les particules PM10,

particules fines de diamètre inférieur à 10 micromètres, sont les polluants qui servent de référence, avec l'ozone (O_3). Nous renvoyons le lecteur à www.londonair.org.uk et à www.airparif.asso.fr pour plus d'informations. Selon la norme française en vigueur depuis 2010, la valeur moyenne annuelle pour le dioxyde d'azote ne devrait pas dépasser $40\mu g/m^3$, soit environ 40 parties par milliard (ppb) pour nos données. En moyenne horaire, elle devrait se situer en dessous de $200\mu g/m^3$. Pour les particules $PM10$, la valeur moyenne annuelle limite est identique à celle du dioxyde d'azote. Et la moyenne journalière ne devrait pas dépasser $50\mu g/m^3$ avec un seuil d'alerte à $80\mu g/m^3$. Notre objectif est défini alors par plusieurs étapes :
- prédire avec le moins d'erreurs les concentrations futures du dioxyde d'azote et de particules fines PM10 à partir des mesures des autres variables ;
- interpréter les résultats, par exemple le dépassement des valeurs recommandées ;
- détecter et évaluer des interactions et tendances dans les données.
Pour simplifier le problème, nous enlevons les valeurs manquantes des données et découpons la date en 4 variables correspondant aux années, mois, jours et heures d'observation. On a alors $n = 42524$ et $d = 12$, et la première variable à expliquer est la concentration de dioxyde d'azote (NO_2). Nous utilisons le protocole défini plus haut pour analyser les données. Cependant, pour éviter les temps de calcul trop longs, seul un résultat est calculé pour chaque algorithme. Sauf pour les écarts importants, la comparaison entre modèles peut être discutée.

	Random Forests	ExtRaTrees	GBM	GBM (optimized)
Mean squared error	55.13	-	264.08	48.5

	SVM	randomGLM	GLMnet	CART	Random Uniform Forests
Mean squared error	55.74	137.42	137.42	130.4	51.78

TABLE 3.3 – Regression (open air data) : NO_2 concentration (in $\mu g/m^3$). n = 42524, p = 12, test sample = 50%. No cross-validation. No tuning.

En considérant la racine carrée de l'erreur quadratique, la concentration de NO_2 est prédite avec un écart moyen d'environ $\pm 8\mu g/m^3$ pour les meilleurs modèles, avec les paramètres par défaut. Les méthodes linéaires comme glmnet ont des erreurs assez larges et comparables à CART. Notons que randomGLM renvoie les mêmes résultats que glmnet : dans les principes du Bagging, l'instabilité des modèles de base est généralement un atout et, pour les problèmes de régression, tend à être une nécessité à cause de la corrélation, souvent présente dans les méthodes ensemblistes. Le Boosting illustre les progrès réalisables en optimisant les paramètres. Notons que l'implémentation des *Extremely Randomized Trees* renvoie une erreur et fonctionne difficilement lorsque le nombre d'observations est important.

```
Random Uniform Forests 'openair' data summary (NO2 concentration):

Out-of-bag (OOB) evaluation
Mean of squared residuals: 53.3898

OOB residuals:
      Min        1Q    Median      Mean        3Q       Max
-63.46000  -3.42900   0.15080  -0.01051   3.70400  57.32000

Variance explained: 88.78%

Theorical (Breiman) bounds:
Theorical prediction error: 51.68169
Upper bound of prediction error: 51.70311
Mean prediction error of a tree: 123.9709
Average correlation between trees residuals: 0.4171
Expected squared bias (experimental): 0.000507
```

Pour la régression, on dispose, comme dans le modèle linéaire simple, du pourcentage de variance expliqué par le modèle. La variance expliquée par le modèle est plutôt importante, ce qui permet d'espérer un niveau d'interprétation suffisant. On remarque la corrélation importante entre les arbres ainsi que l'erreur quadratique moyenne d'un arbre de décision uniformément aléatoire (123.97), un peu plus basse que celle de CART (130.4). Le biais estimé est limité et sa réduction n'apporte que des gains marginaux. L'erreur OOB fournit une borne supérieure de l'erreur de prédiction. Cette borne est valide dès que la moyenne (49.15) et la variance de l'estimateur OOB (404.74) sont supérieures (condition suffisante) à celles du modèle sur les données de test (48.75 et 395.70). Les erreurs théoriques énoncées par Breiman constituent les informations les plus importantes :
- l'erreur de prédiction théorique est estimée à 51.68, ce qui est très proche de l'erreur quadratique moyenne (51.78) et indique que les paramètres de la forêt sont optimaux.
La précision importante des estimateurs est principalement due à la grande quantité de données.
- La corrélation entre les résidus des arbres (0.4171) est le principal facteur de réduction de l'erreur de prédiction.

Pour comprendre l'influence des autres variables sur la concentration de dioxyde d'azote, nous commençons par illustrer leur importance :

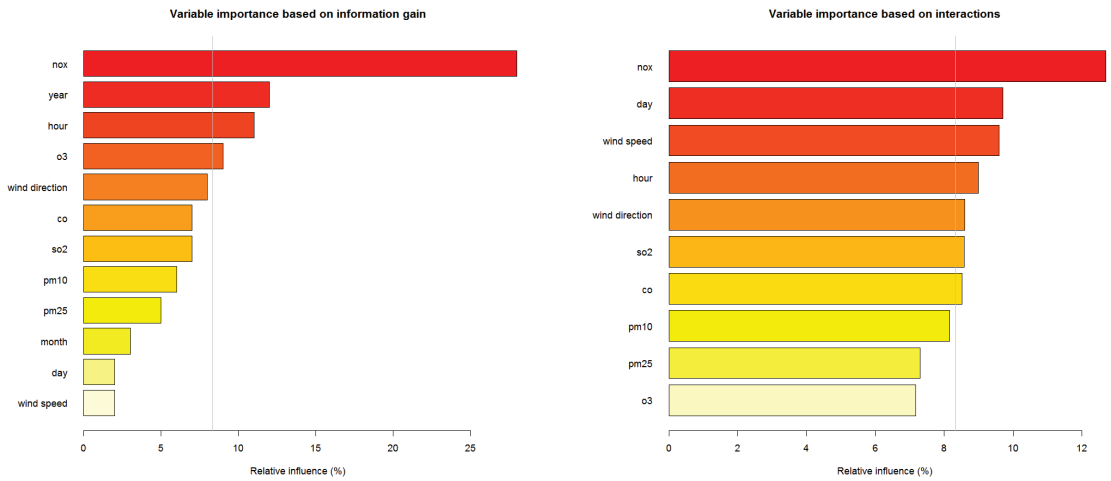

FIGURE 3.3 – Importance globale et locale des variables pour la concentration de dioxyde d'azote dans les données openair

Le NO_x (oxyde d'azote) est naturellement associé au dioxyde d'azote comme polluant et apparaît comme la variable la plus importante. L'autre polluant principal est l'ozone (O_3). On note que l'année et l'heure de la journée apparaissent comme des facteurs de variation de la concentration de NO_2. Les autres variables semblent moins importantes et il est plus utile de les analyser dans un second temps.

Le second graphique donne l'importance des variables en fonction de leurs interactions avec les autres et de leurs capacités prédictives. On retrouve l'oxyde d'azote, ainsi que le jour de mesure de la concentration alors qu'il n'a que peu d'importance globale. Dans un tel cas, cela peut signifier que la variable interagit beaucoup avec les autres variables mais n'a pas de valeur prédictive particulière. La vitesse du vent est dans le même cas. Ce type de comparaison permet de mieux situer les facteurs explicatifs. Par exemple, nous pouvons nous intéresser aux niveaux de concentration dépassant certains seuils :

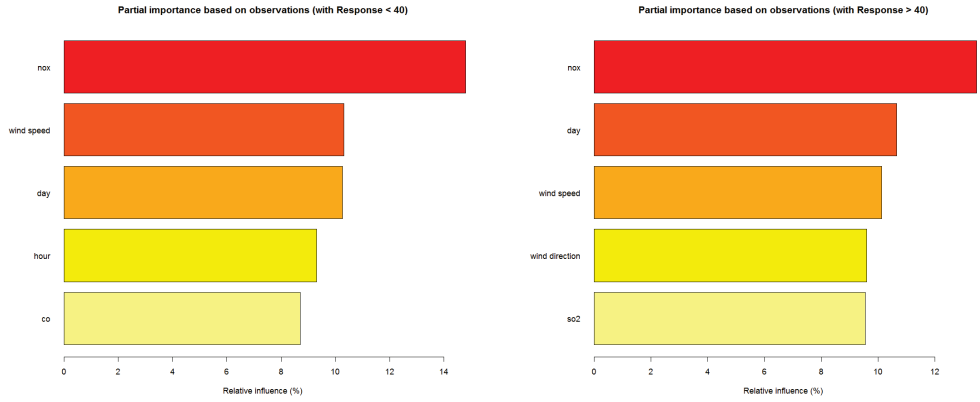

FIGURE 3.4 – Importance partielle des variables selon le seuil de concentration de dioxyde d'azote dans les données openair.

Le graphique d'importance partielle confirme l'oxyde d'azote, comme principal facteur de baisse et de hausse du niveau de dioxyde d'azote. Le monoxyde de carbone et certaines heures de la journée contribuent également à faire décroitre le niveau, tandis que le dioxyde de soufre (SO_2) et la direction du vent contribuent à le faire augmenter. Notons que les effets observés ici sont des effets moyens. Une fois identifiés les polluants importants (oxyde d'azote, ozone, dioxyde de soufre) et les facteurs atmosphériques (direction du vent) ou temporels (heure de la journée), on peut s'intéresser à leurs effets marginaux ou à ceux d'autres variables, grâce à leurs dépendances partielles avec la variable à expliquer.

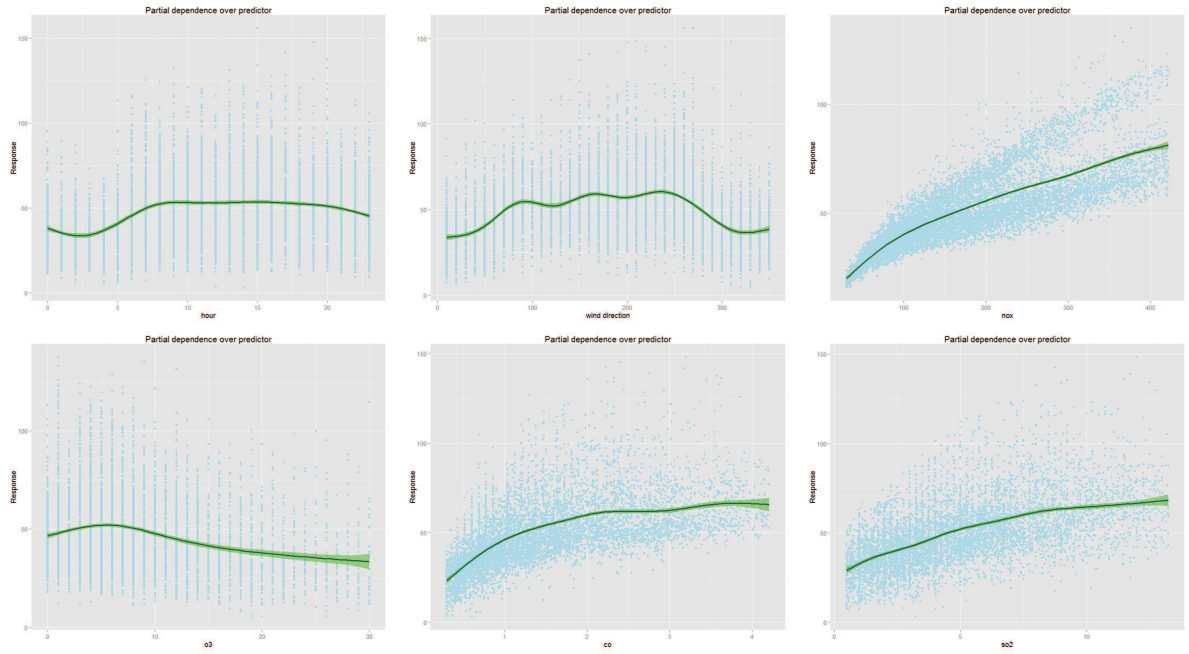

FIGURE 3.5 – Dépendance partielle de la concentration de dioxyde d'azote relativement à six variables explicatives.

Le premier graphique indique la dépendance partielle entre le niveau de dioxyde d'azote et l'heure de la journée. Les heures nocturnes sont celles pendant lesquelles la pollution est la moins importante. Le reste du temps, la pollution reste constante. Dans le second graphique, on observe que la direction du vent, vers l'Ouest ou le Nord, a un effet sur la réduction du niveau de pollution. L'importance et la dépendance partielles permettent de définir au plus près les facteurs de variation. Le monoxyde de carbone joue un rôle réducteur par rapport au niveau moyen, en dessous d'1 ppm (1 mg/m^3). L'augmentation du niveau d'ozone entraine une baisse du niveau de pollution. Toutefois, il convient d'en observer plus attentivement les valeurs : les pics de pollution sont moins nombreux pour de grandes valeurs du niveau d'ozone. On note enfin la relation linéaire entre oxyde d'azote (NO_x) et dioxyde d'azote (NO_2). Notons que la dépendance partielle est utilisable aussi bien sur l'échantillon d'entraînement que de test (dans ce cas, on explique les valeurs prédites par le modèle) et permet de caractériser le rôle explicatif d'une variable relativement à son importance. En utilisant la règle de décision $\bar{g}^{(B)}_{\mathcal{P},ev}$, la prédiction de pics de pollution peut être également interprétée.

Pour les particules de moins de 10 micromètres (PM10), nous reprenons la régression :

	Random Forests	ExtRaTrees	GBM	GBM (optimized)
Mean squared error	118.35	-	295.56	88.43

	SVM	randomGLM	GLMnet	CART	Random Uniform Forests
Mean squared error	158.7	139.66	139.76	169.77	90.29

TABLE 3.4 – Regression (open air data) : PM10 concentration (in $\mu g/m^3$). n = 42524, p = 12, test sample = 50%. No cross-validation. No tuning.

Pour la prédiction de la concentration de particules fines, l'écart moyen entre une prédiction et sa réalisation est de $\pm 10 \mu g/m^3$ avec les paramètres par défaut. Le problème est, ici, plus difficile que pour la prédiction de la concentration de NO_2. Les forêts aléatoires de Breiman sont moins performantes dans le cas de la régression, du fait du tirage *bootstrap* des observations. Le *subsampling* fonctionne mieux et l'utilisation de toutes les données permet de gagner encore plus en précision. On note que CART est un peu moins performant que les méthodes linéaires (GLMnet, randomGLM) et que les SVM rencontrent aussi des difficultés. Le *gradient boosting* fait partie des meilleurs choix.
Le détail des résultats des forêts uniformément aléatoires est présenté ci-dessous :

```
Random Uniform Forests 'openair' data summary (PM10 concentration):

Out-of-bag (OOB) evaluation
Mean of squared residuals: 90.5781

OOB residuals:
      Min        1Q    Median      Mean        3Q       Max
-452.6000   -2.3220    0.6667    0.1429    3.3870  311.5000

Variance explained: 79.59%

Theorical (Breiman) bounds:
Theoorical prediction error: 86.88465
Upper bound of prediction error: 87.80064
Mean prediction error of a tree: 194.1738
Average correlation between trees residuals: 0.4522
Expected squared bias (experimental): 1.6e-05

Test set
Mean of squared residuals: 90.29559
```

On note que l'erreur OOB est un bon estimateur de l'erreur de test, tandis que les propriétés théoriques indiquent comment se décompose l'erreur de prédiction : l'erreur quadratique moyenne d'un arbre de décision uniformément aléatoire est de 194.17 ; l'agrégation des arbres fonctionne bien grâce à une corrélation des résidus inférieure à 0.5, laquelle

réduit l'erreur du même facteur. Le *post-processing* permet de la réduire un peu plus (88.21), à un niveau proche de la limite supérieure de l'erreur de prédiction théorique. Précisons que le post-processing n'utilise pas l'échantillon de test. Ce sont les informations OOB qui permettent sa réalisation. Ces dernières sont fondamentales pour la compréhension et l'analyse théorique et opérationnelle des problèmes d'apprentissage statistique. Par exemple, en analysant les résidus (OOB residuals) de l'erreur OOB, l'écart moyen entre la prédiction et la réalisation de la concentration de particules fines PM10 est de $\pm 4\mu g/m^3$ dans plus de la moitié des cas. On retrouve une analyse identique sur l'échantillon de test (écart de $\pm 3\mu g/m^3$ dans plus de la moitié des cas).

L'analyse des variables indique que les dioxyde et oxyde d'azote et les particules fines de moins de 2.5 micromètres, sont les principaux facteurs de la présence de PM10 (particules fines de moins de 10 micromètres). La dépendance partielle permet de valider cet effet :

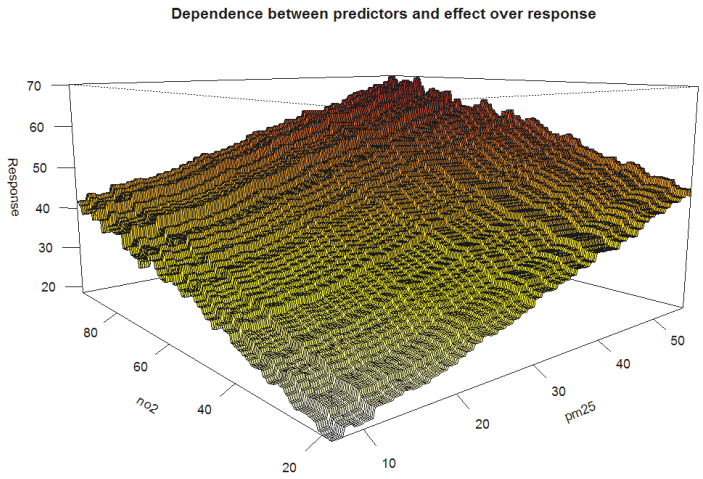

FIGURE 3.6 – Dépendance partielle de la concentration de particules fines PM10($\mu g/m^3$) relativement à celle du dioxyde d'azote et à celle des particules fines de moins de 2.5 micromètres.

La représentation en trois dimensions permet de visualiser la co-influence des variables les plus significatives sur la variable à expliquer. Nous avons exclu les valeurs atypiques du graphique. La dépendance entre le dioxyde d'azote et les particules PM2.5 est linéaire ; leur co-influence est également linéaire sur la concentration de PM10. On peut détailler l'effet sur un plan :

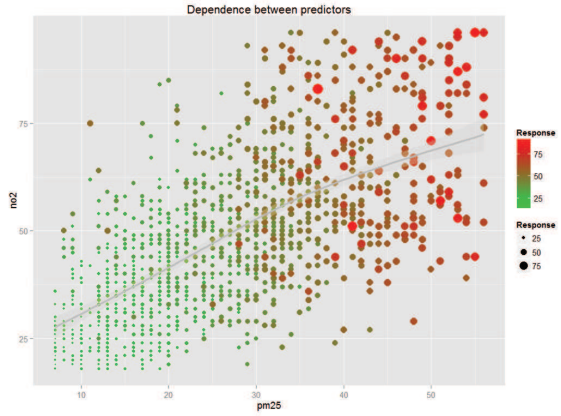

FIGURE 3.7 – Dépendance partielle de la concentration de particules fines PM10($\mu g/m^3$) relativement à celle du dioxyde d'azote et à celle des particules fines de moins de 2.5 micromètres.

Dans une représentation plus classique, on observe une intensité de plus en plus importante de la concentration de PM10, à mesure que les deux variables explicatives prennent des valeurs élevées.

iv) `Grande dimension`

Dans le cadre de la classification, nous nous intéressons à des jeux de données en grande dimension. L'intérêt réside ici moins dans les performances prédictives des algorithmes que dans leur capacité à réduire massivement la dimension du problème sans dégrader (ou peu) les performances, tout en expliquant le phénomène observé. Pour illustrer notre propos, nous considérons les données, mettant en jeu des milliers de gènes ("[DNA] microarray data"), dans l'aide au diagnostic et au traitement de maladies. Pour unifier le cadre de travail, nous utilisons le package R *datamicroarray* (Ramey, 2012) qui recueille de multiples jeux de données réelles en grande dimension pour des maladies génétiques, ou identifiables grâce aux gènes, comme le cancer du poumon, la leucémie, la maladie de Huntington,...
Sous R, le package s'installe ainsi :

```
require (devtools) || install.packages("devtools")
install_github("datamicroarray", "ramey")
library(datamicroarray)
describe_data() # pour visualiser les différents jeux de données
```

Notre échantillon, voir Burczynski et al. (2006), se compose de données pour lesquelles on souhaite différencier la maladie de Crohn de la colite ulcéreuse (Ulcerative Colitis) et des patients non atteints par une de ces deux maladies. Pour cela, chaque observation est constituée de mesures de milliers de gènes et associée à une classe représentée soit par le maladie de Crohn (Crohn's disease), soit par la colite ulcéreuse (Ulcerative Colitis), ou par aucune des deux (normal). On a $n = 127$, la taille de l'échantillon, $p = 22283$, le nombre de variables dont on mesure l'expression, et 3 classes. Nous accédons aux données (elles sont nommées du nom du premier auteur des articles associés) et les utilisons sans aucun retraitement selon notre protocole défini précédemment :

```
data("burczynski", package = "datamicroarray")
X = burczynski$x
Y = burczynski$y
```

L'échantillon de test comporte 35 cas de la maladie de Crohn, 13 cas de colite ulcéreuse et 16 cas normaux. A des fins de reproductibilité (et comme il y a peu d'observations), nous indiquons la liste des indices de l'échantillon de test :

```
test data index :
121  36 102 125  28  80  71  66  60  83 124 123   4  98  84  19  46  42  35
106 127  55  22 113  96 104  49  61  50  75  21 112   6  54  70  97   7 116
 79  51 108  99 111  91   1 120  14  43  34 117  94  16  44  56 119  81 109
126 118 107  32  53  24 114
```

Pour l'aide au diagnostic médical, le pire cas est celui d'une erreur de diagnostic pour la maladie. Le patient est alors soit déclaré sain, soit diagnostiqué pour une autre maladie que celle dont il est réellement atteint, et la conséquence est un risque de décès. Néanmoins, il nous faut d'abord nous intéresser à l'erreur de test des algorithmes, soit leur capacité à bien différencier les 3 classes, puis à l'erreur de diagnostic, soit la (non-)capacité à spécifier correctement la maladie incriminée. Pour rendre plus simple notre propos et comme le nombre de cas de la maladie de Crohn est le plus important, nous voulons que chaque algorithme détecte un maximum de cas pour le diagnostic de la maladie de Crohn. La mesure de ce nombre est notée *Crohn's disease diagnosis sensitivity rate* et définie ainsi :

$$Crohn's\ disease\ diagnosis\ sensitivity\ rate = \frac{Crohn's\ disease\ correctly\ classified\ samples}{All\ Crohn's\ disease\ samples}.$$

Toutefois, cette mesure n'est pas suffisante, car il suffit de diagnostiquer tous les cas comme atteints par la maladie pour ne pas faire d'erreurs. L'algorithme doit donc être également précis, c'est-à-dire avoir un pourcentage important de réussite relativement aux cas qu'il identifie comme atteints par la maladie. Nous notons cette mesure *Crohn's disease diagnosis precision rate* :

$$Crohn's\ disease\ diagnosis\ precision\ rate = \frac{Crohn's\ disease\ correctly\ classified\ samples}{Crohn's\ disease\ samples\ predicted\ by\ the\ model}.$$

Un résumé de ces deux indicateurs peut se comprendre ainsi :
le premier indicateur mesure la capacité de dépistage alors que le second mesure la pertinence du diagnostic ou encore l'efficacité de traitement de la maladie (si un traitement est/était fonctionnel pour tous les malades). Nous reportons d'abord les résultats pour chaque algorithme et pour l'ensemble des variables.

	RF	ExtRaTrees	GBM	GBM (optimized)	SVM
Test error	0.2968	0.2656	0.25	0.2968	0.3437
Crohn's disease diagnosis sensitivity rate	0.7714	0.8	0.8	0.7142	0.7428
Crohn's disease diagnosis precision rate	0.7942	0.8236	0.9063	0.8383	0.7428

	randomGLM	GLMnet	CART	Random Uniform Forests
Test error	-	0.1875	0.3593	0.2656
Crohn's disease diagnosis sensitivity rate	-	0.8	0.7142	0.7428
Crohn's disease diagnosis precision rate	-	0.9433	0.8	0.8966

TABLE 3.5 – Classification (Crohn's disease data) : Crohn's disease diagnosis. n = 127, p = 22283, test sample = 50%. No cross-validation (best out-of 10 trials). No tuning (except number of trees set to 500).

Avec leurs paramètres par défaut, la plupart des algorithmes renvoient un taux d'erreur similaire alors que l'échantillon d'entraînement (63 cas) est assez petit. CART produit des résultats moins bons que les méthodes ensemblistes. GLMnet semble très efficace pour ce type de données en grande dimension. Un diagnostic complet nécessite cependant l'identification des gènes les plus importants pour la maladie. Celle-ci permet un traitement plus adapté et plus simple mais également une meilleure connaissance de la problématique. Pour cela, la sélection et l'interprétation des variables les plus influentes est un préalable. Comme l'échantillon d'entraînement est petit et le nombre de variables très important, l'importance globale ne nous fournit pas d'informations suffisamment exploitables. A la place, nous nous intéressons aux interactions des variables et à la présence des plus significatives dans l'identification des cas de la maladie de Crohn :

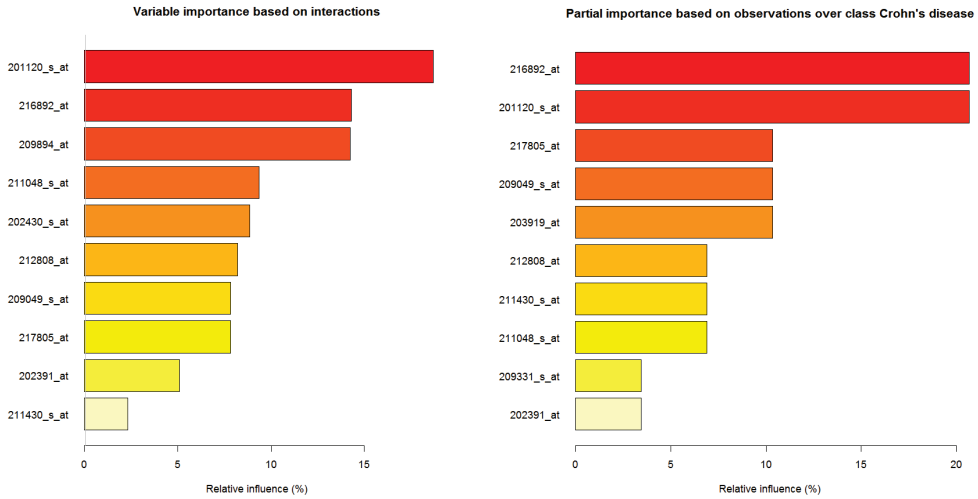

FIGURE 3.8 – Importance locale des variables et importance partielle pour la maladie de Crohn.

L'importance locale permet d'obtenir les interactions les plus importantes et désigne donc les variables qui interviennent le plus dans toutes les situations confondues (patients malades ou non). Pour identifier spécifiquement la maladie de Crohn, l'importance partielle (second graphique) est l'argument essentiel. Deux gènes ont une influence supérieure à 40% de l'influence totale de tous les gènes. L'importance partielle indique, ici, que les deux variables identifiées sont plus importantes que dans les deux autres classes. Pour préciser leur niveau d'importance, on peut, par exemple, comparer avec l'importance partielle pour chacune des autres classes et mesurer la différence d'influence. Précisons néanmoins que les variables sont liées à l'échantillon d'entraînement et de test et qu'elles peuvent changer fortement si l'échantillon n'est pas suffisamment grand. Nous observons ensuite comment, du point de vue du modèle, sont réparties les classes du problème :

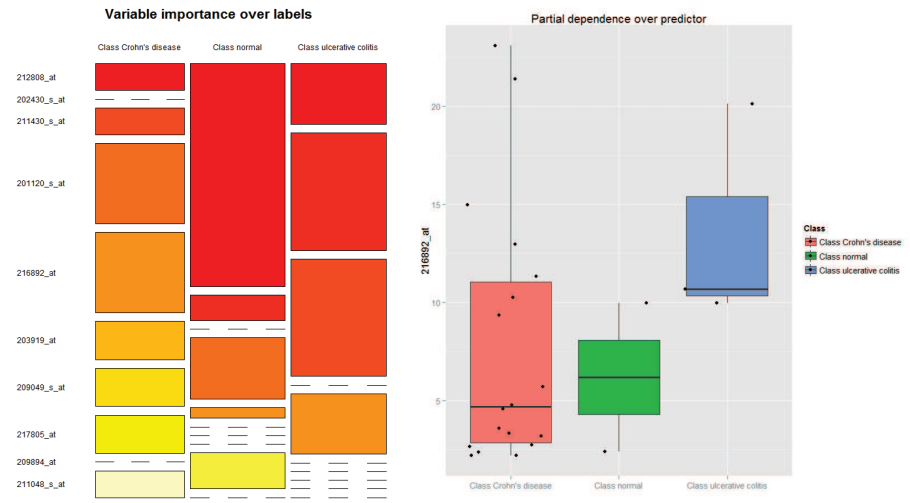

FIGURE 3.9 – Répartition des gènes importants par classe et dépendance partielle du gène identifié comme le plus influent dans le diagnostic de la maladie de Crohn.

Le premier graphique donne la répartition des variables pour chaque classe, tandis que la dépendance partielle, pour le gène "216892_at", en spécifie les détails. L'intérêt de cette représentation est de fournir un résumé de l'effet des variables importantes non plus sur les données mais sur les classes (prédites par le modèle) du problème. Certaines variables peuvent être importantes pour une classe mais pas pour d'autres. Ici, les variables d'intérêt sont celles qui permettent de mieux identifier la maladie de Crohn tout en ayant un effet marginal sur les deux autres classes. Notons que le gène identifié a une influence également chez les patients non malades (ce qui peut expliquer les erreurs de diagnostic). Nous visualisons ensuite les interactions entre les variables les plus influentes :

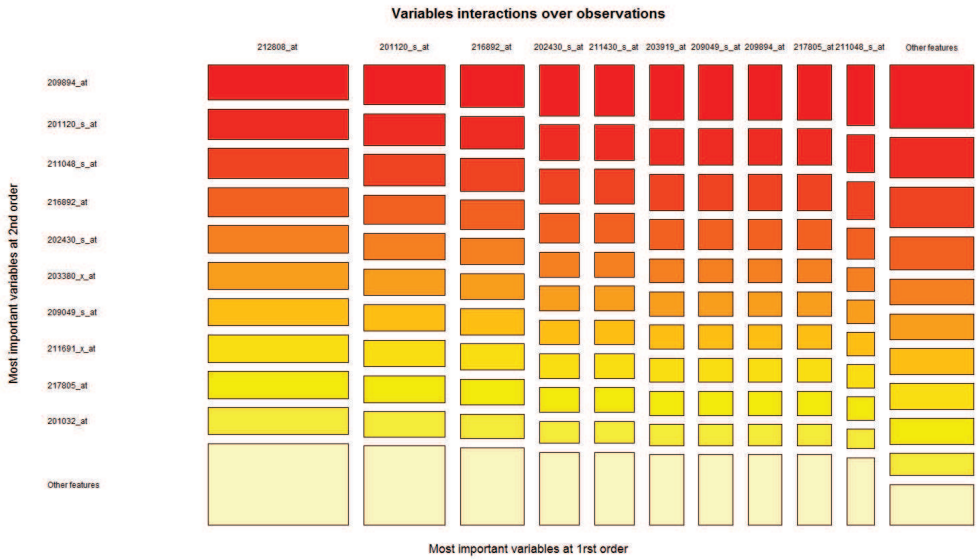

FIGURE 3.10 – Interactions entre toutes les variables.

Les interactions fournissent plusieurs grilles de lecture.
- La surface totale de la mosaïque est le plan de représentation (dont l'aire vaut 1) de l'influence des variables au premier et second ordre.
- La taille des rectangles correspond à l'intensité de la relation entre deux variables : la longueur est l'interaction au premier ordre et la largeur, celle au second ordre.
- La couleur d'un carré correspond à l'influence de la variable sur le phénomène observé, relativement à une autre variable (les couleurs sombres correspondent aux couples de variables les plus influents).
- Chaque direction représente un niveau d'importance. La direction horizontale correspond au facteur le plus influent.
- Pour chaque facteur, les variables sont ordonnées par ordre d'importance décroissant.
- Les variables d'influence faible sont regroupées sous le terme "Others features".

L'aspect le plus intéressant de l'aire d'interaction est le pouvoir explicatif de l'addition des variables les plus influentes. Le graphique indique ici que les interactions de 16 variables permettent d'expliquer la grande majorité de la relation entre les classes du problème et l'expression des variables.

Il nous faut alors vérifier que l'interprétation proposée correspond bien à la réalité. L'échantillon est de petite taille et il existe un risque pour que l'explication par les variables ne soit pas robuste. La difficulté, pour ce type de données, est que l'on dispose rarement d'échantillons importants. L'objectif est donc moins l'interprétation d'un faible nombre de variables, que la pertinence relativement aux prédictions. Peut-on, en sélectionnant moins de variables (gènes), améliorer les prédictions de manière significative ? Le dilemme est généralement le suivant : choisir les variables sélectionnées par le modèle et augmenter significativement la précision. La contrepartie est alors que l'on ne détecte pas tous les cas de maladie. Ou bien choisir un plus grand nombre de variables et détecter un grand nombre de cas. En contrepartie, les erreurs de diagnostic deviennent plus nombreuses.

La première option a l'avantage de préserver un lien fort entre les gènes identifiés et leur participation aussi bien dans un diagnostic précis que dans un traitement éventuel de la maladie. Nous la considérons donc en priorité. Un groupe de 10 gènes est identifié par l'importance partielle pour une précision maximale du diagnostic.

Les forêts aléatoires de Breiman, le gradient Boosting et CART proposent également la sélection de variables. Dans les deux premiers modèles, il faut néanmoins un échantillon de validation pour trouver le nombre de variables optimal. En son absence, nous avons identifié ce nombre en utilisant les labels de l'échantillon de test comme référence pour la maximisation des performances. Précisons que la sélection de variables dans les forêts uniformément aléatoires utilise à la fois les données d'entraînement et de test, mais pas les classes de ces dernières. 70 gènes sont identifiés dans les forêts aléatoires de Breiman. Pour le gradient boosting, l'optimisation est plus délicate et nous avons finalement retenu le même nombre. CART identifie automatiquement un groupe de 12 gènes. L'échantillon de test est inchangé :

	Random Forests	ExtRaTrees*	GBM	GBM (optimized)	SVM*
Test error	0.1406	0.2343	0.2343	0.2031	0.2656
Crohn's disease diagnosis sensitivity rate	0.8571	0.7142	0.8571	0.8857	0.7142
Crohn's disease diagnosis precision rate	0.9375	0.9615	0.8823	0.8611	0.9615

	GLMnet*	CART	Random Uniform Forests
Test error	0.1875	0.3593	0.2188
Crohn's disease diagnosis sensitivity rate	0.7714	0.6857	0.7714
Crohn's disease diagnosis precision rate	0.9642	0.8	0.9642

TABLE 3.6 – Classification (Crohn's disease data) : Crohn's disease diagnosis. n = 127, p $\in [10, 70]$, test sample = 50%. No cross-validation (best out-of 10 trials). Tuning (variable selection).

Les algorithmes marqués par un astérisque n'ont pas de méthode spécifique de sélection de variables et ont été tous utilisés avec la méthode sélectionnée par la forêt uniformément aléatoire. La sélection de variables accroît les performances prédictives de tous les modèles, à la différence près que les variables choisies ne sont pas nécessairement les mêmes : 50% des gènes sélectionnés par les forêts uniformément aléatoires sont présents parmi ceux sélectionnés par les forêts aléatoires de Breiman. Ces dernières génèrent l'erreur de test la plus faible et le meilleur compromis entre précision et exhaustivité, à condition de choisir le bon nombre de gènes sans référence aux labels de l'échantillon de test. Le principal avantage apporté par la sélection de variables opérée est une bien meilleure séparation entre maladie de Crohn et colite ulcéreuse. Dans le cas des forêts uniformément aléatoires, les patients non diagnostiqués pour la maladie de Crohn, alors qu'ils en sont atteints, sont considérés comme non malades. Bien que cela soit à leur désavantage, ils ne sont, malgré tout, pas confondus avec des malades atteints de colite ulcéreuse.

Cette observation nous a conduit à déterminer le nombre minimal de variables permettant une séparation optimale entre maladie de Crohn, colite ulcéreuse et patients non malades. Le processus est le suivant :
- les *interactions* entre tous les gènes influents fournissent la mesure de leur pouvoir explicatif relativement aux interactions de la totalité des gènes. Leur visualisation (figure 3.10) indique qu'un petit nombre de variables explique la grande majorité des interactions.
- Nous utilisons ensuite l'*importance partielle* pour déterminer les variables influentes de chaque classe et nous vérifions qu'elles correspondent à celles dont les interactions sont importantes. Nous identifions alors un groupe de 12 variables.
- Comme les classes sont déséquilibrées, nous leur assignons des poids. Les forêts uniformément aléatoires implémentent plusieurs méthodes de rééquilibrage des classes, en particulier la méthode *class reweighting* proposée par Chen et al. (2004). Son inconvénient est qu'il faut assigner les poids manuellement. Nous utilisons alors l'échantillon de

test ainsi que ses labels pour trouver les poids optimaux.
- Afin de valider la méthode, nous définissons un nouvel échantillon d'entraînement et de test et optimisons à nouveau les poids à partir de ceux définis précédemment, en utilisant cette fois-ci les données OOB et la borne de risque de Breiman. Le résumé de la méthode et les résultats sur le nouvel échantillon de test sont présentés ci-dessous :

```
Features selection = partial importance over each class + interactions +
class reweighting + OOB error & Breiman's bound minimization +
validation on new test sample.

Random Uniform Forests summary on Crohn's disease :

new test sample: 50%
class reweighting = {Crohn's disease: 1, normal: 0.4, ulcerative colitis: 0.9}
Total number of features: 12
number of features tried for each tree, with replacement: 20
number of trees: 1000

Out-of-bag (OOB) evaluation
OOB estimate of error rate: 14.29%
OOB error rate bound (with 1% deviation): 19.13%

OOB confusion matrix:
                    Reference
Prediction           Crohn's disease normal ulcerative colitis class.error
  Crohn's disease                 26      1                  3      0.1333
  normal                           0     19                  1      0.0500
  ulcerative colitis               0      4                  9      0.3077

Theorical (Breiman) bounds
Prediction error (expected to be lower than): 23.49%
Upper bound of prediction error: 34.13%
Average correlation between trees: 0.0688
Strength (margin): 0.476
Standard deviation of strength: 0.278

Test set
Error rate: 9.38%

Confusion matrix:
                    Reference
Prediction           Crohn's disease normal ulcerative colitis class.error
  Crohn's disease                 30      2                  1      0.0909
  normal                           2     16                  0      0.1111
  ulcerative colitis               1      0                 12      0.0769

Geometric mean: 0.9069
```

Notons que le rééquilibrage de classes n'est pas une nécessité et peut se discuter. Son principal avantage est sa capacité à permettre un dépistage plus important, avec un effet marginal sur la précision du diagnostic. Son inconvénient est qu'il faut l'adapter pour chaque nouvel échantillon d'entraînement, lorsque le nombre d'observations est faible. A des fins de reproductibilité, nous indiquons la liste des variables sélectionnées et les résultats des algorithmes qui bénéficient de cette sélection :

```
features selection : {"201120_s_at" "216892_at" "203919_at" "209049_s_at"
"217805_at" "211048_s_at" "211430_s_at" "212808_at" "202391_at"
"209331_s_at" "202430_s_at" "209894_at"}
```

	Random Forests*	ExtRaTrees*	SVM*
Test error	0.1093	0.125	0.125
Crohn's disease diagnosis sensitivity rate	0.9090	0.8529	0.9090
Crohn's disease diagnosis precision rate	0.9677	0.9677	0.9375

TABLE 3.7 – Classification (Crohn's disease data) : Crohn's disease diagnosis. n = 127, p = 12, (new) test sample = 50%. No cross-validation (best out-of 10 trials). Tuning (* : variable selection based on random uniform forests model, number of trees set to 1000).

Les SVM profitent le plus de la sélection de variables effectuée par les forêts uniformément aléatoires. Le boosting, CART ou les méthodes linéaires ont de moins bons résultats que ceux présentés ci-dessus. La comparaison entre algorithmes est toujours difficile car les paramètres par défaut tiennent souvent compte de contraintes moins dépendantes des performances prédictives que du temps de calcul ou de la généricité. L'aspect important relève plutôt de la capacité de chaque modèle à s'adapter au problème. En grande dimension, les méthodes ensemblistes ont l'avantage de ne nécessiter aucun réglage particulier pour aboutir à des résultats corrects. Elle permettent surtout une sélection des variables qui améliore sensiblement les performances, notamment lorsqu'on génère un grand nombre d'arbres pour sa mise en place. Dans le cas des forêts uniformément aléatoires, plusieurs méthodes permettent de choisir les variables les plus influentes et d'expliquer leurs effets. La principal avantage est qu'ici, le choix des variables peut être effectué aussi bien par la visualisation que par le calcul, avec peu d'efforts. Il aboutit à des combinaisons intéressantes, par exemple, en y associant des modèles déterministes comme les SVM.
Les contraintes d'optimisation de ces derniers sont alors fortement réduites.

vi) `Apprentissage incrémental et non-stationnarité`

L'exemple proposé ici est plus détaillé que les cas précédents et essaye de montrer la souplesse des modèles ensemblistes, en particulier dans leur capacité à s'adapter à des contraintes opérationnelles fortes. Lorsque le nombre d'observations tend à devenir très important ou lorsqu'elles arrivent en flux ou par périodes, la complexité des calculs augmente considérablement. Pour nombre d'algorithmes, cette complexité est hors d'atteinte principalement du fait des ressources requises. Un cadre plus général est celui qui y associe la non-stationnarité de la distribution des observations à chaque (ou sur plusieurs) période(s). La consistance statistique ou les bornes de risque ne sont alors plus valides et l'algorithme doit, de plus, trouver un moyen de s'adapter. Dans ce cas, le modèle est incrémental ou *online*, en opposition au mode *offline* qui requiert, pour l'apprentissage, l'ensemble des exemples disponibles jusqu'à la période courante. Dans le cas incrémental, le modèle s'auto-actualise en utilisant uniquement les données de la période courante. Le cadre online est un cas particulier dans lequel chaque nouvel exemple conduit à une actualisation du modèle. Pour une présentation claire et très complète de l'apprentissage incrémental et online en présence de non-stationnarité, nous renvoyons le lecteur à Gama et al. (2010).

Nous définissons le paradigme suivant afin de simplifier notre point de vue.
Nous considérons des données synthétiques en utilisant le protocole défini en *i)* et procédons à la simulation qui suit.
- La durée totale d'évaluation du modèle couvre 5 périodes, $T_1, T_2, ..., T_5$.
- Au début de la période T_1, le modèle reçoit un échantillon d'apprentissage et un échantillon de test.
- A chaque incrément de période, l'échantillon de test devient l'échantillon d'entraînement et un nouvel échantillon de test est reçu.
- A la période $T_i, i > 1$, l'hypothèse d'un trop grand volume de données empêche l'algorithme d'utiliser toutes les données disponibles jusque-là pour l'apprentissage. Il a donc uniquement à sa disposition les informations de la période T_i.
- Si l'algorithme est incrémental, les informations des périodes précédentes figurent dans son modèle et sont mises à jour avec chaque nouvel échantillon.
- Si l'algorithme est offline, il ne peut pas actualiser son modèle et seul l'échantillon de la période courante sert à évaluer les données de test.

Notons que l'hypothèse d'un gros volume de données est une contrainte, de fait, dans de nombreux cas réels. Précisons donc qu'ici, n est petit à chaque période, mais que les mêmes contraintes demeurent. Par exemple, à la période T_{1000}, un algorithme offline, dans le cas stationnaire, devrait traiter $n \times 1000$ observations (contre n, dans le cas incrémental). Dans le cas non-stationnaire, un algorithme offline, outre les temps de calcul, présente le risque de ne pouvoir prendre en compte des changements de distribution.

Nous procédons en plusieurs étapes :
- nous générons 5 sous-échantillons du couple (X, Y). Les paramètres de la distribution de chaque sous-échantillon sont aléatoires. La relation entre X et Y est la même que dans la procédure définie en *i)*.

- Puis, les deux premiers sous-échantillons sont mélangés pour former un échantillon A.
- Selon la même procédure, les deux derniers forment un second échantillon B.
- A la période T_1, nous prenons une partie de A et B pour former les exemples d'entraînement et une autre partie pour les données de test. Ces dernières deviennent l'échantillon d'entraînement de la période T_2, et une partie de chacun des échantillons A et B est à nouveau générée pour les exemples d'entraînement. Nous recommençons le processus jusqu'à la période T_3.
- A la période T_4, une petite partie du 5e sous échantillon est rajoutée à l'échantillon de test et à la période T_5 l'échantillon de test ne comprend plus que des exemples de A.

Pour chaque échantillon d'apprentissage et de test, $n = 1000$ et $d = 100$.
A des fins de reproductibilité, nous fournissons le code R qui génère l'ensemble des données :

```
T = 5; n = 10000;  p = 100
# create a list for storing all data
X = Y = vector("list", T)

# simulation of time slices
for (i in 1:T)
{
   # at each time slice, the distribution is evolving
   ZX = simulationData(n,p);  ZX = fillVariablesNames(ZX)

   epsilon1 = runif(n,-1,1); epsilon2 = runif(n,-1,1)

   rule = 2*(ZX[,1]*ZX[,2] + ZX[,3]*ZX[,4]) + epsilon1*ZX[,5] + epsilon2*ZX[,6]
   ZY = ifelse(rule > mean(rule), 1, 0)

   X[[i]] = ZX; Y[[i]] = ZY
}

# five subsamples are generated
X.sample1 = X[[1]]; Y.sample1 = Y[[1]]
X.sample2 = X[[2]]; Y.sample2 = Y[[2]]
X.sample3 = X[[3]]; Y.sample3 = Y[[3]]
X.sample4 = X[[4]]; Y.sample4 = Y[[4]]
X.sample5 = X[[5]]; Y.sample5 = Y[[5]]

# the mixture of each group of two subsamples gives the final distributions
X.A = rbind(X.sample1, X.sample2); Y.A = c(Y.sample1, Y.sample2)
randomIdx = sample(2*n, 2*n)
X.A = X.A[randomIdx,]; Y.A = Y.A[randomIdx]

X.B = rbind(X.sample3, X.sample4); Y.B = c(Y.sample3, Y.sample4)
randomIdx = sample(2*n, 2*n)
X.B = X.B[randomIdx,]; Y.B = Y.B[randomIdx]
```

```
#T1
# train sample
X1 = rbind(X.A[1:900,], X.B[1:100,])
Y1 = as.factor(c(Y.A[1:900], Y.B[1:100]))

# test sample
X2 = rbind(X.A[901:1000,], X.B[101:1000,])
Y2 = as.factor(c(Y.A[901:1000], Y.B[101:1000]))

#T2
X1 = X2; Y1 = Y2
X2 = rbind(X.A[1001:1800,], X.B[1001:1200,])
Y2 = as.factor(c(Y.A[1001:1800], Y.B[1001:1200]))

#T3
X1 = X2;  Y1 = Y2
X2 = rbind(X.A[1801:2000,], X.B[1201:2000,])
Y2 = as.factor(c(Y.A[1801:2000], Y.B[1201:2000]))

#T4
X1 = X2; Y1 = Y2
X2 = rbind(X.A[2001:2100,], X.B[2001:2800,], X.sample5[1:100,])
Y2 = as.factor(c(Y.A[2001:2100], Y.B[2001:2800], Y.sample5[1:100]))

#T5
X1 = X2;  Y1 = Y2
X2 = X.A[2101:3100,]
Y2 = as.factor(c(Y.A[2101:3100]))

# Offline data T1 to T5 in one sample
X1 = rbind(X.A[1:2100,], X.B[1:2800,], X.sample5[1:100,])
Y1 = as.factor(c(Y.A[1:2100], Y.B[1:2800], Y.sample5[1:100]))
```

Nous illustrons ci-dessous le changement de distribution entre l'échantillon d'entraînement et celui de test pour la variable $X^{(1)}$ et les périodes T_1, T_3 et T_5.

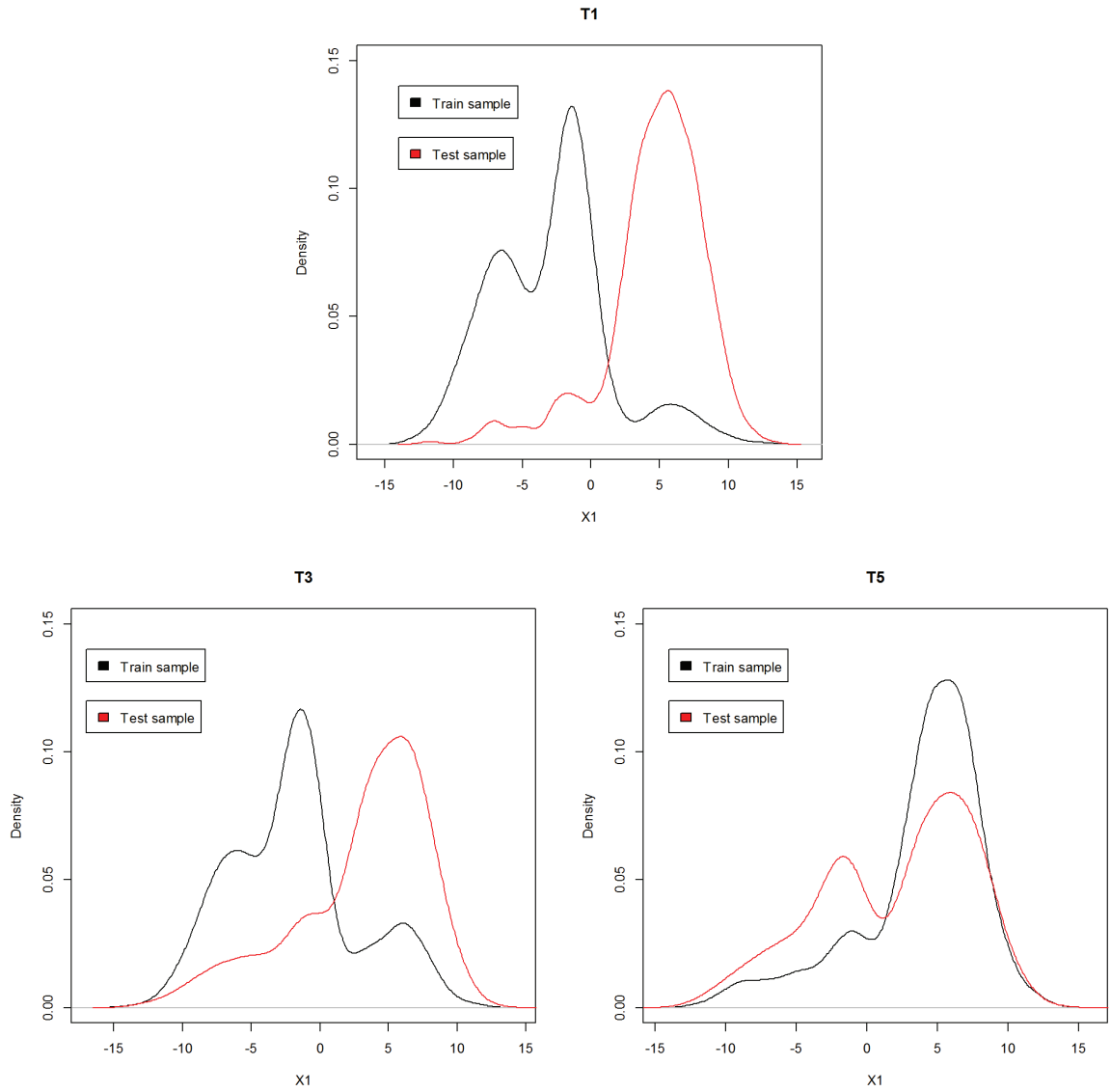

FIGURE 3.11 – Densité de la variable $X^{(1)}$ durant les période T_1, T_3 et T_5.

La distribution évolue à chaque période et à l'intérieur de chacune, l'échantillon d'entraînement se comporte différemment de l'échantillon de test. De plus, la relation liant les données à leurs étiquettes (non représentée) est susceptible d'évoluer également. Dans ce type de situation, ni l'erreur OOB, ni la validation croisée ne permettent d'estimer l'erreur de test.

L'apprentissage incrémental a plusieurs objectifs :
- il s'agit d'abord de vérifier dans quelle mesure les données des périodes précédentes contribuent à l'amélioration des performances sur la période courante.
- Dans un second temps, de confirmer l'adaptation des modèles en cas de changement abrupt de distribution.

Dans le cas des forêts uniformément aléatoires incrémentales, nous utilisons la règle de décision $\bar{g}^{(T)}_{\mathcal{P},inc}$. L'hypothèse principale est l'indépendance entre tous les arbres de la forêt. Une fois construit, aucun arbre n'est modifié. Cependant un ou plusieurs arbres peuvent être supprimés à chaque période. L'aspect incrémental de la forêt est caractérisé par plusieurs propriétés :
- tous les arbres peuvent *voter*, même s'ils n'ont pas été construits pour la période à évaluer. Par exemple, au moment de l'évaluation de l'échantillon de test de la période T_5, les arbres construits à la période T_1 peuvent voter.
- La règle de décision de la forêt détermine de manière autonome le résultat final. Plus clairement, l'algorithme décide de la manière dont doivent être combinés les votes de tous les arbres pour prendre la décision finale. Cela évite d'avoir à spécifier un modèle explicite de pondération des votes.
- Le temps de calcul pour l'apprentissage ne dépend que de la taille de l'échantillon de la période courante, T_i, et le volume de données ne présente alors plus de contraintes pour le modèle. Cependant, le temps de calcul pour la prédiction croît linéairement avec le nombre d'arbres. Il suffit alors de fixer une limite à ce dernier pour contrôler l'ensemble des temps de calcul.

Nous reprenons le protocole défini au début de cette section et les mêmes algorithmes auxquels nous rajoutons leurs versions incrémentales lorsqu'elles existent (ce qui est le cas pour *randomForest* et *gbm*). Nous ajoutons également la régression logistique au comparatif, afin de mieux saisir la diversité des modèles linéaires. Pour des raisons liées aux temps de calcul, seul le premier résultat de chaque algorithme est conservé. La forêt uniformément aléatoire (incrémentale) est notée *rUF* (*rUF (incremental)*) et les forêts aléatoires (incrémentales) de Breiman, *RF* (*RF incremental*). Pour chaque période nous reportons l'erreur de test de chaque modèle. La dernière colonne est l'application du modèle à toutes les données disponibles depuis T_1. Elle n'est fournie qu'à titre de comparaison ; dans un cadre opérationnel avec un grand nombre d'observations depuis la première période, son temps de calcul serait prohibitif ou inaccessible et le modèle offline ne serait, clairement, pas utilisé. Hormis pour cette colonne, tous les modèles utilisent donc, pour l'apprentissage, les données disponibles durant la période T_i.

	T_1	T_3	T_5	$T_5(offline)$
RF	0.402	0.201	0.423	0.339
RF (incremental)	-	0.133	0.355	-
ExtRaTrees	0.4	0.232	0.361	0.345
GBM	0.501	0.436	0.478	0.476
GBM (optimized)	0.295	0.152	0.425	0.311
GBM incremental (optimized)	-	0.213	0.348	-
SVM	0.481	0.283	0.485	0.352
Logistic regression	0.491	0.406	0.53	0.459
GLMnet	0.495	0.41	0.53	0.463
CART	0.359	0.216	0.431	0.371
rUF	0.337	0.158	0.41	0.316
rUF (incremental)	-	0.096	0.325	-

TABLE 3.8 – Test error. Classification with non-stationarity : artificial data. Incremental learning. t = $1, 2...5$. n = 2000, p = 100, at each time slice. Test sample = 50%. No cross-validation. No tuning.

Les méthodes linéaires sont les moins adaptées, lorsque les données deviennent complexes, qu'elles sont disponibles de manière périodique et que peu de variables sont influentes. Elles ont cependant l'avantage d'être rapides. Les méthodes de régularisation, comme le *LASSO* utilisé dans *GLMnet*, ne permettent pas d'améliorations. Parmi les modèles non linéaires, *CART* est plus souple que SVM lorsque les données sont peu nombreuses, tandis que leurs performances tendent à se rapprocher lorsque les observations augmentent fortement. La période T_5 (offline), pour laquelle toutes les données disponibles depuis le début sont prises en compte pour l'apprentissage, illustre cette convergence. Sur cet exemple, les modèles ensemblistes donnent les meilleurs résultats. Le *gradient boosting (GBM)* réduit fortement l'erreur de test, à condition de trouver les meilleurs paramètres de réglage du modèle. Sa version incrémentale semble être une alternative lorsque les données changent de manière plus brutale (période T_5). Les différentes variantes de forêts aléatoires fournissent la meilleure adaptation au problème, en particulier dans leurs versions incrémentales. Dans ce cas, elles donnent des résultats proches, avec les paramètres par défaut, d'un modèle optimisé prenant en compte l'ensemble des données (GBM, T_5 (offline)).

Dans le cas des forêts uniformément aléatoires incrémentales, une optimisation simple consiste à comparer, d'une période à l'autre, les distributions des variables les plus influentes de l'échantillon d'entraînement et de test. Si les variations sont importantes, l'apprentissage de la période courante utilise alors un nombre d'arbres différent. Après l'évaluation, un certain nombre d'arbres est ensuite supprimé de manière aléatoire. De manière plus générale, la gestion du nombre d'arbres, à construire à chaque période, est le seul paramètre nécessaire à contrôler lorsque les données sont en grand nombre et arrivent par périodes fixes. Lorsque ces dernières sont variables, la distribution des variables les plus influentes peut, éventuellement, conduire à une optimisation supplémentaire.

vi) Matrices creuses et classes déséquilibrées

Nous terminons l'illustration des possibilités des forêts uniformément aléatoires par le cas des données comportant de nombreux zéros. L'apprentissage peut alors devenir plus complexe. Nous considérons la classification et montrons ci-dessous le comportement des forêts uniformément aléatoires face à l'état de l'art. Pour cela, nous générons un échantillon synthétique à partir du code défini en *i)* puis, nous imputons aléatoirement la valeur 0 à environ 65% des données. Notons que les labels sont générés après que les zéros sont imputés. Le protocole de test ne change pas sauf le nombre d'arbres de la forêt uniformément aléatoire, $B = 500$.

Nous fournissons, d'abord, le code R suivant génère les données et les labels :

```
n = 1000;  p  =  100
X = simulationData(n,p)
X = fillVariablesNames(X)
epsilon1  = runif(n,-1,1)
epsilon2  = runif(n,-1,1)
s = 1 - exp(-1)

# create and impute zeros
zeros = cbind(sample(n,floor((2*n*p)/(3*s)), replace = TRUE),
sample(p, floor((2*n*p)/(3*s)), replace = TRUE))
X[zeros] = 0

# create and apply rule
rule = 2*(X[,1]*X[,2]  +  X[,3]*X[,4])  +  epsilon1*X[,5]  +  epsilon2*X[,6]
Y = ifelse(rule > mean(rule), 1,0)
```

Les matrices creuses possèdent plusieurs particularités.
- Dans certains cas, le nombre de zéros a plus d'influence sur l'apprentissage que leur proportion. Plus particulièrement, l'augmentation du niveau de *sparsité* (proportion de zéros ou d'une même valeur dans la matrice) détériore les performances beaucoup moins vite lorsque le nombre d'exemples s'accroît sans que la proportion de zéros ne change.
- Deux cas sont typiques. Dans le premier, les zéros sont intrinsèques aux données. Ils correspondent, par exemple, à une absence de choix pour une variable catégorielle, que l'on peut considérer comme un choix nul. Dans le second cas, les zéros sont la conséquence de la modélisation de toutes les possibilités, lorsqu'on souhaite prendre en compte l'ensemble des variables potentiellement pertinentes pour un problème. Un effet possible de ces deux situations est leur association avec une répartition des classes déséquilibrée. Une classe tend alors à être majoritaire et influence la classification.
- Lorsque la dimension du problème grandit, des outils de réduction de dimension, comme la décomposition en valeur singulières (SVD), sont utilisés pour mieux prendre en compte les données. Dans certains cas, cette opération est rendue difficile par la taille de la matrice. Dans le cas des arbres de décision, le problème est traité nativement par un critère d'arrêt dès qu'une sous-matrice des données initiales compte la même valeur pour tous ses éléments. Dans le cas des forêts uniformément aléatoires, ce principe est étendu. L'al-

gorithme évalue et élimine toutes les variables dont les valeurs sont constantes dès que certaines conditions sont réunies. Une conséquence de cette capacité est l'absence de nécessité à réduire la matrice pour l'algorithme. A la place, le sur-échantillonnage, avec remise, des observations peut conduire à l'amélioration des performances. Nous illustrons les résultats des différents algorithmes ci-dessous. Environ 65% des observations ont pour valeur 0 et la classe majoritaire représente environ 80% des cas.

	Random Forests	ExtRaTrees	GBM	GBM (optimized)
Test error	0.112	0.122	0.188	0.076

	SVM	randomGLM	GLMnet	CART	Random Uniform Forests
Test error	0.188	0.174	0.198	0.108	0.076

TABLE 3.9 – Classification : artificial, sparse and imbalanced data. n = 1000, p = 100. Test sample = 50%. No cross-validation (best out-of 10 trials). No Tuning.

Nous notons d'abord que CART est très performant lorsque la sparsité est associée à un déséquilibre des classes. Dans le cas sparse et en présence de bruit, les paramètres par défaut des forêts aléatoires de Breiman (et des ExtRaTrees) ne sont plus adaptés et il est préférable d'augmenter le nombre de variables à prendre en compte pour la construction de chaque arbre. En le doublant (de 10 à 20), l'erreur de test est réduite à 0.072 pour chacun des modèles. Les modèles linéaires semblent en retrait sur ce type de problématique. Afin de mesurer la progression des performances lorsque le nombre d'exemples augmente nous avons doublé la taille de l'échantillon et retenu les meilleures modèles.

	Random Forests	ExtRaTrees	SVM	GBM (optimized)	CART	Random Uniform Forests
Test error	0.05	0.157	0.169	0.013	0.019	0.014

TABLE 3.10 – Classification : artificial, sparse and imbalanced data. n = 2000, p = 100. Test sample = 50%. No cross-validation (best out-of 10 trials). No Tuning.

Il convient, tout d'abord, de noter que les paramètres par défaut ne sont pas adaptés pour les problématiques liées aux matrices creuses et en présence de bruit important. A nouveau, les forêts aléatoires de Breiman et les extRaTrees ont des erreurs similaires aux meilleurs modèles lorsqu'on modifie les paramètres.

Le doublement du nombre d'observations permet de diviser l'erreur par 5. Dans le cas des méthodes linéaires, la situation est plus délicate. Par exemple, le classifieur naïf de Bayes fournit une erreur de test de 16.7%, la régression logistique de 13.6%, identique à GLMnet. De manière plus générale, les arbres de décision semblent particulièrement adaptés aux données sparse. Lorsque le nombre d'observations est important, ce sont les méthodes qui semblent le mieux en profiter.

3.7 Discussion

Comme alternative aux forêts aléatoires de Breiman, les forêts uniformément aléatoires ont plusieurs particularités. Elles présentent nativement une structure incrémentale. En particulier lorsque les données présentent un caractère périodique ou temporel, la forêt uniformément aléatoire *s'auto-agrège* et sa règle de décision prend en compte les nouvelles et les anciennes informations. De manière plus caractéristique, les forêts uniformément aléatoires ne partagent avec la version de référence que la même structure d'arbre et d'agrégation. La règle de décision est, ici, plus simple ; il n'y a pas de recherche locale du point de coupure optimal et le point de vue est plus axé sur la dimension. En limitant le nombre d'hypothèses et en nous appuyant sur les travaux de Devroye et al. (1996), puis de Biau et al. (2008), nous établissons la consistance des forêts uniformément aléatoires dans le cas de la classification. Cette propriété ouvre la voie à de nouveaux travaux sur la vitesse de convergence de l'algorithme. Les forêts uniformément aléatoires héritent de toutes les propriétés de la version de référence et les implémentent. En particulier, les bornes de risque de Breiman sont particulièrement utiles et constituent, à notre sens, un fondement pour la compréhension des forêts aléatoires, tout en permettant de contrôler l'erreur de prédiction du modèle. Les bornes de risque sont étroitement associées à la règle de décision Out-of-bag (OOB) qui en constitue la contrepartie.

Deux difficultés sont généralement associées aux forêts aléatoires : la compréhension de leur mécanisme et l'interprétation des résultats fournis. Dans le premier cas, le mécanisme peut être vu comme une méthode de Monte Carlo, dont le paradigme est la loi des grands nombres. Chaque arbre de décision est, en quelque sorte, une "variable aléatoire" dont on peut contrôler la structure intrinsèque et dont la réalisation est une approximation de l'espérance conditionnelle de la variable à prédire. Lorsque les arbres sont indépendants et qu'il y en a suffisamment, la meilleure approximation que l'on puisse obtenir est alors donnée par la loi des grands nombres. Pour cette raison, les arbres doivent être aussi peu corrélés que possible. Une façon d'y arriver est d'accentuer leur caractère aléatoire de façon à laisser leur variance s'accroître. Pour cela, les arbres doivent être profonds, réduisant par la même occasion le biais de la forêt. Dans le cas de la régression, la corrélation est le principal facteur de réduction de l'erreur de prédiction : la variance de la forêt est approximativement le produit de la variance moyenne des arbres et de la corrélation entre les erreurs résiduelles. Dans le cas de la classification, l'effet de la corrélation est plus complexe, elle n'agit pas directement sur la variance, et le second terme dans la réduction de l'erreur de prédiction est, ici, la marge entre observations bien classées et mal classées. De manière synthétique, pour la classification, les arbres doivent avoir, majoritairement, de bonnes capacités prédictives alors que pour la régression, cette aptitude est moins importante et le point essentiel y est la corrélation.

Généralement, l'interprétation des résultats se révèle plus difficile avec une forêt aléatoire qu'avec un arbre de décision. *L'importance des variables* proposée par Breiman est une première étape de l'interprétation, mais elle n'est généralement pas suffisante car elle n'indique pas comment les variables importantes influencent la variable à prédire. Les forêts uniformément aléatoires implémentent plusieurs outils dont trois, au moins, facilitent l'exploitation des résultats :

- l'importance partielle, qui permet d'identifier les variables explicatives contribuant le plus aux variations de la variable à expliquer ;
- les interactions, qui fournissent une version granulaire de la variance expliquée par le modèle, en mesurant les dépendances mutuelles entre toutes les variables ;
- la dépendance partielle, inspirée des travaux de Friedman, qui détaille l'effet marginal de chaque (couple de) variable(s) explicative(s) sur la variable à expliquer. D'autres outils, comme la *matrice de proximités*, peuvent être utilisés pour une interprétation, finalement, beaucoup plus poussée que celle des arbres de décision ou des modèles linéaires.

Du point de vue applicatif, les performances des forêts uniformément aléatoires sont similaires à celles de la version de référence et le faible nombre de contraintes pesant sur la construction des arbres de décision est une ouverture à de futures améliorations. il nous semble que le choix et le construction des régions de chaque arbre sont un des processus les plus importants pour une meilleure compréhension des forêts aléatoires. De même, l'extrapolation et la prédiction de valeurs extrêmes sont de plus en plus essentiels aux méthodes prédictives et les lacunes des forêts (uniformément) aléatoires dans ce domaine peuvent être comblées en utilisant des outils comme la dépendance partielle, dont nous avons essayé de montrer les possibilités et l'intégration naturelle qu'elle apporte pour l'extrapolation.

Dans l'état de l'art actuel, la méthode générale de construction des arbres repose sur des combinaisons aléatoires des données et des variables afin de générer un maximum de diversité. Il nous semble que la construction d'arbres de décision hétérogènes, sans modification de la règle de décision mais avec des variations dans les conditions d'arrêt ou dans la construction des régions, pourrait fournir de nouveaux développements. De même, l'étude de la vitesse de convergence permettrait une analyse plus fine de l'avantage procuré par la forêt sur les arbres. Dans le cas de la régression, une piste de recherche intéressante est l'analyse de la corrélation entre les résidus des arbres, laquelle est généralement importante et constitue la principale limite à la réduction de l'erreur de prédiction.

Chapitre 4

Fraude aux cotisations sociales et situation financière des entreprises

Dans le cadre de la fraude aux cotisations sociales, nous nous intéressons au travail dissimulé et à l'impact de la situation financière de l'entreprise sur son niveau de fraude. Sous cet angle, et à partir d'un grand échantillon constitué de données synthétiques et réelles, nous mettons en évidence l'existence d'un lien avec des variables spécifiques, et en nombre limité, de l'activité de l'entreprise et observons plusieurs résultats : la capacité d'autofinancement, le besoin de financement du cycle d'exploitation et le rendement du chiffre d'affaires sont les facteurs économiques qui contribuent le plus fortement à la propension à la fraude. Pour les entreprises concernées, la solvabilité est généralement plus faible que la moyenne, tandis que la liquidité n'est pas décisive. L'influence de la situation financière concerne une minorité d'entreprises, mais elle est spécifique aux cas les plus importants de fraude. Nous montrons, empiriquement, que le niveau de fraude, sauf pour quelques exceptions, possède une relation non linéaire avec les variables économiques significatives et illustrons un processus de filtrage de données, par ces mêmes variables, améliorant sensiblement les capacités des modèles prédictifs dans la détection de la fraude.

4.1 Introduction

La Sécurité sociale est le principal garant des prestations sociales rendues aux personnes physiques. Ces prestations couvrent les accidents, les allocations familiales, la maladie, la retraite ou encore le chômage. La garantie des versements ou services associés nécessite des sources de financement suffisantes. Ces dernières proviennent essentiellement des cotisations sociales versées par les salariés et les employeurs. La déclaration, et le paiement, des cotisations est de la responsabilité des entreprises (ou employeurs). Elle apparaît, sur le bulletin de paie de chaque salarié, sous la forme d'un certain nombre de lignes correspondant à des catégories de cotisation et identifiant les sommes versées au titre des cotisations sociales. L'équilibre entre cotisations versées et prestations reçues est l'enjeu fondamental du modèle de Sécurité sociale. Un nombre suffisant de salariés et d'entreprises est un corollaire de sa pérennité. L'assurance que les cotisations dues sont bien celles reçues est tout aussi essentiel. Nous nous intéressons ici à cette problématique. Celle liée aux prestations constitue un autre cadre que nous ne détaillerons pas.

Afin de saisir l'enjeu des cotisations sociales, nous présentons d'abord quelques données. La loi de financement de la Sécurité sociale (LFSS) prévoit 460 Mds d'euros de recettes pour l'année 2013. Les dépenses sont estimées à 470 Mds d'euros. Dans les recettes, les cotisations sociales des entreprises représentent 329 Mds d'euros. Le solde s'explique par l'existence d'autres recettes qui ne sont pas, au sens strict, des cotisations sociales versées par des entreprises. Les prestations sociales adossées sont estimées, pour cette même année 2013, à 340 Mds d'euros.

La question de l'assurance du modèle est alors celle-ci : existe-t-il des moyens garantissant que les cotisations sociales estimées (ou demandées) seront bien celles versées par les entreprises ? La réponse est négative. Sa raison essentielle peut se résumer en trois phases :
- le principe des cotisations sociales repose sur un modèle purement déclaratif ;
- l'unique manière, à ce jour, de garantir les montants est de tous les contrôler ;
- le nombre d'entreprises et la complexité de la législation rendraient le coût des contrôles prohibitif.

Ce dernier élément constitue un point d'entrée dans la compréhension du modèle. Environ 1 200 000 entreprises (avec au moins un salarié) sont enregistrées en France et 5% d'entre elles ont plus de 49 salariés. La grande majorité verse les cotisations dues, mais les contrôles effectués montrent qu'un certain nombre d'entreprises ne les versent que partiellement ou pas du tout. L'absence de contrôles exhaustifs ne permet de fournir que des estimations du montant total des irrégularités aux cotisations sociales. En 2007, ce montant était estimé à un minimum annuel de 6 Mds d'euros (rapport du Conseil des prélèvements obligatoires, 2007). Soit, un peu moins de 2% des cotisations sociales versées par les entreprises chaque année. Si le pourcentage semble faible, on peut remarquer que l'absence d'irrégularités suffirait, au moins, à rétablir la moitié du déficit annuel de la Sécurité sociale. Comprendre et analyser les irrégularités (ou la fraude) aux cotisations sociales représente alors un intérêt fondamental.

Notre propos porte essentiellement (à partir des données issues des bilans comptables, des déclarations de cotisation et des résultats des contrôles) sur l'influence des variables économiques de l'activité de l'entreprise quant à sa propension à frauder. Nous détaillons tout d'abord les mécanismes de la déclaration et du contrôle des cotisations. Puis, dans une seconde partie, nous donnons quelques définitions utiles à l'analyse proposée. En particulier, nous y indiquons comment la complexité de la législation rend beaucoup plus difficile la maîtrise du phénomène de fraude. Nous présentons ensuite les données utilisées pour l'analyse et le pré-traitement appliqué pour prévenir les effets de bord. Les deux sections suivantes constituent la principale partie du document : nous définissons rapidement les modèles utilisés pour l'analyse puis identifions les variables économiques les plus influentes et proposons un point de vue sur leurs effets. Dans l'avant-dernière partie, nous indiquons une application importante de l'utilisation de facteurs économiques dans le cadre d'un modèle prédictif et montrons comment ce type de modèle, non linéaire, permet d'expliquer la relation entre la situation financière et la propension à frauder. Dans la dernière partie, nous présentons nos conclusions.

4.2 Le recouvrement des cotisations sociales et leur contrôle

Le réseau des URSSAF (Union de Recouvrement des cotisations de Sécurité Sociale et d'Allocations Familiales) est l'organisme de collecte des cotisations salariales et patronales. Il existe, du moins cette réforme est-elle en cours au moment où nous écrivons ces lignes, une URSSAF par région. Du point de vue des URSSAF, les entreprises, ainsi que les professions libérales et travailleurs indépendants, sont les cotisants.

Le principe du recouvrement est déclaratif : chaque cotisant indique, de manière périodique, le montant de ses cotisations patronales et salariales ainsi qu'un certain nombre d'éléments nécessaires au traitement de la déclaration. A cette dernière est adossé un certain nombre d'informations qui caractérisent la relation du cotisant à l'URSSAF.
Ces informations concernent l'effectif de l'entreprise, sa masse salariale, ou encore sa date d'immatriculation, son secteur d'activité,... Une fois les montants déclarés, le système informatique les ventile automatiquement dans des catégories appropriées - assurance vieillesse, chômage, maladie, retraite, accidents, transport, CSG/CRDS,... Chacune est le résultat d'un taux d'application et d'une assiette de cotisation, pour un effectif concerné. Ces catégories abondent, par leurs montants, la branche Prestations de la Sécurité sociale et ses différentes caisses, comme l'Assurance maladie ou les Allocations familiales. Depuis 2013, les URSSAF se chargent également de recouvrer les cotisations (de l'Assurance) chômage.
Plusieurs régimes (de cotisation) se côtoient dans les URSSAF dont le principal, et de loin le plus important, est le Régime Général, celui des entreprises avec, au moins, un salarié. Dans le reste du document et sauf mention contraire, les chiffres cités font allusion au Régime Général.

`Les taux et assiettes de cotisation`

Pour donner un point de vue clair, nous abordons ces deux paramètres par l'exemple d'une entreprise avec plusieurs salariés. Chacun d'eux voit, écrit sur son bulletin de paie, un certain nombre de lignes dont chacune correspond à une catégorie de cotisation spécifique. Il y figure (en colonnes) le montant des cotisations salariales et patronales. Avec un taux d'application pour la part salariale et un autre pour la part patronale. L'autre paramètre de la cotisation est son assiette, soit la partie du salaire à laquelle s'applique le taux de cotisation. Ici encore, il y a une part salariale et une autre patronale. Par exemple, un salarié qui effectue des heures supplémentaires va bénéficier d'un taux et d'une assiette spécifiques, car ils ne sont pas les mêmes que pour le nombre d'heures habituel. Pour simplifier la compréhension du processus, nous ne distinguons plus la partie (du montant des cotisations) patronale de la partie salariale et les ajoutons l'une à l'autre. Dans le cas de la fraude, les deux sont éludées par minoration ou non-déclaration.
Pour chaque salarié de l'entreprise, au moins deux catégories sont toujours déclarées et représentent la partie la plus importante des cotisations recouvrées par l'URSSAF :
- la catégorie dite *Cas général* ;
- la Contribution Sociale Genéralisée (CSG) et la Contribution pour le Remboursement de la Dette Sociale (CRDS).

Le *Cas général* sert à abonder les prestations sociales des *branches* Maladie, Maternité, Invalidité Décès, Vieillesse, Solidarité et Allocations familiales. Elle a un taux d'application de 20.95% du salaire brut (part déplafonnée), en 2011. Le Cas général possède également une part plafonnée (le plafonds étant défini par la Sécurité sociale), qui abonde l'Assurance Vieillesse (les retraites de base), dont le taux est de 15.25% (en 2012, il était de 14.95%). Son taux devrait augmenter progressivement dans les années à venir (15.45% en 2016). La CSG-CRDS s'appliquait, en 2011, à 97% des revenus d'activités (salaire, primes et autres avantages assimilables à un salaire) avec un taux de 8%. Son assiette est passée à 98.25% en 2013. Jusqu'en 2010, l'URSSAF ne recouvrait pas les cotisations pour la branche Chômage dont le taux est de 6.4%, en 2011, et a également varié. En les excluant, chaque entreprise devrait donc verser, au minimum et hors mesures de réduction, environ 44.1% de cotisations sociales (salariales et patronales) sur chaque salaire brut.

La réalité est plus nuancée.
Tout d'abord, le taux de cotisation change au-dessus de certains plafonds et la cotisation associée s'ajoute à la part plafonnée, calculée avec un autre taux et une autre assiette. Il existe également un peu plus de 900 catégories, dont plus de 300 ont été utilisées en 2011 par les seules entreprises d'Île-de-France. Une (grande) entreprise peut donc, théoriquement, déclarer un grand nombre de catégories différentes. Dans l'exemple considéré, si certains salariés ont des contrats spécifiques, ils peuvent prétendre à des abattements de cotisation, par le biais d'une catégorie dédiée. Si l'entreprise est en zone franche urbaine, elle peut également bénéficier de mesures de réduction. Certains taux, comme ceux liés aux accidents du travail ou à la retraite complémentaire, dépendent du secteur d'activité de l'entreprise ou du statut du salarié. Notons également que les salaires jusqu'à 1.6 fois le SMIC, peuvent bénéficier d'une mesure de réduction, sur les cotisations patronales uniquement, de sorte que le taux de cotisation effectif peut passer sous le taux minimal. En 2013, le seuil et l'effectif, pour cette mesure, ont été modifiés et segmentés. Ainsi, le delta avec le taux de cotisation minimal peut dépasser, en valeur absolue, 15%.

La législation prévoit, en fait, un maximum de situations possibles ainsi que les assiettes, taux, effectifs concernés et conditions d'application pour chaque catégorie. En 2009, le taux global de cotisation (ratio entre la somme totale des cotisations et la somme des salaires bruts) est d'environ 33% (il est identique en 2011). Le taux de cotisation constitue la principale source de variabilité des cotisations sociales. Toujours en 2009, 95% des entreprises avaient un taux de cotisation compris entre 17% et 50%. Le taux moyen était de 37%.

`Le contrôle`

Le caractère déclaratif simplifie le processus d'enregistrement des cotisations dès lors que l'entreprise a déterminé leur nature pour chaque catégorie et salarié. En contrepartie, l'exactitude des montants et leur justification ne peuvent être garanties autrement que par un contrôle des cotisations. En effet, le peu d'information nécessaire au calcul d'un montant d'une cotisation, taux et assiette, est démultiplié par le nombre important de catégories et de leurs conditions d'application. Selon le secteur d'activité, la zone géographique ou encore le nombre de salariés, le salaire versé, le type de contrat ou le

nombres d'heures travaillées, la part relative d'une même catégorie de cotisation peut fortement varier d'une entreprise à une autre (et d'un salarié à l'autre).

Du point de vue de l'URSSAF, la mission de recouvrement s'accompagne naturellement de la mission de contrôle de ces cotisations. Pour la région Ile-de-France qui représente, environ, 20% de l'ensemble des cotisations reçues, cette mission donne lieu, pour près du demi-million d'entreprises installées, à la possibilité d'un contrôle portant sur les trois dernières années de cotisation et celles de l'année en cours au moment du contrôle. Le rôle de ce dernier est, non seulement, de vérifier la conformité des entreprises à la législation, mais également de lutter contre les phénomènes engendrant des distorsions de concurrence ainsi que leurs conséquences. Ce rôle peut se résumer à la détection des irrégularités aux cotisations sociales et à la lutte contre le travail dissimulé. Dans le premier cas, l'action porte sur la capacité systématique à mettre à jour un montant erroné de cotisations déclarées. Lorsque les cotisations sont omises, de manière volontaire, pour une partie ou pour tous les salariés d'une entreprise, le terme employé est celui de travail dissimulé. En 2011, et pour la France entière, environ 220 millions d'euros de redressements ont été notifiés par les inspecteurs de l'URSSAF dans le cadre du travail dissimulé, sur plus de 7500 actions de contrôle.

L'omission des cotisations est rendue possible par le coût représenté par le contrôle éventuel de toutes les entreprises. Dans un tel cas, une partie conséquente des ressources de la Sécurité sociale serait affectée au seul contrôle des cotisations, en contrepartie d'une absence de travail dissimulé. A ce jour, le coût d'un contrôle est faible relativement aux sommes redressées. L'arbitrage doit donc se faire, pour l'URSSAF, entre un montant potentiellement important d'omissions de cotisation, et un coût de récupération de ces montants qui doit rester négligeable. Pour le cotisant dissimulant volontairement des salariés, la possibilité d'un contrôle, associée à d'éventuelles pénalités et sanctions coûteuses, est la contrepartie du profit engendré par l'absence de cotisations à régler.

La fraude pour travail dissimulé intentionnel constitue un délit et un arsenal de sanctions est généralement prévu contre les entreprises et leurs dirigeants, lorsque le caractère intentionnel est prouvé. Ces sanctions sont d'abord financières - en 2013, l'amende prévue pour les personnes morales est de 225 000 euros. Mais aussi administratives avec une exclusion des marchés, et aides, publics, une fermeture provisoire de l'établissement ayant servi à commettre l'infraction, la possibilité pour les dirigeants de se voir notifier une interdiction de gérer, un remboursement de certaines mesures de réductions ou d'exonérations ayant bénéficié à l'entreprise.

Cependant, le caractère intentionnel doit être prouvé. S'il ne l'est pas, l'entreprise ne rembourse alors que les sommes dissimulées pour la période sur laquelle porte le contrôle. De plus, le nombre d'entreprises (environ 1 200 000) est largement supérieur au nombre de contrôles (environ 165 000 en 2011, dont environ 7500 dans le seul cadre du travail dissimulé). Dans le meilleur des cas, la probabilité pour une entreprise quelconque d'être contrôlée ne dépasse alors pas 15%. L'arbitrage à réaliser est donc explicite et parmi les hypothèses qui y conduisent, figure la santé financière de l'entreprise. En quelque sorte, l'opportunité permise par le recours au travail dissimulé pourrait être rendue admis-

sible par les difficultés économiques de l'entreprise. Bien que cette hypothèse ne soit pas unique, elle peut être analysée à travers l'historique des données traduisant la relation de l'URSSAF au cotisant. Les retards et absence de paiements, le taux de cotisation, les majorations de retard sont autant de signaux qui peuvent, indirectement, montrer un lien avec la fraude. La mise en évidence de la dégradation de l'activité de l'entreprise dans les phénomènes de fraude est, elle, analysée à travers les variables économiques, par exemple, celles ayant trait à la liquidité, la solvabilité ou encore à la rentabilité. Ces variables économiques ne sont cependant pas simples à observer. Dans leur relation à l'URSSAF, les entreprises n'ont pas d'obligation à déclarer leur situation économique et celle-ci n'y est pas enregistrée. De nombreuses entreprises ne sont pas, non plus, tenues de publier leur bilan comptable. Notons cependant deux éléments importants :
- le recours au travail dissimulé peut se faire à l'insu des salariés et leur porte toujours tort, puisqu'une dissimulation de cotisations a un impact, généralement non immédiat, sur la protection sociale dont ils bénéficient. Cet impact peut être particulièrement prononcé lorsque le travail dissimulé procède de la gestion habituelle de l'entreprise.
- Certains procédés de dissimulation des cotisations sont assez sophistiqués pour passer totalement inaperçus des salariés.

Sur, environ, 1500 entreprises contrôlées par l'URSSAF d'Île-de-France en 2009 dans le cadre de la lutte contre le travail dissimulé, nous avons pu recueillir le bilan comptable de près de 900 d'entre elles. Puis, nous avons généré, à partir de leurs données financières, un échantillon synthétique de plus de 9000 entreprises. Nous montrons dans la suite, et à travers l'analyse de ces données, l'influence des variables économiques sur la propension à frauder.

4.3 Définitions

Cette partie regroupe deux types de définitions.
- Le premier est destiné à faciliter la compréhension des indicateurs de l'URSSAF.
- Le deuxième donne une notion plus commune à certains termes utilisés tout au long du document.

Bien qu'évoquée à plusieurs reprises, nous définissons clairement la fraude aux cotisations sociales et la distinguons des irrégularités aux cotisations sociales. La fraude aux cotisation sociales se définit comme la non-déclaration ou la minoration, volontaire, des cotisations sociales dues par l'entreprise (le cotisant). Ce caractère volontaire est soit supposé, soit effectif. Dans le cas d'irrégularités, le caractère volontaire n'existe pas ou ne peut être démontré.

La fraude est, donc, un cas particulier des irrégularités aux cotisations sociales.

Nous discutons de la fraude dans le cadre du travail dissimulé ou dans un cadre assimilable. Les deux principaux indicateurs utilisés par l'URSSAF pour évaluer les irrégularités aux cotisations sociales sont la fréquence de redressement et le taux de redressement.
Les redressements, en valeur, sont les montants consécutifs aux erreurs de déclaration

(ou de paiement) répertoriées par les inspecteurs du contrôle. Nous indiquons, à dessein, le terme "erreurs" car un redressement peut être en faveur de l'URSSAF, ou en faveur du cotisant. Dans le premier cas, le redressement est dit positif et il correspond à une irrégularité. Dans le second, il est dit négatif.

Pour une entreprise contrôlée, un redressement a pour valeur la somme des montants comptables correspondant au nombre d'erreurs et/ou irrégularités relevées par l'URSSAF. Dans le cas de la fraude, il n'y a généralement pas d'erreurs et le montant des redressements ne correspond qu'à des irrégularités volontaires.

La fréquence de redressement correspond au rapport entre le nombre de redressements effectués (en faveur de l'URSSAF et/ou en faveur des cotisants) et le nombre d'entreprises contrôlées.

Le taux global de redressement correspond au rapport entre la somme des montants redressés (en faveur de l'URSSAF et/ou des cotisants) et la somme des montants contrôlés. Lorsque la somme des montants redressés n'inclut que les redressements en faveur de l'URSSAF, le taux est appelé taux de redressement débiteur.

Du point de vue des termes spécifiques et pour simplifier notre propos, nous discutons uniquement de montants nets, soit la différence entre les montants redressés en faveur de l'URSSAF et ceux redressés au profit du cotisant. De même, la fréquence de redressement retenue tout au long du document est celle du nombre de redressements en faveur de l'URSSAF, relativement au nombre d'entreprises contrôlées. Nous n'y ferons que rarement allusion. Comme le caractère volontaire de la fraude est, par définition, explicite, la quasi-totalité des sommes redressées dans le cadre du travail dissimulé le sont en faveur de l'URSSAF.

Une problématique du caractère explicatif des variables économiques est la définition de la situation financière de l'entreprise. Comme les données disponibles ne concernent qu'un seul bilan, nous supposons que la situation financière est définie, au minimum, relativement aux autres entreprises du même échantillon et pour un nombre de variables limité. Ces variables sont celles liées au cycle d'exploitation ; plus elles sont nombreuses, plus la situation de l'entreprise est simple à définir en choisissant un sous-ensemble de variables approprié. Il convient, toutefois, de s'abstraire d'une quelconque définition et, à la place, nous nous limitons à une *caractérisation de la situation financière*.

Une caractérisation des cotisations sociales

Une manière d'appréhender le problème de la fraude et de son contrôle consiste à tenter de caractériser les cotisations sociales. Une entreprise verse pour chaque catégorie de cotisation et pour chaque salarié un montant défini par une assiette et un taux d'application. Par souci de simplification et pour des raisons opérationnelles, les données les plus facilement accessibles sont celles ayant trait à l'entreprise, et non à chaque salarié. Chaque catégorie de cotisation versée peut être vue comme le résultat d'une assiette, d'un taux et d'une proportion d'effectif appliqués à la masse salariale de l'entreprise.

Nous supposons n entreprises, correspondant à l'ensemble des entreprises enregistrées par la Sécurité sociale. L'entreprise $i, 1 \leq i \leq n$, verse, au maximum, p cotisations de catégories différentes. p est le nombre maximal de catégories existantes. Une cotisation qui n'est pas versée par l'entreprise, par exemple si cette dernière n'est pas concernée par la catégorie, vaut simplement 0.
Notons M_i, la masse salariale de l'entreprise i, et $C_{i.}$ la somme des cotisations qu'elle verse.
Alors, pour n'importe quelle cotisation et pour n'importe quelle entreprise, C_{ij}, le montant de la cotisation j versée par l'entreprise i, est défini par :

$$C_{ij} = M_{i} a_{ij} u_{ij} t_{ij}, \tag{4.1}$$

avec $a_{ij} > 0, t_{ij} \in [-1, 1], u_{ij} \in [0, 1]$.

a_{ij} est le coefficient d'assiette de cette cotisation, soit le rapport entre l'assiette de cotisation et le salaire brut,
u_{ij} est la proportion d'effectif à laquelle s'applique la cotisation,
t_{ij} est le taux d'application de la cotisation j pour l'entreprise i ; le taux peut être négatif lorsque l'entreprise bénéficie d'une mesure de réduction.

La cotisation totale versée par l'entreprise i s'écrit alors :

$$C_{i.} = \sum_{j=1}^{p} C_{ij}.$$

Et pour l'ensemble des entreprises, le montant total des cotisations, C, vaut :

$$C = \sum_{i=1}^{n} C_{i.} = \sum_{i=1}^{n} \sum_{j=1}^{p} M_{i} a_{ij} u_{ij} t_{ij}.$$

La relation (4.1) est exacte lorsqu'il n'existe pas de fraude. Pour chaque entreprise et chaque cotisation, C_{ij} est le montant déclaré et enregistré tel quel par l'URSSAF. La masse salariale est également déclarée et le système d'information reconstitue le triplet (a_{ij}, u_{ij}, t_{ij}) à partir de l'intitulé du type de cotisation pour chaque salarié (l'assiette et le taux sont définis par la législation).

Lorsqu'il y a fraude, le système procède exactement de la même façon et considère que la relation (4.1) n'est pas modifiée, car il n'a aucun moyen de vérifier si le changement d'un paramètre est justifié. L'entreprise, elle, interprète plus librement la législation, en particulier lorsque les conditions d'application sont absconses. Ce cas est courant et conduit généralement à des irrégularités non volontaires. Les conditions d'application (par exemple, les parts plafonnée et déplafonnée) génèrent des valeurs de paramètre qui ne peuvent pas être déterminées à priori. D'autres phénomènes, comme la rotation des salariés, font varier, de manière légitime, les taux et assiettes ainsi que la masse salariale. De manière plus explicite, ce sont les entreprises qui déterminent et calculent leurs propres cotisations sur la base de la législation en vigueur. Théoriquement, une entreprise peut donc indiquer un montant C_{ij} complètement arbitraire. Le quadruplet $(M_{i}, a_{ij}, u_{ij}, t_{ij})$

est alors un vecteur aléatoire et il existe une multitude de combinaisons permettant de leurrer le système. Dans le cas du travail dissimulé, la méthode la plus simple est de considérer que pour une catégorie de cotisation j donnée, $u_{ij} = 0$. La cotisation n'a alors pas d'existence pour l'URSSAF et l'entreprise a comme gain un pourcentage de sa masse salariale. Du fait de la structure des cotisations, deux catégories (le Cas général et la CSG) obligatoires pour toutes les entreprises et représentant plus de 80% de l'ensemble des montants versés, la dissimulation de cotisations ne concerne généralement que des montants relativement faibles rapportés à la masse salariale de l'entreprise, sauf lorsque la masse salariale est elle-même partiellement dissimulée. L'importance des sommes non recouvrées par l'URSSAF est avant tout liée au nombre d'entreprises concernées par les irrégularités ou la fraude.

Du point de vue de l'URSSAF, la situation est alors la suivante :
- le montant de cotisations (C_{ij}) et la masse salariale (M_i) déclarés sont des variables aléatoires dont on observe les réalisations pour chaque entreprise et chaque cotisation.
- L'assiette, le taux et la proportion d'effectif sont des paramètres non observés, et non déclarés, car leur calcul dépend de la masse salariale et des conditions d'application de la cotisation.

Lorsqu'une fraude existe, la relation (4.1) n'est, en général, pas une égalité et le système informatique de l'URSSAF vérifie la différence entre la somme des cotisations reçues et leur valeur théorique (le produit des paramètres et de la masse salariale) qui est inconnue. Pour cela, il applique les valeurs par défaut (ou celles connues antérieurement) des paramètres d'assiette et de taux à la masse salariale déclarée et à l'effectif concerné par la cotisation à vérifier. Si l'entreprise n'a pas respecté l'assiette et le taux fixés, alors des écarts de cotisation sont décelés et une régularisation (ou une justification) lui sont demandés. Si la masse salariale et/ou l'effectif sont minorés, les écarts de cotisation sont beaucoup plus difficiles à trouver. Une alternative consiste à mesurer les variations de masse salariale sur plusieurs périodes et à contrôler les entreprises pour lesquelles, des valeurs importantes existent. Cependant, ces variations font naturellement partie du cycle économique d'une entreprise et les faux-positifs peuvent alors être nombreux. Les modèles probabilistes, bien que rarement utilisés, fournissent une approximation du problème des écarts de cotisation en ne considérant que les couples (C_{ij}, M_i) ou plus précisément les variables, aléatoires, C_{ij}/M_i. Leur distribution conditionnelle (selon qu'une fraude est observée ou non) mène alors à une distinction entre les cas les plus probables et les moins probables de fraude. Une alternative (moins efficiente) est le croisement de données (ou fichiers) associés à des seuils de décision. Pour les irrégularités, cette méthode est couramment utilisée. Cependant, pour la grande majorité des cas de travail dissimulé, ce sont des signalements ou les échanges avec les partenaires de l'URSSAF qui permettent un taux de détection important.

En 2011, plus de 75% des entreprises contrôlées en France (près de 80% en Île-de-France), dans le cadre de la lutte contre le travail dissimulé, ont été redressées. Ce taux était de 55% (en Île-de-France, et au plus haut depuis 2006) dans le cas des contrôles comptables d'assiette, lesquels constituent la principale forme de vérification des cotisations ainsi que la principale source de redressements (en volume et valeur).

4.4 Données et protocole

Nous présentons tout d'abord et de manière générique les données brutes, puis nous détaillons la manière dont elles ont été traitées pour obtenir l'échantillon de travail. Pour plus de clarté, les variables sont citées et brièvement décrites en fin de section, puis détaillées en annexe. Les données économiques ont été recueillies manuellement, soit pour une entreprise à la fois, et fournies par un prestataire de l'URSSAF d'Île-de-France. Les autres données sont issues de l'enregistrement des déclarations de cotisation effectuées par les entreprises auprès de l'URSSAF.

4.4.1 Données brutes

L'échantillon de données réelles est initialement constitué de 923 entreprises de la région Île-de-France contrôlées en 2009, dans le cadre de la lutte contre le travail dissimulé [1]. 24 variables économiques explicatives sont enregistrées ; leurs valeurs sont issues de la situation économique et financière de chaque entreprise.

Pour chacune, nous disposons également, pour chaque variable, de la valeur moyenne mesurée pour les entreprises du même secteur d'activité.

La variable à expliquer, soit le niveau de fraude, est représentée par le rapport entre les montants (nets) de redressement effectués par l'URSSAF, dans le cadre de sa politique de contrôle du travail dissimulé, et la masse salariale de l'entreprise.

i) Par obligation légale, une entreprise ne peut être contrôlée que sur ses trois derniers exercices comptables ainsi que sur l'année en cours au moment du contrôle. L'ensemble des entreprises a été contrôlé en 2009, dont 80% après le premier trimestre. Les exercices comptables des entreprises sont essentiellement (aux deux tiers) ceux de cette même année. La masse salariale retenue est donc celle des trois derniers exercices comptables et de l'année 2009 (ainsi que les autres variables enregistrées par l'URSSAF). Elle est ensuite annualisée.

Une première hypothèse de travail est l'inertie de la fraude et des variables économiques : Il n'y a pas de sauts dans les variations de situation économique des entreprises et il n'y a pas de sauts dans les variations du niveau de fraude.

ii) L'échantillon initial représente environ 60% de toutes les entreprises contrôlées en Île-de-France, en 2009, dans le cadre du travail dissimulé. Environ 80% ont été redressées pour un montant net en faveur de l'URSSAF. Cette proportion est d'environ 90% pour les données de l'échantillon. La forte proportion d'entreprises contrôlées et redressées s'explique par les partenariats féconds entre l'URSSAF et d'autres organismes (gendarmerie, inspection du Travail,...) et par le signalement des salariés victimes.

1. Ces entreprises sont sélectionnées essentiellement à partir de la coopération entre l'URSSAF et les organismes partenaires de la lutte contre le travail dissimulé. Il n'y a, en particulier, pas ou peu de sélection basée sur les données.

iii) La faible proportion d'entreprises contrôlées et non-redressées (2%) est essentiellement due à la difficulté de trouver des entreprises présentant un bilan comptable (même partiel) consultable de leurs activités. Pour constituer l'échantillon initial, la recherche de bilan a porté sur 1300 entreprises (parmi près de 1500 d'entre elles, contrôlées).

iv) Les entreprises sont issues de 210 sous-classes d'activité (code NAF - Nomenclature d'Activités Française de l'INSEE) et sont regroupées en 55 secteurs d'activité (divisions NAF). 6 secteurs d'activité regroupent 70% des données. Nous les citons, par ordre décroissant d'importance, ci-après.
- Restauration
- Travaux de construction spécialisés
- Commerce de détail, à l'exception des automobiles et des motocycles
- Commerce de gros, à l'exception des automobiles et des motocycles
- Autres services personnels
- Construction de bâtiments

`Filtrage, données manquantes et pré-traitement`

v) - La difficulté à obtenir des bilans comptables complets génère environ 28% de valeurs manquantes, nulles ou incohérentes (rentabilité excessive, solvabilité négative,...).
- De manière identique, environ 30% des variables mesurant la moyenne des différents secteurs d'activité présentent des valeurs incohérentes. Ces valeurs correspondent aux queues de distribution pour chaque variable. Il n'est pas possible d'éliminer ces valeurs sous peine de réduire fortement la taille de l'échantillon. Nous procédons alors de la manière suivante.
- Dans la matrice correspondant aux valeurs moyennes des différents secteurs d'activité, nous filtrons toutes les données sous le quantile d'ordre 0.1 ou au-dessus du quantile d'ordre 0.9. Les valeurs filtrées sont considérées comme des données manquantes et un algorithme d'imputation de valeurs se charge de reconstruire la matrice complète.
- La même procédure est répétée pour les valeurs manquantes ou incohérentes des données réelles.

vi) Aux variables économiques est ajoutée une partie des variables (effectif, masse salariale, durée de vie, retards de paiement, taux de cotisation,...) enregistrées par l'URSSAF dans le cadre de la déclaration de cotisations sociales des entreprises. Cet ajout correspond à une deuxième hypothèse de travail : *les données économiques seules ne suffisent pas à caractériser la fraude dans le cadre du travail dissimulé.*

vii) Quatre variables sont reconstituées ou construites : les productivités du travail et du capital, la compliance et le taux de cotisation principal. Les données sont ensuite filtrées de façon à prendre en compte plusieurs spécificités que nous détaillons ci-après :
- une entreprise peut posséder plusieurs établissements ;
- quelques entreprises sont enregistrées avec une masse salariale ou un effectif nuls ;
- certaines variables économiques ont une corrélation très élevée (> 0.9) ou présentent des valeurs atypiques.
- Certains redressements sont très importants et peuvent influencer l'inférence.

viii) Un dernier filtre est appliqué sur les données afin de limiter l'influence de certaines variables et préserver la cohérence :
- les variables économiques sont essentiellement des ratios et lorsque ce n'est pas le cas, nous divisons l'ensemble des observations dont l'unité est monétaire par la masse salariale de l'entreprise considérée ;
- Les variables exprimées en jours voient leurs valeurs divisées par 365 ;
- Les effectifs et la masse salariale sont supprimés des données.

Après pré-traitement des données, nous obtenons un échantillon de 814 entreprises et 43 variables, dont 24 de nature économique. La variable à expliquer est le rapport entre le montant net des redressements et la masse salariale. Cette variable est équivalente à une mesure du niveau de fraude. Plus il est élevé, plus la fraude est importante. Dans le cas des grandes entreprises, un niveau de fraude faible peut être associé à des montants redressés élevés, certains dépassant le million d'euros. Ces entreprises (de plus de 250 salariés) présentent une masse salariale se situant, généralement, au-dessus de 10 millions d'euros et tendent à être systématiquement contrôlées par l'URSSAF.

`Variables économiques`

"MARGE BRUTE D'AUTOFINANCEMENT"	"COUVERTURE du BFR"	"COUVERTURE des IMMOS NETTES"
"COUVERTURE du CA"	"SOLVABILITE"	"INDEPENDANCE FINANCIERE"
"RENTABILITE ECONOMIQUE"	"RENTABILITE FINANCIERE"	"RENTABILITE COMMERCIALE"
"CONTRIBUTION DU CAPITAL"	"CONTRIBUTION DE LA VA"	"LIQUIDITE IMMEDIATE"
"LIQUIDITE GENERALE"	"LIQUIDITE REDUITE"	"ENDETTEMENT"
"CAPACITE DE REMBOURSEMENT"	"FINANCEMENT DES STOCKS"	"PRODUCTIVITE DE L'ACTIF"
"DUREE CLIENT"	"DUREE FOURNISSEUR"	"POIDS MASSE SALARIALE"
"PRODUCTIVITE"	"PRODUCTIVITE DU TRAVAIL"	"PRODUCTIVITE DU CAPITAL"

Nous indiquons de façon succincte les acronymes cités ci-dessus :"CA" représente le chiffre d'affaires, "BFR" est le besoin en fonds de roulement, "VA" est la valeur ajoutée.

`Variables de la déclaration de cotisation`

"DUREE DE VIE"	"NB ETS"	"NB ORIG. DEBIT NON RENS."
"MONTANT DEBIT"	"NB CTP EXONERATION"	"NB REMISES SUR MAJORATION"
"MONTANT REMISES SUR MAJORATION"	"MONTANT ECARTS"	"PENALITES"
"NB RETARDS"	"NB TAXATIONS D'OFFICE"	"MONTANT TAXATIONS D'OFFICE"
"NB DEMANDES DELAIS"	"NB DELAIS ACCEPTES"	"% VERSEMENT DEMAT."
"DERNIER CTRL"	"NB CCA"	"COMPLIANCE"
"Tx COTISATION NET"		

Nous donnons une brève description de la situation déclarative d'une entreprise : "NB ETS" représente le nombre d'établissements de l'entreprise. Le "débit" est la somme des montants de déclaration dus par l'entreprise et non réglés. Chaque débit correspond à une catégorie déclarative, par exemple les heures supplémentaires ou encore la CSG, nommée "CTP" (code-type de personnel). Ces "CTP" peuvent être des "exonérations" de cotisations et peuvent conduire, lorsque les "retards" sont importants, à des "majorations" de cotisation ou des "taxations d'office". Une entreprise peut demander un ou plusieurs "délais" de paiement et peut régler, de plus en plus fréquemment, ses cotisations par voie dématérialisée ("DEMAT."). La date du dernier contrôle ("CTRL") et le

nombre de contrôles comptables d'assiette ("NB CCA") effectués par les inspecteurs de l'URSSAF servent, éventuellement, à déterminer le profil de l'entreprise. Tout comme la "compliance", mesure de sa conformité à respecter ses obligations légales, et le taux de cotisation net estimé de la principale catégorie obligatoire.

4.4.2 Echantillon synthétique

La présence quasi systématique d'un niveau de fraude positif dans les données réelles altère la complétude du caractère explicatif des variables économiques. La quasi indisponibilité de données d'entreprises contrôlées et non-redressées est principalement due à trois facteurs :
- un taux de détection élevé ($> 75\%$) du travail dissimulé ;
- la difficulté de trouver un bilan comptable publié et à jour (pour la construction de l'échantillon, au moins, 30 % des entreprises n'avait pas de bilan disponible) ;
- l'absence de bases de données communes entre les données économiques des entreprises et leur déclaration de cotisations.

Pour y pallier, nous définissons alors un échantillon synthétique d'entreprises auquel nous attribuons un niveau de fraude aléatoire.

`Protocole`

On souhaite créer un grand échantillon représentatif des entreprises constitué de données synthétiques. Nous tirons aléatoirement, et proportionnellement à chacun des six secteurs d'activité les plus représentatifs, toute nouvelle observation.

- Nous générons tout d'abord les secteurs d'activité en les tirant uniformément selon leur répartition dans les données réelles. Ainsi, la distribution des secteurs d'activité est la même dans les données synthétiques que dans les données réelles.

Pour chacun des secteurs d'activité :
- la variable économique correspondante a pour moyenne celle du secteur d'activité ;
- sa variance est celle de la variable dans l'échantillon des données réelles.

Ces paramètres sont associés à une distribution de loi gaussienne pour le tirage de chacune des réalisations d'une variable.

De la même manière :
- une variable de la déclaration de cotisations a pour moyenne celle de la totalité des entreprises d'Île-de-France, pour le même secteur d'activité ;
- sa variance est celle de la variable dans l'échantillon des données réelles, pour le même secteur d'activité.

Nous générons ensuite un niveau de fraude aléatoire :
- pour la moitié des entreprises de l'échantillon synthétique, le niveau de fraude est nul ;
- pour les autres, le niveau de fraude est tiré selon la loi exponentielle, de paramètre la moyenne des niveaux de fraude observés (dans l'échantillon initial) sous la médiane.

Pour chaque variable de l'échantillon synthétique, ses réalisations sont alors un mélange des réalisations de la variable relativement à chaque secteur d'activité. Notons qu'un niveau de fraude positif pour la moitié des entreprises se justifie par le fait que nous disposons d'un modèle prédictif qui peut détecter la fraude avec une probabilité de succès minimale, supérieure à 50%.

- Afin d'éviter les valeurs atypiques ou incohérentes, les observations (pour l'échantillon synthétique uniquement) inférieures au quantile d'ordre 0.1 ou supérieures à celui d'ordre 0.9, sont exclues. L'échantillon final est le mélange des données réelles et de l'échantillon synthétique.

4.4.3 Statistiques

Nous présentons un résumé des données à travers quelques statistiques. L'échantillon final est constitué de 10 000 entreprises (dont 92% sont issues de l'échantillon synthétique).
Nous indiquons à nouveau les secteurs d'activité les plus représentés :
- Restauration
- Travaux de construction spécialisés
- Commerce de détail, à l'exception des automobiles et des motocycles
- Commerce de gros, à l'exception des automobiles et des motocycles
- Autres services personnels
- Construction de bâtiments

`Fraude et niveau de fraude`

Le montant moyen des redressements effectués par l'URSSAF (données initiales) est d'environ 27 500 euros par entreprise (médiane de 8300 euros) avec une grande dispersion (écart-type de 100 000 euros). La masse salariale annuelle moyenne est de 515 000 euros (médiane de 66 000 euros). Même si l'action de contrôle est, à l'origine, le travail dissimulé, seuls 60% des redressements le sont effectivement pour cette raison. Le reste est essentiellement de la minoration d'assiette de cotisation.

Niveau de fraude	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Données réelles	-0.005	0.01	0.027	0.069	0.074	0.865
Données synthétiques	0	0	0	0.006	0.008	0.123
Données synthétiques + réelles	-0.005	0	0.001	0.011	0.01	0.865

TABLE 4.1 – Niveau de fraude dans les données réelles (entreprises contrôlées par l'URSSAF), les données synthétiques et l'échantillon final.

Le niveau de fraude dans l'échantillon synthétique est, en moyenne, 10 fois plus petit que dans les données réelles, essentiellement car il est, dans la pratique, inconnu et estimé entre 1 et 3% de la masse salariale de toutes les entreprises, au niveau national. Les niveaux de fraude négatifs correspondent à des contrôles ayant entrainé des remboursements de cotisation aux entreprises.

`Indicateurs`

Pour les variables de la déclaration de cotisations, nous illustrons deux indicateurs potentiels de l'analyse descriptive :
- La compliance est capacité à se conformer aux obligations de déclaration. Au-dessus d'un indice de 0.5, la conformité est considérée comme correcte. Elle est définie relativement à l'obligation faite par la législation aux entreprises de justifier les mesures de réduction (ou les écarts de taux relativement au taux d'application usuel d'une cotisation) dont elles bénéficient. En particulier, elles doivent alors déclarer l'effectif concerné par la mesure de réduction. La compliance se mesure alors en calculant le rapport entre le nombre de catégories de cotisation justifiées, augmenté d'une unité, et le nombre total de catégories déclarées. Si l'entreprise ne bénéficie d'aucune mesure de réduction, sa compliance vaut alors 0.5.
- Le taux de cotisation estimé pour le Cas général, que nous comparons à son niveau théorique de 36.1% (en 2009). Le taux de cotisation estimé est un taux reconstitué à partir des assiettes (plafonnée et déplafonnée), des cotisations versées et de la masse salariale déclarée.

Pour les variables économiques, plusieurs d'entre elles fournissent un aperçu de la situation financière dans l'échantillon des entreprises. Nous illustrons ci-dessous quelques indicateurs ainsi que la distribution de certaines des variables de l'échantillon initial et de l'échantillon synthétique.

Variables	Données réelles	Echantillon synthétique
compliance	0.48 (0.15)	0.48 (0.08)
Taux de cotisation principal	0.30 (0.08)	0.27 (0.05)
Marge brute d'autofinancement	0.36 (1.48)	0.39 (0.32)
Solvabilité	0.43 (0.42)	1.73 (0.46)
Indépendance financière	0.65 (0.18)	0.56 (0.12)
Rentabilité commerciale	0.05 (0.04)	-0.04 (0.07)
Rentabilité économique	0.08 (0.08)	0.05 (0.05)
Liquidité générale	2.09 (1.27)	5.44 (0.75)
Endettement	0.87 (0.69)	1.64 (0.79)
Capacité de remboursement	68 (96)	38 (34)
Productivité (en milliers d'euros)	116 (66)	229 (447)
Productivité du travail	7.19 (5.50)	7.88 (6.37)
Productivité du capital	8.51 (44)	2.72 (4.55)
Poids de la masse salariale	0.84 (0.14)	0.75 (0.09)
Niveau de fraude	0.07 (0.11)	0.006 (0.01)

TABLE 4.2 – Moyenne et écart-type, entre parenthèses, de quelques variables.

Nous détaillons la distribution de quelques unes des variables présentées ci-dessus dans les graphiques qui suivent. La définition de l'ensemble des variables est donnée en annexe. Notons que la moyenne dans l'échantillon synthétique est celle de la population de toutes les entreprises, soit la moyenne pondérée des secteurs d'activité, pour chaque variable considérée. De manière générale, la distribution des variables dans l'échantillon synthétique dépend du poids des secteurs d'activité représentés.

Quatre points peuvent être notés dans les statistiques représentées :
- La solvabilité [2] est, en moyenne, trois fois plus importante dans la population de toutes les entreprises que dans celles contrôlées.
- Cette tendance est similaire pour la liquidité générale [3].
- L'endettement [4] et la capacité de remboursement [5] présentent également de grandes différences. L'endettement est deux fois plus petit dans les données réelles, tout comme la capacité de remboursement. Pour cette dernière, plus la valeur est élevée moins la capacité de remboursement est importante. Plus spécifiquement, les entreprises réellement contrôlées ont moins de dettes à long terme (relativement à leurs fonds propres) mais plus de dettes bancaires (relativement à leur capacité d'autofinancement) que la moyenne des entreprises.
- Le dernier point est lié à la productivité [6]. Elle est deux fois moins importante dans les données réelles. Bien que la productivité du capital [7] soit également bien plus élevée, sa valeur moyenne est surtout due à de nombreuses valeurs extrêmes (5% des données).

2. rapport entre les capitaux propres et l'ensemble des dettes
3. ratio entre l'actif circulant net (stocks, créances et valeurs mobilières de placement) et les dettes à court terme
4. ratio entre les dettes à long terme et les capitaux propres
5. ratio entre les dettes bancaires et la capacité d'autofinancement
6. ratio entre le chiffre d'affaires et l'effectif
7. ratio entre le chiffre d'affaires et le coût du capital

Sans ces dernières, la productivité du capital est, en moyenne, similaire à celle des entreprises d'Île-de-France.

Globalement, l'équilibre financier (solvabilité) et la capacité de remboursement sont plus fragiles dans les entreprises réellement contrôlées, tout comme la liquidité. Leur productivité est également plus faible. A l'inverse, leur endettement est moins important.

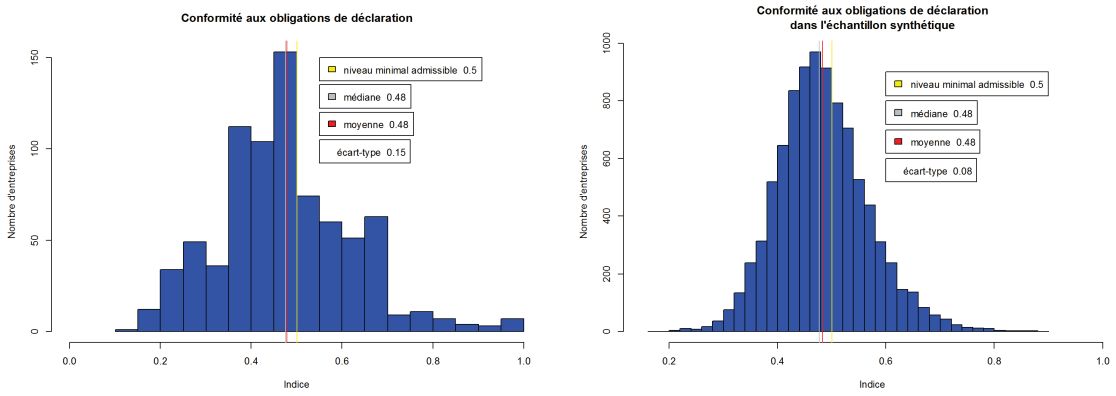

FIGURE 4.1 – Niveau de conformité aux obligations de déclaration de cotisations dans les données réelles et synthétiques.

Pour un niveau de conformité admissible de 0.5, près de la moitié de l'échantillon présente un déficit d'informations sur la nature de ses déclarations. L'indice moyen est de 0.48 contre 0.53 pour l'ensemble des entreprises d'Île-de-France. La conformité moyenne des entreprises de l'échantillon synthétique est similaire à celle des entreprises redressées pour fraude, tandis que la dispersion mesurée, 0.08 contre 0.15 dans les données réelles, est presque deux fois moins élevée.

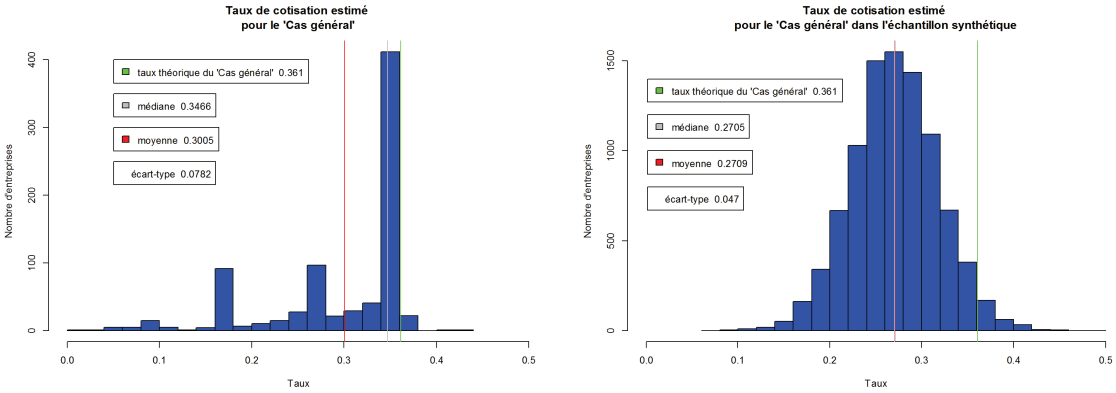

FIGURE 4.2 – Répartition du taux de la principale catégorie de cotisation.

Le Cas général est la principale catégorie de cotisation. Elle est versée par toutes les entreprises mais, comme cela est usuellement le cas pour les cotisations sociales, supporte des exceptions. Généralement s'y ajoute la CSG (8%), les cotisations Chômage (6.4%),

les Accidents du travail (taux variable), éventuellement les Retraites complémentaires (taux variable), et les Mesures de réduction (taux négatif variable) pour obtenir le taux effectif de cotisation de l'entreprise. A posteriori, les taux de cotisation en dessous du niveau théorique posent la question de leur répartition selon les entreprises et de l'impact sur les recettes attendues par l'URSSAF. Dans l'échantillon synthétique, le taux moyen de cotisation pour le "Cas général" (27.09%) est plus petit que celui observé dans les cas de fraude réels (30.05%). La différence peut s'expliquer par une meilleure utilisation des mesures de réduction de cotisations associées à des salaires moyens proches du SMIC, lesquelles permettent la réduction du taux moyen et justifient alors l'écart avec le taux théorique attendu. Notons que le taux global (toutes cotisations incluses) est, en moyenne, de 37% pour les entreprises d'Île-de-France, en 2009.

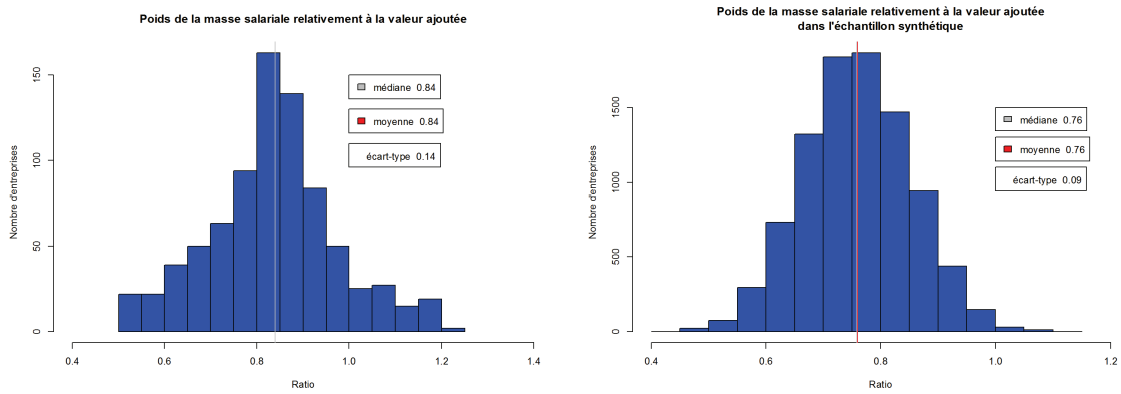

FIGURE 4.3 – Poids de la masse salariale.

Le poids de la masse salariale est un peu plus faible dans les données réelles, avec une dispersion plus importante. Pour ces entreprises, la masse salariale représente, en moyenne, un peu plus de 80% de la valeur ajoutée (au dénominateur du poids de la masse salariale). Cette dernière correspond à la marge commerciale augmentée de la valeur de la production (vendue, stockée et immobilisée) et diminuée de la consommation en provenance de tiers. Environ 20% des entreprises voient leur masse salariale absorber la totalité de la valeur ajoutée créée (poids de la masse salariale supérieur à 1). Cette proportion est inférieure à 1% pour les entreprises de l'échantillon synthétique.

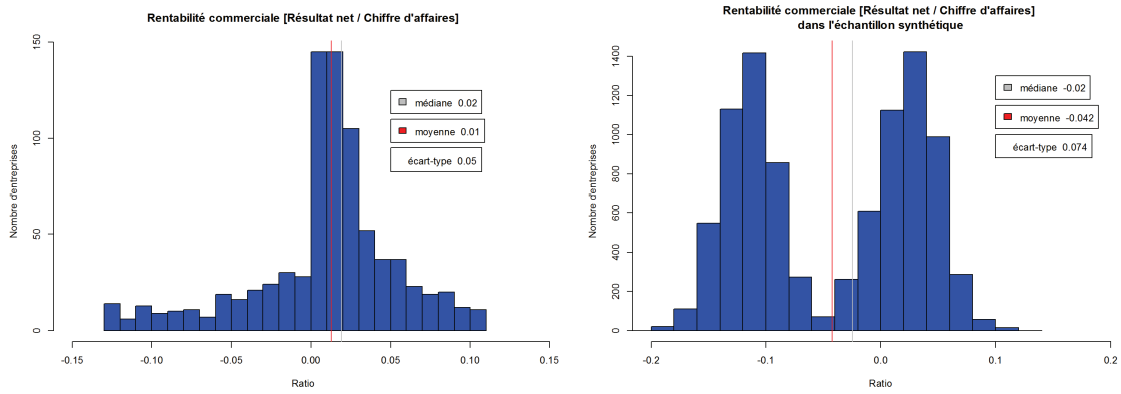

FIGURE 4.4 – Rentabilité commerciale.

La rentabilité commerciale, dans l'échantillon synthétique, possède deux modes qui représentent les différences de rentabilité entre secteurs d'activité. En particulier, dans certains sous-secteurs de la restauration, fortement représentés dans les données, la rentabilité commerciale moyenne pourrait avoir été négative durant l'année 2009.

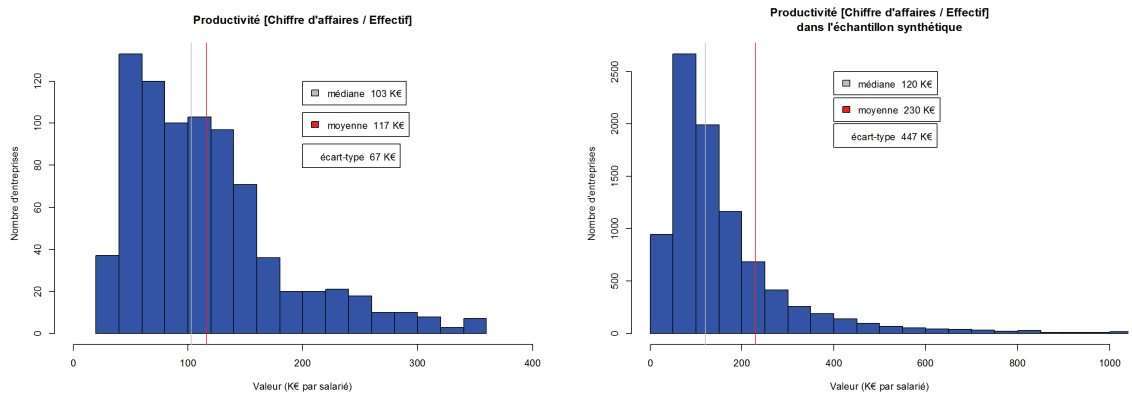

FIGURE 4.5 – Productivité.

La différence de productivité moyenne, entre les données réelles et celles de l'échantillon synthétique, est assez importante, bien que les médianes soient proches. L'écart s'explique, en partie, par une plus grande proportion de fortes valeurs de productivité, au filtrage appliqué pour éviter les valeurs incohérentes dans les données initiales et à l'effectif (en moyenne, environ 19 salariés contre 6 salariés dans l'échantillon synthétique).

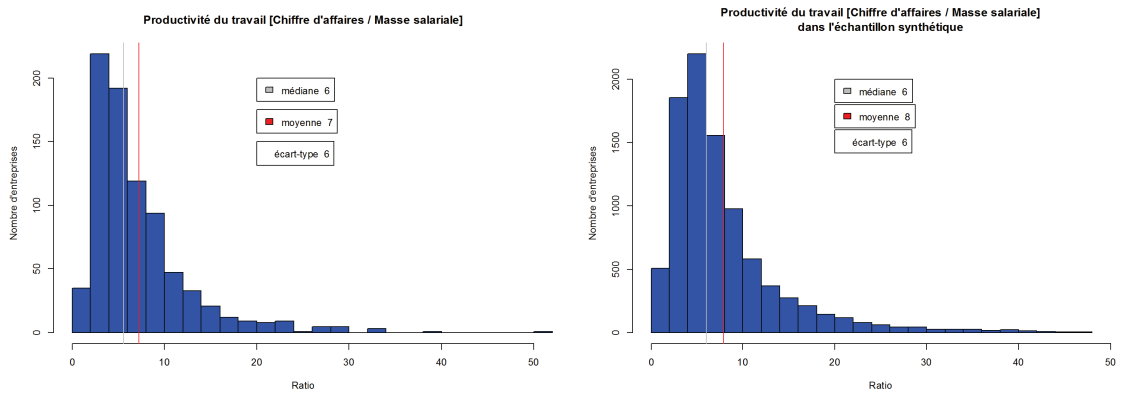

FIGURE 4.6 – Productivité du travail.

Dans la continuité du graphique précédent, la masse salariale pèse peu sur le chiffre d'affaires, en moyenne moins de 15%. La connexion entre la productivité et la productivité du travail peut être établie de la manière suivante : le coût du travail (relativement au chiffre d'affaires) dans les données réelles ne représente une charge importante que pour une minorité d'entreprises. Pour toutes les autres, ce coût est similaire à celui des entreprises de l'échantillon synthétique et leur problème fondamental est celui de l'accroissement de la productivité des salariés.

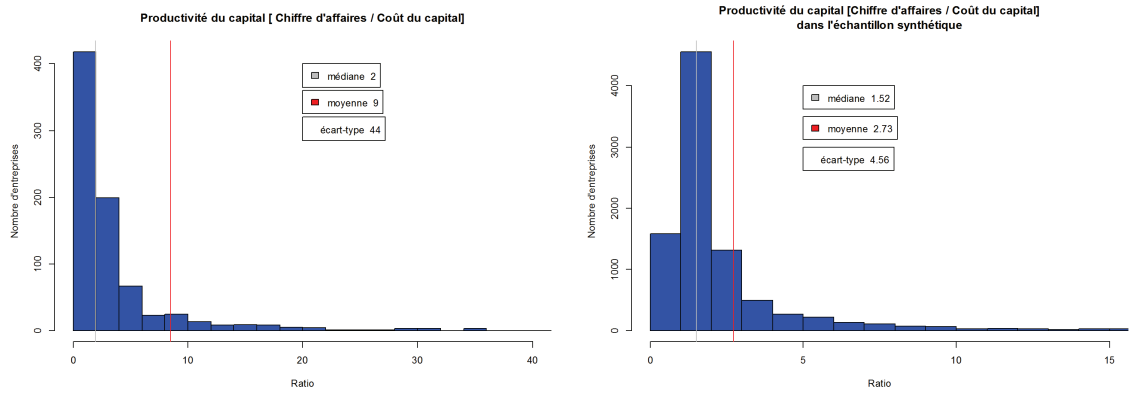

FIGURE 4.7 – Productivité du capital.

Le coût du capital est défini, ici, comme la somme des coûts de production (hors coût du travail) augmentés de la différence entre le chiffre d'affaires et la valeur de la production. Le coût du capital pénalise les moins-values lorsque la valeur de la production est surévaluée. Notons qu'il est reconstitué à partir d'autres variables économiques.

La productivité du capital est très sensiblement inférieure à celle du travail dans l'échantillon synthétique. Les salariés constituent, à priori, le principal moteur de croissance des entreprises. Dans les données réelles, la productivité du capital est importante seulement en moyenne, principalement du fait de la grande dispersion observée. Dans les deux échantillons, les médianes sont proches et dans, environ, 10% des entreprises (et 15% dans l'échantillon synthétique), la productivité du capital est inférieure à 1.

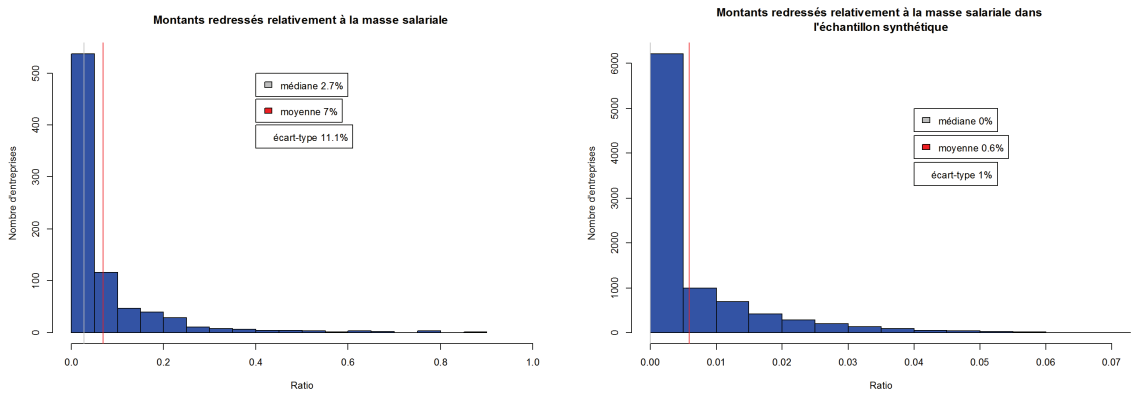

FIGURE 4.8 – Niveau de fraude.

Le niveau de fraude dans l'échantillon synthétique, aléatoire, est en moyenne faible (0.6%) puisque la moitié des entreprises ont un niveau nul, par construction. En moyenne, la masse salariale annuelle (somme des salaires bruts) est de 172 000 euros (médiane de 115 000 euros) pour un effectif moyen de 6 salariés. Précisons que le niveau de fraude est défini relativement à la masse salariale et aux cotisations cumulées des trois dernières années avant le contrôle. Un peu plus de 20% des entreprises ont un niveau de fraude dépassant 1% ; pour les entreprises réellement redressées dans le cadre de la lutte contre le travail dissimulé, cette proportion est de 75%. Dans l'échantillon synthétique, lorsque le niveau de fraude est positif, il représente en moyenne 1.2% de la masse salariale annuelle, soit environ 6000 euros par entreprise redressée, pour les trois années de cotisations déclarées, avec une médiane de 2700 euros. Dans les données réelles, le niveau de fraude est plutôt important. Il correspond à 28 200 euros (en moyenne) par entreprise, soit 7% de la masse salariale annuelle. Toutefois, seules 15% des entreprises fraudent pour un montant supérieur au montant moyen d'un redressement. Le montant médian d'un redressement, 8500 euros, est plus représentatif de la fraude.

Cette observation est une constante de la lutte contre la fraude : un petit nombre d'entreprises constitue la grande majorité de l'ensemble des montants redressés.

Sur les 22 millions d'euros de redressements de notre échantillon de données réelles, 75% ont été générés par 126 entreprises (15%). Un certain nombre d'entreprises fraudent de manière très importante ($> 10\%$ de la masse salariale annuelle) tandis que la majorité se situe à moins de 5%. Le niveau de fraude a, également, un impact (quasi) linéaire sur la marge des entreprises. En particulier, le poids de la masse salariale dans la valeur ajoutée diminue proportionnellement à l'augmentation du niveau de fraude.
En matière d'emplois, le montant moyen de redressement par entreprise représente environ un emploi sur un an, payé au salaire minimum. L'effectif moyen est de 19 salariés. L'effectif médian est de 4 salariés et 75% des entreprises emploient moins de 8 salariés. La masse salariale annuelle est, en moyenne, de 535 000 euros (médiane de 68 000 euros) et la durée d'existence moyenne d'une entreprise est de 12 ans.

Synthèse

Dans le cas des données réelles, nous résumons une partie des informations obtenues : les salaires pèsent sur la valeur ajoutée[8] des entreprises (en moyenne près de 90%), beaucoup moins sur le chiffre d'affaires (environ 10%), et la moitié d'entre elles présentent une rentabilité commerciale positive ($> 2\%$). Le poids de la masse salariale sur la valeur ajoutée indique que les réductions de coût peuvent avoir un effet important lorsqu'elles portent sur les salaires ou l'effectif. Cet effet est concomitant avec des efforts effectués au niveau de la productivité du travail, lesquels se traduisent par un grand facteur d'écart entre le chiffre d'affaires et la masse salariale. Cette dernière est cependant celle déclarée ; le niveau de productivité et le poids de la masse salariale correspondent donc à ceux consécutifs à la fraude aux cotisations sociales. A coûts constants, lorsque l'entreprise ne peut pas augmenter sa marge en accroissant son chiffre d'affaires, le choix le plus avantageux est de réduire sa masse salariale sans réduire la productivité du travail. La minoration des cotisations sociales est un moyen d'y parvenir avec un risque faible. Ce moyen est d'autant plus efficace que la productivité du travail est supérieure à celle du capital. Cette situation est plus naturelle lorsque l'entreprise rencontre ou anticipe des difficultés. Cependant, même des entreprises avec un niveau de rentabilité du chiffre d'affaires (rentabilité commerciale) positif ont recours à la fraude. Pour ces entreprises, plusieurs hypothèses peuvent être faites, comme la probabilité d'être contrôlé, la complexité de la législation, l'impossibilité de diminuer le coût du capital ou encore une difficulté temporaire comme un manque de liquidités.

4.5 Modèles

Nous supposons, Y, la variable traduisant le niveau de fraude et $X = (X^{(1)}, X^{(2)}, ..., X^{(p)})$, le vecteur aléatoire représentant les variables explicatives du phénomène. Pour expliquer ce niveau, nous supposons que sa relation avec les variables est linéaire.
Nous posons alors :

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(m, \sigma^2 I),$$

avec β le vecteur des paramètres associés aux variables explicatives,
et ϵ, un bruit gaussien de paramètres $(m, \sigma^2 I)$ traduisant l'incertitude que nous avons sur le phénomène.
Les paramètres sont estimés à partir des observations de l'échantillon. L'inférence sur Y permet d'obtenir un estimateur $\hat{Y}$ défini par :

$$\hat{Y} = X\hat{\beta},$$

avec

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

où X' est la transposée de X.
Pour évaluer le caractère significatif d'une variable explicative, un test de Student est

8. Somme des salaires et cotisations, des impôts dus à l'Etat et d'une part restante à l'entreprise qui constitue l'excédent brut d'exploitation (EBE)

réalisé pour chaque coefficient $\beta_j, 1 \leq j \leq p$. Une variable $X^{(j)}$ est significative si $\beta_j \neq 0$ avec une grande probabilité.
On teste alors l'hypothèse $H_0 : \beta_j = 0$, contre $H_1 : \beta_j \neq 0$.
Sous H_0,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 (X'X)^{-1}_{j,j}}} \sqrt{n-p-1} \sim \mathcal{T}_{n-p-1},$$

où $\mathcal{T}_{n-p-1}$ est la loi de Student à $n-p-1$ degrés de liberté.

Nous calculons (sous le logiciel de calcul statistique R) la p-value du test, soit la probabilité de rejeter à tort l'hypothèse H_1. Plus elle est petite, plus le paramètre est probablement différent de 0, et plus la variable associée est significative.

La régression linéaire a comme inconvénients de ne capturer ni les effets non linéaires d'un phénomène, ni les effets locaux d'une variable. De plus, la *significativité* d'une variable est liée à des hypothèses qui peuvent ne pas être vérifiées dans la pratique. Nous souhaitons bénéficier d'un point de vue sur ces effets et les visualiser afin d'en permettre une interprétation immédiate. Pour cela nous faisons appel à un modèle non linéaire, les forêts uniformément aléatoires, variante des forêts aléatoires de Breiman (2001). Ces dernières ont l'avantage d'être non paramétriques, de ne pas nécessiter d'hypothèses sur les données et de saisir l'influence des variables explicatives. Une forêt aléatoire fournit également un point de vue local, grâce aux dépendances partielles entre variable explicative et variable à expliquer. Cela revient à observer l'effet d'une variable explicative sur la variable à expliquer, sachant la distribution de toutes les autres.
Avant de détailler le modèle, nous redéfinissons d'abord la relation entre Y et X.
On suppose :

$$Y = f(X) + \epsilon,$$

où f est une fonction quelconque des données et ϵ est aléatoire. Notons que cette relation est purement abstraite. Dans la pratique, nous supposons seulement qu'il existe une relation entre les variables explicatives et la variable à expliquer et qu'on souhaite estimer Y, puis le prédire pour des futures observations (inconnues du modèle) de X.
Une forêt aléatoire est un ensemble d'arbres de décision et son estimateur est donné par la moyenne des estimateurs des arbres sous-jacents. Leur idée générale consiste à considérer que la fonction f est aléatoire aussi bien par les variables prises en entrée, que par les observations qui les constituent. Une manière de reconstituer f est alors la génération d'un grand nombre de modèles approximant chacun, et de manière aléatoire, f, puis leur moyenne, comme une méthode de Monte Carlo. Considérons un arbre de décision binaire. Il correspond à une structure algorithmique qui partitionne de manière récursive l'espace formé par les variables X. A chaque étape (noeud) du partitionnement, l'espace est divisé en deux régions. On procède ainsi pour chaque nouvelle région créée, jusqu'à ce qu'un ou plusieurs critères d'arrêt soient atteints. Une région est définie comme une sous-partition dont la construction repose sur des règles spécifiques à l'arbre de décision. Une région qui ne peut plus être partitionnée est une feuille de l'arbre. La règle de décision de l'arbre attribue une unique valeur à chaque feuille, définie par la moyenne (dans le cas de la régression) des observations de Y dans cette région.
Pour préciser notre propos, nous l'illustrons avec un exemple simple : supposons un

arbre qui ne contienne que deux régions (en plus de la partition de départ constituée de toutes les données). Pour simplifier, on suppose également que X est réduit à une seule variable et que x est fixé. La première région est constituée de toutes les observations pour lesquelles $X > x$, et la seconde de toutes les observations pour lesquelles $X \leq x$. La règle de décision g, de l'arbre, est définie par :

$$g_n(x) = \begin{cases} \frac{1}{\sum_{i=1}^n \mathbf{I}_{\{X_i > x\}}} \sum_{i=1}^n Y_i, \ si \ X_i > x, pour \ tout \ i \in [1, n], \\ \\ \frac{1}{\sum_{i=1}^n \mathbf{I}_{\{X_i \leq x\}}} \sum_{i=1}^n Y_i, \ sinon. \end{cases}$$

g est un estimateur de $\mathbf{E}(Y|X)$, la meilleure approximation de Y qu'on puisse obtenir. Elle est construite de sorte que la distance $\sum_{i=1}^n (Y_i - g_n(X_i))^2$ soit minimale. Dans notre exemple, g n'est cependant pas sans biais et sa variance n'est pas minimale. Pour obtenir un estimateur sans biais non linéaire, on laisse le nombre de régions s'accroître jusqu'à ce qu'il ne reste plus qu'une observation par région. On obtient alors un estimateur de Y plus général, défini par :

$$g_{\mathcal{P}}(x) = g_{\mathcal{P}}(x, R) = \frac{1}{\sum_{i=1}^n \mathbf{I}_{\{X \in R\}}} \sum_{i=1}^n Y_i \mathbf{I}_{\{X_i \in R\}}, \ x \in R.$$

$g_{\mathcal{P}}$ est la règle de décision d'un arbre de décision uniformément aléatoire,
R est la région de l'arbre à laquelle appartient l'observation x,
$\mathcal{P}$ est la partition de l'ensemble des observations telles que $R \in \mathcal{P}$.

Nous illustrons un tel arbre ci-dessous :

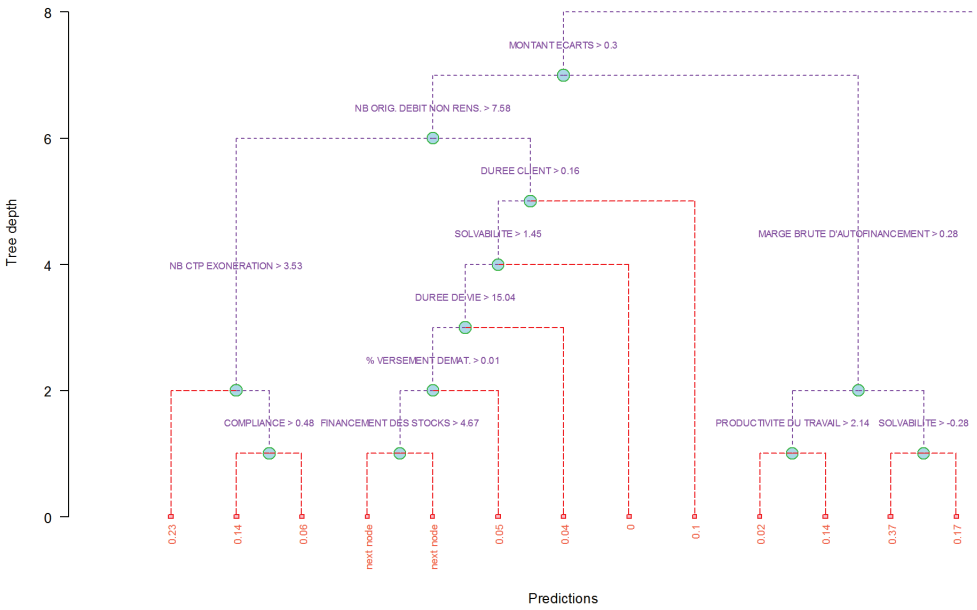

FIGURE 4.9 – Une partie d'un arbre de décision uniformément aléatoire.

Le graphique représente une partie d'un arbre de décision sur l'échantillon entier (données synthétiques et réelles) dans lequel chaque noeud (région) possède deux branches. La variable, au-dessus du noeud, dont les réalisations sont supérieures à une valeur (tirée selon la loi Uniforme sur le support de cette variable dans la région courante) voit ces dernières être partitionnées dans la branche de droite selon le même principe récursif. Les réalisations de valeur inférieure sont partitionnées dans la branche de gauche. Au moment de l'inférence sur de nouvelles observations, l'arbre est parcouru jusqu'à l'atteinte d'une région terminale. Dans notre exemple, en considérant le noeud le plus haut, si le montant des écarts de cotisation (relativement à la masse salariale) de l'entreprise envers l'URSSAF est supérieur à 0.3, alors on explore la marge brute d'autofinancement. Si sa valeur (relativement à la masse salariale) est supérieure à 0.28, la solvabilité est explorée. Sinon, on explore la productivité du travail. Ces deux variables mènent à des régions terminales pour lesquelles on prend une décision. Cette dernière correspond aux branches en rouge et les valeurs associées sont celles du niveau de fraude calculé par l'estimateur g. Les valeurs "next node" correspondent à des niveaux de profondeur supplémentaires de l'arbre, lequel peut être très large et profond.

L'inconvénient des arbres de décision est leur grande variance, due aux nombreuses possibilités aboutissant à une prise de décision. Cet inconvénient est tourné en avantage avec les forêts aléatoires, qui créent une nouvelle règle de décision, $\bar{g}_{\mathcal{P}}^{(B)}$, par la construction de nombreux arbres, permettant ainsi de fortement réduire la variance de l'estimateur lorsque les arbres sont peu dépendants les uns des autres. Un estimateur de Y est $\bar{g}_{\mathcal{P}}^{(B)}$, défini par :

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = \frac{1}{B}\sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(x),$$

où B est le nombre d'arbres de décision nécessaires à l'estimation de Y,
et $g_{\mathcal{P}}^{(b)}(x)$, le b-ème arbre associé à l'observation x.

Dans l'analyse menée, les forêts aléatoires servent d'abord à visualiser de manière immédiate, l'effet d'une variable explicative sur le niveau de fraude. Cet aspect est développé dans la section 4.7. De manière alternative, il est également possible de mener une régression avec le même niveau d'*interprétabilité* qu'un modèle linéaire. En particulier, une manière compatible de l'effectuer est de s'intéresser au graphique de dépendance partielle entre la variable à expliquer et chaque variable explicative. Le graphique de dépendance partielle donne une description de l'effet marginal d'une variable explicative sur la variable à expliquer. L'idée principale de sa mesure est la capture du triplet $\{y, x, j\}$ de chaque région terminale (de chaque arbre). Comme les arbres sont aléatoires, la représentation de Y en fonction de $X^{(j)}$ ne devrait pas être caractéristique, sauf s'il existe une relation spécifique entre les deux variables. La dépendance partielle présente plusieurs intérêts :
- elle permet d'isoler l'effet d'une variable sur la nature du problème observé ;
- la co-influence de deux variables explicatives peut être visualisée et leur dépendance mutuelle mesurée ;
- la dépendance partielle capture aussi bien les effets linéaires que non linéaires ;
- elle est assimilable à la fois comme un modèle descriptif ou prédictif. Dans ce dernier cas, la connexion entre l'interprétation et la prédiction permet une meilleure généralisation de l'effet mesuré.

4.6 Expérimentations et résultats

Notre analyse porte sur le lien entre le niveau de fraude et l'échantillon entier (échantillon synthétique + données réelles), l'évaluation des résultats et l'interprétation permise par les variables économiques. Nous souhaitons, en particulier, mettre en évidence la relation entre situation financière de l'entreprise et propension à la fraude (soit la tendance mesurant le passage d'une absence de fraude à un niveau de fraude positif ou bien, l'accroissement du niveau de fraude relativement au niveau de fraude inconditionnel dans le modèle). Ce processus peut se résumer de la manière suivante :
- les variables significatives fournissent le support de la relation avec le niveau de fraude ;
- dans un second temps, à partir de ce groupe de variables, la situation financière est décrite (indépendamment du niveau de fraude) ;
- puis, les variables compatibles, à la fois, avec une situation financière spécifique (par exemple une rentabilité négative associée à une forte productivité) et une propension à la fraude, sont sélectionnées et leur influence, sur le niveau de fraude, mesurée.
Finalement, la validation de la relation est effectuée sur la base de deux éléments :
i) la caractérisation de la situation financière permet-elle une différenciation des entreprises ? Plus précisément, a-t-on une propension à la fraude plus importante lorsqu'on observe des représentations spécifiques de la situation financière ?
ii) Peut-on évaluer la nature, le sens et l'intensité de l'influence des variables économiques ?

Notre analyse débute par les résultats de la régression linéaire entre le niveau de fraude et les variables, pour l'ensemble de l'échantillon, puis se poursuit par l'évaluation du modèle non linéaire sur les données, qui complète et étend le point de vue.

Données synthétiques + réelles : relation avec le niveau de fraude

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.48874 -0.01131 -0.00197  0.00736  0.68625

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       6.113e-02  5.474e-03  11.166  < 2e-16 ***

Equilibre financier
MARGE BRUTE D'AUTOFINANCEMENT     5.533e-03  6.052e-04   9.143  < 2e-16 ***
COUVERTURE du BFR                -2.199e-03  4.849e-04  -4.535 5.82e-06 ***
COUVERTURE des IMMOS NETTES      -1.607e-03  4.446e-04  -3.613 0.000304 ***
COUVERTURE du CA                 -1.714e-02  3.266e-03  -5.248 1.57e-07 ***
SOLVABILITE                      -1.069e-02  1.028e-03 -10.402  < 2e-16 ***

Profitabilité
RENTABILITE COMMERCIALE          -1.580e-02  8.692e-03  -1.817 0.069209 .
CONTRIBUTION DU CAPITAL           1.598e-02  2.797e-03   5.715 1.13e-08 ***
CONTRIBUTION DE LA VA             1.431e-02  4.002e-03   3.577 0.000349 ***

Liquidité
LIQUIDITE IMMEDIATE              -3.880e-03  5.135e-04  -7.556 4.53e-14 ***
LIQUIDITE GENERALE               -1.456e-03  3.725e-04  -3.909 9.32e-05 ***
LIQUIDITE REDUITE                -1.576e-03  4.910e-04  -3.210 0.001330 **

Endettement
CAPACITE DE REMBOURSEMENT        -1.609e-05  7.335e-06  -2.194 0.028291 *

Productivité
PRODUCTIVITE DE L'ACTIF           1.384e-03  5.791e-04   2.389 0.016900 *
DUREE CLIENT                      2.292e-02  8.395e-03   2.730 0.006342 **
POIDS MASSE SALARIALE            -8.520e-03  3.419e-03  -2.492 0.012710 *
PRODUCTIVITE                      5.519e-03  1.822e-04  30.290  < 2e-16 ***
PRODUCTIVITE DU TRAVAIL          -7.678e-04  5.870e-05 -13.079  < 2e-16 ***

Variables de la déclaration de cotisations
DUREE DE VIE                     -2.864e-04  4.928e-05  -5.811 6.40e-09 ***
NB ORIG. DEBIT NON RENS.         -1.843e-04  6.177e-05  -2.983 0.002858 **
MONTANT DEBIT                     3.934e-04  4.182e-05   9.406  < 2e-16 ***
NB CTP EXONERATION               -4.821e-04  1.054e-04  -4.574 4.85e-06 ***
NB REMISES SUR MAJORATION         9.951e-04  2.877e-04   3.459 0.000543 ***
MONTANT REMISES SUR MAJORATION -1.114e+00  2.455e-01  -4.537 5.78e-06 ***
MONTANT ECARTS                    1.029e-01  3.287e-03  31.297  < 2e-16 ***
PENALITES                         7.801e-01  1.980e-01   3.940 8.19e-05 ***
NB RETARDS                        2.703e-04  7.064e-05   3.826 0.000131 ***
NB TAXATIONS D'OFFICE             7.759e-04  2.421e-04   3.205 0.001353 **
MONTANT TAXATIONS D'OFFICE       -6.695e-02  2.957e-03 -22.641  < 2e-16 ***
NB DEMANDES DELAIS               -2.176e-03  5.437e-04  -4.001 6.34e-05 ***
DERNIER CTRL                      3.746e-05  8.877e-06   4.219 2.47e-05 ***
COMPLIANCE                        8.361e-03  3.546e-03   2.358 0.018408 *
Tx COTISATION NET                -1.651e-02  6.007e-03  -2.748 0.006002 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02938 on 9956 degrees of freedom
Multiple R-squared: 0.3908,     Adjusted R-squared: 0.3882
F-statistic: 148.5 on 43 and 9956 DF,  p-value: < 2.2e-16
```

Pour une plus grande lisibilité, seules les variables significatives sont mentionnées. Le niveau de significativité est donné dans le tableau par la p-value, $Pr(>|t|)$.
Le niveau de fraude initial spécifié par le modèle est de 6.11% et les signes des coefficients indiquent une diminution ou une augmentation de ce niveau selon l'interprétation et les valeurs prises par les variables. L'explication fournie par les coefficients est, toutefois, tempérée par leurs valeurs, dont les ordres de magnitude indiquent que l'effet global sur le niveau de fraude est probablement le résultat d'une combinaison de facteurs. Le pourcentage de variance expliqué par la régression (39%) est limité par les relations non linéaires présentes dans les données. Nous traitons cet aspect plus loin.

Précisons le cadre d'interprétation général du modèle linéaire.
Un coefficient négatif indique une tendance à la baisse du niveau de fraude, lorsque la variable explicative prend des valeurs élevées. A l'inverse, lorsque la variable prend des valeurs faibles mais positives, la tendance de la fraude est un rapprochement vers son niveau inconditionnel, soit une absence d'effet spécifique de la variable. Lorsque les valeurs de la variable sont (de plus en plus) négatives, la tendance de la fraude est une augmentation de son niveau.
Un coefficient positif indique une tendance à la hausse du niveau de fraude, lorsque la variable prend des valeurs élevées. A l'inverse, lorsqu'elle prend des valeurs faibles mais positives, la tendance de la fraude est un rapprochement vers son niveau inconditionnel. Lorsque ces valeurs deviennent (de plus en plus) négatives, la tendance de la fraude est une diminution de son niveau.

4.6.1 Effet des variables

70% des variables économiques (17 sur 24) se révèlent significatives. Parmi celles de la déclaration de cotisations 15, sur 19, sont significatives. Toutefois, en dernier lieu, seule la propension à la fraude nous intéresse et seules 15 variables (dont 7 de nature économique) contribuent à la faire augmenter.

Dans le cas des variables de la déclaration de cotisations, nous notons deux facteurs présentant une relation avec le niveau de fraude :
- les écarts de cotisation de l'entreprise envers l'URSSAF et la politique de recouvrement de cette dernière, représentée par les taxations d'office, les pénalités et les remises sur majoration de cotisations accordées ;
- le taux de cotisation.

En considérant la propension à frauder, la politique de recouvrement identifie à la fois les facteurs d'influence d'un niveau de fraude en augmentation mais également ceux qui contribuent à une réduction. Des anomalies dans le paiement des cotisations, en particulier des écarts de cotisation, traduisent une propension à frauder plus importante ; les pénalités qui en découlent sont le prolongement de la situation de l'entreprise envers l'URSSAF et il est probable qu'un contrôle ait lieu, car ces variables sont enregistrées pendant les trois années qui précèdent un éventuel contrôle. Une politique de recouvrement plus spécifique, par le biais de taxations d'office, l'acceptation de délais de paiement ou les remises sur majoration de cotisations, tend à réduire le niveau de fraude.

Les deux dernières variables (la conformité aux obligations de déclaration et le taux de cotisation) semblent présenter une contradiction avec le niveau de fraude. Dans le cas de la *compliance*, cela s'explique par le fait qu'une majorité d'entreprises ne remplit pas ses obligations de conformité à la législation. Les signaux émis par cette variable sont alors bruités. Les entreprises qui ont un taux de cotisation élevé sont, elles, celles qui semblent respecter le mieux la législation.

`Variables économiques les plus significatives`

Dans le cas des variables économiques, nous ne nous référons pas, pour le moment, au niveau de fraude et distinguons d'abord les plus significatives, identifiées par les p-values les plus faibles. Ces variables font d'abord référence à l'équilibre financier et à la profitabilité de l'entreprise. La solvabilité, la marge brute d'autofinancement [9] (capacité d'autofinancement) et le ratio de couverture du chiffre d'affaires [10] sont les variables les plus importantes de l'équilibre financier. Ces variables traduisent les capacités financières, et d'exploitation, à long terme des entreprises. Dans ce même groupe, la durée client [11], à l'origine de variations de trésorerie lorsque l'entreprise recouvre ses créances tardivement, constitue également une variable d'intérêt.

La rentabilité commerciale, ainsi que les contributions du capital [12] et de la valeur ajoutée [13], constituent un second groupe de variables significatives et influencent à la fois les situations de court et de long terme. En particulier pour les deux dernières, la capacité d'autofinancement se retrouve au numérateur.

Les dettes à court terme (via la liquidité) et les facteurs de productivité constituent le dernier groupe des variables les plus significatives.

La situation financière a trait au cycle d'exploitation courant (dettes à court terme, chiffre d'affaires, stocks et créances, rentabilité, liquidité, productivité) et de plus long terme (actif total, solvabilité, capitaux propres, autofinancement,...). Des difficultés peuvent s'expliquer, par exemple, par des dettes d'exploitation importantes et/ou une rentabilité négative. A contrario, une liquidité et une solvabilité élevées, associées à une forte productivité, indiquent un équilibre favorable à l'activité d'exploitation de l'entreprise. Nous poursuivons l'analyse en structurant le propos relativement aux termes d'analyse financière (cf annexe) des entreprises.

9. somme du bénéfice net et des dotations aux amortissements et provisions
10. ratio entre le fonds de roulement (différence entre capitaux permanents et actifs immobilisés) annualisé et le chiffre d'affaires
11. délai moyen de paiement des clients
12. ratio de la capacité d'autofinancement annuelle aux capitaux permanents
13. ratio de la capacité d'autofinancement à la valeur ajoutée (somme de la marge commerciale et de la différence entre la valeur de la production et des coûts de production...)

`Equilibre financier, profitabilité, liquidité, endettement, productivité`

i) Nous considérons d'abord l'équilibre financier de l'entreprise, représenté dans les variables significatives par les cinq premières variables. L'équilibre financier de l'entreprise présente un point de vue sur la situation à long terme (> 1 an) à travers les capitaux propres et permanents, l'actif immobilisé et l'ensemble des dettes. Une entreprise en situation favorable, sous ce point de vue, possède alors des ratios de couverture élevés, un bon niveau de solvabilité et une grande capacité d'autofinancement. Cela se traduit par des capitaux propres importants et/ou peu de dettes à long terme. Des difficultés peuvent apparaître sous le signe d'une dette élevée associé à un niveau de capitaux propres insuffisant.

ii) La profitabilité s'exprime à travers la rentabilité et la capacité d'autofinancement. Une rentabilité négative oblige l'entreprise à devoir se financer à nouveau ou à puiser dans ses réserves pour la poursuite de son cycle d'exploitation. Une rentabilité positive indique que les dettes à court terme peuvent être couvertes, au moins en partie, par les bénéfices accumulés durant l'exercice. La capacité d'autofinancement est la contrepartie de la présence des dettes à long terme et des fonds propres de l'entreprise. Plus elle est importante, plus l'entreprise peut se financer facilement à long terme. Par exemple, des dettes à long terme ne pénalisent pas l'entreprise si sa capacité d'autofinancement est importante. De même, cette dernière peut être faible et sans répercussion sur l'entreprise, si celle-ci n'a que peu de dettes.

iii) La liquidité est la capacité de l'entreprise à faire face à ses dettes à court terme. Plus elle est élevée, plus l'entreprise dispose de ressources pour le cycle d'exploitation courant et moins sa situation financière paraît délicate.

iv) L'endettement n'est représenté dans le modèle que par l'incidence de la capacité de remboursement (des dettes bancaires), laquelle est liée à la capacité d'autofinancement (à son dénominateur) et traduit donc en partie l'influence de l'équilibre financier sur l'activité de l'entreprise.

v) Les cinq dernières variables économiques sont liés aux différents facteurs de productivité de l'entreprise. La durée client fait référence aux délais moyens de paiement des clients. Des délais importants peuvent fragiliser la trésorerie de l'entreprise. Les ratios de productivité sont tous liés au chiffre d'affaires (au numérateur). Des valeurs élevées traduisent une couverture suffisante des coûts et de l'actif de l'entreprise par le chiffre d'affaires. Fondamentalement, les facteurs de productivité constituent le moteur de croissance de l'entreprise.

Chacun des facteurs décrits ci-dessus présente un lien partiel avec la situation financière selon son intensité dans le cycle économique de l'entreprise. Toutefois, la combinaison de plusieurs d'entre eux est généralement plus adaptée à une description cohérente de la situation financière. Avant de la caractériser, nous précisons la manière dont les variables économiques interagissent avec le niveau de fraude.

`Relations avec le niveau de fraude`

Les coefficients de la régression fournissent un degré d'analyse plus précis sur la nature de la relation entre les variables économiques et la propension à la fraude. Nous souhaitons obtenir un point de vue plus explicite sur l'effet des variables significatives les plus importantes.

- Dans le cas du facteur productivité, l'augmentation du délai moyen de paiement des clients, la productivité des salariés et la productivité de l'actif[14] sont les variables dont les niveaux élevés sont en relation avec la propension à la fraude. Pour ces deux dernières, cela semble contradictoire. Une forte productivité des salariés traduit, toutes choses égales par ailleurs, une augmentation du chiffre d'affaires ou une réduction du nombre de salariés. Lorsque la fraude est avérée, une productivité des salariés importante peut alors être l'effet de la dissimulation de salariés. Dans le cas de la productivité de l'actif, cet effet est beaucoup moins évident : *Il est cependant plausible que la dissimulation de salariés ne soit pas uniquement la volonté d'une réduction de coûts, mais celle d'un rendement plus important du chiffre d'affaires.*

A contrario (et toujours pour le facteur productivité), un poids de la masse salariale plus grand contribue à réduire le niveau de fraude, ce qui est cohérent puisque le poids des cotisations sociales est, alors, lui aussi en augmentation. Cette interprétation souffre cependant d'une exception qui peut devenir la règle : si la valeur ajoutée se réduit (par exemple par une baisse de la marge commerciale ou une augmentation des coûts de production, hors travail), un arbitrage peut être réalisé entre réduction de la masse salariale et augmentation des marges. Dans ce cas, le poids de la masse salariale doit être interprété avec plus de précaution. Nous omettons, ici, la productivité du travail traitée plus spécifiquement dans la section qui suit.

- Les facteurs de liquidité et de rentabilité traduisent les contraintes immédiates, en matière de financement, qui peuvent modifier la propension de l'entreprise à frauder. A mesure que la rentabilité commerciale se dégrade, en particulier quand elle est négative, le modèle indique une propension à la fraude plus grande. L'entreprise subit alors des pertes financières et le recours au travail dissimulé traduit une manière de les amoindrir ou un changement de stratégie. Les deux autres facteurs de rentabilité, les contributions du capital et de la valeur ajoutée, traduisent une augmentation du niveau de fraude avec celle de la capacité d'autofinancement. *Nous notons, ici, qu'une grande capacité d'autofinancement peut être en contradiction avec une rentabilité commerciale négative. Cette situation est possible lorsque la fraude concerne des entreprises aux profils de rentabilité très différents. Dans le cas des facteurs de rentabilité, la propension à la fraude traduit une préoccupation purement financière.*

Dans le cas de la liquidité, que nous résumons à la liquidité générale, la propension à la fraude est une fonction croissante des dettes à court terme. Moins elles sont importantes (et plus la liquidité augmente), plus le niveau de fraude baisse.

14. ratio entre le chiffre d'affaires et l'actif comptable

- Pour les situations de plus long terme, l'équilibre financier est le facteur de référence. La solvabilité et l'ensemble des ratios de couverture ont une relation décroissante avec la propension à la fraude. Plus l'équilibre financier est assuré, moins le recours au travail dissimulé semble nécessaire. L'exception est, à nouveau, la capacité d'autofinancement dont les grandes valeurs tendent à accroître la propension à la fraude.

A court terme, la fraude est profitable à la fois par des gains de productivité et des gains financiers. Dans le premier cas, la réduction du coût du travail ou l'augmentation du rendement du chiffre d'affaires, ainsi que les variations de trésorerie (à travers les délais de paiement), constituent les trois éléments qui favorisent le recours à la fraude.
Dans le second, une rentabilité commerciale négative traduit un cycle d'exploitation délicat et les gains financiers issus du travail dissimulé peuvent servir à compenser les pertes d'exploitation. Les différentes variables liées à la productivité, ainsi que la rentabilité commerciale, sont toutes liés au chiffre d'affaires. Ce dernier apparaît comme la variable la plus influente de la propension à frauder.
A plus long terme, la variable décisive est la capacité d'autofinancement. Elle intervient dans la réalisation de l'équilibre financier, en contribuant au financement ou au remboursement des dettes à long terme. Elle permet également, lorsque les dettes sont faibles, de renforcer les capitaux propres de l'entreprise.

Dans les paramètres générés par le modèle linéaire, l'endettement ne semble pas avoir d'effet sur le niveau de fraude sauf, marginalement, dans le cas de la capacité de remboursement. *En d'autres termes, dans la propension à la fraude, la perte (ou les variations) de revenus d'exploitation joue un rôle beaucoup plus décisif que l'accroissement des dettes. Cela se traduit par l'absence d'influence de la solvabilité et de la liquidité dans l'augmentation du niveau de fraude. Toutefois, les dettes ont un rôle important et quand l'entreprise est en mesure d'y faire face par de nombreuses ressources (fonds propres, bénéfices, trésorerie, stocks et créances), la tendance donnée par le modèle est une réduction du niveau de fraude.*

Finalement, du point de vue linéaire, nous pouvons résumer la relation entre la propension à la fraude et les variables économiques à travers quatre éléments :
- les variations de trésorerie ;
- les pertes d'exploitation ;
- le rendement du chiffre d'affaires ;
- les variations de la capacité d'autofinancement.

Ces différents éléments résument des situations financières différentes dont nous souhaitons connaître l'impact global sur le niveau de fraude.

4.6.2 Caractérisation de la situation financière

Lorsque la rentabilité est négative, la solvabilité et la liquidité à de faibles niveaux, l'entreprise, manifestement, ne peut pas être dans une situation confortable durant l'exercice courant. On peut ainsi rajouter d'autres variables économiques significatives qui permettent de confirmer un tel point de vue, de le nuancer ou de l'infirmer. Cette méthode

a l'inconvénient de n'être satisfaisante que pour un petit nombre d'entreprises et d'agréger des variables qui agissent différemment sur l'activité de l'entreprise. Nous souhaitons avoir un lien explicite entre différentes représentations de la situation financière et le niveau de fraude. C'est le cas d'une rentabilité commerciale négative, synonyme de pertes financières, reliée à une augmentation du niveau de fraude. Dans le même temps, la capacité d'autofinancement ou les variables de productivité nous indiquent que les pertes d'exploitation ne sauraient être exclusives. Deux options sont possibles :
i) identifier plusieurs variables dont les effets marginaux sont, de manière systématique, une propension à la fraude plus importante. Dans ce cadre, on s'intéresse alors à la somme de leurs influences locales sachant un niveau de fraude strictement positif et pour différents seuils de ce dernier. Nous adoptons ce point de vue dans la section plus loin, par l'utilisation d'un modèle non linéaire.
ii) La seconde option, développée dans les lignes qui suivent, propose une analyse de l'influence globale des variables significatives. Pour cela, nous caractérisons la situation financière de la façon qui suit.

1 - Dans une première étape, nous considérons l'ensemble E des catégories économiques qui définissent la situation financière d'une entreprise quelconque,
où $E = \{équilibre financier, profitabilité, liquidité, endettement, productivité\}$.
2 - Nous supposons que toute caractérisation de la situation financière implique, au minimum, une variation de revenus, à court ou long terme, pour l'entreprise. Nous nous intéressons à la situation financière de l'entreprise sachant le niveau de fraude dans le modèle.
3 - La caractérisation de la situation financière est alors définie par l'appartenance d'une variable significative à une des catégories de E, si :
- elle est liée à une augmentation du niveau de fraude dans le modèle ;
- elle entraîne une variation de revenus.
La première condition est nécessaire. Une absence de la seconde entraîne l'affectation de la variable comme élément secondaire de toutes les catégories.
4 - Chaque catégorie non vide est alors constituée d'au moins deux variables significatives et constitue un sous-groupe de la situation financière.
5 - Finalement, la caractérisation de la situation financière est l'agrégation des sous-groupes générés et nous mesurons son influence globale sur la propension à la fraude, à travers chacune des variables sélectionnées.

L'avantage de ce procédé est l'observation d'un plus grand nombre d'entreprises conjointement à un respect des paramètres du modèle, de la structure globale des variables économiques et du niveau de fraude aléatoire. On souhaite surtout éviter les incohérences liées à la présence de variables dont les réalisations sont contradictoires, peu corrélées, ou redondantes, relativement au niveau de fraude. Elles sont donc, dans un premier temps, traitées séparément, alors que leur interprétation n'a lieu qu'une fois leur fusion, dans le même ensemble, réalisée.

Le modèle linéaire produit sept variables significatives d'une relation entre un accroissement du niveau de fraude et des facteurs économiques :

- le délai moyen de paiement des clients ;
- la contribution du capital (dont le numérateur est la capacité d'autofinancement) ;
- la rentabilité commerciale, lorsqu'elle est négative ;
- la contribution de la valeur ajoutée (dont le numérateur est la capacité d'autofinancement) ;
- la capacité d'autofinancement, lorsqu'elle est positive ;
- la productivité des salariés (dont le numérateur est le chiffre d'affaires) ;
- la productivité de l'actif (dont le numérateur est le chiffre d'affaires).

Nous procédons en plusieurs étapes binaires pour former des groupes d'entreprises susceptibles d'avoir un niveau de fraude moyen plus important que celui observé dans tout l'échantillon, mais également une probabilité d'observation d'un niveau positif, plus grande.

a) La capacité d'autofinancement, lorsqu'elle est importante, n'est pas cohérente avec une rentabilité commerciale négative. Lorsque les deux variables sont de signe opposé, nous sommes, par construction, en présence d'entreprises avec des profils (en termes de revenus) nécessairement différents. Précisons que la rentabilité commerciale doit être négative pour être associée à un accroissement du niveau de fraude et que la capacité d'autofinancement doit être, de même, positive. Ces deux variables doivent alors être traitées séparément.

b) Comme la capacité d'autofinancement est présente dans les contributions du capital et de la valeur ajoutée, ces dernières ne sont pas, non plus, compatibles avec une rentabilité commerciale négative. Avec la contribution du capital, nous mesurons plus spécifiquement la réduction des capitaux permanents (somme des fonds propres, provisions et dettes à long terme), au dénominateur de la variable. Et avec la contribution de la valeur ajoutée, nous mesurons la réduction de la valeur ajoutée.

Les points a) et b) aboutissent donc à la construction de quatre groupes mesurant des aspects distincts de l'équilibre financier et de la rentabilité tous liés à des variations de revenus.

c) Le délai moyen de paiement des clients (lorsqu'il augmente) affecte la trésorerie de l'entreprise et les variations résultant de longs délais sont des situations de très court terme (au maximum un trimestre, dans nos données). Le délai moyen de paiement des clients constitue alors le support d'un cinquième groupe d'entreprises pour lesquelles on peut observer l'effet de cette variable sur le niveau de fraude.

Nous avons, à cet instant, cinq concepts correspondant à des situations financières spécifiques : le premier est lié aux pertes financières du cycle d'exploitation, le second à l'équilibre financier, les deux suivants à la profitabilité à court et long terme, et le dernier aux variations de trésorerie.

d) Il reste à affecter le facteur productivité. Les variables qui le constituent sont liées à la mesure du chiffre d'affaires de l'entreprise relativement à son actif total, et à son

effectif ; le rendement du chiffre d'affaires détermine le niveau d'efforts consentis par l'entreprise pour maintenir ou accroître son activité. La productivité de l'actif définit un sixième (et dernier) groupe qui mesure les variations du chiffre d'affaires.

La variable productivité (de l'effectif) est la seule variable qui n'est pas liée à une variation de revenus. Cette particularité est due aux raisons suivantes : dans le modèle linéaire, on peut noter que la productivité du travail (rapport entre le chiffre d'affaires et la masse salariale) présente une relation décroissante avec le niveau de fraude. Plus elle augmente, plus le niveau de fraude tend à se réduire. Or, au dénominateur de ce dernier, nous avons également la masse salariale. Dans le modèle, l'accroissement du chiffre d'affaires contribue donc à une diminution de la fraude. Dans le cas de la productivité (de l'effectif), dont le coefficient est positif, une augmentation du niveau de fraude est alors plus probablement due à une réduction de l'effectif (ou à une dissimulation) qu'à un accroissement du chiffre d'affaires. La variation de la variable productivité n'est alors pas liée, lorsqu'il y a fraude, à une variation de revenus effective mais artificielle. Elle est aussi la seule variable qui fait intervenir l'effectif, dont la minoration du nombre est la cause principale du travail dissimulé. A chacun des six sous-groupes constitués, nous associons alors la variable productivité.

La caractérisation de la situation financière proposée consiste donc à considérer différents états de variation de revenus pour l'entreprise, relativement à la propension à frauder et aux efforts entrepris pour accroître le rendement du chiffre d'affaires (relativement aux ressources disponibles).

`Influence de la situation financière sur le niveau de fraude`

Comparativement à des efforts de productivité plus importants que dans la moyenne des entreprises, nous observons l'influence sur le niveau de fraude des entreprises des facteurs suivants :
- la capacité d'autofinancement ;
- les pertes et revenus d'exploitation ;
- les capitaux propres et dettes à long terme ;
- les délais de paiement des clients et les variations de trésorerie qui en découlent.

Les facteurs observés sont, naturellement, les mêmes que précédemment mais, pour la mesure du niveau de fraude, nous nous intéressons aux entreprises pour lesquelles ils présentent des valeurs supérieures à la moyenne. Le choix de la moyenne, comme seuil inférieur d'observation des variables économiques est arbitraire, mais reste cohérent avec le modèle. Plus les variables prennent des valeurs s'écartant de la moyenne, plus probable est l'accroissement du niveau de fraude. Notons qu'au moment de l'agrégation des variables, l'ensemble de la distribution de chacune d'elles (conditionnellement à une propension à la fraude plus importante) est capturé, sauf dans le cas de la productivité qui apparaît alors comme une variable fixée. Pour chaque sous-groupe, nous mesurons également et à posteriori, les valeurs moyennes de la solvabilité, la liquidité, la capacité de remboursement, la productivité du travail et le taux de cotisation.

Afin d'en faciliter la lecture, nous illustrons tous les sous-groupes définis précédemment par un tableau synthétique des variables.

	Durée client	Rentabilité commerciale	Contribution de la valeur ajoutée	Contribution du capital
Moyenne	0.15	-0.07	0.21	0.26
Médiane	0.14	-0.07	0.20	0.24

	Marge brute d'autofinancement	Productivité de l'actif	Productivité
Moyenne	0.86	2.58	3.44
Médiane	0.73	2.50	2.55

TABLE 4.3 – Tableau synthétique des variables significatives (pour des valeurs supérieures à la moyenne ou négatives) d'une propension à la fraude dans le modèle linéaire, conditionnellement à une productivité également supérieure à la moyenne.

Les entreprises dont la propension à la fraude peut être importante se retrouvent dans un ou plusieurs cas de figure des variables du tableau. Cependant, à ce stade, la tendance relativement à la moyenne des entreprises ne peut être établie. En effet, chaque variable correspond à un modèle spécifique qui ne présente pas nécessairement de lien avec les autres. L'indication principale, ici, est la présence potentielle d'un niveau de fraude plus élevé, en moyenne, pour chacune des variables prises une à une. Par exemple, dans le cas de la marge brute d'autofinancement, le niveau de fraude moyen est de 30% supérieur au niveau de fraude observé pour toutes les entreprises. Dans le cas de la rentabilité commerciale il ne l'est que de 15%. Obtenir un point de vue global, à la fois sur les variables et sur la propension à la fraude nécessite d'assembler les modèles.

4.6.3 Synthèse

Assembler les modèles de base revient à considérer le point de vue suivant : le modèle ensembliste contient toutes les observations capturées par chacun des modèles de base (associés à une ou plusieurs variables spécifiques). Il présente alors l'avantage de pouvoir évaluer toute la distribution d'une variable significative, là où n'importe quel modèle de base n'en évalue qu'une partie. Dans l'agrégation de modèles, seule la productivité n'est pas totalement capturée. Mais c'est une variable dont le modèle linéaire indique une tendance à l'augmentation de la fraude à mesure que sa valeur augmente. Finalement, dans l'agrégation de modèles, nous observons l'influence globale de toutes les variables significatives sur le niveau de fraude, sachant que la productivité est supérieure à la moyenne.

Deux niveaux d'analyse sont abordés ici :
- dans le premier, nous comparons les moyennes dans le modèle avec celles de toutes les entreprises de l'échantillon, pour les mêmes variables. Nous obtenons ainsi une estimation des variables économiques discriminantes dans l'analyse du niveau de fraude.
- Dans le second, nous vérifions que les variables retenues sont effectivement reliées à un niveau de fraude plus important que la moyenne et une probabilité d'observation de la fraude, également, plus grande. En particulier, nous souhaitons connaître le nombre de cas de fraude réels expliqués par le modèle.

	Entreprises sélectionnées par le modèle d'agrégation	Ensemble des entreprises	p-value
Marge brute d'autofinancement	0.54	0.39	< 2.2e-16
Contribution du capital	0.14	0.16	2.9e-12
Contribution de la VA	0.12	0.15	< 2.2e-16
Rentabilité commerciale	0	-0.04	< 2.2e-16
Durée Client	0.10	0.09	0.05964
Productivité de l'actif	2.13	2.12	0.458
Productivité	3.41	1.6	< 2.2e-16

	Entreprises sélectionnées par le modèle d'agrégation	Ensemble des entreprises	p-value
Solvabilité	1.3	1.62	< 2.2e-16
Liquidité générale	4.96	5.17	6.501e-13
Capacité de remboursement	43	41	0.04583
Productivité du travail	14	8	< 2.2e-16
Couverture du BFR	1.57	1.26	< 2.2e-16
Couverture du CA	0.23	0.19	< 2.2e-16
Poids de la masse salariale	0.78	0.77	1.709e-08
Pénalités	0.0019	0.0017	0.00164
Montants écarts (de cotisation)	0.20	0.16	< 2.2e-16
Taux de cotisation (principal)	0.27	0.27	0.2669

	Entreprises sélectionnées par le modèle d'agrégation	Ensemble des entreprises	p-value
Niveau de fraude (moyenne)	1.83%	1.10%	2.579e-10

TABLE 4.4 – Valeurs moyennes (et p-value du rejet à tort de l'hypothèse, H_1, d'une différence de moyennes) selon la caractérisation de la situation financière (premier tableau), puis à posteriori (second tableau), face à l'ensemble des entreprises de l'échantillon.

Dans la première partie du tableau ci-dessus, nous avons la caractérisation de la situation financière tandis que la seconde partie est observée à posteriori (une fois la caractérisation spécifiée).

Dans la première partie du tableau, la caractérisation de la situation financière confirme la capacité d'autofinancement comme un facteur important de la propension à la fraude. Comme elle est au numérateur des contributions du capital et de la valeur ajoutée, on s'attend à ce que ces variables aient des moyennes supérieures à celles de l'ensemble des entreprises ; ce qui n'est pas le cas. Il est alors possible qu'une réduction des capitaux permanents, en particulier des fonds propres, joue un rôle (marginal) dans la propension à la fraude. Dans le cas de la contribution de la valeur ajoutée, ce rôle est probablement plus important et correspond à une réduction de la marge commerciale ou à une augmentation des coûts de production. Les pertes d'exploitation, à travers la rentabilité commerciale, ne sont pas un facteur de fraude. Tout comme le délai moyen de paiement des clients. Ces deux variables agissent à la marge et s'accompagnent probablement d'autres facteurs, comme toutes les variables représentées.

Une fois la situation financière fixée, la seconde partie du tableau fait apparaître des différences sensibles dans le cas de la solvabilité, la couverture du besoin en fonds de roulement [15] et la productivité du travail. Cette dernière est, cependant, corrélée (0.5) avec la productivité, ce qui explique probablement sa valeur moyenne. De manière spécifique, la solvabilité est, en moyenne, 20% plus faible dans les entreprises pour lesquelles la propension à la fraude est importante. L'ensemble des dettes (au dénominateur de la solvabilité) a donc un poids dans la propension à la fraude. A l'opposé, les dettes à court terme (à travers la liquidité) n'influencent probablement pas cette propension. La couverture du besoin en fonds de roulement est, en moyenne 25% plus forte dans les entreprises dont la caractérisation de la situation financière indique un niveau de fraude plus grand. Cette variable montre, pour des grandes valeurs, la présence d'une trésorerie importante ou d'un faible besoin en fonds de roulement. Cette situation est un paradoxe, également observé pour d'autres variables : certaines entreprises possède des ressources plus importantes que la moyenne et une propension à la fraude également plus élevée.

`Les facteurs de l'influence de la situation financière sur la propension à la fraude`

Nous retenons alors 8 variables dans le lien de la propension à la fraude avec la situation financière, que nous pouvons regrouper sous les facteurs suivants : la capacité d'autofinancement, les fonds propres, les dettes à long terme, le rendement du chiffre d'affaires et le besoin de financement du cycle d'exploitation.

La capacité d'autofinancement distingue trois variables : la marge brute d'autofinancement, la contribution du capital et la contribution de la valeur ajoutée.
Les fonds propres sont en partie financés par la capacité d'autofinancement et s'expriment également par la contribution du capital et la solvabilité.
Les dettes à long terme distinguent la solvabilité comme variable d'intérêt.
Le rendement du chiffre d'affaires est l'expression de la productivité, de la productivité du travail et de la contribution de la valeur ajoutée.
Le besoin de financement du cycle d'exploitation est observé à travers le ratio de couverture du chiffre d'affaires et la couverture du besoin en fonds de roulement.

L'ensemble des ces facteurs précise une représentation de la situation financière et une propension à la fraude plus importante que la moyenne. Ils ne nous disent pas comment distinguer les entreprises dont la fraude pourrait être plus importante et fournissent plutôt un état général des conditions de l'activité économique qui peuvent faire émerger une fraude plus fréquente et d'intensité plus grande. Nous notons que, ni les pertes d'exploitation strictes, ni les dettes à court terme, n'apparaissent comme des facteurs de la propension à la fraude.

Parmi les facteurs retenus, un élément critique est la disponibilité de ressources de financement tant à court qu'à long terme. Le processus qui les lie à la propension à la fraude est complexe mais détermine fondamentalement l'arbitrage entre l'absence et la

15. ratio entre le fonds de roulement (différence entre les capitaux permanents et les actifs immobilisés) et le besoin en fonds de roulement

présence de fraude. La capacité d'autofinancement, le besoin en fonds de roulement ou encore les capitaux propres en sont des exemples. De notre point de vue, l'interaction de ces ressources avec les dettes à long terme et le processus de production constitue un thème central dans la compréhension de la propension à la fraude, lorsque la situation financière y contribue fortement.

Du point de vue des variables de la déclaration de cotisation, les pénalités et les écarts entre cotisations attendues par l'URSSAF et cotisations reçues (relativement à leur masse salariale) présentent un différentiel d'environ +25% relativement à l'ensemble des entreprises. Pour celles dont la propension à la fraude est plus grande relativement à la situation financière, les écarts de cotisation, notifiés avant un contrôle, représentent en moyenne 20% (contre 16% pour l'ensemble des entreprises) de la masse salariale. Ces écarts sont assimilables à des contestations, fondées ou non, de la part des entreprises sur les cotisations dues à l'URSSAF. Le taux de cotisation de la principale catégorie est en moyenne identique, quelle que soit la propension à la fraude considérée. Lorsque le travail est dissimulé, le taux de cotisation n'a aucune raison d'être différent de la situation habituelle car les salariés concernés n'ont alors tout simplement pas d'existence.

Le deuxième niveau d'analyse est la mesure du niveau et de la proportion de fraude, lorsque les entreprises sont sélectionnées selon le modèle proposé. Nous observons alors la fraude sachant la caractérisation de la situation financière.

	Nombre	Proportion de fraude	Niveau de fraude moyen
Entreprises sélectionnées par le modèle d'agrégation	2739	56.44%	1.83%

TABLE 4.5 – Caractérisation de la situation financière et fraude observée.

Sur les 5392 entreprises dont le niveau de fraude est positif, le modèle en sélectionne, environ, 51% pour lesquelles il est suffisamment probable que la situation financière présente un lien avec le niveau de fraude. Cette sélection est ensuite éprouvée face à la réalité : sur 2739 entreprises choisies par le modèle, 56% présentent effectivement un niveau de fraude positif. Le point d'intérêt est l'évaluation du niveau de fraude (1.83%) qui est 70% plus élevé que le niveau observé pour tout l'échantillon (1.1%). *La situation financière n'explique que partiellement l'ensemble des cas, mais constitue le principal facteur des niveaux de fraude les plus élevés.*
Pour illustrer ce résultat, nous considérons le niveau de fraude dans les données réelles une fois la situation financière déterminée.

	Proportion dans les données réelles	Niveau de fraude moyen dans les données réelles
Entreprises sélectionnées par le modèle d'agrégation	39.92%	10.94%

TABLE 4.6 – Caractérisation de la situation financière et fraude observée dans les données réelles.

Environ 40% des cas de fraude dans les données réelles sont expliqués par la situation financière et sont associés à un niveau de fraude moyen de plus de 10% (de la masse salariale). En l'absence d'explication par les variables économiques, ce niveau de fraude est largement inférieur à 1%. La caractérisation de la situation financière de l'entreprise permet ainsi la sélection des cas de dissimulation de cotisations les plus importants. Notons que ce résultat éclaire ceux constatés dans la lutte contre la fraude, pour lesquels plus de 75% des montants redressés sont le fait de moins de 20% des entreprises contrôlées et redressées. La modélisation par les variables économiques en permet la généralisation : dès que le niveau de fraude est important, la plus grande partie en est explicable par la situation financière de l'entreprise et conduit à des montants de redressement également élevés. Une application opérationnelle est donnée plus loin : les cas de fraude et d'irrégularités les plus nombreux concernent des niveaux faibles ($< 1\%$ de la masse salariale) et sont très difficilement détectables, en partie parce qu'ils sont dus à des erreurs des entreprises. L'utilisation de variables économiques sur les cas de fraude réels (et volontaires) permet de définir un seuil en-dessous duquel ces variables perdent leur caractère significatif. L'application d'un modèle prédictif pour toutes les irrégularités aux cotisations sociales au-dessus du seuil permet alors d'accroître fortement la précision du modèle et de généraliser la prédiction aux cas plus importants.

`Variables économiques et données réelles`

Nous illustrons ici la manière dont les variables économiques influencent le niveau de fraude dans le cas des données réelles :

Données réelles : relation avec le niveau de fraude

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.50926 -0.03170 -0.00529  0.01999  0.34723

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.345e-01  3.642e-02   3.694 0.000236 ***
MARGE BRUTE D'AUTOFINANCEMENT   1.381e-02  1.820e-03   7.587 9.48e-14 ***
INDEPENDANCE FINANCIERE        -3.025e-02  1.772e-02  -1.707 0.088250 .
CAPACITE DE REMBOURSEMENT      -4.782e-05  2.845e-05  -1.681 0.093231 .
FINANCEMENT DES STOCKS         -4.891e-03  2.035e-03  -2.404 0.016467 *
PRODUCTIVITE DE L'ACTIF         9.076e-03  3.571e-03   2.541 0.011245 *
DUREE CLIENT                    1.027e-01  4.800e-02   2.139 0.032733 *
PRODUCTIVITE                    2.343e-03  5.636e-04   4.158 3.57e-05 ***

MONTANT DEBIT                  -5.153e-02  2.036e-02  -2.531 0.011566 *
NB REMISES SUR MAJORATION       1.972e-03  1.144e-03   1.723 0.085263 .
MONTANT REMISES SUR MAJORATION -2.350e+00  9.732e-01  -2.415 0.015975 *
MONTANT ECARTS                  5.044e-01  1.838e-02  27.446  < 2e-16 ***
PENALITES                      -1.385e+00  5.182e-01  -2.672 0.007693 **
NB RETARDS                     -1.448e-03  3.557e-04  -4.071 5.16e-05 ***
MONTANT TAXATIONS D'OFFICE     -3.224e-01  3.562e-02  -9.050  < 2e-16 ***
NB DEMANDES DELAIS             -6.893e-03  2.108e-03  -3.269 0.001127 **
DERNIER CTRL                    1.884e-04  7.789e-05   2.419 0.015810 *
NB CCA                          2.939e-04  1.481e-04   1.984 0.047638 *
Tx COTISATION NET              -1.530e-01  3.912e-02  -3.911 0.000100 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06791 on 770 degrees of freedom
Multiple R-squared: 0.6487,     Adjusted R-squared: 0.6291
F-statistic: 33.06 on 43 and 770 DF,  p-value: < 2.2e-16
```

Les principales variables identifiées précédemment sont à nouveau présentes avec des coefficients encore plus grands : les variations de trésorerie (durée client), la capacité d'autofinancement et le rendement du chiffre d'affaires (à travers les deux variables de productivité) sont les principaux facteurs qui affectent, à la hausse, le niveau de fraude. L'ajout de données synthétiques permet la recherche d'autres facteurs moins significatifs dans les données réelles sans altérer ceux déjà présents. Dans la régression présentée ci-dessus, seule l'indépendance financière[16] et le financement des stocks[17] apparaissent comme "nouvelles" variables, sans que leurs coefficients n'indiquent une relation avec la propension à la fraude. Toutefois, l'élément le plus notable est l'absence de tout facteur, ou presque, lié aux dettes ou aux pertes d'exploitation de l'entreprise. Comme nous le verrons dans la partie qui suit, cet aspect est bel et bien présent mais n'est pas capturé par les modèles linéaires.

16. ratio entre les capitaux propres et les capitaux permanents
17. ratio entre les dettes aux fournisseurs et les stocks

4.7 Extensions

L'aspect explicatif des variables économiques est utile à la validation d'hypothèses sur la fraude, notamment lorsqu'elles paraissent paradoxales comme dans le cas d'une capacité d'autofinancement élevée. D'autres applications sont possibles, comme l'amélioration des capacités des modèles prédictifs ou la relation à des aspects spécifiques des cotisations sociales, par exemple lorsque celles-ci sont caractérisées par des modèles génératifs. Nous présentons dans les parties qui suivent divers aspects de ces possibilités.

4.7.1 Modèles non linéaires et importance des variables économiques

Les modèles non linéaires apportent deux avantages à la compréhension du problème de la fraude :
- leurs capacités prédictives sont généralement meilleures ;
- Dans certains cas, ils sont non paramétriques, ce qui permet de bénéficier d'une modélisation sans aucune hypothèse. En contrepartie, l'interprétation des variables peut être plus délicate.

`Variables les plus influentes et interactions`

Nous montrons, ici, comment la relation entre la propension à la fraude et la situation financière de l'entreprise est prise en compte dans le cadre des forêts uniformément aléatoires[18]. Nous avons énoncé dans précédemment un bref descriptif de ce modèle et nous exposons principalement les divers résultats permettant d'interpréter en un minimum d'étapes la problématique posée. Dans un premier temps, nous reportons les résultats de la modélisation de façon à les comparer au modèle linéaire.

18. les forêts uniformément aléatoires sont des modèles ensemblistes pour la classification, la régression, les systèmes de recommandation, ... Elles sont dérivées du Bagging (Breiman, 1996) et des forêts aléatoires (Breiman, 2001).

```
Random  uniform  forest on 'artificial + real fraud' dataset (summary):

Out-of-bag (OOB) evaluation
Mean of squared residuals: 6e-04

OOB residuals:
       Min         1Q     Median       Mean         3Q        Max
-0.4924000 -0.0020240  0.0049750  0.0007953  0.0069730  0.3533000

Variance explained: 60.43%

Theorical (Breiman) bounds
Theorical prediction error: 0.000532
Upper bound of prediction error: 0.000538
Mean prediction error of a tree: 0.00127
Average correlation between trees residuals: 0.4235
Expected squared bias (experimental): 0
```

Le résumé du modèle permet de produire un premier résultat sur notre échantillon. Relativement au modèle linéaire de départ, le pouvoir prédictif est presque deux fois plus important, avec un pourcentage de variance expliqué de 60%. Sur les données non vues par le modèle, dites Out-of-Bag ou OOB (Breiman, 2001) et équivalentes à un échantillon de test (sur environ 30% des données), l'erreur de prédiction est 20% inférieure que précédemment (0.0244 contre 0.0293, sur l'échantillon d'apprentissage, dans le modèle linéaire). Dans la suite, nous abordons l'influence des variables du point de vue de la visualisation. L'intérêt de cette dernière réside dans la transition entre le modèle et la relation entre variables économiques et niveau de fraude, qui s'effectue en trois étapes :
- l'importance locale qui mesure les interactions entre toutes les variables ;
- L'importance partielle qui transcrit l'influence des variables sur les niveaux de fraude positifs et en produit les plus importantes ;
- La dépendance partielle qui permet de décrire l'évolution de la variable à expliquer en fonction de n'importe quelle variable influente.

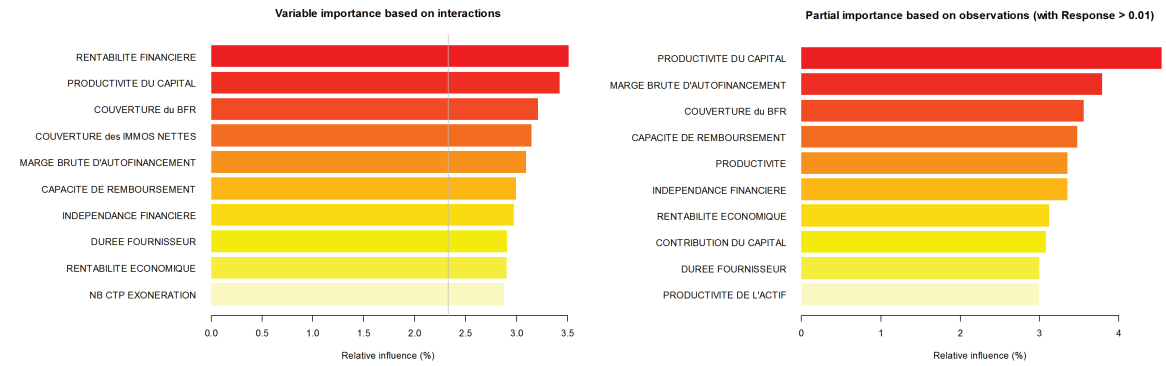

FIGURE 4.10 – Importance locale et importance partielle des variables sur le niveau de fraude dans une forêt uniformément aléatoire.

Le premier graphique est la mesure d'*importance locale* des variables. Elle est calculée en mesurant à la fois les interactions entre les variables et leur fréquence d'apparition dans les prédictions du modèle. Le second est la mesure d'*importance partielle*. Elle ne tient compte que de la fréquence d'apparition des variables dans les prédictions pour toutes les observations et pour un seuil de niveau de fraude supérieur à un seuil, ici 1%. Notons que ces types de mesure sont facilités par le caractère ensembliste des forêts uniformément aléatoires. 500 modèles de base, appliqués chacun à une perturbation de l'échantillon, sont générés pour construire le modèle complet.

Le principal élément fourni par les mesures d'importance locale et partielle est la nature de l'influence des variables économiques. Elle s'explique essentiellement par leurs interactions qui sont plus importantes, relativement au niveau de fraude, que celles entre les variables de la déclaration de cotisations. Dans une forêt uniformément aléatoire, les interactions sont additives et leur somme explique la totalité du phénomène observé. Naturellement, elles sont en grand nombre et l'importance locale permet d'en dégager les plus décisives. Prise individuellement, chaque variable économique n'a qu'une influence faible et seule leur co-influence permet d'expliquer le poids de la situation financière dans la relation au niveau de fraude. Sur les 10 variables que nous avons retenues, 8 sont de nature économique et expliquent environ 25% de toutes les interactions dans le modèle. Ce résultat permet de comprendre les raisons pour lesquelles la situation financière n'influence le niveau de fraude que partiellement. Il montre également pourquoi la détection de la fraude, dans le cas du travail dissimulé, est difficile. En particulier, les variables intrinsèques à la déclaration de cotisations ne permettent pas, non plus, de distinguer et d'expliquer les cas de fraude.
Dans la continuité du modèle linéaire, les cas de fraude les plus importants donnent un relief plus précis de la situation financière. L'importance partielle fournit une première analyse et nous retrouvons la majorité des variables discriminantes obtenues précédemment, en particulier la marge brute d'autofinancement et la productivité. Le modèle non linéaire introduit toutefois de nouvelles variables comme la productivité du capital, l'indépendance financière et la rentabilité économique. Afin de mieux les situer, nous observons leur comportement lorsque le seuil de niveau de fraude change :

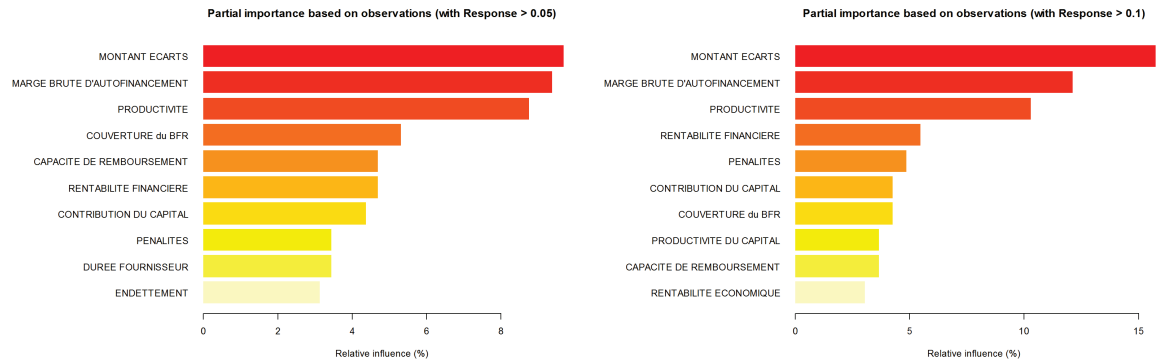

FIGURE 4.11 – Importance partielle des variables lorsque le niveau de fraude devient important.

En conditionnant les variables à l'accroissement du niveau de fraude et en restreignant le nombre de variables influentes à 10 au maximum, les effets locaux sont mis en évidence de manière plus prononcée. *Nous retrouvons quatre variables économiques présentes pour tous les niveaux de fraude entre 1% et (plus de) 5%. Elles caractérisent la situation financière des entreprises à travers l'équilibre financier (marge brute d'autofinancement et couverture du BFR), l'endettement (capacité de remboursement) et à la productivité des salariés.* Les pertes d'exploitation explicites (rentabilité économique[19]) ont, ici aussi, un rôle plus marginal, ce qui correspond aux observations faites sur les données réelles.

`Remarques:` La capacité d'autofinancement exprime en partie des pertes d'exploitation, lorsqu'elle est négative. la disparition de la productivité du capital à mesure que le niveau de fraude augmente indique que cette variable a surtout une influence lorsque le niveau de fraude reste faible. Cette remarque est également valable pour la rentabilité financière et pour toutes les variables qui ne sont pas persistantes avec l'augmentation du niveau de fraude.

L'importance locale détermine si la somme des interactions entre variables suffit à expliquer globalement la relation entre situation financière et propension à la fraude. Ici, aucune variable n'interagit avec les autres pour plus de 4% de l'ensemble des interactions identifiables par le modèle, ce qui signifie que la situation financière n'a pas d'effet global sur le niveau de fraude.
L'importance partielle détermine les variables influentes lorsque le niveau de fraude augmente. Ici, seul un petit nombre de variables économiques a un effet (local) sur la propension à la fraude. La productivité échappe à ce scénario par son absence dans les interactions entre variables. En ce sens, elle constitue une variable critique car elle indique une corrélation avec les niveaux de fraude positifs et une absence de relation lorsque le niveau de fraude est nul ou très faible (cette dernière situation est identifiable en utilisant à nouveau l'importance partielle). En d'autres termes, une trop grande productivité est un signal de fraude non bruité relativement à la distribution des autres variables explicatives.

19. ratio entre l'excédent brut d'exploitation (marge commerciale augmentée des subventions d'exploitation et diminuée des salaires et impôts) et le total du bilan

`Effets locaux des variables influentes`

De manière générale, le modèle non linéaire fournit des indications sur la nature et l'importance du bruit inhérent aux données et toute la difficulté réside dans son isolation. Pour cela, nous introduisons la dépendance partielle dont l'objectif est la mesure du niveau de fraude relativement à une variable explicative sachant la distribution de toutes les autres variables.

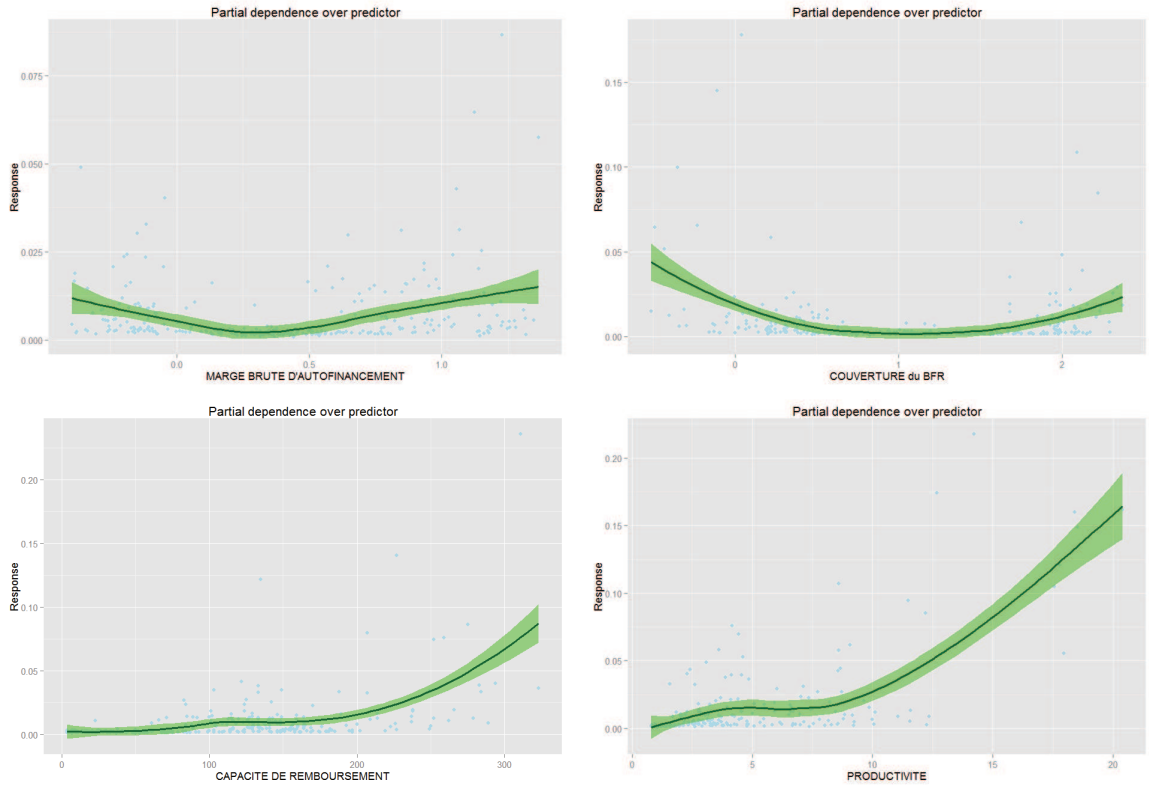

FIGURE 4.12 – Dépendances partielles et effet sur le niveau de fraude.

La variable *Response* désigne (de manière générique) le niveau de fraude. La dépendance partielle traduit l'effet marginal d'une variable explicative sur le niveau de fraude. L'effet non linéaire apparaît clairement sur les deux premiers graphiques et la tendance linéaire de la productivité avec le niveau de fraude est sensible pour des grandes valeurs de la variable. La capacité de remboursement suit cette même tendance avec des valeurs largement au-dessus de sa moyenne (41). Précisons que le nombre absolu d'observations (points en bleu) dans chaque graphique compte beaucoup moins lorsque l'effet des variables est local (dans ce cas, on en observe peu relativement à la totalité) que leur nombre relativement à l'absence ou à la présence de fraude.

- Le premier niveau d'interprétation de la dépendance partielle est l'ordre d'importance de la variable. Ici, seule la marge brute d'autofinancement est mesurée au second ordre ; cela est équivalent à considérer que la situation financière de l'entreprise est d'abord caractérisée par l'existence d'une autre variable (inconnue) plus spécifique que celle dont

on mesure l'influence. Disposer de plusieurs ordres permet une compréhension plus fine du problème. Pour les variables mesurées au premier ordre, la mesure de dépendance partielle signifie que chacune, par sa valeur ou sa plage de valeurs, a un effet primordial sur la propension à la fraude.

- La dépendance partielle permet également de mesurer la co-influence d'un couple de variables ainsi que leur dépendance en termes de corrélation. Ainsi, la dépendance dans le modèle entre les quatre variables est faible et la corrélation linéaire ne dépasse jamais 0.1, sauf celle entre la productivité et la marge brute d'autofinancement (0.28).

- Le troisième niveau d'analyse se situe dans les valeurs prises par les variables explicatives. Dans le cas de la capacité d'autofinancement, la propension à la fraude augmente au-dessus du 3^e quartile (0.6) et lorsque la variable prend des valeurs négatives, soit lorsque l'activité génère potentiellement des pertes d'exploitation. Le niveau de fraude est, alors, en moyenne, respectivement, de 1.08% et 1.31%. Les valeurs de la capacité d'autofinancement comprises entre 0 et 0.6 correspondent à 70% de toutes les observations de la variable et constituent la plage sur laquelle la capacité d'autofinancement n'est pas discriminante de la propension à la fraude.

La couverture du BFR, que l'on peut exprimer comme le rapport de la trésorerie au besoin en fonds de roulement, augmenté d'une unité, est dans le même cadre d'analyse. Une propension à la fraude en augmentation est liée aux quantiles extrêmes (soit les quantiles d'ordre 0.05 et d'ordre 0.8) de la variable. Ils correspondent respectivement à une trésorerie ou à un besoin en fonds de roulement négatifs (couverture du BFR inférieure à 0.5) ou bien à une trésorerie supérieure à la moitié du besoin en fonds de roulement (couverture du BFR supérieure à 1.5). Précisons que le besoin en fonds de roulement (BFR) est la mesure des ressources de l'entreprise nécessaires au financement de son cycle d'exploitation. Pour 75% des observations de la couverture du BFR, la dépendance partielle n'indique aucune relation avec la propension à la fraude. La couverture du BFR influence la propension de l'entreprise à frauder lorsque la trésorerie devient critique ou, paradoxalement, lorsque d'importantes ressources financières de court terme sont disponibles. A la différence de la capacité d'autofinancement, elle nécessite une réaction immédiate de l'entreprise et peut constituer une motivation explicite de la fraude. On peut remarquer que les capacités non linéaires étendent largement l'analyse faite dans le cadre de la régression linéaire, pour laquelle le coefficient associé à la couverture du BFR est négatif. Ici, la dépendance partielle permet l'analyse sur toute la distribution de la variable.

La capacité de remboursement est la troisième variable décisive. Elle est liée aux dettes bancaires (numérateur) et à la capacité d'autofinancement (dénominateur). Plus les dettes bancaires sont élevées (ou la capacité d'autofinancement faible), plus la propension à la fraude s'accroît. Les quantiles de queue de distribution sont ici encore prépondérants et expliquent les difficultés auxquelles peut faire face l'entreprise. Par exemple, une capacité de remboursement de 100 exprime un niveau de couverture des dettes bancaires par la capacité d'autofinancement de 1%. Même si ces dettes sont à long terme, elles exigent un remboursement périodique dont l'impact peut altérer les ressources de l'entreprise. Surtout, elles peuvent limiter des financements supplémentaires pour le cycle d'exploitation.

La productivité est la variable pour laquelle l'influence sur le niveau de fraude est mise en évidence le plus rapidement car elle prend en compte l'effectif. La propension à la fraude augmente de manière importante pour des valeurs de productivité supérieures à 2.5 (quantile d'ordre 0.85), soit environ un rendement du chiffre d'affaires par salarié de 280 000 euros en tenant compte de tout l'échantillon. Dans les données réelles de fraude, la productivité est alors supérieure à 175 000 euros. Les niveaux de productivité élevés résultent soit d'un processus de production très performant, soit d'une dissimulation de salariés.

Lorsque le niveau de fraude est important, la productivité l'est donc aussi. L'inverse n'est généralement pas valide et, comme dans le cas des autres variables décisives, caractériser la situation financière relativement à la propension à la fraude nécessite habituellement deux ou plusieurs variables influentes. La dépendance partielle nous évite en partie une telle sélection en intégrant la distribution de toutes les variables.

Ainsi, on peut, pour n'importe quel échantillon prenant en compte les mêmes secteurs d'activité que ceux définis pour l'analyse, généraliser la relation entre situation financière et propension à la fraude. C'est une différence fondamentale avec le modèle linéaire : dans ce dernier les coefficients sont déterminés par l'échantillon. Dans les forêts uniformément aléatoires, la situation financière est inférée pour (presque) tous les paramètres (des arbres de décision aléatoires sous-jacents). Un nouvel échantillon ne nécessite pas de calibration de paramètres et son inférence est simplement une généralisation du modèle.

`Non linéarité`

L'aspect le plus paradoxal de la relation entre le niveau de fraude et la représentation de la situation financière est son caractère non linéaire. Dans la majorité des variables économiques influentes, en particulier celles incluant la capacité d'autofinancement, la dépendance partielle montre que le niveau de fraude moyen est important sur les queues de distribution et faible ou nul au centre. Illustrons cela avec la dépendance partielle entre la rentabilité économique et le niveau de fraude :

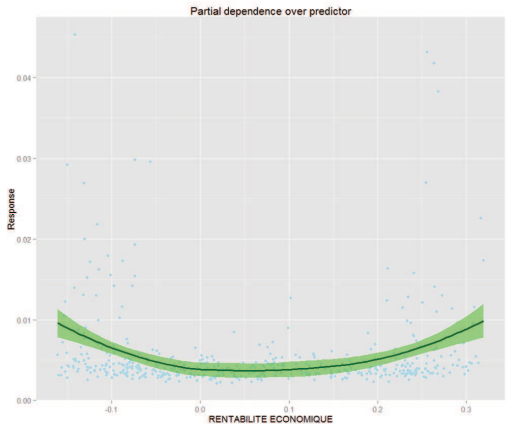

FIGURE 4.13 – Dépendance partielle de la rentabilité économique et du niveau de fraude.

La rentabilité économique correspond à l'excédent brut d'exploitation (EBE) rapporté au total du bilan. Lorsqu'il est négatif, l'entreprise subit des pertes d'exploitation. Ci-dessus, plus les pertes sont importantes (relativement au bilan) plus le niveau de fraude moyen est important. A mesure que la rentabilité économique s'améliore, le niveau de fraude moyen baisse pour atteindre un plateau (environ 0.4%) assimilable à une absence ou à une très faible fraude. Pour des valeurs élevées de rentabilité économique, avec des excédents d'exploitation dépassant 20% du bilan et largement au dessus-de la moyenne (6%) de toutes les entreprises, le niveau de fraude s'accroît à nouveau. Plus simplement, à bilan identique, les entreprises subissant d'importantes pertes d'exploitation et celles bénéficiant de larges excédents ont un niveau de fraude comparable, alors que leurs situations sont totalement différentes.

Dans le premier cas, des difficultés dans le cycle d'exploitation peuvent mener l'entreprise à de sévères complications sans une réaction rapide. Compte tenu des niveaux de perte et de la fraude, le recours à cette dernière est une possibilité parmi d'autres dans la résorption des pertes. Elle a l'avantage d'être immédiatement mesurable et de dégager des liquidités, en contrepartie d'un redressement alourdi de pénalités et sanctions si la fraude est avérée. Généralement, l'arbitrage se fait entre un risque à court terme potentiellement fatal pour l'activité et un risque (de redressement) dans un temps indéterminé qui n'est critique que si la situation de l'entreprise ne change pas entretemps.

Dans le cas d'une rentabilité économique importante, l'arbitrage a lieu entre un gain certain et un risque dont la réalisation, elle, ne l'est pas. La fraude entraîne des excédents encore plus élevés et peut jouer également un rôle dans une stratégie commerciale (par exemple, un contrôle plus granulaire des prix de vente) alors que le risque de redressement n'empêche pas l'entreprise de réaliser de nouveaux excédents.

`Co-dépendances`

Afin de distinguer la manière dont interagissent les variables influentes sur la propension à la fraude, nous illustrons ci-dessous quelques mesures de dépendance partielle par couple de variables.

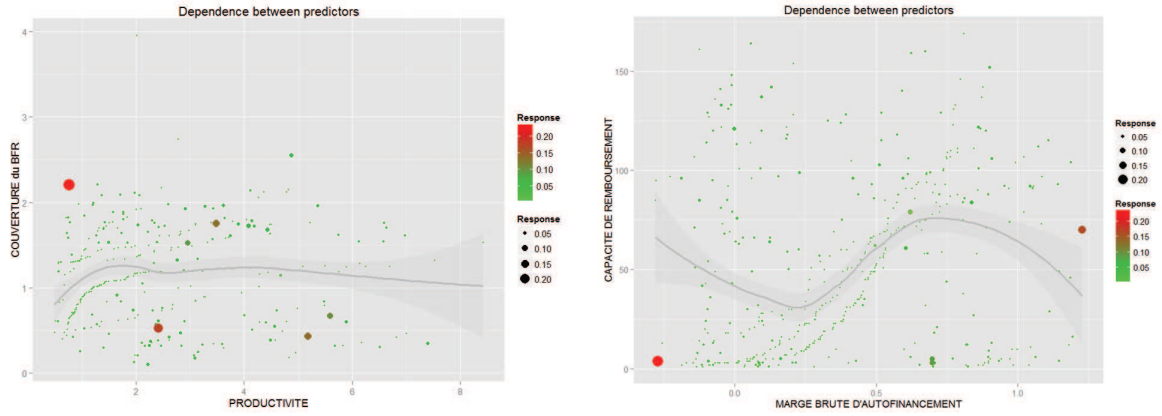

FIGURE 4.14 – Co-dépendances partielles et effet sur le niveau de fraude.

Dans chacun des graphiques, la répartition dans le plan du niveau de fraude (points en vert et en rouge) est quasiment uniforme, et la mesure de dépendance, conditionnellement au niveau de fraude, entre chaque couple de variables explicatives (courbe en gris) ne varie que très localement dans le cas de la productivité et de la couverture du BFR. Pour les capacités d'autofinancement (entre 0 et 0.6) et de remboursement (entre 25 et 75), une certaine dépendance est observable mais elle ne concerne que les niveaux de fraude les plus faibles, tandis que tout autour se répartissent les cas importants correspondant aux queues de distribution des variables. La co-dépendance permet de justifier le caractère additif de chaque variable influente. Pour une entreprise quelconque, le niveau de fraude sera d'autant plus important qu'une combinaison additive de la capacité d'autofinancement, de la couverture du BFR, de la capacité de remboursement ou de la productivité sera présente.

`Liquidité et solvabilité`

Nous illustrons également la solvabilité et la liquidité qui sont les deux facteurs qui, avec la couverture du BFR, décrivent les contraintes de financement et de capacité de l'entreprise à rembourser ses dettes à court ou long terme.

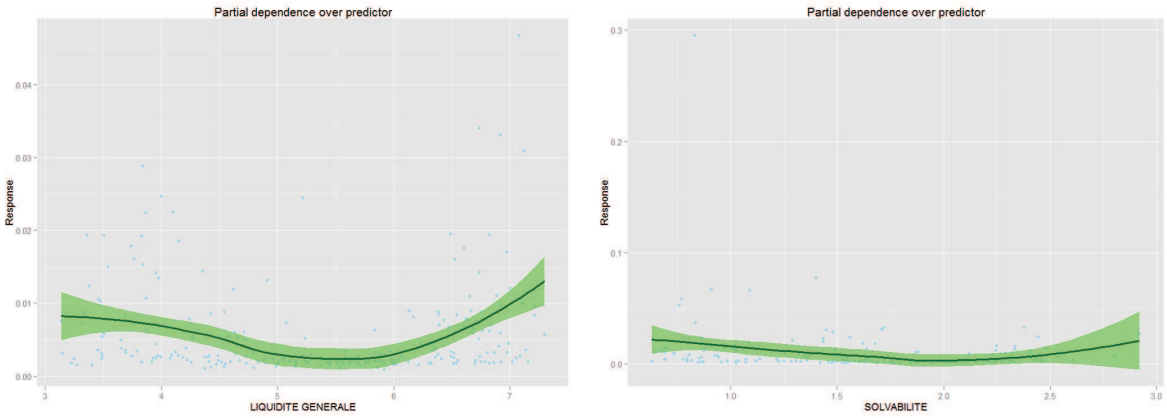

FIGURE 4.15 – Dépendances partielles de la liquidité générale et de la solvabilité et effet sur le niveau de fraude.

La liquidité prolonge le paradoxe observé pour certaines variables. Le niveau de fraude diminue à mesure que la liquidité augmente et, pour les grandes valeurs de la variable, s'accroît à nouveau. A contrario, une forte solvabilité coïncide avec une décroissance du niveau de fraude. Ces variables ne sont cependant pas retenues dans le modèle comme influentes et leur éventuel caractère explicatif n'est cohérent qu'une fois la situation financière caractérisée.

Du point de vue des dettes, ce ne sont donc pas les capacités de l'entreprise à y faire face qui comptent mais son besoin de financement immédiat (à travers la couverture du BFR) et sa capacité de remboursement des dettes bancaires relativement à son niveau d'autofinancement.

4.7.2 Seuils et modèles prédictifs

Une application de la présence de variables économiques est la détermination de seuils pour les modèles prédictifs. Supposons que nous souhaitions prioriser les cas les plus importants de fraude, nous pouvons, soit tenter de minimiser l'erreur quadratique, soit opter pour une classification selon que l'entreprise aura fraudé ou non. Dans la pratique, moins de 10 000 entreprises (sur 1 200 000) sont contrôlées, à ce jour, dans le cadre de la lutte contre le travail dissimulé. Bien que le taux de détection soit important ($> 75\%$) une problématique posée est de savoir comment détecter des cas plus importants sans recourir à une augmentation massive de contrôles. Pour cela, nous avons mesuré le niveau de fraude en dessous duquel les variables économiques perdaient leur caractère significatif. Sous ce seuil, le niveau de fraude peut être considéré comme du bruit, sans possibilité (ou très difficilement) de caractériser la situation financière relativement à la propension à frauder, ou de détecter une entreprise qui aura eu recours au travail dissimulé. En-dessous du seuil

il n'y a donc pas fraude. Le point de vue est celui de la classification. Nous recherchons alors, de manière empirique, un seuil en-dessous duquel nous considérons qu'il n'y a pas de fraude, si les variables caractérisant la situation financière ne sont pas significatives. En contrepartie, au-dessus, les variables retenues pour le modèle complet doivent rester significatives.

`Données synthétiques + réelles : niveau de fraude < 0.6%`

```
Residuals:
       Min         1Q     Median         3Q        Max
-0.0072649 -0.0008399 -0.0007061 -0.0001323  0.0055226

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2.317e-03  4.055e-04   5.713 1.16e-08 ***
COUVERTURE des IMMOS NETTES  -9.001e-05  3.160e-05  -2.849  0.00441 **
LIQUIDITE IMMEDIATE          -6.540e-05  3.562e-05  -1.836  0.06637 .
LIQUIDITE GENERALE           -8.117e-05  2.501e-05  -3.246  0.00118 **
POIDS MASSE SALARIALE         4.697e-04  2.318e-04   2.027  0.04275 *

DUREE DE VIE                  8.914e-06  3.539e-06   2.519  0.01180 *
MONTANT TAXATIONS D'OFFICE   -5.198e-04  2.474e-04  -2.101  0.03564 *
% VERSEMENT DEMAT.            3.164e-04  1.022e-04   3.096  0.00197 **
COMPLIANCE                   -3.984e-04  2.403e-04  -1.658  0.09742 .

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001549 on 6445 degrees of freedom
Multiple R-squared: 0.02261,    Adjusted R-squared: 0.01609
F-statistic: 3.467 on 43 and 6445 DF,  p-value: 1.998e-13
```

Un seuil de 0.6%, correspondant à une réduction de 30% du nombre d'entreprises dont le niveau de fraude est positif, élimine toutes les variables économiques significatives qui contribuent à une propension à la fraude plus importante. En dessous d'un niveau de fraude de 0.6% de la masse salariale, il n'est alors pas possible d'établir une relation entre situation financière et niveau de fraude en considérant un échantillon constitué d'entreprises avec un niveau de fraude positif ou nul. Au-dessus de 0.6%, seule la rentabilité commerciale n'est plus significative parmi les variables retenues dans le modèle complet.

Dans la pratique, cela permet d'envisager la classification comme un outil complémentaire d'analyse, voire comme l'outil primaire de prédiction. L'avantage du modèle linéaire est, ici, une mise en oeuvre rapide du filtrage pour les algorithmes de détection et nous avons généralisé cette méthode (en utilisant un seuil de 0.9%), à la détection des irrégularités aux cotisations sociales pour l'ensemble des entreprises d'Île-de-France avec, comme résultat, une bien meilleure *précision*[20] de la détection (+7%).

On peut appliquer cette méthode aux cas de travail dissimulé et mieux mettre en évidence l'apport des variables économiques. Comparons la précision de la classification selon l'utilisation ou non d'un seuil pour la définition du caractère frauduleux. Lorsqu'il n'y a pas

20. rapport entre le nombre de cas positifs (fraude) correctement prédits par l'algorithme et le nombre de cas positifs prédits par l'algorithme

de seuil, il y a tout simplement fraude dès que le niveau de cette dernière est positif. Les cas de fraude sont alors plus nombreux, mais ils sont un frein à la précision des modèles prédictifs. Nous utilisons comme modèle la forêt uniformément aléatoire et l'évaluation se fait par les données Out-Of-Bag (OOB), lesquelles ne participent pas à la prédiction du niveau de fraude.

Seuil du niveau de fraude	Erreur de prédiction	Précision	Nombre de cas de fraude détectés	Nombre total de cas de fraude
0	0.4629	0.5748	2931	5392
0.006	0.2923	0.8203	753	3511

TABLE 4.7 – Filtrage par seuil et classification selon l'absence ou la présence de fraude par une forêt uniformément aléatoire dans l'échantillon de données réelles et synthétiques.

Dans le tableau ci-dessus, on considère que la fraude n'est observable qu'à partir d'un certain niveau de fraude (le seuil). L'évaluation fournit une *erreur de prédiction*, correspondant au nombre d'entreprises identifiées, à tort, comme ayant fraudé (relativement au nombre total d'entreprises). La précision définit la capacité du modèle à détecter, avec une grande exactitude, les cas de fraude. Généralement, on ajoute la *sensibilité*[21] comme second (ou premier, selon le problème) critère de qualité pour le modèle. Dans ce cas, la capacité à trouver toute la fraude prime.

Le seuil défini par les variables économiques permet d'améliorer fortement la précision (delta de +25%) en contrepartie d'une baisse de la sensibilité (delta de -33%). Toutefois, dans le cas de la détection de la fraude, la précision est le paramètre le plus important :
- détecter toute (ou une grande partie de) la fraude est un idéal contraint par les ressources physiquement disponibles ;
- détecter (et surtout rendre générique la détection) les cas les plus importants est le coeur de la lutte contre le travail dissimulé et a généralement plus d'impact.

Plus spécifiquement, il est possible de construire un processus de détection guidé par des variables économiques, la précision et la sensibilité :
- dans un temps initial, un seuil est défini grâce aux variables économiques, puis la précision est évaluée.
- Puis, ce seuil évolue en fonction de l'évaluation faite des variables économiques, tout comme la précision est transférée vers la sensibilité (pour les problèmes difficiles, moins de précision implique généralement plus de sensibilité).
Dans le détail, on peut remarquer que le seuil déterminé à l'aide des variables économiques laisse de côté plus de 80% des entreprises avec un niveau de fraude positif.

Cette inefficacité apparente est, en fait, le principal effet pratique des variables économiques. La majorité des cas de fraude sont involontaires, ce qui explique leurs faibles valeurs et la distribution, de loi exponentielle, du niveau de fraude réellement observé. Pour ces mêmes raisons, une partie des contrôles de la lutte contre le travail dissimulé

21. rapport entre le nombre de cas positifs correctement prédits et le nombre de cas positifs

aboutit à un redressement pour minoration d'assiette et non comme travail dissimulé explicite. L'utilisation de variables économiques permet de conceptualiser ce cadre et d'y effectuer une recherche quasi exhaustive des cas de fraude les plus importants.

Nous illustrons ci-dessous une comparaison des montants de redressements récupérables selon que l'on détermine ou non un seuil du niveau de fraude.

Seuil	Niveau de fraude moyen des cas de fraude détectés	Niveau de fraude moyen des cas de fraude non détectés	Montant total de la fraude détectée
0	1.60%	0.59%	10 039 577 euros
0.006	6.25%	0.59%	5 283 542 euros

TABLE 4.8 – Filtrage par seuil et montants des redressements dans l'échantillon mélangeant les données réelles et synthétiques.

Lorsqu'un seuil est utilisé, le niveau de fraude moyen est quatre fois plus important. En d'autres termes, le montant moyen d'un redressement est multiplié par un facteur 4. Il n'y a, de plus, pas de conséquence sur la fraude non trouvée puisque son niveau moyen (0.59%) reste identique. Logiquement, le montant total de la fraude est moindre que lorsqu'il n'y a pas de seuil (et qu'on cherche alors à détecter toute la fraude). Toutefois, il n'est que deux fois plus petit (5 millions d'euros) alors que le filtrage par seuil sélectionne quatre fois moins d'entreprises.

Le principal apport des variables économiques est un meilleur positionnement du seuil à partir duquel on peut détecter beaucoup plus facilement les entreprises frauduleuses. En contrepartie d'une perte de puissance, le rendement de la détection est fortement accru et permet surtout la détection de (presque) tous les cas de fraude importants.

L'élément le plus intéressant est la généricité de la méthode. Elle s'applique aussi bien au travail dissimulé qu'à toutes les irrégularités aux cotisations sociales. On peut la connecter au résultat empirique suivant : pour les variables à expliquer dont la valeur 0 est la plus fréquente, l'erreur de prédiction d'un modèle prédictif (ensembliste) pour les cas positifs (de fraude) est très souvent importante, à cause du déséquilibre avec les cas négatifs et de l'absence d'informations, dans les données, suffisamment pertinentes pour une meilleure discrimination entre les cas. Une manière abordable de l'analyser est le traitement spécifique des valeurs nulles. Ici, l'analyse des variables économiques en est un paradigme, par l'exploitation de leur caractère explicatif pour produire un modèle prédictif moins puissant mais plus précis.

4.7.3 Une modélisation des cotisations sociales

Nous terminons l'exploration des possibilités offertes par les variables économiques sur un point de vue plus expérimental. Dans le modèle linéaire, le taux de cotisation est une variable significative et possède une relation décroissante avec le niveau de fraude. Plus une entreprise paierait un taux de cotisation global élevé, moins elle serait susceptible de recourir à la fraude. Cela n'est pas le cas. Le taux de cotisation a une relation non

linéaire avec le niveau de fraude : ce dernier est important à la fois pour de faibles taux de cotisation et pour les taux les plus importants. Nous l'illustrons à nouveau dans la dépendance partielle du niveau de fraude au taux de cotisation.

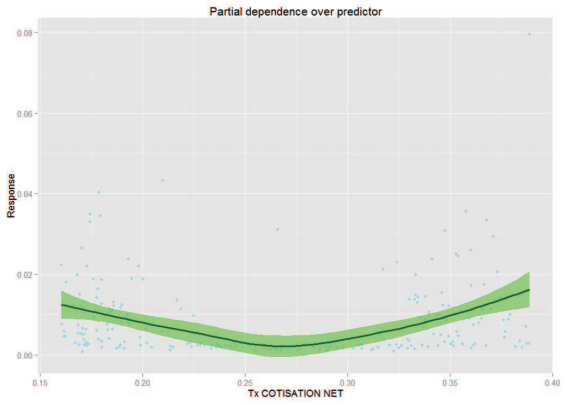

FIGURE 4.16 – Dépendance partielle du niveau de fraude au taux de cotisation principal.

Le taux de cotisation global (sur le graphique, nous illustrons le taux de cotisation principal) correspond à la quantité $\frac{C_{i.}}{M_i}$ pour l'entreprise i, soit le rapport entre le montant total des cotisations versées, $C_{i.}$ et la masse salariale, M_i. Mais, ni la cotisation déclarée, ni la masse salariale ne sont théoriquement exactes. Nous montrons dans les lignes qui suivent comment générer des contreparties pour le taux de cotisation global. Le modèle proposé s'affranchit de la validité des paramètres de cotisation au profit de leur cohérence, relativement à la présence ou à l'absence de fraude.

Chaque entreprise déclare, au moins, deux types de cotisation que sont le *Cas général* et la CSG. Notons τ, le taux de cotisation correspondant au premier cas. Nous supposons alors que l'entreprise paie une part fixe de cotisation et une part variable, de sorte que :

$$C_{i.} = F_{i.} + V_{i.},$$

où F est la partie fixe, et la plus importante, de la cotisation et V, la partie variable.

Comme nous n'avons pas de certitude sur les cotisations et la masse salariale, nous supposons que n'importe quelle entreprise est susceptible de frauder. On a alors pour l'entreprise i :

$$F_{i.} = \tau_i \left(1 - \sum_{j=1}^{d} \lambda_{ij} a_{ij}\right) M_i, \quad 0 \leq \lambda_{ij} < 1, \quad d = p - 1.$$

Lorsqu'il n'y a pas fraude, λ_j vaut théoriquement 0, et la partie fixe est le produit du taux, τ_i, et de la masse salariale. Le coefficient d'assiette, a_{ij}, vaut 1 et tous les salariés paient la cotisation. Lorsqu'il y a une fraude ou une incohérence, $\lambda_j > 0$ et la partie fixe de la cotisation est minorée de toutes les cotisations éludées. Cette minoration ne prend explicitement en compte que le coefficient d'assiette, tandis que λ_j intègre à la fois le taux et la proportion d'effectif de chaque cotisation éludée.

Nous définissons ensuite la part variable de la cotisation. On a :

$$V_{i.} = M_i \sum_{j=1}^{d} a_{ij} u_{ij} t_{ij}.$$

Pour rappel, u et t correspondent, respectivement, à la proportion d'effectif et au taux de cotisation. La part variable n'intègre pas une éventuelle minoration de cotisations, déjà transcrite dans le produit $a\lambda$. On souhaite mesurer le couple $(a_{.j}, \lambda_{.j}), 1 \leq j \leq d$, pour toutes les entreprises.

Nous considérons alors une entreprise et une cotisation quelconques et omettons leurs indices pour plus de simplicité. Il nous faut une mesure préalable des taux et des proportions d'effectif de chaque cotisation. Pour la cotisation du *Cas général*, nous disposons de la décomposition du taux :

$$\tau = \tau_d + \frac{M_p}{M_d} \tau_p,$$

où M_p et M_d sont les masses salariales plafonnée et déplafonnée,
τ_p et τ_d les taux associés.

Pour la mesure de u, il n'y a aucun moyen de la garantir, mais elle peut être générée. La seule condition dont il faut s'assurer est qu'elle permette de reconstruire chaque cotisation. On commence par construire la borne supérieure de u. On suppose :

$$u \leq \alpha \sqrt{\frac{|V|}{M}},$$

$$u_{max} = \underset{\alpha}{\texttt{max}}\ (u),$$

et

$$\alpha = \sqrt{\frac{1}{t}} > \sqrt{\frac{1}{t_{CSG}}}.$$

t_{CSG} est le taux de cotisation de la CSG. Comme tous les salariés paient cette cotisation, la proportion d'effectif vaut nécessairement 1 pour la CSG. De plus, M doit alors être récrit car la CSG a pour assiette 98.25% (en 2013) des revenus d'activité (dont la masse salariale). Nous supposons que M intègre cette contrainte. En remarquant que $\texttt{max}\ (u) = \texttt{max}\ (u^2)$ et que $u \geq u^2$, on pose alors $x = V$ et on en déduit :

$$u(x) \in [u^2(x), u_{max}]$$

avec

$$u_{max} = \sqrt{\frac{1}{t_{CSG}}} \sqrt{\frac{|V|}{M}}.$$

Un estimateur de $u(x)$ est donné par :

$$u^*(x) = \begin{cases} \frac{1}{2}(u^2(x) + u_{max}), si\ \alpha > \sqrt{\frac{1}{t_{CSG}}} \\ 1,\ sinon. \end{cases}$$

La proportion d'effectif est positive ou nulle, plus petite ou égale à 1 et cohérente avec la CSG. De plus, nous disposons de bornes supérieures et inférieures.

De la même manière, on définit t de manière générative. On pose :

$$|t| \leq \sqrt{\frac{|V|}{uM}},$$

et on en déduit un estimateur :

$$t^*(x) = \begin{cases} sign(x)\frac{1}{2}\sqrt{\frac{|x|}{u^*(x)M}}, si\ x < x_{CSG} \\ \\ sign(x)\sqrt{\frac{x_{CSG}}{M}},\ sinon. \end{cases}$$

La génération de t est cependant partielle et se limite aux cas où le taux d'application de la cotisation est non calculable ou n'est pas donné. Le choix de t est effectué de telle sorte que la mesure de réduction (donc avec un taux négatif) la plus importante dans le modèle ait un taux, en valeur absolue, très proche du taux observé dans la réalité.

Le point essentiel dans ce modèle est que les valeurs effectives de (u, t) n'ont pas d'intérêt intrinsèque. Seule leur cohérence pour la reconstruction des cotisations nous est utile et elles n'interviennent que pour la mesure de (λ, a). Comme les garanties sur les paramètres sont difficiles à obtenir, on s'intéresse plus précisément à la distribution de (λ, a) conditionnellement à la présence ou à l'absence de fraude. La cohérence est alors assurée. Il n'est pas possible, bien sûr, de détecter aussi simplement la fraude à partir des distributions conditionnelles, mais il peut être, par exemple, intéressant d'analyser les distributions de (λ, a) relativement à la caractérisation de la situation économique. Pour trouver les valeurs admissibles (λ, a), on peut remarquer que :

$$\begin{cases} C_{.j} = M\left(\tau\left(1 - \sum_{j=1}^{d} \lambda_{.j} a_{.j}\right) + a_{.j} u_{.j} t_{.j}\right) \\ \\ \sum_{j=1}^{p} C_{.j} = \tau M\left(1 - \sum_{j=1}^{d}\left(\lambda_{.j} - \frac{u_{.j} t_{.j}}{\tau}\right) a_{.j}\right). \end{cases} \tag{4.2}$$

En posant $\lambda_{.j} \equiv \lambda$, le système d'équations possède une solution explicite pour toutes les cotisations $j, 1 \leq j \leq p$, d'une entreprise quelconque. Soit $Y = \{$ fraude, absence de fraude $\}$, la distribution de choix pour la cotisation $C_{.j}$ est celle de $(\lambda, a_{.j})|Y$. En utilisant les résultats obtenus grâce aux variables économiques, lesquelles permettent d'exposer les niveaux de fraude les plus significatifs, les distributions conditionnelles fournissent un outil dans la compréhension du mécanisme de fraude au sein de la structure des cotisations.
Notons deux autres manières d'analyser le problème. La première fait appel à l'imputation par un des nombreux modèles disponibles, dont les forêts (uniformément) aléatoires. On peut, par exemple, imputer les paramètres des cotisations pour lesquelles aucune fraude n'a été constatée, puis généraliser le modèle à tout type d'entreprise. Une seconde manière consiste à utiliser des algorithmes évolutionnaires (éventuellement couplés à des algorithmes d'apprentissage automatique) pour approximer la relation (4.2).

4.8 Discussion

La fraude aux cotisations sociales constitue un sujet d'analyse important, particulièrement lorsque le niveau de déficit entre recettes et prestations sociales devient chronique. De notre point de vue, la question essentielle est celle de la garantie du montant des cotisations versées par les entreprises. Cette garantie est, à ce jour, inexistante. Principalement pour des questions de ressources, il apparaît difficile de limiter les phénomènes de fraude par des méthodes classiques (croisement de données et signalements essentiellement). Parallèlement, nous notons que le contrôle de la fraude fiscale semble dépasser les questions de ressources (le budget de la Sécurité sociale est comparativement plus important que l'ensemble des recettes fiscales) et fait dorénavant partie des revenus participant du budget de l'Etat (plus de 2 Mds d'euros attendus en 2014). Ce parallèle permet de positionner la question de la fraude, et, plus généralement, des irrégularités aux cotisations sociales.

L'analyse proposée donne un point de vue empirique de la question, lorsque le niveau de fraude présente une relation avec une caractérisation possible de la situation financière de l'entreprise. Nous montrons l'existence d'un nombre de variables économiques limité, en relation avec la propension à la fraude. Ces variables économiques ne déterminent ni la fraude, ni même la santé financière intrinsèque de l'entreprise. En effet, nous n'avons à disposition qu'un seul exercice comptable pour chaque entreprise. De plus, la présence des variables de la déclaration de cotisations est une nécessité sans laquelle le caractère significatif est beaucoup plus complexe à évaluer. Pour l'analyse, un grand échantillon, mêlant des données synthétiques avec un niveau de fraude aléatoire, et des cas réels de fraude (60% des cas redressés en 2009), est généré. Puis un premier modèle, linéaire, met en évidence des facteurs économiques typiques de la situation financière des entreprises relativement à la propension à frauder. La capacité d'autofinancement, le besoin de financement du cycle d'exploitation et la productivité apparaissent comme les facteurs essentiels de la relation entre propension à la fraude et la situation financière de l'entreprise. Environ 40% des cas de fraude peuvent être expliqués par cette relation. En analysant les données plus en détail, le lien avec la propension à frauder apparaît, généralement, non linéaire et possède une forme en U au regard d'une partie des variables influentes. Cette analyse étend la modélisation linéaire et permet de situer les cas de fraude. La propension à la fraude s'exprime dans les queues de distribution des variables et la situation financière correspond essentiellement à quatre facteurs :
- l'équilibre financier résumé par la capacité d'autofinancement ;
- les variations de trésorerie (couverture du besoin en fonds de roulement) ;
- l'endettement à travers la capacité de remboursement des dettes bancaires ;
- les variations du chiffre d'affaires, résumées par la productivité des salariés.

Le point le plus important dans la compréhension de la fraude est le caractère local et additif des facteurs. L'aspect local est déterminé par la faible influence individuelle de chaque facteur tandis que leur combinaison est justifiée par les faibles interdépendances qui les caractérisent. Malgré le caractère local de la contribution des facteurs économiques, une explication des cas de fraude les plus importants est fournie par une représentation de la situation financière. Plus le niveau de fraude augmente, plus la probabilité d'une

combinaison de plusieurs facteurs économiques explicatifs est grande. La relation entre propension à la fraude et variables économiques fournit un point de vue explicatif (par le nombre de variables significatives) qualitatif (par leur spécificité) mais aussi quantitatif (par l'amélioration des capacités des modèles prédictifs) sur la situation financière des entreprises. En inclure un grand nombre dans ce point de vue est une étape naturelle mais plus complexe, et passe nécessairement par l'assistance de modèles prédictifs. En nous référant à nouveau à la lutte contre la fraude fiscale et aux recettes envisagées, la lutte contre la fraude aux cotisations sociales semble subir les mêmes enjeux.

4.9 Annexe et définition des variables

Nous présentons quelques informations supplémentaires sur les entreprises, la définition complète des variables utilisées et l'état de la régression linéaire sur les données réelles (en incluant l'ensemble des variables), sur l'échantillon final (données réelles + synthétiques)

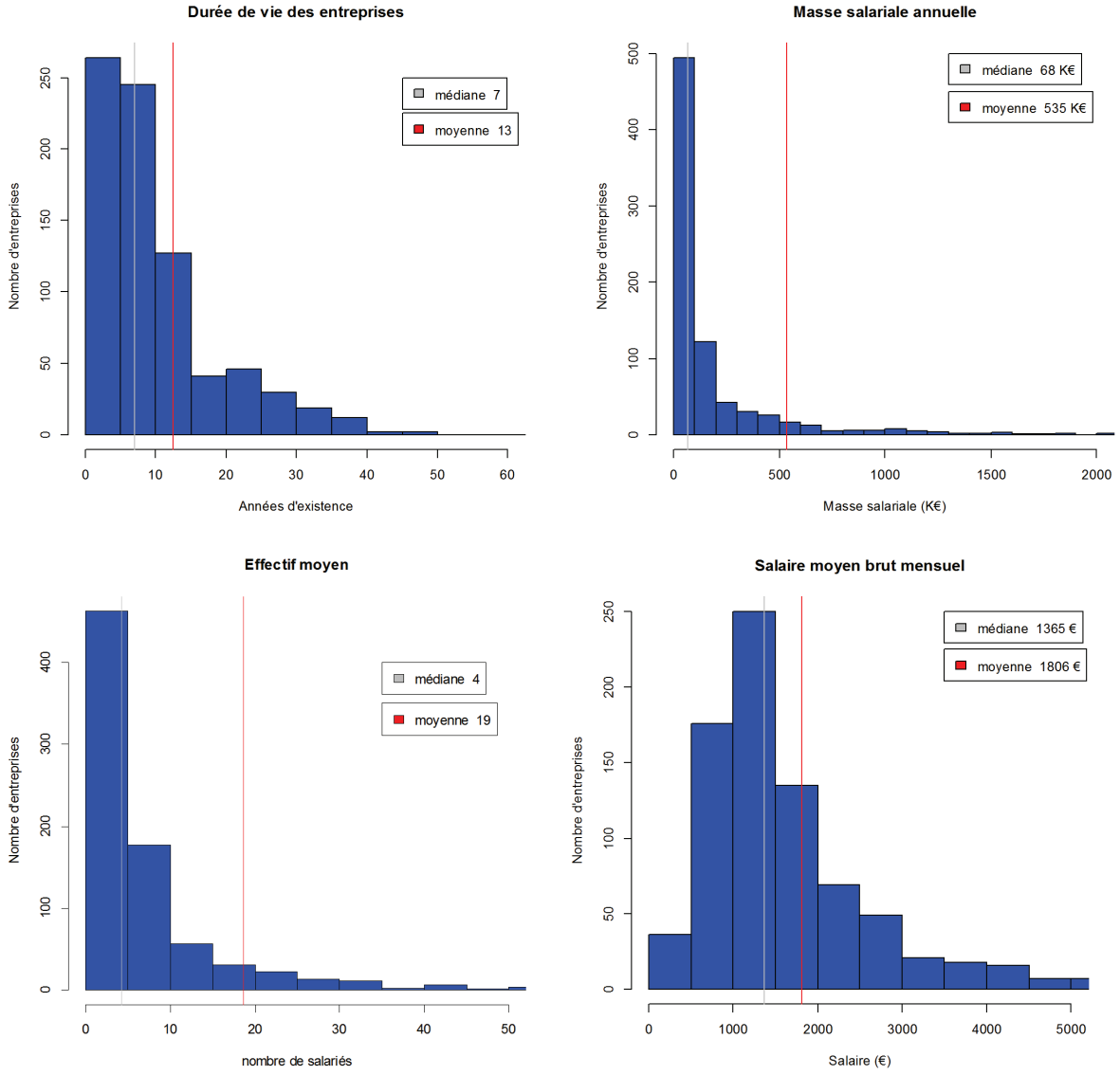

FIGURE 4.17 – Quelques caractéristiques des entreprises contrôlées en 2009, dans le cadre de la lutte contre le travail dissimulé.

La masse salariale déclarée est plutôt dispersée et la majorité des entreprises contrôlées sont des petites entreprises, conformément à leur représentation dans la population. L'effectif enregistré est également celui déclaré et se traduit par une même distribution que la masse salariale. Le salaire moyen mensuel subit une compression vers le salaire minimum, et en deçà, signe d'une tension sur l'activité économique des entreprises ou de salaires

structurellement faibles dans les secteurs d'activités considérés. Certains salaires ($> 3\,000$ euros) paraissent importants au regard d'éventuelles difficultés. Toutefois, cette observation est contrainte par le fait qu'un ajustement des salaires à l'activité économique n'est ni systématique, ni la norme.

`Définition des variables économiques:`

Pour la facilité de lecture, les variables sont définies dans une forme compacte. La plupart d'entre elles sont la somme d'autres variables du cycle d'exploitation de l'entreprise, dont nous ne gardons que les plus importantes ou celles les plus proches du sujet. Les variables sont notées dans le même ordre que celui généré par le fournisseur des données.

`Equilibre financier`
Marge brute d'autofinancement (ou, à peu de choses près, capacité d'autofinancement) : somme du résultat net et des dotations aux amortissements et provisions... La variable est exprimée relativement à la masse salariale.

Couverture du BFR : ratio entre le fonds de roulement (différence entre les capitaux permanents et les actifs immobilisés) et le besoin en fonds de roulement.

Couverture des immos nettes : ratio entre les capitaux permanents (fonds propres, provisions, impôts différés, dettes à long terme) et les immobilisations nettes (éléments du patrimoine de l'entreprise comme les brevets, les usines, les machines,...).

Couverture du CA : ratio entre le fonds de roulement (annualisé) et le chiffre d'affaires.

Solvabilité : ratio entre les capitaux propres et l'ensemble des dettes.

Indépendance financière : ratio des capitaux propres aux capitaux permanents.

`Profitabilité`
Rentabilité économique : ratio de l'excédent brut d'exploitation (marge commerciale augmentée des subventions d'exploitation et diminuée des salaires et impôts) au total du bilan.

Rentabilité financière : ratio du résultat net aux capitaux propres.

Rentabilité commerciale : ratio du résultat net au chiffre d'affaires.

Contribution du capital : ratio de la capacité d'autofinancement annuelle aux capitaux permanents.

Contribution de la VA : ratio de la capacité d'autofinancement à la valeur ajoutée (somme de la marge commerciale et de la différence entre la valeur de la production et les coûts de production...).

`Liquidité`
Liquidité immédiate : ratio entre les disponibilités (trésorerie) et les dettes à court terme (dettes du cycle d'exploitation, à moins d'un an).

Liquidité générale : ratio entre l'actif circulant net (stocks, créances et valeurs mobilières de placement) et les dettes à court terme.

Liquidité réduite : ratio entre la somme des disponibilités et des créances réelles, et les dettes à court terme.

`Endettement`
Endettement : ratio entre les dettes à long terme et les capitaux propres.

Capacité de remboursement : ratio entre les dettes bancaires et la capacité d'autofinancement.

Financement des stocks : ratio entre les dettes aux fournisseurs et les stocks.

`Productivité`
Productivité de l'actif : ratio entre le chiffre d'affaires et l'actif comptable.

Durée client : ratio entre l'encours clients et le chiffre d'affaires ; équivalent au délai moyen de paiement des clients.

Durée fournisseur : délai moyen des crédits accordés par les fournisseurs.

Poids masse salariale : ratio entre la masse salariale et la valeur ajoutée.

Productivité : ratio entre le chiffre d'affaires et la masse salariale, pondéré par (l'inverse de) l'effectif. Equivalent à l'effet de levier produit par les salariés. Pour la productivité dans un sens plus classique, la masse salariale vaut 1.

Productivité du travail : ratio entre le chiffre d'affaires et la masse salariale.

Productivité du capital : ratio entre le chiffre d'affaires et le coût du capital estimé (variable reconstituée). Le coût du capital est défini, ici, comme la somme des coûts de production (hors coût du travail) augmentés de la différence entre le chiffre d'affaires et la valeur de la production.

`Définition des variables de la déclaration de cotisation:`

Durée de vie : durée d'existence de l'entreprise.

Nb ets : nombre d'établissements (pour une entreprise qui a des filiales).

Nb orig. débit non rens. : nombre d'origines de l'écart entre montants attendus de cotisations et montants recouvrés, non renseignés (équivalent au nombre de catégories de cotisation pour lesquels des écarts de cotisations sont observés).

Montant débit :montant de la dette de l'entreprise vis-à-vis de l'URSSAF.

Nb CTP exonération : nombre de catégories de cotisation pour lesquelles l'entreprise bénéficie d'une mesure de réduction.

Nb remises sur majoration : nombre de remises sur majoration de cotisations.

Montant écarts : montant des écarts de cotisation.

Pénalités : pénalités imputées à l'entreprise par l'URSSAF.

Nb retards : nombre annuel de retard de cotisations (selon le cas, elles sont dues mensuellement ou trimestriellement).

Nb taxations d'office : nombre de taxations forfaitaires établies par l'URSSAF (lorsqu'elle possède préalablement une information sur le statut déclaratif de l'entreprise).

Montant taxations d'office : montant annuel taxé forfaitairement.

Nb demandes délais : nombre de demandes de délais de paiement.

Nb délais acceptés : nombre de demandes de délais acceptés.

% versement demat. : pourcentage de versements de cotisation effectuées par voie électronique.

Dernier ctrl : délai écoulé depuis le dernier contrôle effectué par les inspecteurs de l'URSSAF (au minimum 3 ans, au maximum la différence entre l'année courante et 1900).

Nb CCA : nombre de contrôles comptables d'assiette (le type de contrôle le plus répandu) effectués par les inspecteurs de l'URSSAF.

compliance : indice (entre 0 et 1) de conformité de l'entreprise à respecter ses obligations déclaratives. Non enregistrée par l'URSSAF et basée sur un modèle théorique.

Tx cotisation net : taux de cotisation net (après déduction des mesures de réduction et dérogatoires) estimé. Non enregistré par l'URSSAF et basé sur un modèle théorique.

`Statistiques de l'échantillon final : données synthétiques + données réelles.`

Toutes les variables économiques sont des ratios. La marge brute d'autofinancement et la productivité sont exprimées en proportion de la masse salariale, les durées client et fournisseur en années. Les valeurs pour les variables de la déclaration de cotisation, hormis celles intrinsèquement non monétaires, sont exprimées relativement à la masse salariale.

```
MARGE BRUTE D'AUTOFINANCEMENT COUVERTURE du BFR COUVERTURE des IMMOS NETTES COUVERTURE du CA
Min.   :-5.1922               Min.   :-1.0600   Min.   :0.160               Min.   :-0.33425
1st Qu.: 0.1588               1st Qu.: 0.8443   1st Qu.:3.907               1st Qu.: 0.03559
Median : 0.3541               Median : 1.1289   Median :4.309               Median : 0.14833
Mean   : 0.3905               Mean   : 1.2548   Mean   :4.162               Mean   : 0.18885
3rd Qu.: 0.5659               3rd Qu.: 1.4413   3rd Qu.:4.667               3rd Qu.: 0.28992
Max.   :32.3559               Max.   : 5.2301   Max.   :7.572               Max.   : 0.93243

 SOLVABILITE     INDEPENDANCE FINANCIERE RENTABILITE ECONOMIQUE RENTABILITE FINANCIERE
Min.   :-0.200   Min.   :0.06459         Min.   :-0.17000       Min.   :-0.5695
1st Qu.: 1.281   1st Qu.:0.48328         1st Qu.: 0.02145       1st Qu.: 0.1992
Median : 1.704   Median :0.56418         Median : 0.05733       Median : 0.2992
Mean   : 1.626   Mean   :0.56938         Mean   : 0.05594       Mean   : 0.2943
3rd Qu.: 1.987   3rd Qu.:0.64821         3rd Qu.: 0.09028       3rd Qu.: 0.3974
Max.   : 3.420   Max.   :1.08741         Max.   : 0.26000       Max.   : 0.9500

RENTABILITE COMMERCIALE CONTRIBUTION DU CAPITAL CONTRIBUTION DE LA VA LIQUIDITE IMMEDIATE
Min.   :-0.20149        Min.   :-0.2433         Min.   :-0.28540      Min.   :0.050
1st Qu.:-0.11080        1st Qu.: 0.0829         1st Qu.: 0.09987      1st Qu.:3.157
Median :-0.01272        Median : 0.1575         Median : 0.15855      Median :3.673
Mean   :-0.03798        Mean   : 0.1565         Mean   : 0.15292      Mean   :3.430
3rd Qu.: 0.02788        3rd Qu.: 0.2316         3rd Qu.: 0.21586      3rd Qu.:4.036
Max.   : 0.13852        Max.   : 0.7200         Max.   : 0.43950      Max.   :5.776

LIQUIDITE GENERALE LIQUIDITE REDUITE  ENDETTEMENT      CAPACITE DE REMBOURSEMENT
Min.   :0.300      Min.   :0.090     Min.   :-0.6147   Min.   :  0.00
1st Qu.:4.817      1st Qu.:3.768     1st Qu.: 0.9450   1st Qu.: 13.00
Median :5.380      Median :4.220     Median : 1.5342   Median : 29.00
Mean   :5.169      Mean   :4.038     Mean   : 1.5770   Mean   : 40.62
3rd Qu.:5.872      3rd Qu.:4.626     3rd Qu.: 2.1943   3rd Qu.: 55.00
Max.   :8.722      Max.   :7.071     Max.   : 4.0316   Max.   :507.00

FINANCEMENT DES STOCKS PRODUCTIVITE DE L'ACTIF  DUREE CLIENT       DUREE FOURNISSEUR
Min.   :0.570          Min.   :0.2516          Min.   :5.023e-05   Min.   :0.01629
1st Qu.:3.882          1st Qu.:1.7737          1st Qu.:6.459e-02   1st Qu.:0.12088
Median :4.609          Median :2.1085          Median :8.700e-02   Median :0.14872
Mean   :4.628          Mean   :2.1173          Mean   :9.963e-02   Mean   :0.15090
3rd Qu.:5.450          3rd Qu.:2.4735          3rd Qu.:1.279e-01   3rd Qu.:0.17864
Max.   :8.737          Max.   :4.1779          Max.   :3.006e-01   Max.   :0.33151

POIDS MASSE SALARIALE  PRODUCTIVITE     PRODUCTIVITE DU TRAVAIL PRODUCTIVITE DU CAPITAL
Min.   :0.4238        Min.   : 0.0014   Min.   : 0.782          Min.   :  0.0076
1st Qu.:0.6980        1st Qu.: 0.7384   1st Qu.: 3.906          1st Qu.:  1.1537
Median :0.7632        Median : 1.0439   Median : 5.988          Median :  1.5445
Mean   :0.7660        Mean   : 1.5947   Mean   : 7.825          Mean   :  3.1999
3rd Qu.:0.8316        3rd Qu.: 1.7424   3rd Qu.: 9.399          3rd Qu.:  2.5118
Max.   :1.2100        Max.   :85.6905   Max.   :51.469          Max.   :776.2540
```

```
 DUREE DE VIE       NB ETS       NB ORIG. DEBIT NON RENS. MONTANT DEBIT      NB CTP EXONERATION
Min.   : 1.00   Min.   : 1.000   Min.   : 0.00            Min.   : 0.01988   Min.   :  0.000
1st Qu.: 7.00   1st Qu.: 1.000   1st Qu.: 8.00            1st Qu.: 0.41186   1st Qu.:  5.000
Median :10.00   Median : 1.000   Median :11.00            Median : 0.49487   Median :  7.000
Mean   :10.58   Mean   : 1.136   Mean   :11.31            Mean   : 5.58290   Mean   :  7.416
3rd Qu.:13.00   3rd Qu.: 1.000   3rd Qu.:14.00            3rd Qu.: 0.62999   3rd Qu.:  9.000
Max.   :71.00   Max.   :66.000   Max.   :87.00            Max.   :66.05372   Max.   :351.000

NB REMISES SUR MAJORATION MONTANT REMISES SUR MAJORATION MONTANT ECARTS     PENALITES
Min.   : 0.00             Min.   :0.0000000             Min.   :0.0000   Min.   :0.0000000
1st Qu.: 1.00             1st Qu.:0.0007322             1st Qu.:0.0579   1st Qu.:0.0008077
Median : 1.00             Median :0.0013276             Median :0.1208   Median :0.0013709
Mean   : 1.08             Mean   :0.0014869             Mean   :0.1636   Mean   :0.0016720
3rd Qu.: 1.00             3rd Qu.:0.0019924             3rd Qu.:0.2284   3rd Qu.:0.0021875
Max.   :58.00             Max.   :0.0821956             Max.   :3.0844   Max.   :0.0854745

  NB RETARDS      NB TAXATIONS D'OFFICE MONTANT TAXATIONS D'OFFICE NB DEMANDES DELAIS
Min.   :  0.000   Min.   : 0.000        Min.   :0.00000            Min.   : 0.0000
1st Qu.:  3.000   1st Qu.: 1.000        1st Qu.:0.06679            1st Qu.: 0.0000
Median :  5.000   Median : 1.000        Median :0.12190            Median : 0.0000
Mean   :  5.854   Mean   : 1.168        Mean   :0.21313            Mean   : 0.2427
3rd Qu.:  8.000   3rd Qu.: 2.000        3rd Qu.:0.26853            3rd Qu.: 0.0000
Max.   :209.000   Max.   :65.000        Max.   :1.61058            Max.   :13.0000

NB DELAIS ACCEPTES % VERSEMENT DEMAT.  DERNIER CTRL          NB CCA           COMPLIANCE
Min.   : 0.0000    Min.   : 0.00000   Min.   :  0.2971   Min.   :   0.000   Min.   :0.1250
1st Qu.: 0.0000    1st Qu.: 0.08613   1st Qu.: 57.7500   1st Qu.:   1.000   1st Qu.:0.4251
Median : 0.0000    Median : 0.18130   Median : 83.7500   Median :   1.000   Median :0.4753
Mean   : 0.4906    Mean   : 0.23357   Mean   : 80.8607   Mean   :   1.822   Mean   :0.4815
3rd Qu.: 1.0000    3rd Qu.: 0.31335   3rd Qu.:104.7549   3rd Qu.:   2.000   3rd Qu.:0.5326
Max.   :11.0000    Max.   :29.75299   Max.   :200.9346   Max.   :1650.000   Max.   :1.0000

Tx COTISATION NET
Min.   :0.0195
1st Qu.:0.2409
Median :0.2722
Mean   :0.2733
3rd Qu.:0.3070
Max.   :0.4847
```

Statistiques des données réelles : entreprises contrôlées en 2009 au titre du travail dissimulé.

```
MARGE BRUTE D'AUTOFINANCEMENT COUVERTURE du BFR COUVERTURE des IMMOS NETTES COUVERTURE du CA
Min.   :-5.1922               Min.   :-1.0600   Min.   :0.160               Min.   :-0.33425
1st Qu.: 0.0313               1st Qu.: 0.5225   1st Qu.:0.880               1st Qu.:-0.05959
Median : 0.1520               Median : 0.8399   Median :1.538               Median : 0.02740
Mean   : 0.3676               Mean   : 0.7813   Mean   :1.822               Mean   : 0.02278
3rd Qu.: 0.4320               3rd Qu.: 1.0700   3rd Qu.:2.306               3rd Qu.: 0.10137
Max.   :32.3559               Max.   : 2.5500   Max.   :4.584               Max.   : 0.39178

 SOLVABILITE      INDEPENDANCE FINANCIERE RENTABILITE ECONOMIQUE RENTABILITE FINANCIERE
Min.   :-0.2000   Min.   :0.2200          Min.   :-0.17000       Min.   :-0.2600
1st Qu.: 0.1200   1st Qu.:0.5404          1st Qu.: 0.03551       1st Qu.: 0.0900
Median : 0.3350   Median :0.6372          Median : 0.08000       Median : 0.2134
Mean   : 0.4357   Mean   :0.6455          Mean   : 0.07510       Mean   : 0.2387
3rd Qu.: 0.6075   3rd Qu.:0.7400          3rd Qu.: 0.12000       3rd Qu.: 0.3305
Max.   : 1.8000   Max.   :0.9900          Max.   : 0.26000       Max.   : 0.9500

RENTABILITE COMMERCIALE CONTRIBUTION DU CAPITAL CONTRIBUTION DE LA VA LIQUIDITE IMMEDIATE
Min.   :-0.130000       Min.   :-0.2100         Min.   :-0.27000      Min.   :0.0500
1st Qu.: 0.001017       1st Qu.: 0.0700         1st Qu.: 0.04000      1st Qu.:0.2800
Median : 0.019390       Median : 0.1521         Median : 0.09000      Median :0.7634
Mean   : 0.013062       Mean   : 0.1699         Mean   : 0.08595      Mean   :0.9070
3rd Qu.: 0.030644       3rd Qu.: 0.2400         3rd Qu.: 0.15000      3rd Qu.:1.1276
Max.   : 0.110000       Max.   : 0.7200         Max.   : 0.35000      Max.   :3.5325

LIQUIDITE GENERALE LIQUIDITE REDUITE  ENDETTEMENT      CAPACITE DE REMBOURSEMENT
Min.   :0.300      Min.   :0.0900    Min.   :-0.1700   Min.   :  2.00
1st Qu.:1.212      1st Qu.:0.5825    1st Qu.: 0.3100   1st Qu.:  5.00
Median :1.830      Median :1.1959    Median : 0.8378   Median : 34.00
Mean   :2.090      Mean   :1.3510    Mean   : 0.8593   Mean   : 67.78
3rd Qu.:2.611      3rd Qu.:1.8508    3rd Qu.: 1.1780   3rd Qu.: 80.00
Max.   :5.321      Max.   :4.1395    Max.   : 2.8200   Max.   :507.00

FINANCEMENT DES STOCKS PRODUCTIVITE DE L'ACTIF  DUREE CLIENT     DUREE FOURNISSEUR
Min.   :0.570          Min.   :0.720           Min.   :0.01644   Min.   :0.03562
1st Qu.:2.862          1st Qu.:1.370           1st Qu.:0.08493   1st Qu.:0.09041
Median :3.540          Median :1.870           Median :0.10137   Median :0.12877
Mean   :3.420          Mean   :1.918           Mean   :0.11004   Mean   :0.13435
3rd Qu.:4.348          3rd Qu.:2.270           3rd Qu.:0.13151   3rd Qu.:0.15890
Max.   :6.020          Max.   :3.920           Max.   :0.27671   Max.   :0.33151

POIDS MASSE SALARIALE  PRODUCTIVITE     PRODUCTIVITE DU TRAVAIL PRODUCTIVITE DU CAPITAL
Min.   :0.5100        Min.   : 0.0014   Min.   : 0.782          Min.   :  0.0076
1st Qu.:0.7600        1st Qu.: 0.4926   1st Qu.: 3.638          1st Qu.:  1.3500
Median :0.8400        Median : 1.2887   Median : 5.559          Median :  1.9430
Mean   :0.8399        Mean   : 3.3483   Mean   : 7.186          Mean   :  8.5173
3rd Qu.:0.9139        3rd Qu.: 3.3494   3rd Qu.: 8.942          3rd Qu.:  3.8066
Max.   :1.2100        Max.   :85.6905   Max.   :51.469          Max.   :776.2540
```

```
DUREE DE VIE       NB ETS       NB ORIG. DEBIT NON RENS. MONTANT DEBIT     NB CTP EXONERATION
Min.   : 2.00   Min.   : 1.000   Min.   : 0.00            Min.   :0.02134   Min.   :  0.000
1st Qu.: 5.00   1st Qu.: 1.000   1st Qu.: 7.00            1st Qu.:0.38133   1st Qu.:  4.000
Median : 7.00   Median : 1.000   Median :13.00            Median :0.49649   Median :  7.000
Mean   :12.52   Mean   : 1.247   Mean   :13.49            Mean   :0.49869   Mean   :  7.941
3rd Qu.:14.00   3rd Qu.: 1.000   3rd Qu.:16.00            3rd Qu.:0.61570   3rd Qu.: 10.000
Max.   :71.00   Max.   :66.000   Max.   :87.00            Max.   :0.88148   Max.   :351.000

NB REMISES SUR MAJORATION MONTANT REMISES SUR MAJORATION MONTANT ECARTS      PENALITES
Min.   : 0.000            Min.   :0.000000              Min.   :0.00000   Min.   :0.0000000
1st Qu.: 0.000            1st Qu.:0.000000              1st Qu.:0.01830   1st Qu.:0.0000000
Median : 0.000            Median :0.000000              Median :0.06118   Median :0.0000000
Mean   : 1.673            Mean   :0.001420              Mean   :0.13790   Mean   :0.0009933
3rd Qu.: 2.000            3rd Qu.:0.001561              3rd Qu.:0.18801   3rd Qu.:0.0004180
Max.   :58.000            Max.   :0.082196              Max.   :3.08437   Max.   :0.0854745

NB RETARDS     NB TAXATIONS D'OFFICE MONTANT TAXATIONS D'OFFICE NB DEMANDES DELAIS
Min.   :  0.00   Min.   : 0.000        Min.   :0.00000            Min.   : 0.000
1st Qu.:  3.00   1st Qu.: 0.000        1st Qu.:0.00000            1st Qu.: 0.000
Median :  6.00   Median : 0.000        Median :0.00000            Median : 0.000
Mean   : 10.12   Mean   : 1.631        Mean   :0.05070            Mean   : 0.769
3rd Qu.: 12.00   3rd Qu.: 2.000        3rd Qu.:0.03939            3rd Qu.: 1.000
Max.   :209.00   Max.   :65.000        Max.   :1.61058            Max.   :13.000

NB DELAIS ACCEPTES % VERSEMENT DEMAT.   DERNIER CTRL        NB CCA           COMPLIANCE
Min.   : 0.0000    Min.   : 0.000000   Min.   :  3.00   Min.   :   0.000   Min.   :0.1250
1st Qu.: 0.0000    1st Qu.: 0.000000   1st Qu.: 11.00   1st Qu.:   0.000   1st Qu.:0.3725
Median : 0.0000    Median : 0.003738   Median : 84.00   Median :   0.000   Median :0.4750
Mean   : 0.2543    Mean   : 0.283743   Mean   : 72.17   Mean   :   4.688   Mean   :0.4789
3rd Qu.: 0.0000    3rd Qu.: 0.320025   3rd Qu.:111.00   3rd Qu.:   2.000   3rd Qu.:0.5850
Max.   :11.0000    Max.   :29.752993   Max.   :111.00   Max.   :1650.000   Max.   :1.0000

Tx COTISATION NET
Min.   :0.0195
1st Qu.:0.2664
Median :0.3466
Mean   :0.3005
3rd Qu.:0.3600
Max.   :0.4292
```

Données réelles : régression entre le niveau de fraude et l'ensemble des variables.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.50926 -0.03170 -0.00529  0.01999  0.34723

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.345e-01  3.642e-02   3.694 0.000236 ***
MARGE BRUTE D'AUTOFINANCEMENT   1.381e-02  1.820e-03   7.587 9.48e-14 ***
COUVERTURE du BFR               3.690e-04  3.821e-03   0.097 0.923091
COUVERTURE des IMMOS NETTES    -3.360e-03  2.659e-03  -1.264 0.206694
COUVERTURE du CA                2.460e-03  2.511e-02   0.098 0.922010
SOLVABILITE                    -1.357e-03  7.067e-03  -0.192 0.847758
INDEPENDANCE FINANCIERE        -3.025e-02  1.772e-02  -1.707 0.088250 .
RENTABILITE ECONOMIQUE         -2.214e-02  4.056e-02  -0.546 0.585272
RENTABILITE FINANCIERE         -9.799e-03  1.318e-02  -0.743 0.457480
RENTABILITE COMMERCIALE         3.567e-03  7.208e-02   0.049 0.960542
CONTRIBUTION DU CAPITAL        -3.201e-03  1.910e-02  -0.168 0.866934
CONTRIBUTION DE LA VA          -7.880e-03  3.281e-02  -0.240 0.810294
LIQUIDITE IMMEDIATE             3.371e-03  3.720e-03   0.906 0.365161
LIQUIDITE GENERALE              3.598e-03  2.907e-03   1.238 0.216232
LIQUIDITE REDUITE              -2.275e-03  3.985e-03  -0.571 0.568152
ENDETTEMENT                    -7.979e-05  4.393e-03  -0.018 0.985512
CAPACITE DE REMBOURSEMENT      -4.782e-05  2.845e-05  -1.681 0.093231 .
FINANCEMENT DES STOCKS         -4.891e-03  2.035e-03  -2.404 0.016467 *
PRODUCTIVITE DE L'ACTIF         9.076e-03  3.571e-03   2.541 0.011245 *
DUREE CLIENT                    1.027e-01  4.800e-02   2.139 0.032733 *
DUREE FOURNISSEUR              -2.945e-02  4.039e-02  -0.729 0.466032
POIDS MASSE SALARIALE          -1.774e-02  2.632e-02  -0.674 0.500594
PRODUCTIVITE                    2.343e-03  5.636e-04   4.158 3.57e-05 ***
PRODUCTIVITE DU TRAVAIL         3.784e-04  5.522e-04   0.685 0.493372
PRODUCTIVITE DU CAPITAL        -1.245e-05  5.585e-05  -0.223 0.823627

DUREE DE VIE                   -5.495e-05  2.432e-04  -0.226 0.821287
NB ETS                         -1.710e-03  3.585e-03  -0.477 0.633578
NB ORIG. DEBIT NON RENS.        1.767e-04  2.916e-04   0.606 0.544696
MONTANT DEBIT                  -5.153e-02  2.036e-02  -2.531 0.011566 *
NB CTP EXONERATION             -6.446e-04  7.490e-04  -0.861 0.389762
NB REMISES SUR MAJORATION       1.972e-03  1.144e-03   1.723 0.085263 .
MONTANT REMISES SUR MAJORATION -2.350e+00  9.732e-01  -2.415 0.015975 *
MONTANT ECARTS                  5.044e-01  1.838e-02  27.446  < 2e-16 ***
PENALITES                      -1.385e+00  5.182e-01  -2.672 0.007693 **
NB RETARDS                     -1.448e-03  3.557e-04  -4.071 5.16e-05 ***
NB TAXATIONS D'OFFICE          -4.169e-04  1.030e-03  -0.405 0.685735
MONTANT TAXATIONS D'OFFICE     -3.224e-01  3.562e-02  -9.050  < 2e-16 ***
NB DEMANDES DELAIS             -6.893e-03  2.108e-03  -3.269 0.001127 **
NB DELAIS ACCEPTES              2.520e-03  3.890e-03   0.648 0.517245
% VERSEMENT DEMAT.              2.671e-03  8.002e-03   0.334 0.738618
DERNIER CTRL                    1.884e-04  7.789e-05   2.419 0.015810 *
NB CCA                          2.939e-04  1.481e-04   1.984 0.047638 *
COMPLIANCE                     -3.498e-02  2.427e-02  -1.441 0.149863
Tx COTISATION NET              -1.530e-01  3.912e-02  -3.911 0.000100 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06791 on 770 degrees of freedom
Multiple R-squared: 0.6487,     Adjusted R-squared: 0.6291
F-statistic: 33.06 on 43 and 770 DF,  p-value: < 2.2e-16
```

Données synthétiques + réelles: régression entre le niveau de fraude et l'ensemble des variables

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.48874 -0.01131 -0.00197  0.00736  0.68625

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     6.113e-02  5.474e-03  11.166  < 2e-16 ***
MARGE BRUTE D'AUTOFINANCEMENT   5.533e-03  6.052e-04   9.143  < 2e-16 ***
COUVERTURE du BFR              -2.199e-03  4.849e-04  -4.535 5.82e-06 ***
COUVERTURE des IMMOS NETTES    -1.607e-03  4.446e-04  -3.613 0.000304 ***
COUVERTURE du CA               -1.714e-02  3.266e-03  -5.248 1.57e-07 ***
SOLVABILITE                    -1.069e-02  1.028e-03 -10.402  < 2e-16 ***
INDEPENDANCE FINANCIERE         6.434e-04  2.561e-03   0.251 0.801678
RENTABILITE ECONOMIQUE         -7.114e-03  5.887e-03  -1.208 0.226937
RENTABILITE FINANCIERE          2.606e-03  2.080e-03   1.253 0.210354
RENTABILITE COMMERCIALE        -1.580e-02  8.692e-03  -1.817 0.069209 .
CONTRIBUTION DU CAPITAL         1.598e-02  2.797e-03   5.715 1.13e-08 ***
CONTRIBUTION DE LA VA           1.431e-02  4.002e-03   3.577 0.000349 ***
LIQUIDITE IMMEDIATE            -3.880e-03  5.135e-04  -7.556 4.53e-14 ***
LIQUIDITE GENERALE             -1.456e-03  3.725e-04  -3.909 9.32e-05 ***
LIQUIDITE REDUITE              -1.576e-03  4.910e-04  -3.210 0.001330 **
ENDETTEMENT                    -5.666e-04  5.702e-04  -0.994 0.320441
CAPACITE DE REMBOURSEMENT      -1.609e-05  7.335e-06  -2.194 0.028291 *
FINANCEMENT DES STOCKS         -4.461e-05  3.261e-04  -0.137 0.891203
PRODUCTIVITE DE L'ACTIF         1.384e-03  5.791e-04   2.389 0.016900 *
DUREE CLIENT                    2.292e-02  8.395e-03   2.730 0.006342 **
DUREE FOURNISSEUR              -1.141e-02  7.250e-03  -1.574 0.115540
POIDS MASSE SALARIALE          -8.520e-03  3.419e-03  -2.492 0.012710 *
PRODUCTIVITE                    5.519e-03  1.822e-04  30.290  < 2e-16 ***
PRODUCTIVITE DU TRAVAIL        -7.678e-04  5.870e-05 -13.079  < 2e-16 ***
PRODUCTIVITE DU CAPITAL         8.668e-06  2.224e-05   0.390 0.696744

DUREE DE VIE                   -2.864e-04  4.928e-05  -5.811 6.40e-09 ***
NB ETS                          1.217e-04  7.353e-04   0.166 0.868525
NB ORIG. DEBIT NON RENS.       -1.843e-04  6.177e-05  -2.983 0.002858 **
MONTANT DEBIT                   3.934e-04  4.182e-05   9.406  < 2e-16 ***
NB CTP EXONERATION             -4.821e-04  1.054e-04  -4.574 4.85e-06 ***
NB REMISES SUR MAJORATION       9.951e-04  2.877e-04   3.459 0.000543 ***
MONTANT REMISES SUR MAJORATION -1.114e+00  2.455e-01  -4.537 5.78e-06 ***
MONTANT ECARTS                  1.029e-01  3.287e-03  31.297  < 2e-16 ***
PENALITES                       7.801e-01  1.980e-01   3.940 8.19e-05 ***
NB RETARDS                      2.703e-04  7.064e-05   3.826 0.000131 ***
NB TAXATIONS D'OFFICE           7.759e-04  2.421e-04   3.205 0.001353 **
MONTANT TAXATIONS D'OFFICE     -6.695e-02  2.957e-03 -22.641  < 2e-16 ***
NB DEMANDES DELAIS             -2.176e-03  5.437e-04  -4.001 6.34e-05 ***
NB DELAIS ACCEPTES             -2.831e-04  4.888e-04  -0.579 0.562474
% VERSEMENT DEMAT.             -1.558e-03  1.506e-03  -1.035 0.300803
DERNIER CTRL                    3.746e-05  8.877e-06   4.219 2.47e-05 ***
NB CCA                          3.397e-05  4.068e-05   0.835 0.403739
COMPLIANCE                      8.361e-03  3.546e-03   2.358 0.018408 *
Tx COTISATION NET              -1.651e-02  6.007e-03  -2.748 0.006002 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02938 on 9956 degrees of freedom
Multiple R-squared: 0.3908,     Adjusted R-squared: 0.3882
F-statistic: 148.5 on 43 and 9956 DF,  p-value: < 2.2e-16
```

Chapitre 5

Résultats

5.1 Introduction

Dans ce dernier chapitre, nous introduisons l'ensemble des objets qui permettent de détecter et d'évaluer (en unités monétaires) les irrégularités aux cotisations sociales. La connexion entre outils théoriques et opérationnels prend une place importante dans les lignes qui suivent et nous indiquons tous les choix qui nous ont mené aux résultats proposés. L'algorithme développé dans le troisième chapitre est le moteur essentiel de la détection et de l'évaluation des montants de redressements. Nous commençons par résumer les données utilisées pour rendre opérationnelle la détection. Puis, nous introduisons les mesures statistiques et les propriétés théoriques qui constituent le socle de la pertinence de n'importe quel modèle. Nous résumons ensuite les forêts uniformément aléatoires à travers leur principe, les propriétés théoriques qu'elles proposent et les mécanismes algorithmiques offerts. Nous insistons en particulier sur la transition entre garanties théoriques, validation empirique et garanties opérationnelles. Nous terminons ce chapitre par les résultats réels obtenus par le modèle et l'évaluation de l'ensemble des déclarations de cotisations des entreprises d'Île-de-France pour l'année 2013. Le lecteur pressé peut se rendre directement à cette partie, considérée comme le résultat essentiel de cette thèse : *les irrégularités aux cotisations sociales sont, en majorité, involontaires. Elles sont en bien plus grand nombre que ce qui est, à ce jour, observé et représentent des sommes beaucoup plus importantes.*

5.2 Les données d'apprentissage et de test

Comme nous l'avons indiqué dans le second chapitre, le contrôle d'une entreprise n'est possible que sur les trois dernières années de cotisations ainsi que sur celles de l'année en cours au moment du contrôle. Par exemple, pour les entreprises qui seront contrôlées en 2014, seules les cotisations des années 2011, 2012, 2013 et celles en cours en 2014, pourront être vérifiées. Ce principe est le même chaque année et, toujours dans le second chapitre, nous avons indiqué le processus de récupération des données.

Pour plus de clarté, nous considérons maintenant les années 2010 à 2013 dans tout le reste du document. Notre objectif ultime est une évaluation complète des déclarations de

cotisation de toutes les entreprises d'Île-de-France, pour les contrôles à réaliser en 2013. Les années 2010 à 2012 sont celles dont les déclarations auront (auraient) été contrôlées. Cependant, au moment où le modèle doit effectuer l'évaluation (à la fin 2012), seules les déclarations de 2010 et 2011 sont disponibles. Lorsque celles de 2012 le deviennent, dans le courant de l'année qui suit, le modèle peut, à nouveau, être actualisé.

Avant d'évaluer les entreprises, il nous faut d'abord savoir dans quelle mesure le modèle est capable *d'apprendre* la relation liant les irrégularités (par leur nombre et leurs montants) aux déclarations. Il nous faut également pouvoir donner des garanties sur la qualité de cet apprentissage. Pour effectuer ces deux tâches, nous avons alors besoin des contrôles (et des résultats) effectués par l'URSSAF en 2010 et en 2011 sur les déclarations de ces mêmes années. La procédure est la suivante :
i) les déclarations et les résultats des contrôles sont divisés en deux sous-échantillons aléatoires ;
ii) le premier sous-échantillon est *l'échantillon d'apprentissage*, à partir duquel le modèle est construit ;
iii) une fois l'apprentissage effectué, nous soumettons au modèle, pour évaluation, le second sous-échantillon, *l'échantillon de test*, et comparons les résultats du modèle à ceux des contrôles. Dans l'échantillon de test, le modèle n'a jamais connaissance des résultats des contrôles. Ce procédé est équivalent à une évaluation opérationnelle d'entreprises ; la seule différence est leur nombre, plus petit que dans une situation réelle.

Une fois ces étapes franchies, nous devons alors fournir les garanties sur la capacité de généralisation du modèle, en particulier lorsque le nombre d'entreprises à évaluer devient important.
- Nous souhaiterions disposer d'échantillons d'apprentissage et de test aussi grands que possible. Pour cela, nous pouvons par exemple nous intéresser aux contrôles des années précédentes. Cela n'est, en général, pas possible. Les déclarations des années précédant l'année 2010 ne sont pas contrôlées en 2013 et bien que nous disposions des résultats des contrôles effectués alors, il n'y a pas d'assurance que la relation entre ces déclarations et les résultats n'ait pas évolué. Le risque pris est une obsolescence du modèle par une présence trop importante de données dont la distribution n'aurait pas grand chose en commun avec celle que nous souhaitons évaluer. Nous verrons par la suite comment lever cette contrainte.
- Dans chaque échantillon, nous disposons donc d'entreprises contrôlées et redressées et d'entreprises contrôlées et non redressées. Le modèle doit être capable de distinguer les entreprises pour lesquelles la probabilité d'existence d'une irrégularité est faible et celles pour lesquelles elle est élevée.

`Les irrégularités et le montant net redressé`

Une fois les deux échantillons constitués, nous distinguons les irrégularités en utilisant le seuil défini dans le chapitre précédent. En d'autres termes, du point de vue du modèle, une irrégularité existe pour une entreprise, si le montant net redressé lors du contrôle est supérieur à 0.9% de la masse salariale des années contrôlées, à condition que cette dernière ne soit pas trop importante. D'autres règles s'ajoutent. Par exemple, les montants redressés

ne doivent pas être trop faibles. La détermination exacte d'une irrégularité est effectuée automatiquement par l'algorithme de traitement générique (présenté dans le second chapitre) lequel a l'objectif d'assurer au moteur de détection la meilleure information. Cela se traduit notamment par une reconnaissance complète de tous les motifs de redressement d'une entreprise contrôlée et par une notation qui spécifie le degré d'importance d'une irrégularité. Présentons un exemple. Dans de nombreux cas, une entreprise contrôlée et redressée l'est pour plusieurs motifs (relatifs aux différentes catégories de cotisation) qui conduisent, pour chacun, à une somme redressée spécifique. Par exemple, une entreprise pourra être redressée (d'un montant en faveur de l'URSSAF) pour une irrégularité sur la CSG, sur les heures supplémentaires et, dans le même temps, redressée d'un montant en sa faveur pour une mesure de réduction sous-estimée. Dans cet exemple, il y a alors trois motifs de redressement équivalant à trois montants dont un en faveur de l'entreprise. Leur somme donne le montant net (comptable) redressé. Pour qu'une irrégularité existe, il faut alors que ce montant net soit en faveur de l'URSSAF. Dans la pratique, 40% des déclarations de cotisation redressées en 2011 (et correspondant donc chacune à un ou plusieurs motifs de redressements) l'ont été pour un montant (par déclaration) inférieur ou égal à 1000 euros. Or, le modèle doit également être capable d'estimer le montant net redressé, pour chaque entreprise dont la probabilité d'existence d'une irrégularité serait plus grande que 0.5.

`Variables économiques et effet de seuil`

Lorsqu'une irrégularité s'avère être une fraude, l'analyse empirique menée dans le chapitre précédent, montre que la situation financière des entreprises est un critère explicatif important. Si près de la moitié des redressements le sont pour des sommes relativement faibles, alors l'hypothèse la plus simple est la présence de nombreuses erreurs dans les déclarations des entreprises. L'analyse, à travers les variables économiques, illustre une propriété empirique importante : pour un montant total de redressements quasi-similaire, un modèle qui explique les cas de fraude les plus importants sera plus précis dans la détection de nouveaux cas qu'un autre, qui voudrait uniquement détecter toute la fraude. L'application de ce résultat à nos données est alors motivé par deux éléments :
- toutes les entreprises ne peuvent être contrôlées et le modèle ne doit pas rendre la tâche plus complexe aux inspecteurs du contrôle ;
- en éliminant les montants de redressements trop faibles, la recherche des montants plus significatifs est rendue plus simple.
Expliciter la propriété énoncée précédemment revient alors à définir un seuil au-delà duquel l'irrégularité existe du point de vue du modèle. L'algorithme de traitement générique l'intègre à son processus de décision et fournit les échantillons d'apprentissage et de test.

Nous faisons, maintenant, appel à la notation introduite dans le premier chapitre. Une irrégularité Y est une variable aléatoire à valeurs dans $\{0, 1\}$. Pour une entreprise $i, 1 \leq i \leq n$, où n est le nombre d'entreprises disponibles, une irrégularité existe si $Y_i = 1$. Les déclarations de cotisations des entreprises sont représentées par le vecteur aléatoire X, à valeurs dans $\mathbb{R}^d$, où d est le nombre de variables disponibles. Notre échantillon d'apprentissage (pour la classification des irrégularités) est noté $D_n^{A_Y}$, et défini par $D_n^{A_Y} = \{(X_i, Y_i), 1 \leq i \leq n\}$. De la même manière, nous notons l'échantillon de test $D_n^{V_Y}$.

Le montant net redressé est représenté par une variable aléatoire, notée R, à valeurs dans $\mathbb{R}$. Un montant net redressé peut être positif, nul ou négatif (s'il est en faveur de l'entreprise). Pour une entreprise i, si $Y_i = 0$, alors $R_i \leq 0$. Sinon, $R_i > 0$. Notre échantillon d'apprentissage (pour l'estimation du montant de chaque irrégularité) est noté $D_n^{A_R}$, et défini par $D_n^{A_R} = \{(X_i, R_i), 1 \leq i \leq n\}$. De la même manière, nous notons l'échantillon de test $D_n^{V_R}$.

Formellement, nous souhaitons donc, pour les irrégularités, estimer la fonction

$$\eta(x) = \mathbf{P}(Y = 1|X = x),$$

où x est une déclaration de cotisations observée.
Nous souhaitons également estimer l'espérance conditionnelle du montant de redressement, $\mathbf{E}(R|(X, Y))$.

Pour une partie de l'*évaluation des performances* de l'algorithme, l'échantillon d'apprentissage comprend 406 entreprises (10% du total) et l'échantillon de test 3663 entreprises. 1065 variables sont retenues.

5.3 Les données de l'expérimentation opérationnelle et de l'évaluation complète

La *validation des résultats* passe généralement par une expérimentation opérationnelle. Pour cette dernière, le modèle effectue ses recommandations pour toutes les entreprises, en fournissant un certain nombre d'informations permettant une décision immédiate. Parmi elles, la probabilité d'existence d'une irrégularité dans la déclaration de cotisations de l'entreprise est l'élément fondamental.

Lorsque le modèle doit être expérimenté dans l'année courante, l'échantillon d'apprentissage correspond aux données et résultats de tous les contrôles portant sur les trois dernières années de cotisation des entreprises dont les déclarations ont été vérifiées.

- La temporalité est importante car seules les déclarations des trois dernières années avant le moment d'un contrôle peuvent être vérifiées. Il faut donc veiller à ce qu'une expérimentation tienne compte des obligations légales régissant le contrôle des cotisations. Pour l'expérimentation opérationnelle, l'échantillon d'apprentissage comprend 12155 entreprises et 1065 variables. 40% des entreprises présentent une irrégularité dans leur déclaration. L'échantillon de test correspond aux entreprises qui seront effectivement contrôlées par les inspecteurs de l'URSSAF, sur la base des recommandations du modèle. Pour valider les résultats de l'algorithme, 500 entreprises ont été recommandées, pour l'année 2012, sur l'unique base de la probabilité d'existence d'une irrégularité dans la déclaration de cotisations de chacune d'elles. 167 entreprises ont été effectivement contrôlées par les inspecteurs de l'URSSAF.

- L'évaluation complète correspond à l'*industrialisation de l'algorithme*. L'apprentissage est effectué de la même manière que dans l'expérimentation. Les recommandations sont

cependant beaucoup plus complètes et, pour chaque entreprise, un peu moins de 20 indicateurs sont fournis. Les inspecteurs du contrôle peuvent alors décider, de la manière dont ils le souhaitent, la priorisation des contrôles à effectuer. L'industrialisation du moteur de détection est traduite de manière explicite en quelques phases : l'algorithme de traitement générique prend en entrée les bases de données et les transforme en matrice de travail pour l'algorithme de détection. Ce dernier effectue l'apprentissage puis évalue toutes les entreprises. Une synthèse est effectuée par le moteur de détection et les recommandations du modèle sont stockées et peuvent être lues par un tableur. Le moteur de détection s'auto-actualise et sauvegarde l'ensemble des informations. L'ensemble des opérations est effectué avec un (seul) logiciel, libre et gratuit, *R*. Le matériel requis est une unique station de travail avec suffisamment de mémoire centrale. L'échantillon d'apprentissage est constitué d'au moins 4069 entreprises contrôlées pour le compte des années 2010 et/ou 2011. Nous notons que la taille de l'échantillon d'apprentissage est relative en cas d'apprentissage incrémental. Il n'y a pas d'échantillon de test puisque toute la population des entreprises est évaluée par l'algorithme, soit les 311 341 entreprises d'Île-de-France qui ont effectué des déclarations de cotisation en 2010 et/ou en 2011.

5.4 Mesures empiriques, définitions et propriétés

Afin d'évaluer les performances de notre modèle et pour une compréhension universelle des résultats, nous fournissons ici des indicateurs statistiques dont le principal intérêt est la visualisation et le résumé de l'information fournie par un modèle. Afin d'éclairer le lecteur, les indicateurs présentés, classiquement utilisés dans les problèmes de classification binaire, sont abordés dans le cadre de la détection des irrégularités aux cotisations sociales. Nous supposons que la règle de décision d'un modèle (classifieur) pour l'évaluation de la présence ou de l'absence d'une irrégularité repose sur l'estimation de la fonction $\eta(x)$. Lorsque l'estimation donne une probabilité supérieure à un seuil, le modèle décide que l'irrégularité est présente.

`Classification, matrice de confusion, précision, sensibilité`

Afin de mesurer la qualité d'un modèle, de nombreux indicateurs sont généralement disponibles et, selon l'objectif de l'analyse, permettent de distinguer différents modèles. Par convention, les irrégularités sont définies comme les cas positifs. Nous rappelons donc notre objectif : *détecter avec la plus grande exactitude le caractère régulier ou irrégulier de la déclaration de cotisations de n'importe quelle entreprise et, le cas échéant, estimer avec la plus grande exactitude le montant de redressement espéré.*
Parallèlement, il nous faut résumer les résultats au regard de la politique de contrôle de l'URSSAF. Dans le cadre de la détection, le point le plus important est la capacité à ne commettre que peu d'erreurs lorsqu'on indique l'existence d'une irrégularité, quitte à ne pas toutes les détecter. Cette capacité est appelée *précision* ; elle correspond à la fréquence des redressements positifs de l'URSSAF. Avant de donner, à nouveau, sa définition, nous indiquons la manière dont sont habituellement présentés les résultats d'une classification binaire ; ici, les classes correspondent à l'absence (classe 0) ou la présence d'une irrégularité (classe 1). La visualisation des résultats est, généralement, associée à

l'échantillon de test, de sorte qu'il soit possible d'apprécier rapidement la qualité du modèle. L'information primaire est fournie par la *matrice de confusion*. Elle est constituée de deux lignes et deux colonnes et nous en présentons un exemple ci-dessous :

		Classe réelle	
		0	1
Classe estimée	0	VN	FN
	1	FP	VP

TABLE 5.1 – Exemple générique d'une matrice de confusion.

La matrice de confusion correspond au croisement entre le nombre de résultats des classes estimées par le modèle et celui des classes réelles. L'absence d'irrégularité correspond à la classe 0 et la présence d'irrégularité à la classe 1.
VN correspond aux *vrais négatifs*, soit le nombre de situations pour lequel le modèle estime qu'il n'y a pas d'irrégularités dans les déclarations de cotisation, et pour lequel cela est effectivement le cas.
FN correspond aux *faux négatifs*, soit le nombre de situations pour lequel le modèle estime qu'il n'y a pas d'irrégularité, et pour lequel cela n'est pas le cas. Ce sont les situations considérées par le modèle, à tort, comme ne présentant pas d'irrégularités.
VP correspond aux *vrais positifs*, soit le nombre de situations pour lequel le modèle estime qu'il y a des irrégularités dans les déclarations de cotisation, et pour lequel cela est effectivement le cas.
FP correspond aux *faux positifs*, soit le nombre de situations pour lequel le modèle estime, à tort, qu'il y a des irrégularités dans les déclarations de cotisation et pour lequel ce n'est pas le cas.

Toutefois, ces différentes valeurs sont présentées en valeur absolue (nombre) et ne rendent pas compte clairement de la classification lorsque la taille de l'échantillon de test augmente. Des indicateurs, en valeur relative, sont alors définis et permettent de comparer différents modèles pour un même échantillon de test. Nous nous intéressons d'abord à la *précision*, notée ici P_r, dont nous rappelons la définition :

$$P_r = \frac{VP}{VP + FP}.$$

La précision est équivalente à la *fréquence des redressements positifs* de l'URSSAF et répond à la question suivante : *avec quel niveau d'exactitude le modèle est-il capable de détecter les irrégularités effectives parmi les cotisations dont il estime qu'elles en présentent ?*

La valeur de la *précision*, comprise entre 0 et 1, donne ce niveau. Par exemple, une précision de 0.5 signifie que sur l'échantillon de test, seule la moitié des irrégularités estimées par le modèle le seront effectivement. Supposons qu'il y ait 100 irrégularités et que le modèle estime qu'il n'y en a que 10, alors seulement 5 seront détectées si les recommandations du modèle sont suivies. Ce dernier doit donc être capable, en plus d'une

grande précision, de savoir déterminer l'existence d'un nombre suffisant d'irrégularités, s'il y en a. Pour cela, on calcule la *sensibilité* (appelée aussi rappel) que nous notons S_e,

$$S_e = \frac{VP}{VP + FN}.$$

La sensibilité répond à la question suivante : *quelle est la capacité relative du modèle à détecter toutes les irrégularités aux cotisations sociales (en l'absence de prise en compte des fausses alarmes) ?*

La valeur de la *sensibilité*, comprise entre 0 et 1, donne cette capacité. Par exemple, sur les 100 irrégularités supposées précédemment, si le modèle a une sensibilité de 0.5, il en détectera 50. La question centrale sera alors de savoir combien d'entreprises il faudra contrôler avant que ces 50 cas ne soient trouvés. Précisons ce point de vue. Pour trouver toutes les irrégularités, il suffit d'affirmer que l'ensemble des déclarations de toutes les entreprises présentent une irrégularité. La sensibilité est alors de 1. Mais pour l'atteindre, il faudrait contrôler toutes les entreprises, ce qui n'est pas envisageable. La sensibilité, seule, ne suffit donc pas et doit toujours être accompagnée, au minimum, du *taux de faux-positifs*, noté FPR, soit le rapport entre le nombre d'irrégularités détectées à tort par le modèle et le nombre total de cas ne présentant aucune irrégularité. Il est donné par :

$$FPR = \frac{FP}{FP + VN},$$

et répond à la question suivante : *quel est le niveau de fausses alarmes du modèle ?*

Plus simplement, un modèle qui sait détecter un grand nombre d'irrégularités, mais avec de nombreuses fausses alarmes, n'a pas d'intérêt. Cela est équivalent à disposer d'un système de sécurité qui se déclencherait de manière intempestive. Le taux de faux-positifs doit donc demeurer aussi bas que possible.

Nous avons donc à notre disposition trois indicateurs, complémentaires, *la précision*, *la sensibilité* et *le taux de faux-positifs*. Lorsqu'un modèle doit être expérimenté pour la détection de nouvelles irrégularités, la *précision* est le critère retenu puisque, par construction, seuls les cas positifs (les irrégularités) nous intéressent alors.

`Classification, visualisation, AUC, F-score`

Une réponse plus globale aux deux questions posées par la précision et la sensibilité est fournie par l'association de ces dernières à des indicateurs encore plus synthétiques.
- Le premier est la *courbe ROC* ("Receiver Operating Characteristic"), représentée sur un plan dont les ordonnées sont les valeurs prises par la *sensibilité* et les abscisses, les valeurs prises par le *taux de faux-positifs*. La courbe ROC répond à la question suivante : *comment évolue la capacité du modèle à détecter les irrégularités lorsqu'on fait varier le seuil différenciant l'absence d'irrégularité de la présence d'irrégularité ?*
Par exemple, un classifieur probabiliste supposera qu'une irrégularité est présente si la probabilité d'observation de cette dernière est supérieure à 0.5 (le seuil). La courbe ROC permet de visualiser la performance du classifieur pour toutes les valeurs possibles de ce seuil. Elle permet également la comparaison de plusieurs modèles.

On lui associe, généralement, un indicateur défini par l'aire sous la courbe ROC, appelée encore *AUC* ("Area Under (ROC) Curve"). L'AUC renvoie une valeur entre 0 et 1 et répond à la question suivante : *quelle est la probabilité pour que le modèle place un cas effectivement positif (une irrégularité) devant un cas négatif (une absence d'irrégularité) lorsqu'une décision doit être prise ?*

Considérons un échantillon d'apprentissage $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$ et un classifieur associé, g. On récrit la *sensibilité* et le *taux de faux-positifs* en fonction de n :

$$S_e(n) = \frac{\sum_{i=1}^{n} \mathbf{I}_{\{g(X_i)=1\}} \mathbf{I}_{\{Y_i=1\}}}{\sum_{i=1}^{n} \mathbf{I}_{\{Y_i=1\}}}, \quad FPR(n) = \frac{\sum_{i=1}^{n} \mathbf{I}_{\{g(X_i)=1\}} \mathbf{I}_{\{Y_i=0\}}}{\sum_{i=1}^{n} \mathbf{I}_{\{Y_i=0\}}}.$$

Pour exprimer l'AUC, nous reformulons la définition de Bradley (1997) :

$$AUC = 1 - \sum_{j} \left\{ (1 - S_e(j)) \Delta FPR + \frac{1}{2} \Delta S_e \Delta FPR \right\},$$

avec $\Delta FPR = FPR(j) - FPR(j-1)$ et $\Delta S_e = S_e(j) - S_e(j-1)$.
Au-delà de la n-ème observation, on prolonge généralement la courbe ROC jusqu'au point $(1, 1)$ et le calcul de l'AUC suppose alors $FPR(n+1) = S_e(n+1) = 1$.

Par exemple, un modèle avec une AUC de 1, ne fait jamais d'erreurs. Il classe tous les cas positifs avec une probabilité égale à 1. L'AUC correspond, en quelque sorte, à la marge de sécurité du modèle lorsqu'il détecte une irrégularité. Un modèle avec une AUC de 0.5 place avec la même probabilité (0.5) un cas effectivement positif devant un cas négatif. Les déclarations de cotisation sont, dans ce cas, classées au hasard.

- La deuxième manière de visualiser les performances d'un modèle est la *courbe de précision-rappel*. Elle transcrit la manière dont décroît la précision du modèle lorsque sa sensibilité (le rappel) augmente. Si on souhaite, par exemple, détecter l'ensemble des irrégularités, la sensibilité tendra vers 1 à mesure que le nombre de contrôles se rapprochera du nombre total d'entreprises. En contrepartie, il sera difficile de maintenir le même niveau de précision, car du point de vue du modèle, une irrégularité est présente (ou existe) avec une certaine probabilité qu'il s'agit d'estimer. Lorsque cette dernière est supérieure à un seuil (par exemple 0.5), le modèle décide que l'irrégularité existe. Par construction, il existe alors probablement des entreprises pour lesquelles la probabilité estimée mène, à tort, à l'existence d'une irrégularité ; ainsi la courbe de précision-rappel est reliée à la seule présence d'irrégularités et plus on veut en détecter, plus les fausses alarmes risquent d'être nombreuses, à moins de maintenir une précision suffisante. On associe, généralement, un autre indicateur synthétique à la courbe de précision-rappel que l'on appelle *F-score*, ou *F-mesure*, et noté F_β. Le F-score mesure, à la fois, la précision et la capacité (sensibilité) d'un modèle à détecter les irrégularités. Il est défini par :

$$F_\beta = (1 + \beta^2) \frac{P_r \times S_e}{\beta^2 P_r + S_e}, \quad avec\ \beta \geq 0,$$

où β est le paramètre correspondant au niveau de priorité accordé à la sensibilité ou à la précision. Par exemple, si $\beta = 1$, alors la sensibilité et la précision ont la même importance. Si $\beta = 0.5$, alors la précision est plus importante que la sensibilité.

Le F-score répond à la question suivante : *quelle est la capacité du modèle à détecter un grand nombre d'irrégularités avec une grande exactitude ?*

Dès que l'on souhaite généraliser un modèle, autrement dit l'utiliser pour détecter le plus grand nombre d'irrégularités avec la plus grande précision, le F-score est un critère intéressant et permet de comparer plusieurs modèles admissibles. Comme mesure du *F-score*, nous utilisons $F_{0.5}$ de façon à maintenir la cohérence avec la précision et ajoutons à cette dernière, l'AUC comme dernier indicateur, dans les résultats que nous indiquons. Nous illustrons également les capacités du modèle par sa *courbe ROC* et sa *courbe de précision-rappel*.

Les indicateurs définis soulèvent plusieurs points :
i) nous n'avons à disposition qu'un échantillon comme support des outils commentés, alors que notre objectif est de fournir un point de vue sur l'ensemble des entreprises ;
ii) la plupart des indicateurs mettent l'accent sur les irrégularités, alors que les déclarations de cotisations sans irrégularités sont majoritaires et devraient donc être analysées ;
iii) en faisant varier, à la hausse, la probabilité au-dessus de laquelle une irrégularité est déclarée présente, la précision augmente, ce qui a un intérêt lorsque le nombre de contrôles à effectuer est limité. Les performances d'un modèle dépendent alors directement du nombre de ces contrôles et les indicateurs définis ici doivent être relativisés.

Nous faisons donc face à deux problématiques qui lient la modélisation et l'expérimentation opérationnelle :
- les indicateurs sont-ils valides pour l'application d'un modèle à la totalité des entreprises ?
- Comment les adapter lors d'une expérimentation sur un petit nombre d'entreprises ?

Pour y répondre, il nous faut nous intéresser à l'*erreur de test* et à sa contrepartie théorique l'*erreur de prédiction* ou *erreur de généralisation*. L'analyse de cette dernière est fondamentale et permet de s'affranchir des limites d'un échantillon pour donner un point de vue sur la population entière des entreprises.

`Erreur de test`

Dans le cadre de la détection des irrégularités aux cotisations sociales, un modèle recommande les cas dont il estime qu'ils présentent tous une irrégularité dans leur déclaration. Pour cela, il lui faut, néanmoins, posséder la capacité de distinguer les cas positifs des cas négatifs. Cette aptitude est appelée *capacité de généralisation* et se mesure à l'aide de *l'erreur de test*, définie par :

$$erreur\ de\ test = 1 - \frac{VP + VN}{VP + FP + VN + FN} = 1 - \frac{VP + VN}{n}.$$

Elle répond à la question suivante : *quel est le niveau d'erreur commis par le modèle lorsqu'il estime l'absence ou la présence d'irrégularités, dans les déclarations de cotisations de toutes les entreprises ?*

Plus l'erreur de test (comprise entre 0 et 1) est petite, plus la capacité de généralisation est importante et plus le modèle sait faire la distinction entre présence et absence d'irrégularités. *L'erreur de test est un indicateur global des capacités d'un modèle dans la pratique. Elle ne dit pas comment est répartie cette capacité.* Un modèle peut avoir une grande capacité à détecter la présence d'irrégularités et une mauvaise à détecter leur absence. Si les irrégularités sont largement majoritaires, l'erreur de test sera, alors, petite. Pour cette raison, il nous faut disposer de propriétés théoriques fortes assurant une erreur de test la plus petite possible, quelle que soit la répartition entre présence et absence d'irrégularités. Si nous en disposons, alors nous pouvons contrôler cette répartition.

L'élément essentiel de l'erreur de test est sa dépendance à n, le nombre d'observations (d'entreprises) évaluées par le modèle. Du point de vue théorique, nous nous intéressons alors à la nature de l'erreur de test, lorsque n tend vers l'infini et nous l'appelons alors *erreur de prédiction* ou *erreur de généralisation*. Les propriétés mathématiques d'un modèle découlent directement de la nature et de l'analyse de cette erreur.
i) Une propriété attendue est que l'erreur de test converge vers la vraie erreur de prédiction du modèle lorsque le nombre d'observations augmente.
ii) Une seconde propriété est la convergence de l'erreur de prédiction vers l'erreur la plus petite possible.
iii) Lorsque le nombre d'observations est fixé, des *bornes de risque* permettent de mesurer l'écart entre l'erreur empirique du modèle (sur un échantillon d'entreprises) et l'erreur qui serait commise si le modèle était expérimenté au niveau opérationnel (sur la totalité ou une partie des entreprises).

Un modèle, armé de ces trois propriétés et des mesures statistiques précédentes, fournit alors des garanties théoriques et opérationnelles solides sur sa capacité à obtenir les résultats annoncés.

Pour les réunir, il nous faut nous assurer d'un modèle suffisamment souple pour supporter un grand nombre de contraintes. Cette souplesse permet, en particulier, de s'affranchir d'hypothèses probabilistes qui peuvent limiter la capacité de généralisation du modèle. Parmi elles, la plus importante et la seule qui ne puisse être levée totalement est la suivante : *Les observations doivent être indépendantes et identiquement distribuées.*
Cette hypothèse, dite *i.i.d.*, implique qu'il n'y ait pas de changement de distribution (ni de ses paramètres) entre l'échantillon d'apprentissage et l'échantillon de test. Ce dernier s'étend potentiellement à l'ensemble des entreprises non contrôlées. L'indépendance implique que l'apparition d'une observation n'influence pas celle d'une autre observation de l'échantillon. Les déclarations de cotisations sont contrôlées sur, au plus, trois années consécutives. Pour la construction de nos échantillons d'apprentissage et de test, nous avons recours aux résultats des contrôles portant sur l'année 2010 et l'année 2011 (et réalisés en 2010, quelques uns, ou en 2011, la majorité). Il n'est pas exclu que la distribution, pour un échantillon d'entreprises non contrôlées, puisse être différente de celle de l'échantillon d'apprentissage. Il n'est pas exclu, non plus, que le choix d'une entreprise à contrôler ne soit pas influencé par le choix d'une autre car une partie des contrôles résulte d'un ciblage. Les expérimentations opérationnelles permettent de le vérifier (et, dans une moindre mesure, les mécanismes algorithmiques ajoutés au processus de détection).

Notons que l'algorithme de traitement générique essaie de maintenir l'absence de biais de sélection lors de l'évaluation d'une entreprise. Seules ses cotisations sociales relativement à sa masse salariale, ainsi que sa relation vis-à-vis de l'URSSAF, nous sont utiles. On ne s'intéresse pas à l'entreprise elle-même.

La majorité des contraintes étant (partiellement) levées sur la détection des irrégularités, il nous faut, avant de présenter le modèle et ses garanties théoriques et pratiques, définir des indicateurs pour l'estimation du montant de redressement. En effet, il convient d'abord de recommander un contrôle lorsque les montants de redressement espérés ne sont pas trop petits, quelle que soit la taille de l'entreprise.

`Régression, erreur quadratique moyenne, intervalle de confiance`

L'estimation du montant de redressement est assimilable à un problème de régression. Lorsque le modèle estime qu'une irrégularité est présente, il calcule dans le même temps le montant du redressement associé. Mais (comme indiqué auparavant) le modèle n'a pas connaissance de la masse salariale de l'entreprise. A la place du montant redressé, nous utilisons alors, et uniquement au moment de la régression, le *niveau d'irrégularité*, défini par le rapport entre le montant redressé et la masse salariale de l'entreprise. De la même manière que pour la détection, l'estimation repose sur des mesures, beaucoup moins nombreuses cependant, qui permettent d'évaluer la précision de l'estimation. La première mesure de référence est l'*erreur quadratique moyenne* (empirique), calculée sur l'échantillon de test, et que nous notons MSE (mean squared error). Elle est définie par :

$$MSE(X, Y, R) = \frac{1}{\sum_{i=1}^{n} \mathbf{I}_{\{Y_i=1\}}} \sum_{i=1}^{n} \left[(R_i - g(X_i))^2 \mathbf{I}_{\{Y_i=1\}} \right],$$

où g est la règle de décision du modèle pour l'estimation du montant redressé, lorsque l'irrégularité Y est présente, et $R \in [-1, 1]$ est le niveau d'irrégularité, défini ainsi uniquement au moment de la régression.
L'erreur quadratique moyenne est une mesure de la distance moyenne séparant l'estimateur de sa vraie valeur. Généralement, on souhaite mesurer un écart autour de la vraie valeur et on utilise alors la racine carrée de l'erreur quadratique, notée $RMSE$ et définie par :

$$RMSE(X, Y, R) = \sqrt{MSE(X, Y, R)}.$$

Cette mesure répond à la question suivante : *quel est l'écart moyen entre le montant estimé d'un redressement et sa vraie valeur ?*
Comme cet écart est mesuré sur l'ensemble des entreprises testées, il nous fournit le risque moyen (en terme d'écart) pris lorsque nous estimons le montant d'une irrégularité. L'erreur quadratique moyenne admet une contrepartie théorique, détaillée dans le troisième chapitre, et qui permet de nous assurer sous certaines conditions :
i) une borne de risque, lorsque le nombre d'observations est fixé, soit une valeur maximale de l'erreur quadratique moyenne empirique ;
ii) la convergence de l'erreur quadratique vers la vraie erreur de prédiction du modèle.
Notons que dans le cas de la régression, nous n'avons pas la convergence de l'erreur de prédiction vers la plus petite erreur possible.

L'inconvénient de l'erreur quadratique moyenne est son caractère global. Nous souhaitons obtenir un intervalle de confiance pour chaque montant de redressement estimé, de façon à ce que dernier soit compris dedans, avec une grande probabilité. Pour cela nous construisons un *intervalle de prédiction bootstrap*. L'appellation bootstrap est liée au fait que le montant de redressement estimé est issu de B tirages avec remise des observations (dit bootstrap) et de B modèles construits pour l'estimer, avec $B > 1$. L'intervalle de prédiction provient alors de la distribution empirique des B estimateurs de chaque montant. Nous revenons plus en détail sur cette procédure au moment de la présentation opérationnelle du modèle.

`Modèle, rendement, score global`

Il reste alors à fournir un indicateur quantitatif des montants redressés, afin de positionner le modèle en termes de *rendement*, puis de lui attribuer un *score global*, mesurant, à la fois, ses capacités de détection et son potentiel de découverte des recettes non récupérées par la Sécurité sociale. Nous rappelons alors les indicateurs définis dans le second chapitre. Le rendement d'un modèle est le montant comptable moyen redressé par entreprise contrôlée, sur la base des recommandations de ce modèle. Notons qu'il est calculé sur le nombre de contrôles réalisés et non sur le nombre de redressements. Ainsi, il exprime également la pertinence d'un modèle dans les recommandations soumises. Le montant comptable moyen par entreprise contrôlée est défini par :

$$\bar{R} = \frac{\sum_{i=1}^{N_C} R_i^+ - \sum_{i=1}^{N_C} R_i^-}{N_C},$$

où R_i^+ correspond à la somme des montants redressés en faveur de l'URSSAF, pour l'entreprise i,
R_i^- correspond à la somme des montants redressés en faveur de l'entreprise, pour l'entreprise i.
Le montant comptable moyen répond à la question suivante : *quel est le montant moyen récupérable à l'issue d'un contrôle effectué selon les recommandations du modèle ?*

Toutefois, lors d'une expérimentation opérationnelle, l'échantillon utilisé doit être représentatif de la population des entreprises. Un modèle qui ne recommande que des entreprises au-delà d'une certaine taille, produira un montant comptable et un rendement biaisé et inexploitable. De même, la sur-représentation d'un secteur d'activité ne pourra conduire à aucune conclusion sur le rendement ou les capacités de généralisation.
Le rendement d'un modèle est généralement suffisant pour témoigner de sa pertinence en comparaison de ce qui existe déjà dans la politique de contrôle de l'URSSAF. Il donne un point de vue aux inspecteurs du contrôle sur les enjeux financiers et permet, par exemple, de prioriser des plans de contrôle. Le point essentiel est la capacité du rendement d'un modèle à éclairer sur le montant total des irrégularités et sur la manière dont il est possible de le récupérer. Nous passons alors d'une optique de détection des irrégularités sur un certain nombre de cas, à une politique de mise en oeuvre complète de la détection du plus grand nombre d'irrégularités et des montants de redressement associés.

Pour estimer le pertinence d'un modèle relativement à ses capacités de détection et à son potentiel économique (le produit de son rendement et du nombre d'entreprises redressées), nous considérons le *score d'importance de la détection* qui mesure simultanément ces deux capacités. Il est noté S_D et nous rappelons sa définition :

$$S_D = \frac{\left(P_r - \frac{1-P_r}{P_r}\right)\left(1 + \frac{M_C}{M_R N_R}\right)\sum_{i=1}^{N_R} R_i}{M_R} \times 100, \tag{5.1}$$

où P_r est la *précision* du modèle,
M_C la masse salariale totale des entreprises contrôlées,
M_R, la masse salariale totale des entreprises redressées,
N_R, leur nombre,
R_i, le montant net redressé pour l'entreprise i.

Le score d'importance répond à la question suivante : *quel est le potentiel économique du modèle sachant ses capacités de détection des irrégularités ?*

Plus le score est important, plus le potentiel économique est important et plus les capacités de détection le sont également. A l'inverse, un modèle avec une capacité de détection trop faible ne pourra voir augmenter son potentiel économique s'il est généralisé. En d'autres termes, son score n'augmentera qu'avec une diminution du nombre de contrôles. Si les capacités de détection sont insuffisantes, le score est négatif. Le score d'importance est, de notre point de vue, un critère important pour évaluer un modèle. Il synthétise un grand nombre d'informations que nous avons traitées jusqu'ici et peut s'appliquer à n'importe quelle problématique associant des enjeux financiers à des problèmes de détection. Lorsque le nombre d'entreprises redressées devient important, le score d'importance admet une limite. On a :

$$S_D \to S_{D_\infty}, si\ N_R \to \infty\ et\ \frac{N_R}{n} \to 0,\ quand\ n \to \infty,$$

avec

$$S_{D_\infty} = \frac{\left(P_r - \frac{1-P_r}{P_r}\right)\sum_{i=1}^{N_R} R_i}{M_R} \times 100. \tag{5.2}$$

La relation (5.2) implique que pour n'importe quel modèle, le potentiel de découverte du montant total des irrégularités n'est limité que par sa *précision*, à la condition que cette dernière ne s'effondre pas lorsque le nombre de contrôles augmente (soit que la fréquence des redressements positifs ne s'effondre pas).
Lorsque $P_r = 1$, et en supposant que toutes les irrégularités soient de la fraude, S_{D_∞} est égal au niveau de fraude aux cotisations sociales. Pour trouver toute la fraude, sans utiliser de ressources gigantesques, il faudrait disposer d'un modèle qui ne fasse jamais d'erreurs. Hors de ce cadre idéal, notre objectif est de fournir les garanties sur P_r qui permettent de maintenir une *précision* élevée à mesure que le nombre de contrôles augmente. Dans la section qui suit, nous discutons du modèle développé pour prendre en compte l'ensemble des contraintes et indicateurs définis jusqu'ici, et de ses propriétés.

5.5 Les forêts uniformément aléatoires comme algorithme d'apprentissage

Dans le troisième chapitre, notre analyse s'est exclusivement portée sur les *forêts uniformément aléatoires*, dont nous avons développé quelques uns des aspects les plus fondamentaux. Nous résumons et étendons ce propos dans les lignes qui suivent et rappelons à nouveau les propriétés théoriques dont nous aurons besoin pour la connexion avec les résultats obtenus. Puis, nous en donnons un exemple étendu et effectuant un comparatif rapide avec d'autres algorithmes sur les données de cotisations sociales.

Les forêts uniformément aléatoires sont un modèle ensembliste d'apprentissage statistique, développé d'abord pour les besoins de la détection des irrégularités aux cotisations sociale, puis étendu à n'importe quel type de problème en classification, régression ou en recommandation ("ranking"). Rappelons la définition des modèles ensemblistes. A l'inverse d'un unique modèle pour traiter un problème, un modèle ensembliste crée plusieurs modèles de base (en général des centaines, voire milliers) sur tout ou sur des parties du problème, puis les agrège (ou les combine) pour traiter le problème final. Deux types de modèles ensemblistes sont généralement utilisés : le *Boosting* (Freund et Schapire, 1997), dans lequel les modèles de base sont construits de manière séquentielle. L'erreur commise par chacun, sur l'estimation de la variable à prédire, est utilisée pour améliorer l'estimation du modèle de base qui suit. Le deuxième type de modèle ensembliste, dont s'inspirent les forêts uniformément aléatoires, est le *Bagging* (Breiman, 1996) dans lequel les modèles de base sont construits indépendamment les uns des autres.

5.5.1 Principe

Une version améliorée du Bagging sont les *forêts aléatoires* (Breiman, 2001) qui en diffèrent par un caractère aléatoire plus poussé et par l'utilisation d'*arbres de décision* comme modèles de base. Les forêts uniformément aléatoires sont une variante des forêts aléatoires de Breiman et en partagent le principe.

i) Dans une forêt aléatoire, pour la construction de chaque arbre de décision, les n observations de l'échantillon d'apprentissage sont tirées, uniformément, avec remise (bootstrap) parmi les n disponibles. Pour chaque arbre, l'échantillon d'apprentissage est donc perturbé et toujours différent de l'échantillon initial. Puis, on tire, sans remise, un nombre v de variables parmi les d variables du problème. v est fixe pour l'ensemble des arbres de décision et, généralement, beaucoup plus petit que d.
ii) Dans une forêt uniformément aléatoire le tirage des observations se fait de la même manière que dans les forêts aléatoires de Breiman, *uniquement dans le cas de la classification*. Dans le cas de la régression et c'est la *première différence*, on tire, sans remise, m observations parmi n, avec $m < n$. Cette méthode est assimilable au *Subagging* (Bühlmann et Yu, 2002) ou sous-échantillonnage sans remise pour la construction de chaque modèle de base parmi B modèles. Puis, on tire, *avec remise*, un nombre v de variables parmi les d variables du problème. v peut donc être bien plus grand que d. C'est la *seconde différence* avec la version de référence.

iii) La construction des arbres de décision, les modèles de base, est également différente dans les deux versions. Nous rappelons brièvement leur définition. Un arbre de décision, est une structure algorithmique qui partitionne de manière récursive l'espace des observations, puis prend une décision lorsque des conditions d'arrêt sont atteintes. Dans les forêts aléatoires de Breiman, le *partitionnement récursif* des arbres de décision est binaire. Au départ, l'espace est divisé en deux régions disjointes et complémentaires. Puis chacune est divisée à nouveau en deux nouvelles régions disjointes. Le processus recommence ainsi jusqu'à ce qu'une ou plusieurs conditions d'arrêt soient rencontrées. Pour identifier une région, il faut alors connaître sa frontière, définie par une variable parmi les v choisies à chaque étape de la construction d'une région, et par un *point de coupure*, une (unique) observation parmi toutes celles demeurant dans la région en cours d'identification. Ces deux aspects nécessitent deux optimisations : une première pour le choix du meilleur point de coupure de chaque variable ; et une seconde pour le choix de la meilleure variable parmi toutes celles candidates.

iv) Dans une forêt uniformément aléatoire, *il n'y a pas d'optimisation du point de coupure. Ce dernier est généré aléatoirement, en tirant un point selon la loi Uniforme sur le support de chaque variable candidate (dont celles répétées). C'est la troisième différence.* Cette caractéristique donne son nom aux *arbres de décision uniformément aléatoires* et, par extension, aux forêts uniformément aléatoires. Pour chaque région et sa région complémentaire, la meilleure variable parmi toutes celles candidates est alors choisie selon un critère dit d'*entropie* (classification), ou à partir la *somme des carrés résiduels* (régression), pour les deux régions. Cette étape constitue la *dernière différence* importante avec la version de référence de Breiman.

v) Lorsque l'arbre de décision ne peut plus être partitionné, une décision est alors prise. Elle correspond au choix de la classe (pour la classification) à attribuer à la région concernée, dite région terminale (ou feuille). Ce choix est effectué grâce à un *vote majoritaire* parmi les instances (les classes) de la variable à prédire présentes dans la région. Dans le cas de la régression, il ne reste, en général, qu'une seule instance. Lorsque ce n'est pas le cas, la moyenne des valeurs de la variable à prédire est choisie.

vi) Un arbre de décision uniformément aléatoire n'impose quasiment aucune contrainte aux données. L'arbre est développé au maximum de ses possibilités et une fois construit, ne peut plus être modifié d'aucune façon. Il ne peut qu'être supprimé.

vii) La forêt uniformément aléatoire est alors construite en générant B arbres, avec $B >> 1$. Sa règle de décision est alors celle du *vote majoritaire des règles de décision de tous les arbres* ou de leur *moyenne* (éventuellement pondérée), dans le cas de la régression. Les forêts uniformément aléatoires ont pour fondement le paradigme de Breiman : *les arbres de décision doivent être sans biais, avoir une grande variance et être peu dépendants les uns des autres.*

viii) La forêt uniformément aléatoire est *nativement incrémentale*. En d'autres termes, il est possible d'assembler une forêt de forêts sans en changer la structure ou les propriétés. Dans la pratique, cela permet de lever les problématiques liées au volume des données

et, par exemple, de traiter l'ensemble des entreprises d'Île-de-France (ou plus) sur une unique station de travail.

5.5.2 Propriétés théoriques et applications

Les forêts uniformément aléatoires héritent des mêmes propriétés que les forêts aléatoires de Breiman. Nous reprenons sous une forme compacte les résultats du troisième chapitre et montrons comment les appliquer à la problématique des irrégularités aux cotisations sociales. Les propriétés théoriques génèrent les garanties fournies par le modèle et immédiatement applicables à la problématique. En particulier, nous insistons sur les données OOB qui permettent de construire des bornes de l'erreur de prédiction et permettent de contrôler cette erreur pour n'importe quel échantillon de test. Dans tout le reste du document, nous désignons les forêts uniformément aléatoires par leur règle de décision $\bar{g}_{\mathcal{P}}^{(B)}$, définie pour la classification (la détection des irrégularités) par :

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = \begin{cases} 1, \ si \ \sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x)=1\}} > \sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x)=0\}} \\ \\ 0, \ sinon. \end{cases} \tag{5.3}$$

Et dans le cas de la régression (l'estimation du montant de chaque irrégularité) par :

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(x), \tag{5.4}$$

où $g_{\mathcal{P}}^{(b)}(x)$ est la règle de décision du $b-$ème arbre uniformément aléatoire, défini sur une partition $\mathcal{P}$ des données, et associée à l'observation x (les informations relatives à la déclaration de cotisations et à la relation de l'entreprise avec l'URSSAF),
et $\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x)=1\}}$ est la fonction indicatrice, à valeurs dans $\{0,1\}$. Elle vaut 1, si $g_{\mathcal{P}}^{(b)}(x) = 1$ et 0 sinon.

***A**)* Dans le cas de la classification, $\bar{g}_{\mathcal{P}}^{(B)}(x)$ est donc un estimateur implicite de la fonction $\eta(x) = \mathbf{P}(Y = 1|X = x)$. La relation (5.3) est équivalente à la relation suivante :

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = 1, \ si \ \frac{1}{B} \sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x)=1\}} > 1/2.$$

Lorsque $B \to \infty$,

$$\frac{1}{B} \sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x)=1\}} \to \mathbf{E}_{\theta} \left\{ \mathbf{I}_{\{g_{\mathcal{P}}(x,\theta)=1\}} \right\} = \mathbf{P}_{\theta} \left(g_{\mathcal{P}}(X,\theta) = 1 | X = x \right). \tag{5.5}$$

La relation (5.5) implique qu'un seuil naturel pour qu'une irrégularité existe est une probabilité d'observation de cette dernière supérieure à 0.5. Ce seuil n'est cependant pas unique et on peut en définir d'autres, en modifiant la forêt uniformément aléatoire, de façon à augmenter le niveau de confiance de chaque estimation.

i) A chaque observation x, nous associons la valeur $y, y \in Y$, mesurant l'absence ou la présence d'une irrégularité. Le risque d'erreur est alors donné par $\mathbf{I}_{\{\bar{g}_{\mathcal{P}}^{(B)}(x) \neq y\}}$.

Pour toutes les observations, la probabilité d'erreur (soit l'*erreur de prédiction ou de généralisation*) est notée $L(\bar{g}_{\mathcal{P}}^{(B)})$, définie par :

$$L(\bar{g}_{\mathcal{P}}^{(B)}) = \mathbf{E}\left\{\mathbf{I}_{\{\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\}}\right\} = \mathbf{P}\left\{\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right\}.$$

Notre échantillon comporte n observations, correspondant aux nombre de contrôles effectués. Nous ne pouvons alors calculer la probabilité d'erreur pour toutes les observations possibles mais seulement pour n d'entre elles. Elle est appelée probabilité d'erreur conditionnelle (l'*erreur de test*) et elle est donnée par :

$$L_n(\bar{g}_{\mathcal{P}}^{(B)}) = \mathbf{P}\left\{\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y | D_n\right\},$$

où D_n est l'échantillon à notre disposition. On note $\widehat{L}_n$, la contrepartie empirique de L_n, définie par :

$$\widehat{L}_n(\bar{g}_{\mathcal{P}}^{(B)}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{g_{\mathcal{P}}^{(B)}(X_i) \neq Y_i\}}.$$

On souhaite savoir comment se comporte $\widehat{L}_n$ face à L. En effet, calculer l'erreur sur l'échantillon de test ne garantit pas qu'elle sera valide lors de l'évaluation de l'ensemble des déclarations de toutes les entreprises. Nous faisons alors appel aux outils introduits par Breiman (2001). Nous notons, mg, la marge, sur tous les arbres de la forêt uniformément aléatoire, entre les observations bien classées et les observations mal classées. Elle est définie par :

$$mg(X,Y) = \frac{1}{B}\left(\sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X)=Y\}}\right) - \frac{1}{B}\left(\sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X)\neq Y\}}\right),$$

et l'*erreur de prédiction* est notée PE^* dans la forêt uniformément aléatoire. Elle est équivalente à L et définie par :

$$PE^* = \mathbf{P}_{\mathbf{X},\mathbf{Y}}\left\{mg(X,Y) < 0\right\}.$$

Plus précisément, si une déclaration est mal classée par le modèle, la marge est négative. L'erreur de prédiction est simplement le passage à la limite du nombre de fois, relativement à toutes les observations, où la marge est négative.
Supposons que chaque arbre soit caractérisé par un paramètre θ et posons
$g_{\mathcal{P}}(X) \overset{def}{=} g_{\mathcal{P}}(X,\theta)$ et $g_{\mathcal{P}}^{(b)}(X) \overset{def}{=} g_{\mathcal{P}}(X,\theta_b)$. Nous avons alors un premier résultat (Breiman, théorème 1.2 (2001)) :

$$lorsque\ B \to \infty,\ PE^* \overset{p.s.}{\to} PE = \mathbf{P}_{\mathbf{X},\mathbf{Y}}\{\mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) = Y) - \mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) \neq Y) < 0\}.$$

Cette propriété nous indique qu'à mesure de l'augmentation du nombre d'arbres de décision, l'erreur calculée sur toutes les observations se rapproche de la *vraie erreur de*

prédiction du modèle, PE, définie sur tous les arbres possibles. De plus, elle admet une borne supérieure. Si nous disposons d'un grand nombre d'observations, alors pour un nombre d'arbres, B, fixé, l'erreur de test sera plus petite qu'une limite formulée explicitement. Une borne supérieure de l'erreur de prédiction de la forêt (uniformément) aléatoire, PE^*, est donnée par (Breiman, théorème 2.3, (2001)) :

$$PE^* \leq \frac{\bar{\rho}(1-s^2)}{s^2}, \ avec\ s > 0. \tag{5.6}$$

$\bar{\rho}$ est la corrélation moyenne entre tous les arbres,
$s = \mathbf{E}_{\mathbf{X},\mathbf{Y}}\{mr(X,Y)\}$, et $mr(X,Y)$ est la limite de $mg(X,Y)$.
Une condition suffisante pour que $s > 0$ est donnée par $\mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) = Y) > 1/2$.

La relation (5.6) entraîne plusieurs conséquences. La première est l'absence de nécessité à construire un grand nombre d'arbres. La seconde, plus implicite, est que l'erreur de test a pour limite supérieure la borne de risque de Breiman. La seule condition nécessaire est que la règle de décision de l'arbre soit un peu plus performante que le hasard, à n fixé. L'aspect le plus intéressant est que la relation (5.6) est valide dans la pratique, pour un nombre d'observations et un nombre d'arbres fixés. Détaillons le processus en nous référant à la détection des irrégularités aux cotisations sociales. Nous disposons d'un échantillon d'apprentissage $D_n^{A_Y} = \{(X_i, Y_i), 1 \leq i \leq n\}$ constitué de l'information contenue dans les déclarations de cotisations et des résultats des contrôles pour n entreprises. Dans la construction de la forêt aléatoire, nous tirons n observations, avec remise, pour la construction de chaque arbre. De l'échantillon initial, un certain nombre d'observations (un peu plus d'un tiers) ne participe donc pas à la construction des arbres, du fait du tirage avec remise. Ces données sont dites OOB et nous en avons détaillé le processus dans le troisième chapitre. La règle de décision OOB peut donc être utilisée comme *estimateur de l'erreur de test* en utilisant uniquement l'échantillon d'apprentissage. Elle est définie par :

$$\bar{g}^{(B)}_{\mathcal{P},oob}(x) = \begin{cases} 1, \ si \ \sum_{b=1}^{B} \mathbf{I}_{\left\{g^{(b)}_{\mathcal{P}}(x)=1\right\}} \mathbf{I}_{\left\{b \in G^-(x,B)\right\}} > \sum_{b=1}^{B} \mathbf{I}_{\left\{g^{(b)}_{\mathcal{P}}(x)=0\right\}} \mathbf{I}_{\left\{b \in G^-(x,B)\right\}} \\ \\ 0, \ sinon. \end{cases}$$

$G^-(x,B)$ est l'ensemble des B arbres n'ayant jamais classé l'observation x lors de l'apprentissage.
$\bar{g}^{(B)}_{\mathcal{P},oob}(x)$ est la règle de décision de la forêt uniformément aléatoire, réduite à B' arbres, $B' < B$. Elle possède les mêmes propriétés que $\bar{g}^{(B)}_{\mathcal{P}}(x)$. Un estimateur de l'erreur de test est donné par :

$$\overline{PE}^{(B)}_{oob} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{\left\{\bar{g}^{(B)}_{\mathcal{P},oob}(X_i) \neq Y_i\right\}}.$$

La relation (5.6) implique que :

$$\overline{PE}^{(B)}_{oob} \leq \frac{\hat{\bar{\rho}}_{oob}(1-\hat{s}^2_{oob})}{\hat{s}^2_{oob}}.$$

Comme application à la détection des irrégularités aux cotisations sociales, l'erreur de prédiction, au moment de l'évaluation de l'ensemble des entreprises, inconnue par définition, tend à être plus petite que la borne de risque de Breiman, estimée à partir des

données OOB de l'échantillon d'apprentissage, si la taille de ce dernier est assez grande. Toutefois, il convient de déterminer dans quelle mesure la taille de l'échantillon d'apprentissage peut influencer la borne de risque. Nous y revenons dans les lignes qui suivent.

ii) La convergence de la probabilité d'erreur conditionnelle L_n vers la "vraie" erreur de prédiction L est aussi dépendante du modèle. Elle n'indique pas si la *plus petite erreur de prédiction possible*, notée L^*, est atteinte lorsque nous évaluons l'ensemble des entreprises et si nous disposons d'un échantillon d'apprentissage assez grand pour cela. Cette notion est appelée *consistance* et permet d'affirmer qu'au fur et à mesure que la taille de l'échantillon d'apprentissage augmente, on se rapproche de la plus petite erreur possible. Les forêts uniformément aléatoires sont des classifieurs consistants. Notons $PE^* \stackrel{def}{=} L_n$, $PE \stackrel{def}{=} L$ et k, le nombre moyen de régions construites par chaque arbre. On a, pour la forêt uniformément aléatoire (proposition 3 (chapitre 3, section 3.3)) :

$$L_n \stackrel{p.s.}{\to} L^*, \ si \ \frac{n}{k^2 \log n} \to \infty, \ quand \ n \to \infty. \tag{5.7}$$

Si le nombre moyen de régions des arbres n'est pas trop important relativement au nombre d'observations, plus la taille de l'échantillon d'apprentissage est grande, plus on se rapproche de la plus petite erreur atteignable, laquelle est indépendante du modèle.
Dans la pratique, le nombre de contrôles est limité et n'augmente que très peu d'une année sur l'autre, ou bien diminue. Pour disposer d'un échantillon dont la taille augmente et pour ne pas atteindre de temps de calculs trop importants, nous faisons alors appel à la *forêt uniformément aléatoire incrémentale* qui permet de répliquer une augmentation de la taille de l'échantillon. Pour cela, chaque nouvel échantillon d'entreprises contrôlées (chaque année) est *appris* par le modèle et *mémorisé*. L'évaluation du modèle sur un échantillon de test est le résultat de l'apprentissage sur l'échantillon courant et de la mémorisation des échantillons précédents.

iii) La troisième propriété, dont nous avons besoin, est liée à la taille des échantillons d'apprentissage et de test. Nous avons deux situations :
- dans l'expérimentation opérationnelle, l'échantillon de test est bien plus petit que l'échantillon d'apprentissage.
- Dans l'évaluation complète, l'échantillon de test est l'ensemble de la population des entreprises.
Lorsque le modèle doit être expérimenté "sur le terrain", l'échantillon d'apprentissage est constitué de toutes les déclarations de cotisations contrôlées pendant une période donnée et l'échantillon de test est constitué d'un certain nombre (petit) de déclarations que le modèle aura évalué comme présentant des irrégularités. Du fait du caractère expérimental et d'autres contraintes opérationnelles, les inspecteurs du contrôle n'en mènent que ce petit nombre. Les résultats obtenus caractérisent, de fait, l'efficacité du modèle. Idéalement l'échantillon de test devrait être plus grand que l'échantillon d'apprentissage, lui même assez grand, afin d'avoir un contrôle optimal de l'erreur de prédiction. Ici, ce n'est pas le cas et nous faisons appel à l'inégalité d'Hoeffding et au corollaire suivant (Devroye, Györfi, Lugosi, théorème 9.1 (1996)) :

Corollaire. Devroye, Györfi, Lugosi, 1996 (corollaire 12.2).
Soit $\mathcal{C}$, une classe arbitraire de règles de décision de la forme $\phi : \mathbb{R}^d \to \{0,1\}$. Soit $\phi_n^* \in \mathcal{C}$, une règle de décision qui minimise le nombre d'erreurs commis sur un échantillon d'apprentissage D_n, parmi toutes les règles de décision de $\mathcal{C}$, soit que :

$$\widehat{L}_n(\phi_n^*) \leq \widehat{L}_n(\phi_n), \text{ pour tout } \phi \in \mathcal{C},$$

avec $\widehat{L}_n(\phi_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{I}_{\{\phi(X_i)\neq Y_i\}}$. Alors, pour n'importe quels n et $\epsilon > 0$,

$$\mathbf{P}\left\{\left|\widehat{L}_n(\phi_n^*) - \mathbf{E}\left\{\widehat{L}_n(\phi_n^*)\right\}\right| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}.$$

Ce résultat est applicable dès qu'on dispose d'un classifieur capable de minimiser l'erreur sur l'échantillon d'apprentissage. Dans le cas des forêts uniformément aléatoires, la règle de décision $\bar{g}_{\mathcal{P}}^{(B)}$ ne commet pas d'erreurs sur D_n. En quelque sorte, le classifieur apprend "par coeur" la relation entre les déclarations de cotisations et les résultats des contrôles sur l'échantillon d'apprentissage. Notons que cela ne garantit pas qu'un tel classifieur soit le meilleur parmi toutes les règles de décision, mais on ne peut pas faire mieux puisque la probabilité d'erreur conditionnelle de la forêt tend vers la vraie erreur de prédiction du modèle.
- Pour un échantillon d'apprentissage de taille n, nous calculons d'abord l'erreur OOB, $\widehat{L}_n(\bar{g}_{\mathcal{P},oob}^{(B)})$. L'application du corollaire nous donne l'écart ϵ qui sépare l'estimation de son espérance mathématique, avec une probabilité inférieure à $2e^{-2n\epsilon^2}$. Ainsi, l'erreur OOB se rapproche de PE^* à mesure que n augmente.
- Dans le cas de l'expérimentation opérationnelle, l'échantillon de test a une taille n' plus petite que n, et nous fixons alors, ϵ, la marge entre l'erreur OOB, calculée sur l'échantillon d'apprentissage, et son espérance mathématique. Cette dernière est inconnue mais demeure la meilleure estimation que l'on puisse avoir de l'erreur au moment du test. L'application du corollaire nous donne la probabilité de réalisation d'une marge d'erreur supérieure à celle que nous avons fixé, pour les n' observations de l'échantillon de test.
- Dans le cas de l'évaluation complète, l'application du corollaire se fait pour les n observations de l'échantillon d'apprentissage, lui-même suffisamment grand pour que la probabilité de réalisation de la marge d'erreur soit proche de zéro. La limite supérieure de l'erreur de prédiction est ensuite estimée par la borne de risque de Breiman, ce qui fournit un niveau de confiance plus élevé dans la non-réalisation de la marge d'erreur.

Dans la pratique, nous posons $\epsilon = 1\%$ et sa probabilité de réalisation est équivalente au risque de dépassement des erreurs de prédiction estimées par le modèle.

B) Dans le cas de la régression, $\bar{g}_{\mathcal{P}}^{(B)}$ est un estimateur du montant de redressement R. Il est défini par la relation (5.4) et l'objectif est le contrôle de l'erreur quadratique moyenne, soit l'espérance des différences (au carré) entre les montants qui sont (ou seront) effectivement redressés lors des contrôles et leurs estimations. L'erreur de prédiction (ou de généralisation) est, ici, l'espérance de l'erreur quadratique. Pour un arbre de décision uniformément aléatoire, l'erreur de prédiction est l'espérance, sur tous les arbres de paramètre θ, de l'erreur quadratique moyenne. Elle est notée $PE(g_{\mathcal{P}}(X,\theta))$, et définie par :

$$PE(arbre) \overset{def}{=} PE(g_{\mathcal{P}}(X,\theta)) = \mathbf{E}_\theta \mathbf{E}_{\mathbf{X},\mathbf{R}}(R - g_{\mathcal{P}}(X,\theta))^2.$$

Pour la forêt uniformément aléatoire, un estimateur de l'erreur quadratique moyenne des B arbres de la forêt, $PE^*(\bar{g}_{\mathcal{P}}^{(B)}(X))$, est défini par :

$$PE^*(forêt) \overset{def}{=} PE^*(\bar{g}_{\mathcal{P}}^{(B)}(X)) = \mathbf{E}_{\mathbf{X},\mathbf{R}}\left(R - \frac{1}{B}\sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(X,\theta_b)\right)^2.$$

Les outils utilisés pour la classification s'adaptent, en partie, au cas de la régression et nous avons (Breiman, théorème 11.1 (2001)) :

$$lorsque\ B \to \infty,\ \mathbf{E}_{\mathbf{X},\mathbf{R}}\left(R - \frac{1}{B}\sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(X,\theta_b)\right)^2 \overset{p.s.}{\to} \mathbf{E}_{\mathbf{X},\mathbf{R}}\left(R - \mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right)^2.$$

L'erreur quadratique moyenne converge alors vers l'espérance de l'erreur quadratique du modèle. Comme dans la classification, nous utilisons l'erreur OOB pour contrôler l'erreur de prédiction. On a :

$$\bar{g}_{\mathcal{P},oob}^{(B)}(x) = \frac{1}{\sum_{b=1}^{B} \mathbf{I}_{\left\{b \in G^-(x,B)\right\}}} \sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(x)\mathbf{I}_{\left\{b \in G^-(x,B)\right\}}$$

et

$$\overline{PE}^*(\bar{g}_{\mathcal{P},oob}^{(B)}(X,\theta)) = \frac{1}{n}\sum_{i=1}^{n}\left(R_i - \bar{g}_{\mathcal{P},oob}^{(B)}(X_i)\right).$$

Sous certaines conditions, l'erreur quadratique moyenne OOB fournit une borne supérieure de l'erreur quadratique moyenne sur les données de test. La plus petite erreur possible n'est, ici, pas certaine d'être atteinte lorsque le nombre d'observations augmente. Dans la pratique, cela n'est pas une nécessité car l'estimation du montant de redressement R, dépend de l'évaluation du modèle selon l'absence ou la présence d'irrégularités. Plus précisément, la quantité que nous cherchons à minimiser pour estimer au mieux les montants de redressement dépend de Y, la variable qui définit l'absence ou la présence d'irrégularités. L'erreur quadratique moyenne est notée $\overline{PE}^*(\bar{g}_{\mathcal{P}}^{(B)}(X,\theta)|Y)$, et définie par :

$$\overline{PE}^*(\bar{g}_{\mathcal{P}}^{(B)}(X,\theta)|Y) = \frac{1}{\sum_{i=1}^{n}\mathbf{I}_{\{Y_i=1\}}}\sum_{i=1}^{n}\left[(R_i - \bar{g}_{\mathcal{P}}^{(B)}(X_i))\mathbf{I}_{\{Y_i=1\}}\right]^2.$$

La dépendance, dans le modèle, du montant de redressement à la présence d'irrégularités permet une meilleure prise en compte des erreurs d'estimation. Il est généralement difficile d'estimer les montants de redressements à cause d'un grand nombre de montants nuls, lorsqu'il n'y a pas d'irrégularités.

`Intervalles de prédiction et de confiance`

Nous souhaitons fournir un intervalle de prédiction satisfaisant du montant de redressement, lorsque la probabilité d'existence d'une irrégularité dépasse le seuil de 0.5.
Les intervalles de prédiction (définis dans le chapitre 3, section 3.5.6) produits par les

forêts uniformément aléatoires sont généralement trop larges. De plus, la distribution des montants de redressement est de type exponentielle : près de la moitié des redressements valent 0, et plus le montant de redressement augmente, moins il y a d'entreprises redressées. L'objectif est, notamment, d'estimer le montant minimum d'un montant de redressement avec une faible probabilité d'erreur. Nous utilisons alors la procédure de construction d'intervalles de prédiction (et de confiance) bootstrap, proposée dans le troisième chapitre :
avec une probabilité approchée $1-\alpha$,

$$R_i \in \left[\hat{q}_{\alpha/2}(\bar{g}_{\mathcal{P}}^{(S)}(X_i)) + z_{\alpha/2}\sqrt{\frac{\widehat{\mathbf{Var}}_{\theta_\mathbf{S}}(g_{\mathcal{P}}(X_i,\theta_S))}{S}}, \hat{q}_{1-\alpha/2}(\bar{g}_{\mathcal{P}}^{(S)}(X_i)) + z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathbf{Var}}_{\theta_\mathbf{S}}(g_{\mathcal{P}}(X_i,\theta_S))}{S}} \right], \tag{5.8}$$

où $z_{\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la loi $\mathcal{N}(0,1)$,
et $\widehat{\mathbf{Var}}_{\theta_\mathbf{S}}(g_{\mathcal{P}}(X_i,\theta_S))$ est la variance empirique de la règle de décision dont les valeurs sont uniques pour X_i, et sont en nombre S.

On obtient également un intervalle de confiance de $\bar{R}$:
avec une probabilité approchée $1-\alpha$,

$$\bar{R} \in \left[\frac{1}{n}\sum_{i=1}^{n} \tilde{q}_{\alpha/2}(g_{\mathcal{P}}(X_i,\theta)), \frac{1}{n}\sum_{i=1}^{n} \tilde{q}_{1-\alpha/2}(g_{\mathcal{P}}(X_i,\theta)) \right], \tag{5.9}$$

avec $\tilde{q}_{\alpha/2}(g_{\mathcal{P}}(X_i,\theta)) = \hat{q}_{\alpha/2}(\bar{g}_{\mathcal{P}}^{(S)}(X_i)) + z_{\alpha/2}\sqrt{\frac{\widehat{\mathbf{Var}}_{\theta_\mathbf{S}}(g_{\mathcal{P}}(X_i,\theta_S))}{S}}$

Dans le cas des irrégularités aux cotisations sociales, ou dans tous les phénomènes pour lesquels les réalisations de la variable à prédire valent 0 dans la majorité des cas, les méthodes de construction d'intervalles de confiance bootstrap les plus performantes, comme BC_a (DiCiccio, Efron, 1996), nécessitent une adaptation aux modèles ensemblistes. La règle de décision $\bar{g}_{\mathcal{P}}^{(B)}$ est la moyenne des B règles de décision, $g_{\mathcal{P}}$, des arbres. La distribution de $\bar{g}_{\mathcal{P}}^{(B)}$ réplique celle de R et l'intervalle de prédiction pour R_i, construit à partir de la distribution des B valeurs de $g_{\mathcal{P}}(X_i,\theta)$ a comme inconvénient d'affecter en borne inférieure les zéros, ou bien des valeurs négatives des niveaux d'irrégularité, un trop grand nombre de fois. La relation (5.8) essaie de contourner ce problème en construisant des intervalles de prédiction plus réalistes. Néanmoins, cela peut ne pas suffire et nous utilisons les informations supplémentaires obtenues grâce à l'analyse des variables économiques effectuée dans le quatrième chapitre. Sous un seuil de 0.9% de la masse salariale contrôlée, nous savons que l'estimation du montant de redressement devient beaucoup plus complexe. Nous considérons, dans la relation (5.8), que la borne inférieure vaut 0 dès qu'elle passe sous le seuil. La relation (5.9) fournit un intervalle de confiance plus large pour $\bar{R}$, mais plus réaliste de l'ensemble des montants de redressement. De plus, lorsqu'on ne considère que les cas d'irrégularités, la relation (5.9) fournit un intervalle de confiance beaucoup plus précis. Nous illustrons la méthode proposée dans les résultats qui suivent et dans l'expérimentation opérationnelle.

5.5.3 Protocole

Pour une présentation claire des résultats, nous résumons nos données et présentons le protocole. Nous souhaitons évaluer l'ensemble des déclarations de cotisation des entreprises d'Île-de-France à travers l'absence ou la présence d'irrégularités, puis estimer, le cas échéant, le montant de redressement associé. Cette évaluation conduit aux recommandations du modèle, fournies à la direction du Contrôle de l'URSSAF. Nous nous intéressons uniquement aux contrôles comptables d'assiette.

`Caractéristiques des données`

Pour les contrôles à effectuer en 2013, le modèle évalue donc les déclarations des trois dernières années de cotisation, 2010, 2011 et 2012 (ce sont celles dont les déclarations seront contrôlées). Comme les données de la dernière année ne sont jamais disponibles avant le second trimestre de l'année qui suit, le modèle utilise des données et des résultats de contrôle des années 2010 et 2011. Toutefois, l'historique disponible est constitué de toutes les données et contrôles effectués entre 2006 et 2011. Deux possibilités existent :
a) utiliser seulement les données en relation avec les années qui seront contrôlées ;
b) utiliser tout l'historique.
Dans le premier cas, on dispose de moins d'exemples pour l'apprentissage ; en contrepartie il n'y a pas de risque que la détection identifie des relations obsolètes (par exemple, suite à un changement de législation) pour le contrôle des cotisations à effectuer. Dans le second cas, ce risque est présent mais il peut être supprimé en testant systématiquement le modèle contre celui établi dans le point a). En contrepartie, l'apprentissage est long (plus de 50 000 contrôles entre 2006 et 2011 et 1065 variables). Les forêts uniformément aléatoires incrémentales (et plus généralement les algorithmes incrémentaux) permettent de tirer avantage des deux possibilités, sans allonger les temps de calcul.

La population des entreprises enregistrées par l'URSSAF, en Île-de-France, est au nombre de 311 241 en 2011 et l'identifiant unique de chacune est son SIREN. 1065 variables sont définies, dont 34 sont des variables virtuelles (initialisées à 0), destinées à accueillir de nouvelles informations, par exemple en cas d'ajout de nouvelles catégories de cotisation. Environ 90% des observations ont pour valeur 0 et 97% des variables sont des catégories de cotisation.

`Apprentissage et validation`

Afin de simplifier le processus, nous ne considérons, d'abord, que l'étape *a)*.
- Nous disposons, pour les années 2010 et 2011, d'un échantillon D_n couvrant les données et les résultats des contrôles effectués pour le compte de ces mêmes années. Cet échantillon est filtré, puis divisé en un échantillon d'apprentissage et un échantillon de test afin de mesurer les capacités d'apprentissage et de généralisation de l'algorithme.
- Puis, un apprentissage complet est effectué sur D_n et les erreurs sont estimées.
- La dernière étape est l'évaluation de toute la population des entreprises, à partir du modèle construit grâce à D_n.

Toutefois, les exigences opérationnelles requièrent une expérimentation systématique. Nous avons donc fourni deux évaluations :
i) une pour l'année 2012, dont un échantillon a été utilisé pour valider le modèle en conditions réelles. Notre échantillon d'apprentissage D_n comportait, en plus, les données et résultats de l'année 2009. C'est *l'expérimentation opérationnelle*.
ii) Et une seconde évaluation, pour l'année 2013. Nous la nommons *évaluation complète*.

Pour la forêt uniformément aléatoire et une évaluation complète, nous avons défini le protocole suivant qui fixe les étapes les plus importantes :
- La présence ou l'absence d'une irrégularité est décidée par le *vote* majoritaire, parmi l'ensemble des arbres de décision uniformément aléatoires. Ce vote dépend d'un seuil, fixé au minimum à 50% *des votes* + 1 *vote* pour la présence d'une irrégularité, et peut varier. Plus précisément, le seuil peut être décidé, de manière autonome, par l'algorithme ou manuellement.
- Comme les irrégularités sont minoritaires, l'algorithme peut éventuellement modifier le poids de chaque vote. Cette étape est indépendante de la précédente.

L'estimation du montant de redressement est effectuée en plusieurs temps.
- L'algorithme effectue un deuxième apprentissage en utilisant les montants redressés (quelles que soient leurs valeurs) de tous les contrôles de l'échantillon.
- Au moment de l'évaluation, une estimation du montant de redressement et un intervalle de confiance sont générés. Plusieurs conditions doivent alors être réunies pour que l'estimation du montant de redressement soit validée.
- Si une irrégularité n'est pas détectée et que la borne inférieure de l'intervalle de confiance est supérieure à un seuil, l'estimation du montant de redressement est validée. Ce cas correspond à des irrégularités difficilement détectables mais de priorité faible.
- Si une irrégularité est détectée et que la borne inférieure de l'intervalle de confiance est plus petite que le seuil, l'estimation du montant de redressement est indiquée mais non validée. Ce type d'irrégularité correspond à des cas flous pour lesquels la prise de décision dépend à la fois de la probabilité d'existence de l'irrégularité et de la priorisation des contrôles.
- Dans les autres cas, le montant de redressement vaut 0, si l'irrégularité n'est pas détectée, et est égal à son estimation si l'irrégularité est détectée.
- Chaque estimation d'un montant de redressement est accompagnée d'un intervalle de confiance.

L'estimation du montant de redressement dépend donc de la détection ou non d'une irrégularité, mais supporte des cas particuliers.

- Nous n'effectuons pas de validation croisée. A la place, l'estimation des erreurs se fait à l'aide des informations OOB et de la borne de risque de Breiman (pour la classification). En plus de leurs propriétés, elles ont l'avantage de demander un temps de calcul non prohibitif (relativement à l'implémentation).

`Paramètres des forêts uniformément aléatoires`

Nous utilisons deux modèles de *l'évaluation complète* fournie pour 2013. Le premier modèle fait appel à la forêt uniformément aléatoire pour les données des années 2010 et 2011. Le second fait appel à la forêt uniformément aléatoire incrémentale et utilise les données des années 2006 à 2009 et le premier modèle. La forêt uniformément aléatoire incrémentale possède la *mémoire*, et les paramètres, de ses apprentissages passés ainsi que celle de l'apprentissage de l'échantillon courant.

Pour l'expérimentation opérationnelle, le modèle incrémental n'a pas été utilisé. De nombreux paramètres et méthodes sont intégrés à l'algorithme. Nous en faisons mention, lorsqu'ils sont utilisés, au moment de la présentation des résultats mais ne les détaillons pas.

`Autres spécificités`

Une fois les données intégrées par le modèle, le passage de l'estimation des performances de l'algorithme à son industrialisation tient lieu de processus primordial et se déroule en trois étapes :
- la réalité opérationnelle indique que moins de 15% de toutes les entreprises sont (et peuvent être) contrôlées chaque année. Pour estimer les performances de l'algorithme, nous nous conformons à cette réalité : l'échantillon d'apprentissage est toujours constitué d'au plus 10% des données et résultats de contrôle de la période courante.

- Au moment de l'expérimentation opérationnelle, l'échantillon d'apprentissage est constitué de toutes les données et résultats de contrôles de la période courante, filtrés par l'algorithme de traitement générique. Typiquement, pour les recommandations du modèle ayant conduit à l'expérimentation, en 2012, l'algorithme de traitement générique a fourni un échantillon d'apprentissage de 12155 entreprises (soit une partie des données et résultats des contrôles de la période 2009-2011) et 1065 variables. L'échantillon de test était beaucoup plus petit. Un second apprentissage a alors été réalisé avec un échantillon de la même taille que l'échantillon de test. Son seul objectif était l'estimation des erreurs qui ne devaient pas être dépassées, avec une grande probabilité, au moment de l'expérimentation.

- La dernière étape est l'évaluation complète de toutes les entreprises (pour l'année 2013). L'échantillon d'apprentissage est constitué de 4069 entreprises (contrôles concernant la période 2010-2011). Toutefois, cette phrase n'est pas tout à fait exacte car nous utilisons la forêt uniformément aléatoire incrémentale pour cette évaluation et environ
60 000 entreprises ont conduit à sa construction. 294 517 entreprises sont évaluées par la forêt uniformément aléatoire incrémentale et 17 387 le sont par l'algorithme de traitement générique pour un total de 311 904 entreprises. Dans cette dernière étape, les forêts uniformément aléatoires permettent de faire un choix entre un modèle incrémental et un modèle statique.

5.5.4 Un exemple de résultats

Nous fournissons ici la sortie complète d'une forêt uniformément aléatoire.
L'échantillon de travail est constitué des données et résultats des contrôles de la période 2009-2011. Sur les 12155 entreprises sélectionnées par l'algorithme de traitement générique, 10% des exemples sont utilisés comme échantillon d'apprentissage et 90% comme échantillon de test.

Les éléments les plus importants des résultats sont les erreurs de prédiction OOB et la borne de risque de Breiman. Elles constituent une borne supérieure de l'erreur de prédiction pour un échantillon de test, quelconque, de taille plus grande que l'échantillon d'apprentissage. Elles ne sont pas optimales car on cherche, avant toute chose, à obtenir des garanties sur les capacités de généralisation.

`Classification et détection des irrégularités`

La sortie ci-dessous est le standard des résultats d'une forêt uniformément aléatoire :
- les paramètres du modèle ;
- les résultats et les erreurs de prédiction OOB ;
- l'estimation des bornes de risques de Breiman ;
- lorsque l'échantillon de test est fourni, son évaluation et les résultats obtenus.

Sortie standard d'une forêt uniformément aléatoire : apprentissage (10%) et test (90%) sur les contrôles (12155) portant sur la période 2009-2011.

```
Call:
randomUniformForest.default(X = X1, Y = as.factor(Y1), xtest = X2,
    ytest = as.factor(Y2), ntree = 1000)

Type of random uniform forest: Classification

                            paramsObject
ntree                               1000
mtry                                1420
nodesize                               1
maxnodes                             Inf
replace                             TRUE
bagging                            FALSE
depth                                Inf
depthcontrol                       FALSE
OOB                                 TRUE
importance                          TRUE
subsamplerate                          1
classwt                            FALSE
classcutoff                        FALSE
oversampling                       FALSE
outputperturbationsampling         FALSE
targetclass                           -1
rebalancedsampling                 FALSE
randomcombination                  FALSE
randomfeature                      FALSE
categorical variables              FALSE
featureselectionrule             entropy

Out-of-bag (OOB) evaluation
OOB estimate of error rate: 23.62%

OOB confusion matrix:
          Reference
Prediction   0   1 class.error
         0 606 153      0.2016
         1 134 322      0.2939

Theorical (Breiman) bounds
Prediction error (expected to be lower than): 25.26%
Upper bound of prediction error: 40.56%
Trees average correlation: 0.0767
Strength (margin): 0.4827
Standard deviation of strength: 0.3074
```

```
Test set
Error rate: 23.53%

Confusion matrix:
          Reference
Prediction    0    1 class.error
         0 5299 1351      0.2032
         1 1223 3067      0.2851

Area Under ROC Curve: 0.7533
F1 score: 0.7044
Geometric mean: 0.751
```

- Le premier élément de la sortie est l'appel effectué à l'algorithme, à travers lequel sont assignés les exemples d'apprentissage et de test.

- Le second présente les paramètres que l'on peut modifier pour optimiser le modèle. Dans la pratique, il est possible de passer au Bagging (en initialisant l'option éponyme) ou à une forêt totalement aléatoire (*randomfeature*). Ici, les paramètres affichés sont ceux par défaut, à l'exception du nombre d'arbres de la forêt (*ntree*) défini à 1000.

- L'erreur de prédiction OOB, assimilable à une estimation (pessimiste) de l'erreur de test (la capacité à distinguer l'absence de la présence d'irrégularités) est de 0.2362. La matrice de confusion renvoie la répartition des résultats de l'évaluation OOB. L'erreur de prédiction sur l'absence d'irrégularités est estimée à 0.2016. Pour ce même échantillon d'apprentissage et le modèle construit, cela est équivalent (pour un échantillon de test de taille supérieure) à dire que le modèle se trompe, au plus, dans 20.16% de tous les cas évalués comme exempts d'irrégularités. La valeur estimée est moins importante que le message implicite : comme il n'y a qu'une mesure, l'estimation peut, bien sûr, varier. Mais les propriétés du modèle garantissent qu'au moment du test, l'erreur de prédiction effective sera proche avec une grande probabilité.

`Remarque:` notons que même si les évaluations sont effectuées avec des échantillons, la théorie (de l'apprentissage) statistique et les propriétés des forêts aléatoires assurent de la validité des résultats, lorsque le nombre d'observations tend vers l'infini, soit pour la population entière des entreprises. En conséquence, les différentes erreurs mesurées le sont du point de vue de l'évaluation par le modèle. Par exemple, dans le cas ci-dessus, on ne cherche pas à savoir dans quelle mesure le modèle sait détecter tous les cas d'absence d'irrégularités. On cherche plutôt, lorsque le modèle est explicitement sollicité, quelle est l'erreur commise. La différence est l'absence d'hypothèse de la deuxième situation.

De la même manière, l'erreur de prédiction sur la présence d'irrégularités est estimée à 0.2939. La précision, P_r, vaut 1 - 0.2939 = 70.61%. La précision est la deuxième mesure de référence (avec l'erreur de test). Pour ce même échantillon d'apprentissage, le taux de succès du modèle pour la détection d'une irrégularité dans la déclaration de cotisations des entreprises devrait être plus grand que 70%, pour un échantillon de test quelconque de taille supérieure, avec une grande probabilité.

- La borne de risque de Breiman (0.2526) est une limite supérieure de l'erreur de test attendue. La corrélation moyenne entre les arbres (0.0767) permet de vérifier que le passage de la théorie à la pratique se fait sans difficultés (les arbres doivent être peu corrélés). La marge (0.4827) résume l'écart moyen entre les observations bien classées (par le modèle) et mal classées, soit la différence entre les probabilités de chacun des évènements.

- Le dernier élément est constitué des résultats sur l'échantillon de test. Son argument essentiel est la non-violation des erreurs OOB au moment de l'évaluation.

`Régression et estimation du montant de redressement`

A la place de l'estimation du montant de redressement, R, nous considérons plutôt le ratio du montant de redressement à la masse salariale déclarée des trois dernières années de cotisation. Du fait des indications données précédemment, le montant de redressement estimé est issu d'une règle plus complexe que son seul calcul par le modèle et dépend à la fois de la détection d'irrégularités, de l'intervalle de confiance associé, de la masse salariale de l'entreprise, et de l'évaluation fournie par l'algorithme de traitement générique.

irregularity level	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Train sample	-34.77%	0	0	0.53%	0.38%	47.94%
Test sample	-32.46%	0	0	0.62%	0.37%	100.8%

TABLE 5.2 – Statistiques du ratio "montant de redressements sur masse salariale" pour les entreprises contrôlées par l'URSSAF d'Île-de-France pour le compte de la période 2009-2011.

Le niveau d'irrégularités représente, en moyenne, 0.6% de la masse salariale des entreprises dont les déclarations de la période 2009-2011 ont été contrôlées.

```
Type of random uniform forest: Regression

Out-of-bag (OOB) evaluation
Mean of squared residuals: 8e-04

OOB residuals:
       Min         1Q     Median       Mean         3Q        Max
-0.4529000 -0.0007411  0.0003787  0.0006594  0.0043940  0.3479000

Variance explained: 15.61%

Theorical (Breiman) bounds:
Theorical prediction error: 0.000621
Upper bound of prediction error: 0.000641
Mean prediction error of a tree: 0.001527
Average correlation between trees residuals: 0.4196
Expected squared bias (experimental): 4e-06

Test set
Mean of squared residuals: 0.000607
```

Comme dans la classification, les erreurs calculées à partir des données OOB sont les plus importantes. L'écart moyen (la racine carrée de l'erreur quadratique moyenne) entre le niveau d'irrégularité estimé et le niveau d'irrégularité réel est de 2.82% ($\sqrt{8 \times 10^{-4}}$). La borne de risque de Breiman (6.21×10^{-4}) estime l'erreur théorique que l'on peut espérer atteindre. La corrélation entre les erreurs résiduelles (0.4196) est plutôt élevée et suggère que seules quelques variables se révèlent décisives. Le biais estimé est assez faible et sa correction ne modifie les performances que marginalement. Notons également que la variance expliquée par le modèle est peu satisfaisante. Cependant, l'estimation du montant de redressement dépend d'autres facteurs non pris en charge par la régression.

5.5.5 Comparaison avec quelques modèles

Nous illustrons les résultats de la détection d'irrégularités à travers quelques algorithmes parmi les plus utilisés. Une seule mesure est effectuée et les échantillons d'apprentissage et de test sont les mêmes que dans la section précédente. L'intérêt est moins la comparaison des modèles que leurs capacités à bien appréhender la nature du problème. Pour les modèles ensemblistes, le nombres d'arbres est fixé à 1000 sauf pour le Bagging (200 arbres), dont l'implémentation R (package *ipred*) consomme une trop grande quantité de mémoire vive. Le modèle $GBM(optimized)$ est le seul à être optimisé, les paramètres des autres modèles étant ceux définis par défaut. Un résumé de certains des algorithmes est présenté dans la dernière partie du troisième chapitre.

Models	Test error	Precision	AUC
Random Forests	0.2505	72.74%	0.7265
ExtRaTrees	0.2351	71.71%	0.7528
Bagging	0.2274	71.76%	0.7641
GBM (default)	0.2873	59.31%	0.7458
GBM (optimized)	0.2220	72.53%	0.7694
CART	0.25	67.92%	0.7455
k Nearest Neighbours (k = 16)	0.3524	60.18%	0.6037
Logistic regression	0.3131	63.96%	0.659
GLMnet (LASSO penalty)	0.3133	63.81%	0.6594
SVM	0.3610	64.75%	0.5734
Random Uniform Forests	0.2353	71.49%	0.7533
Random Uniform Forests (incremental)	0.2009	76.19%	0.7923

TABLE 5.3 – Performances de quelques modèles, avec leurs paramètres par défaut, pour la détection d'irrégularités (contrôles comptant pour la période 2009-2011). Echantillon d'entraînement : 1215 entreprises. Echantillon de test : 10940.

La majorité des algorithmes atteignent des performances équivalentes dans la détection des irrégularités. La *précision* indique le niveau moyen d'exactitude, lorsqu'un modèle détermine une irrégularité. Par exemple, une précision de 75% signifie qu'en moyenne, les recommandations de contrôle proposées par le modèle seront exactes dans 75% des cas. L'*AUC (Area Under (ROC) Curve)* évalue la capacité du modèle à attribuer un score plus important à la présence effective d'une irrégularité qu'à son absence effective, lorsqu'une observation (une déclaration de cotisations) est choisie au hasard. La valeur maximale de l'*AUC* est de 1. Dans ce cas, le classifieur (le modèle) est parfait et détecte toutes les irrégularités. Une valeur de 0.8 indique que la probabilité, d'attribution par le modèle, d'un score plus grand à la présence effective d'une irrégularité est de 80%. A la différence de la précision, l'*AUC* prend en compte toutes les situations pour évaluer les capacités d'un modèle. L'erreur de test mesure l'erreur commise dans la distinction entre présence et absence d'irrégularité. Par exemple, une erreur de test de 0.2 indique que le modèle échoue dans 20% des cas à distinguer correctement l'absence de la présence d'irrégularité. l'erreur de test est, généralement, accompagnée d'autres mesures car elle dépend fortement de la distribution des irrégularités (les cas positifs) dans les déclarations évaluées. En particulier, si la proportion d'irrégularités est petite dans l'échantillon de test, un modèle peut avoir une très bonne erreur de test et échouer, plus largement, à détecter les cas positifs.

Les *SVM* (*Support Vector Machines*) et la méthode des "k plus proches voisins" sont les modèles pour lesquels l'apprentissage est le plus problématique. Pour le type de données soumis, caractérisées par de nombreux zéros ($> 80\%$ des données) et beaucoup de variables (> 1000), l'optimisation des paramètres d'un modèle est une des pistes d'amélioration des performances. Pour le modèle *GBM* (*Gradient Boosting Machines*), l'optimisation est générique (les paramètres proposés ne sont pas spécifiques aux données) et permet des gains importants. De manière générale, nous avons observé que les modèles ensemblistes,

en particulier ceux associés à des arbres de décision, obtenaient les meilleurs résultats sur les données de cotisations sociales. Notons que certaines méthodes, comme le classifieur naïf de Bayes, ne sont pas fonctionnelles car il peut arriver que des variables ne soient constituées que de zéros dans l'échantillon d'apprentissage. L'aspect algorithmique devient alors beaucoup plus fondamental. Cela est, par exemple, le cas lorsqu'on souhaite effectuer un apprentissage incrémental, dont les principaux avantages et particularités sont une réduction importante des temps de calcul et la prise en compte d'un historique de données de façon transparente. Les forêts uniformément aléatoires incrémentales bénéficient de la mémoire des apprentissages précédents et possèdent virtuellement plus d'exemples que les autres algorithmes.

`L'apprentissage incrémental`

Afin de mieux préciser les bénéfices d'un apprentissage incrémental, nous en précisons quelques aspects :
- un historique des exemples sur plusieurs périodes (ou par paquets) doit être disponible.
- à chaque période, la distribution des exemples peut avoir changé ;
- le volume total des données est généralement très important et leur apprentissage ne peut se faire en une passe ;
- la période courante est celle que l'on évalue, et la question posée est celle du choix à faire quant à l'historique des exemples disponibles.

Ce choix influence directement les performances d'un algorithme. Plus il dispose d'exemples, plus il est performant. Deux conditions sont nécessaires : les exemples doivent être identiquement distribués et les capacités de calcul ne doivent pas être limitées. Un algorithme incrémental lève ces contraintes en effectuant l'apprentissage au fur et à mesure de l'arrivée des nouvelles données. En contrepartie, l'algorithme doit être conçu de façon à ce que l'ajout de nouveaux exemples ait approximativement le même effet sur les performances qu'un apprentissage s'effectuant en une seule passe sur le volume total (ou une grande partie) des données. Pour cette raison, peu de modèles disposent d'une version incrémentale et parmi ceux que nous avons donné en exemple, seules les forêts aléatoires de Breiman et le modèle GBM en possèdent une implémentation, à notre connaissance.

Nous illustrons ci-dessous les résultats d'un apprentissage incrémental. L'échantillon de test correspond à 90% des entreprises contrôlées pour le compte de la période 2009-2011. L'échantillon d'apprentissage est constitué des 10% restants et des entreprises contrôlées pour le compte de la période 2008-2010. Pour chaque modèle, le nombre d'arbres est (au total) de 1000 et les paramètres ne changent pas relativement à l'exemple précédent. Nous souhaitons connaître les performances d'un modèle effectuant un apprentissage incrémental par période (2008-2010, puis 2009-2011) comparativement à un apprentissage pour toutes les périodes (dit offline), plus coûteux. L'apprentissage incrémental nécessite d'adapter le seuil (la probabilité) à partir duquel la présence d'une irrégularité est décidée. Autrement, les algorithmes tendent, tous, à privilégier les cas majoritaires. Pour chaque modèle incrémental et pour rendre la comparaison équilibrée, nous corrigeons le seuil de façon à ce que le nombre d'irrégularités dans le modèle corresponde approximativement au nombre réel d'irrégularités dans l'échantillon de test. Notons que cette procédure ne

requiert pas la disponibilité des données de test et peut être automatisée.

Models	Test error	Precision	AUC
Random Forests (offline)	0.2151	73.43%	0.7766
Random Forests (incremental)	0.2256	72.01%	0.7658
GBM (offline, optimized)	0.1869	77.17%	0.804
GBM (incremental, optimized)	0.1982	75.18%	0.795
random uniform forests (offline)	0.1837	78.42%	0.8059
random uniform forests (incremental)	0.2058	74.5%	0.7863

TABLE 5.4 – Apprentissage (incrémental) de la détection d'irrégularités période après période contre apprentissage (offline) en une passe des deux périodes.

La comparaison entre modèles incrémentaux bénéficie moins aux forêts aléatoires de Breiman, alors que le modèle GBM, dans sa version optimisée, en tire le mieux parti. L'intérêt des forêts uniformément aléatoires est essentiellement dû à l'absence d'optimas locaux. La comparaison avec les versions *offline* montre un écart, de 2 à 3 points avec les versions incrémentales et tous les résultats sont, ici, meilleurs que dans le tableau précédent. L'ajout de données supplémentaires améliore bien la qualité de l'apprentissage en contrepartie d'un temps de calcul plus élevé dans le cas *offline*. Il est tentant d'ajouter un historique plus long par le biais des autres périodes disponibles. Dans la pratique, les contraintes sont d'abord liées aux capacités de calcul. Celles-ci peuvent être limitées rapidement, lorsque le modèle est industrialisé et utilisé à sa pleine mesure. Cependant, le risque principal est l'absence de maîtrise de la distribution des exemples. Un paradigme de ce risque sont les mesures de réduction de cotisation qui sont modifiées régulièrement et demeurent une des principales sources d'irrégularités ou d'erreurs. Si des mesures de réduction disparaissent ou voient leurs taux et/ou assiettes modifiés, les informations des périodes trop lointaines peuvent devenir obsolètes et introduire un biais dans la détection. La version incrémentale des forêts uniformément aléatoires réduit ce risque en permettant la comparaison systématique de la *mémoire* (des apprentissages précédents) et de l'apprentissage de la période courante. Une fois construit, un arbre n'est jamais modifié et lorsque la taille de la forêt devient trop importante, un certain nombre d'entre eux peut être supprimé ou sélectionné.

5.6 Résultats en laboratoire

Les résultats validés en laboratoire sont ceux qui servent de référence pour n'importe quelle évaluation comptant pour la même année que celle dont on modélise la détection des irrégularités aux cotisations sociales. Pour l'année 2013, nous disposons, au minimum, des données déclaratives et des résultats des contrôles de la période 2010-2011. L'apprentissage est effectué sur l'échantillon constituant ces exemples. Les différentes erreurs mesurées, grâce aux données OOB, déterminent les garanties fournies par le modèle. A ces mesures sont ajoutés les facteurs qui permettent une meilleure interprétation des résultats. Nous récapitulons les différentes étapes du processus établi en laboratoire.

Trois questions sont fondamentales :
- peut-on fournir des garanties sur le nombre de cas d'irrégularités prédit ?
- Peut-on estimer correctement le montant de redressement moyen pour l'ensemble des irrégularités qui pourraient être détectées sur la base des recommandations du modèle ?
- Peut-on répondre à ces deux questions avant que les contrôles n'aient lieu ?

`Erreurs de référence`

La première étape est la détermination des erreurs. Il s'agit d'estimer pour n'importe quel échantillon de test plus grand que l'échantillon d'apprentissage, les erreurs qui ne devraient pas être dépassées. De cette manière, la situation est fixée : *les erreurs OOB doivent constituer des bornes supérieures des résultats qui seront obtenus au moment de l'expérimentation opérationnelle.* Nous indiquons ci-dessous les différentes erreurs estimées à partir de l'échantillon d'apprentissage constitué des 4069 entreprises contrôlées et sélectionnées pour le compte de la période 2010-2011. Le nombre d'arbres est fixé à 200 et le nombre de variables candidates (paramètre *mtry*) à la construction de chaque région d'un arbre est fixé à 800, pour la classification. Tous les autres paramètres sont ceux par défaut. Dans le cas de la régression, la variable à prédire est le rapport entre le montant net (comptable) redressé et la masse salariale cumulée des deux années de cotisation (2010 et 2011). L'intervalle de confiance bootstrap est calculé à partir de la procédure décrite dans la section 5.5.2.

Résultats: *erreurs de référence estimées pour la détection des irrégularités aux cotisations sociales déclarées en 2010 et 2011 et "contrôlables" en 2013.*

```
Out-of-bag (OOB) evaluation
OOB estimate of error rate: 20.79%

OOB estimate of AUC: 0.7917

OOB confusion matrix:
          Reference
Prediction    0    1 class.error
         0 1883  358      0.1598
         1  488 1340      0.2670

Theorical (Breiman) bounds
Prediction error (expected to be lower than): 20.83%
Upper bound of prediction error: 34.83%
Trees average correlation: 0.071
Strength (margin): 0.5042
Standard deviation of strength: 0.2976
```

Erreurs de référence pour l'estimation des montants de redressement :

```
Out-of-bag (OOB) evaluation
Mean of squared residuals: 9e-04

OOB residuals:
       Min         1Q     Median       Mean         3Q        Max
-0.3956000 -0.0007314  0.0006400  0.0005721  0.0056550  0.2483000

Variance explained: 46.05%

Theorical (Breiman) bounds:
Theorical prediction error: 0.000915
Upper bound of prediction error: 0.000922
Mean prediction error of a tree: 0.001985
Average correlation between trees residuals: 0.4647
Expected squared bias (experimental): 2e-06

OOB 99% Bootstrap Confidence Interval
for the true mean of 'irregularity level': [0.178%, 2.967%]*.
Estimate of the true mean of 'irregularity level': 0.972%

OOB 99% Bootstrap Confidence Interval
for the true mean of positive values of 'irregularity level': [0.896%, 3.665%]**.
Estimate of the true mean of positive values for 'irregularity level': 2.26%
```

```
Prediction intervals :
OOB probability of being over upper bound (99.5%)
of prediction level for the true value of 'irregularity level': 0.109
OOB probability of being under lower bound (0.5%)
of  prediction level for the true value of 'irregularity level': 0.066
```

A la différence des modèles déterministes, les erreurs estimées dans les modèles aléatoires peuvent augmenter ou diminuer pour plusieurs apprentissages du même échantillon. Les erreurs et mesures OOB fournissent des estimations des erreurs attendues au moment du test. Elles sont meilleures que celles de l'exemple précédent, essentiellement grâce à la plus grande taille de l'échantillon. Seule l'erreur quadratique moyenne est moins satisfaisante, bien que la variance expliquée par le modèle augmente de manière importante.

Les erreurs indiquées ici signifient que si nous utilisons explicitement ce modèle, alors pour n'importe quel échantillon de test, de taille supérieure à l'échantillon d'apprentissage et pour la détection d'irrégularités, les erreurs effectivement observées ne seraient supérieures à celles présentées ci-dessus qu'avec une probabilité petite, et connue. Pour l'estimation des montants de redressement, les bornes des intervalles de confiance du niveau d'irrégularités sont celles qui ne devraient être dépassées qu'avec une probabilité faible.

Les points importants dans la régression sont l'intervalle de confiance associé au niveau moyen d'irrégularités (*) et celui associé aux valeurs positives de ce même niveau (**).
- Dans le premier cas, l'intervalle de confiance tient compte des erreurs qui seront commises par le modèle au moment des recommandations de contrôle. C'est donc un intervalle qui mesure le rapport entre le montant de redressement estimé et la masse salariale (cumulée sur deux années, 2010 et 2011) de chaque entreprise recommandée. La moyenne de ce rapport est estimée à 0.972%, ce qui signifie, en simplifiant et du point de vue du modèle, que le montant net des redressements serait, en moyenne, d'environ 1% de la masse salariale (cumulée sur 2 ans) de toutes les entreprises, redressées ou non. Toutefois, une telle généralisation serait hâtive puisque l'évaluation complète des entreprises n'est, à ce stade, pas encore réalisée.

- Dans le second cas, l'intervalle mesure le rapport entre les montants de redressement estimés, pour les entreprises qui présenteraient de manière effective des irrégularités, et leur masse salariale (cumulée sur deux années, 2010 et 2011). Ici, on ne s'intéresse qu'aux succès du modèle. L'estimation du niveau moyen d'irrégularités est alors de 2.26%. Ce point de vue indique qu'en moyenne, que les montants de redressement estimés correspondraient à un peu plus de 2% de la masse salariale des entreprises redressées.

- *L'élément fondamental dans l'estimation de chaque montant de redressement est la borne inférieure de son intervalle de prédiction.* De manière plus précise, nous faisons une distinction entre l'intervalle de confiance d'un moyenne, ici le niveau moyen d'irrégularités, et les intervalles de prédiction pour chaque montant de redressement (estimé pour chaque entreprise). En effet, il convient d'effectuer des recommandations qui se verront largement réalisées, du point de vue des enjeux financiers. Pour cela, on expose chaque niveau d'irrégularité (de chaque entreprise) effectivement observé dans les contrôles avec

sa borne inférieure estimée par le modèle grâce aux données OOB. Malgré un niveau de confiance de 99%, on observe dans 6.6% des cas une valeur du montant de redressement plus petite que la borne inférieure estimée par le modèle. On peut, néanmoins, réduire le pourcentage en consentant beaucoup plus d'efforts à la construction de la forêt. Nous avons comparé notre procédure d'intervalle de prédiction bootstrap à la méthode *bootstrap percentile* et au bootstrap standard, disponibles dans le paquet *boot* du logiciel R : les bornes inférieures y sont dépassées (respectivement) 3 et 2 fois plus fréquemment. Nous avons essayé également la méthode BCa (dans le même paquet), mais elle a systématiquement renvoyé une erreur. Nous indiquons ici que les intervalles de prédiction (pour chaque valeur d'une variable) fournissent des arguments adaptés qui permettent d'étendre et de conforter l'analyse par un intervalle de confiance (pour la moyenne d'une variable).

- Dans le cas de la détection pure (absence ou présence d'irrégularités), l'*erreur de test* estimée est de 20.79%. La *précision* vaut 73.70% (1 - 0.2670) et donne le taux de détection des irrégularités attendu. Plus simplement, la valeur de la précision signifie que le modèle doit être en mesure de prédire les cas d'irrégularités avec un taux de succès, au minimum, d'un peu plus de 73%. L'AUC vaut 0.7917 et estime la probabilité qu'un plus grand score soit attribué à une irrégularité effective qu'à une non-irrégularité effective, si des déclarations de cotisation sont soumises, au hasard, au modèle.

Dans les faits, ces trois mesures sont décisives. Une erreur de test petite indique que le modèle est pertinent sur, au minimum, une partie du problème posé. Une précision importante indique que le modèle génère peu de faux-positifs. Une aire sous la courbe ROC (AUC) élevée signifie que le modèle distingue avec une exactitude du même niveau, l'absence de la présence d'irrégularité, ce qui permet d'envisager sa généralisation.

Les principaux résultats en laboratoire peuvent être résumés en quelques arguments :
- la capacité de détection des irrégularités d'un modèle peut (et doit) être déterminée avant que les contrôles recommandés (par ce même modèle) ne soient effectivement menés.
- Plusieurs méthodes et outils existent, cependant les modèles ensemblistes font partie de ceux présentant des propriétés adaptées. L'optimisation de leurs paramètres est généralement facultative et ils ne requièrent que peu (ou pas) d'hypothèses sur les données.
- La détection des irrégularités aux cotisations sociales se traduit par un problème de classification binaire plus adapté aux modèles non linéaires.
- Les forêts uniformément aléatoires permettent d'améliorer, systématiquement, les résultats des contrôles de l'URSSAF, à la fois par leurs capacités de détection et par la somme des montants nets redressés. Elles complètent et généralisent ces résultats.
- L'évaluation du montant total net de redressements dépasse très largement les sommes redressées chaque année par l'URSSAF et, de notre point de vue, peut être considérée comme un instrument de la réduction du déficit de la Sécurité sociale.

Nous discutons plus précisément des ces deux derniers points dans la dernière partie de la thèse, consacrée à l'évaluation complète de toutes les entreprises d'Île-de-France. Avant de l'aborder, nous nous consacrons dans les lignes qui suivent à la recherche des performances optimales du modèle, puis à l'interprétation des résultats de l'algorithme.

Performances optimales

Nous souhaitons maintenant utiliser l'ensemble des exemples à notre disposition et non plus ceux des années 2010 et 2011. Pour cela, nous considérons dans un unique échantillon toutes les déclarations et résultats de contrôle de 2006 à 2011. Le risque pris, relativement aux résultats obtenus précédemment, est que nous ne savons pas si la relation entre ces contrôles et, par exemple, ceux qui ont été réalisés en 2012 (pour la vérification des déclarations effectuées de 2009 à 2011) n'aura pas changé. Dans la pratique, il convient de le vérifier à plusieurs étapes de la modélisation.

Résultats: *performances optimales des erreurs de référence estimées pour la détection des irrégularités aux cotisations sociales déclarées en 2010 et 2011 et "contrôlables" en 2013.*

```
Out-of-bag (OOB) evaluation
OOB estimate of error rate: 15.49%

OOB estimate of AUC: 0.8387

OOB confusion matrix:
          Reference
Prediction     0     1 class.error
         0 31977  4534      0.1242
         1  4733 18573      0.2031

Theorical (Breiman) bounds
Prediction error (expected to be lower than): 16.05%
Upper bound of prediction error: 31.36%
Average correlation between trees: 0.0776
Strength (margin): 0.5709
Standard deviation of strength: 0.3197
```

La réduction des erreurs est d'environ 25%, tandis que le nombre d'observations est multiplié par 15. Il doit augmenter considérablement pour espérer atteindre des gains importants. Les données sont cependant disponibles et une précision de, presque, 80% dans la détection des irrégularités est atteinte. Surtout, nous n'avons pas constaté de dégradation de performance (sur des échantillons de test) liée à l'utilisation de tout l'historique. De même, la plupart des méthodes profitent de l'augmentation du nombre d'observations.

5.6.1 Importance des variables et visualisation

Une critique souvent associée aux forêts aléatoires est le manque d'outils pour les interpréter. Les exemples sont fournis à l'algorithme qui en construit un modèle puis renvoie un résultat (au moment du test) sans explication sur sa relation avec les variables du problème. Les forêts aléatoires de Breiman proposent des alternatives comme des mesures de l'importance des variables ou de dépendance partielle (entre variables explicatives et

variable à prédire). Les forêts uniformément aléatoires implémentent ces outils et en proposent d'autres, que nous illustrons dans les lignes qui suivent.

`Visualisation des erreurs`

La premier outil est la visualisation de l'erreur de prédiction en fonction du nombre d'arbres défini dans l'algorithme. On peut ainsi savoir comment évolue l'erreur lorsqu'on augmente le nombre de modèles de base.

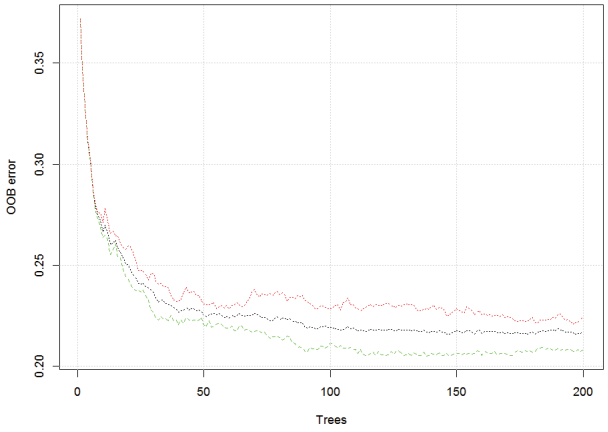

FIGURE 5.1 – Evolution de l'erreur de test OOB de la régression en fonction du nombre d'arbres de décision de la forêt uniformément aléatoire.

Pour la détection des irrégularités, l'erreur OOB décroît à mesure que le nombre d'arbres augmente et se stabilise à partir de 100 arbres. Pour un seul arbre (uniformément aléatoire), l'erreur de prédiction est assez élevée et beaucoup plus, par exemple, que pour un arbre de type CART. Pour la réduire, l'effort porte principalement sur la réduction de la corrélation entre les arbres.

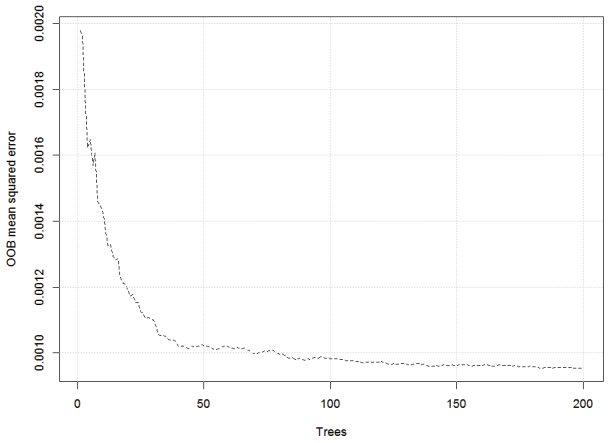

FIGURE 5.2 – Evolution de l'erreur quadratique moyenne OOB de la régression en fonction du nombre d'arbres de la forêt uniformément aléatoire.

Ce principe de *décorrélation* est repris pour la régression. Notons que, comme pour la classification, l'erreur produite par la forêt est d'environ la moitié de celle produite par un arbre, ce qui est fortuit. Nous poursuivons notre illustration en intégrant les différents outils présentés dans la section 5.4.

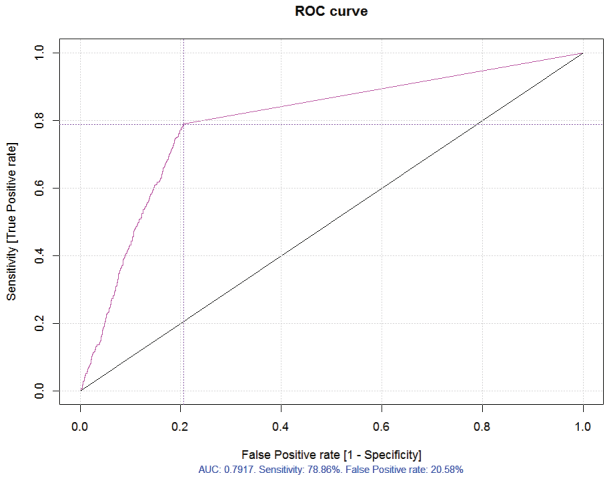

FIGURE 5.3 – Courbe ROC et évaluation OOB du modèle.

La courbe *ROC* montre la capacité du modèle à détecter l'ensemble des irrégularités (sensibilité) en fonction du taux de faux-positifs (le nombre d'irrégularités détectés à tort relativement au nombre total d'irrégularités). La ligne diagonale est l'expression d'un classifieur qui ferait aussi bien que le hasard. Il aurait, en moyenne, une chance sur deux de trouver une irrégularité.

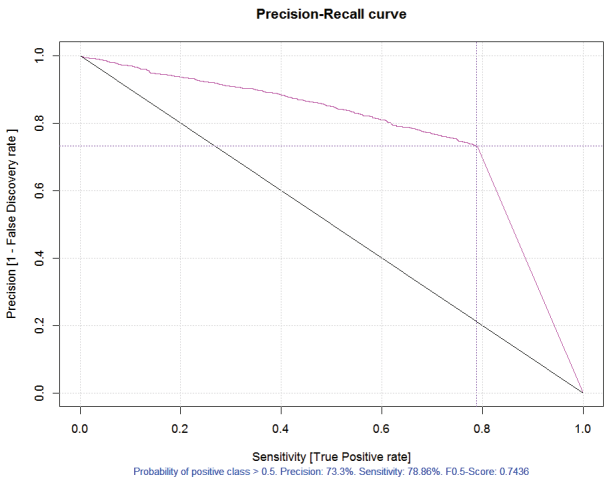

FIGURE 5.4 – Courbe de précision-rappel de la détection des irrégularités aux cotisations sociales.

La courbe de précision-rappel (ou précision-sensibilité) correspond à l'évolution du niveau d'exactitude, moyen, de la seule prédiction des irrégularités en fonction de la capacité du modèle à détecter l'ensemble des irrégularités. Plus le nombre d'irrégularités est grand, plus la précision diminue jusqu'à une limite, représentée par le point correspondant au "coude" de la fonction. Cette limite est la valeur de la précision qui est retenue pour un modèle.

`Importance des variables`

L'influence des variables explicatives est une problématique récurrente dans les modèles prédictifs. Mesurer l'importance des variables permet généralement une meilleure compréhension du problème et peut, dans certaines situations, améliorer les performances. Les forêts uniformément aléatoires proposent plusieurs mesures d'importance.

i) La première méthode mesure l'importance globale des variables explicatives. Elle est fondamentalement liée à l'algorithme et repose sur la fréquence d'observation et l'intensité, relativement à un critère d'optimisation, des variables qui participent à la construction de chaque région d'un arbre. L'importance globale des variables répond à la question suivante : *Quelles sont les variables les plus influentes de la problématique analysée ?* Lorsque le nombre de variables devient important et lorsqu'elles n'ont chacune qu'une influence limitée, l'interprétation nécessite de dépasser le simple cadre de l'importance des variables. Nous ne produisons pas le graphique issu de la mesure d'importance globale mais en indiquons une synthèse : sur la relation liant les irrégularités aux déclarations de cotisations, aucune variable n'a une influence globale supérieure à 3%. Notons que des divergences sur l'intensité de l'influence des variables peuvent apparaître entre différents modèles, en particulier lorsque les variables sont nombreuses.

ii) Nous utilisons alors une seconde méthode, définissant l'importance locale des variables. Elle est issue de leurs interactions et de leurs capacités à différencier l'absence de la présence d'irrégularités. L'importance locale des variables répond à la question sui-

vante : *Quelles sont, à la fois, les variables les plus influentes et celles qui différencient le plus distinctement les résultats escomptés ?*

L'importance locale procure plusieurs avantages :
- elle peut être utilisée sur l'échantillon d'apprentissage ou sur l'échantillon de test ;
- elle est beaucoup moins sensible au nombre de variables ;
- le caractère explicatif du modèle est mieux pris en compte.

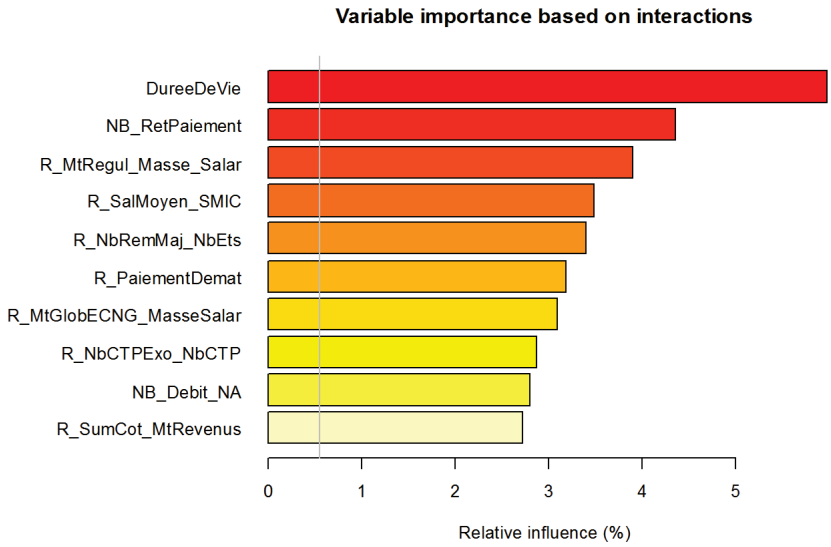

FIGURE 5.5 – Influence (locale) des variables les plus importantes dans la détection des irrégularités aux cotisations sociales.

Le graphique ci-dessus représente les 10 variables dont les interactions avec d'autres sont les plus importantes dans le modèle. Le préfixe "R" indique que la variable considérée est issue d'un ratio de deux variables, séparées dans l'intitulé par un tiret bas "_". La variable la plus influente est la *durée de vie* de l'entreprise. Puis, vient le *nombre de retards de paiement* (NB_RetPaiement) de cotisations, lesquelles sont dues à échéances fixes, mensuelles ou trimestrielles.

Le *montant des régularisations* (R_MtRegul_Masse_Salar) de cotisations dues (relativement à la masse salariale), le niveau du *salaire moyen brut relativement au salaire minimum* (R_SalMoyen_SMIC), et le *nombre de remises sur majorations sur cotisations* relativement au nombre d'établissements de l'entreprise (R_NbRemMaj_NbEts) constituent d'autres variables influentes. L'intensité mesurée peut sembler insuffisante au regard de l'interprétation, cependant l'influence relative indiquée l'est comparativement à plus de 1000 variables candidates.

Les sixième et septième variables sont le *nombre de paiements dématérialisés* (R_PaiementDemat) et le *montant global des écarts entre cotisations attendues et cotisations reçues* (R_MtGlobECNG_MasseSalar). Les trois dernières variables sont liées,

respectivement, au *nombre de mesures de réduction*, relativement au nombre de catégories de cotisation déclarées par l'entreprise (R_NbCTPExo_NbCTP), au *nombre de catégories débitrices d'origine non déterminée* (NB_Debit_NA) et au *montant total des cotisations* de l'entreprise, relativement à l'ensemble des revenus d'activité versés aux salariés (R_SumCot_MtRevenus).

Les interactions totales des 10 variables représentent environ un tiers de celles de l'ensemble. Le trait vertical illustre le niveau en-dessous duquel le niveau d'interaction d'une variable avec les autres serait probablement dû au hasard. Nombre de variables ont d'abord un lien avec la politique de recouvrement de l'URSSAF. Celles mesurant les anomalies dans les paiements des cotisation semblent les plus décisives en matière d'interactions. Un second groupe illustre la situation propre de l'entreprise comme le niveau moyen de rémunération ou la durée de vie. Un dernier groupe fait référence à la nature des cotisations versées, à travers la présence de mesures de réduction et du taux global de cotisation.

iii) Nous n'avons cependant aucune information sur la manière dont ces variables agissent sur l'absence ou la présence d'irrégularités. Pour y répondre, nous introduisons une mesure d'importance des variables relativement à chacune des classes du problème. Elle répond à la question suivante : *parmi les variables influentes, quelles sont les plus discriminantes ?*
En particulier ici, l'information essentielle est l'explication de la présence d'irrégularités.

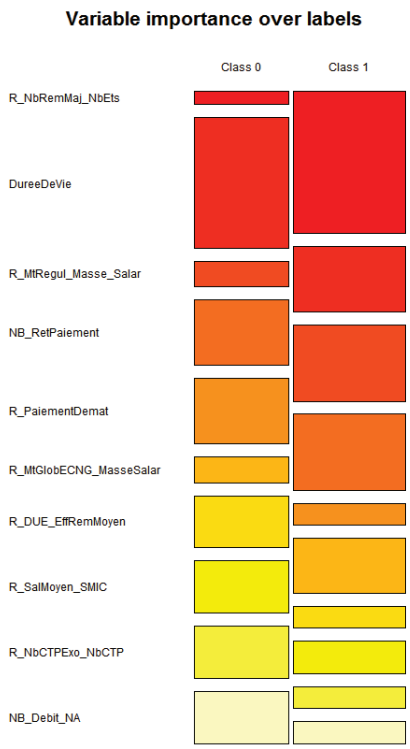

FIGURE 5.6 – Importance des variables relativement à l'absence ou la présence d'irrégularités.

Les classes ("Class 0" et "Class 1") correspondent à l'absence et à la présence d'irrégularités. Les variables sont présentées par ordre décroissant d'influence sur toutes les classes et sont identiques (sauf une) à celles identifiées dans le graphique précédent. Chaque rectangle est proportionnel à l'influence de la variable sur la classe correspondante et l'intensité de la couleur indique l'importance de la variable. Le *nombre de remises sur majorations de cotisations* apparaît, paradoxalement, comme la variable la plus discriminante entre les deux classes. Le *montant des régularisations* et le *montant des écarts entre cotisations attendues et cotisations reçues* sont également dans cette situation. Leur faible influence dans le graphique précédent relativise fortement d'éventuelles conclusions. L'explication principale de l'influence des variables relativement à l'absence ou à la présence d'irrégularités réside dans le petit nombre de variables caractéristiques d'une déclaration présentant des irrégularités. Plus précisément, la présence d'irrégularités s'explique d'abord par des anomalies de cotisations, déterminées par la politique de recouvrement de l'URSSAF, et non par les variations intrinsèques des cotisations versées.

`Dépendances partielles et co-influence`

iv) Nous poursuivons l'analyse en explorant la *dépendance partielle* de chaque classe avec une variable identifiée comme influente. La dépendance partielle représente la relation qui unit une variable à une classe (ou à la variable à prédire, dans le cas de la régression), sachant toutes les valeurs prises par les autres variables du problème. Dans le cas de la classification, elle exprime, à posteriori, la distribution marginale de la variable considérée, relativement à chaque classe de la variable à expliquer. Prenons un exemple. Le *nombre de remises sur majorations sur cotisations* est une variable (relativement) influente et discriminante de la présence ou de l'absence d'irrégularité. Nous souhaitons connaître la traduction de ces deux propriétés (l'influence et le caractère discriminant) pour toutes les observations d'un échantillon (d'apprentissage ou de test). La dépendance partielle donne la plage de valeurs sur laquelle la variable est influente, ainsi que le seuil (ou la zone) à partir duquel elle discrimine chacune des classes. Elle répond à la question suivante : *quelle est la zone d'influence d'une variable et comment se répartit-elle selon la classe considérée ?*

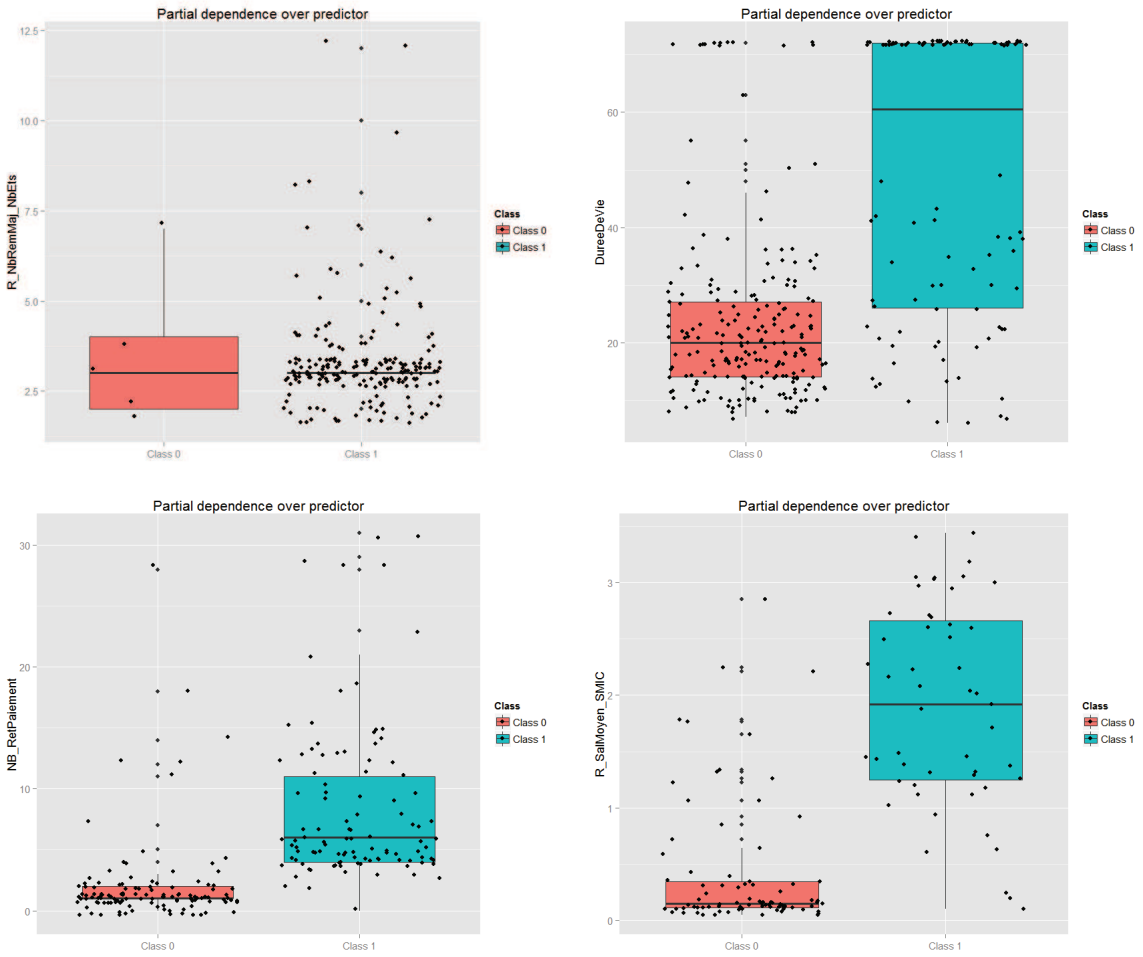

FIGURE 5.7 – Dépendances partielles de quatre variables explicatives relativement à chaque classe du problème.

Le *nombre de remises sur majorations sur cotisations* (R_NbRemMaj_NbEts) est représenté sur le premier graphique (à gauche). Chaque rectangle est une "boite à moustaches" (box plot) allant du premier (bord inférieur horizontal) au troisième quartile de la distribution des observations, entrecoupé de la médiane. Les segments (verticaux) prolongeant les rectangles ont pour extrémités, inférieure et supérieure, les 5^e et 95^e centiles de la distribution. Au-delà, les points correspondent aux valeurs extrêmes ou atypiques.

Le graphique de dépendance partielle du *nombre de remises sur majorations sur cotisations* montre que son caractère discriminant est essentiellement dû à l'existence de nombreuses valeurs atypiques, lorsqu'une irrégularité est présente. Quand ce n'est pas le cas, trop peu d'observations caractérisent la variable. Son influence sur la présence d'irrégularités s'exerce à partir de trois remises accordées (sur les deux dernières années) après que l'entreprise a subi des majorations de cotisations. A contrario, la *durée de vie*, le *nombre de retards de paiement* ou le niveau du *salaire moyen* discriminent les deux classes. Pour chacune des variables, des valeurs élevées sont un signe d'une possible présence d'irrégularités dans la déclaration de l'entreprise. L'interprétation de la dépendance partielle est intimement liée à l'intensité de l'influence locale des variables et le carac-

tère explicatif des variables se traduit ici par une conjonction de leurs effets. En d'autres termes, plus l'influence d'une variable est faible plus on a besoin de variables supplémentaires pour expliquer un même effet en utilisant la dépendance partielle. Cet élément est la principale contribution de la dépendance partielle : la présence d'irrégularités est plus probable lorsqu'il existe une conjonction de niveaux élevés pour les variables liées à la politique de recouvrement de l'URSSAF et semble plus effective pour les entreprises dont la durée de vie est longue.

v) Nous souhaitons, à présent, étudier la dépendance entre les variables explicatives influentes, conditionnellement au fait qu'elles partagent la même classe ou non. Il s'agit de connaître, à la fois, la relation entre les variables explicatives, prises deux à deux, relativement à la modélisation du problème et la manière dont cette dépendance agit sur la présence ou l'absence d'irrégularités. Ce type de dépendance répond à la question suivante : *comment la dépendance entre deux variables explicatives influe-t-elle sur les différentes classes du problème ?*
La dépendance est d'abord calculée à travers toutes les interactions entre les deux variables explicatives, pour chaque observation. On obtient ainsi une mesure d'influence réciproque, équivalente à une corrélation, et la zone d'influence des deux variables sur les classes du problème.

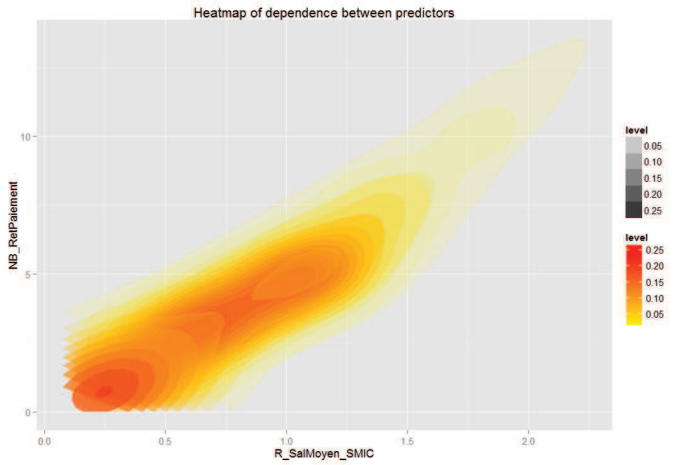

FIGURE 5.8 – Zone d'influence de la dépendance entre le niveau du salaire moyen et le nombre de retards de paiement.

Sur le graphique ci-dessus, nous avons choisi d'illustrer la co-influence du *nombre de retards de paiement* (en ordonnées) et du *salaire moyen* sur l'absence ou la présence d'irrégularités. La mesure de dépendance entre les variables est d'environ 0.13 (pour un maximum de 1). Leur corrélation est donc faible et les zones les plus foncées indiquent là ou leur co-influence est la plus importante. Elle a, surtout, un impact lorsque la dépendance est importante. Lorsque ce n'est pas le cas, la dépendance indique des effets locaux. Pour les mesurer, nous décomposons la zone de co-influence ainsi :
- les classes sont redéfinies selon qu'elles sont identiques ou non ;
- pour chaque variable, la distribution des classes est représentée.

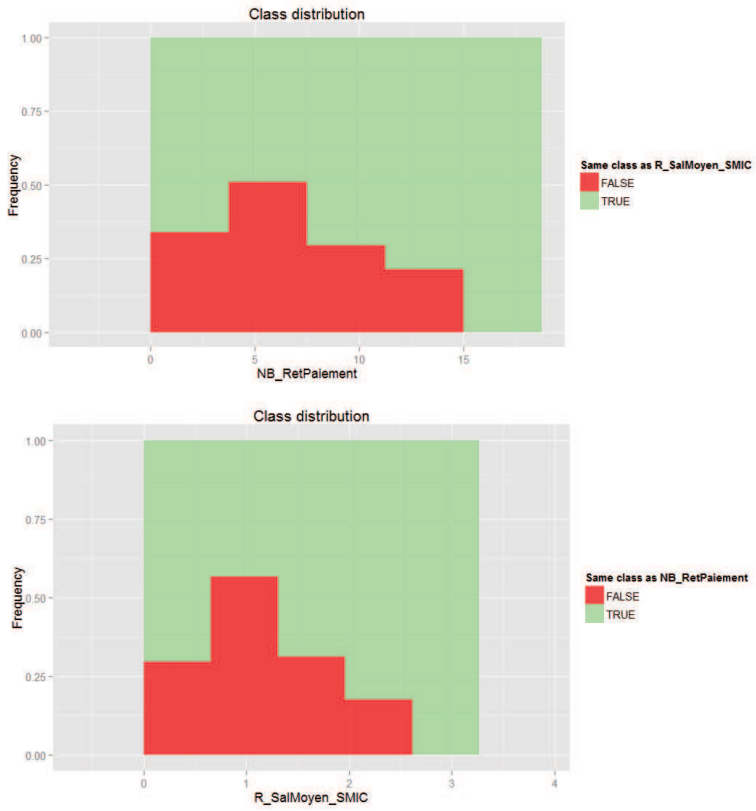

FIGURE 5.9 – Distribution d'une même classe sur la zone de co-influence du nombre de retards de paiement et du niveau de salaire moyen.

Chacun des graphiques identifie les zones ou les variables partagent la même classe et celles où cela n'est pas le cas. La dépendance partielle du *nombre de retards de paiement* obtenue précédemment (figure 5.7) nous indique qu'en dessous de quatre ou cinq retards de paiement, cette variable ne présente généralement pas de liens avec la présence d'irrégularité. La connexion avec les deux graphiques ci-dessus indique une tendance plus probable à l'absence d'irrégularités lorsque le nombre de retards de paiement a une valeur inférieure à quatre et lorsque le niveau du salaire moyen est plus petit que le SMIC. Plus les retards de paiement deviennent importants, plus une irrégularité est probable, en particulier lorsque le niveau du salaire moyen dépasse 1.5 SMIC.

Pour chacune des variables influentes, ce type d'analyse peut être mené et permet d'obtenir les zones dans lesquelles la co-influence de deux variables a un impact sur une classe particulière. Cet impact dépend à la fois de la dépendance entre les deux variables et du niveau d'intensité dans la zone de co-influence. Selon le cas, la conclusion apportée peut être fortement nuancée. Ici, le lien entre nombre de retards de paiement importants et salaire moyen élevé, sur la présence d'irrégularité, n'est pas principal (du fait que chacune des variables n'a qu'une influence faible).

`Interactions`

vi) La somme des effets locaux peut s'exprimer au moyen des interactions mutuelles entre toutes les variables. Les interactions fournissent un résumé de toutes les informations recueillies et répondent à la question suivante : *l'interprétation des variables influentes fournit-elle un cadre général explicatif de la problématique posée ?*
L'importance locale y apporte une première réponse par la somme des influences relatives des variables retenues. Les interactions décomposent cette somme sur toutes les variables et en donnent une représentation visuelle.

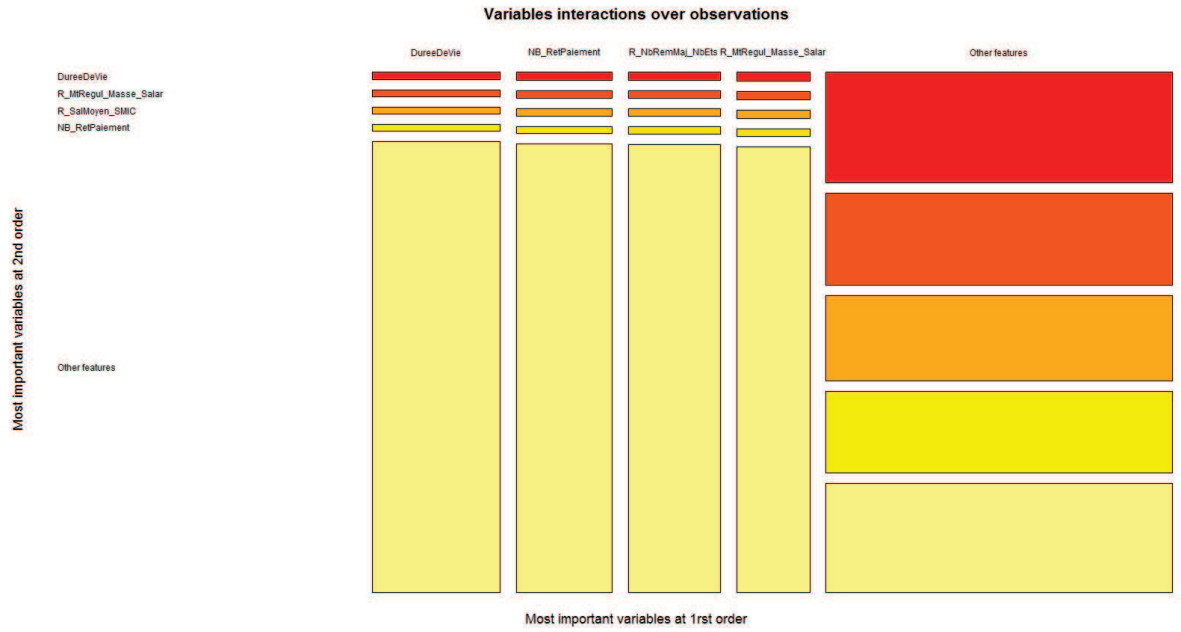

FIGURE 5.10 – Interactions entre toutes les variables.

Le graphique exprime les interactions (traduites par une mesure de dépendance) de toutes les variables au premier et second ordre. Le premier ordre signifie que la présence ou l'absence d'irrégularité peut d'abord s'expliquer par la variable considérée. Au second ordre, on suppose qu'il existe une autre variable, inconnue, expliquant le caractère recherché avant la variable considérée.
- Les variables sont désignées par ordre décroissant d'importance, à chaque ordre. "Others features" est une méta-variable regroupant l'influence de toutes les autres variables que celles mentionnées et l'aire totale du rectangle vaut 1.
- Un sous-rectangle représente la dépendance entre deux variables. Sa taille est proportionnelle à son influence sur le problème et sa couleur indique son degré relatif d'importance.

Les interactions fournissent une synthèse de l'interprétation. La présence d'irrégularités est explicable essentiellement par des anomalies (retards, régularisations, écarts de cotisation,...) dans les cotisations versées. La dépendance limitée entre variables explicatives, la somme de leurs influences et les effets observés dans les cas d'absence d'irrégularités indiquent la présence de situations spécifiques qui ne fournissent pas un cadre général d'explication. Cela est le cas lorsque des retards de paiement en nombre et des mon-

tants de régularisation importants interviennent conjointement à la présence de fortes rémunérations dans l'entreprise. La présence d'irrégularités est alors plus probable. Dans le graphique des interactions, la présence de situations spécifiques est généralement déterminée par le second ordre. Moins les variables y sont influentes (en les comparant à l'ensemble "Others features"), plus il est probable que la nature de la problématique soit issue de la juxtaposition de plusieurs phénomènes. Ici, on remarque en particulier l'absence des catégories de cotisation comme variables explicatives. Elles ont bien une influence, mais celle-ci est diluée par le grand nombre de zéros et de catégories. Une question posée est celle de la perturbation du modèle par ces nombreuses catégories. Nous les avons supprimées et avons mesuré l'importance des variables. Les variations n'ont lieu qu'à la marge et la situation ne change globalement pas. Cette absence de relation forte entre irrégularités et nature (ou montant) des cotisations, hormis par des anomalies, est ce qui caractérise en premier lieu les irrégularités aux cotisations sociales et explique que 40% des entreprises redressées le soient pour un montant inférieur à 1000 euros. Elle explique également la difficulté à détecter plus d'irrégularités par des méthodes plus classiques. Certains des modèles utilisés en apprentissage statistique ont l'avantage de n'être que peu sensibles au bruit contenu dans les données. Dans le cas des forêts uniformément aléatoires, le modèle reconstitue les liens faibles entre données et nature du problème et les ajoute les uns ou autres afin d'en déduire une combinaison aboutissant à une relation forte. La seule détection des irrégularités n'évacue pas la question de l'estimation des montants de redressement, laquelle nécessite la présence de l'ensemble des catégories de cotisation dont on espère un caractère explicatif plus important. Nous reprenons l'analyse précédente et adoptons une formulation plus compacte de la visualisation des résultats et de leur interprétation.

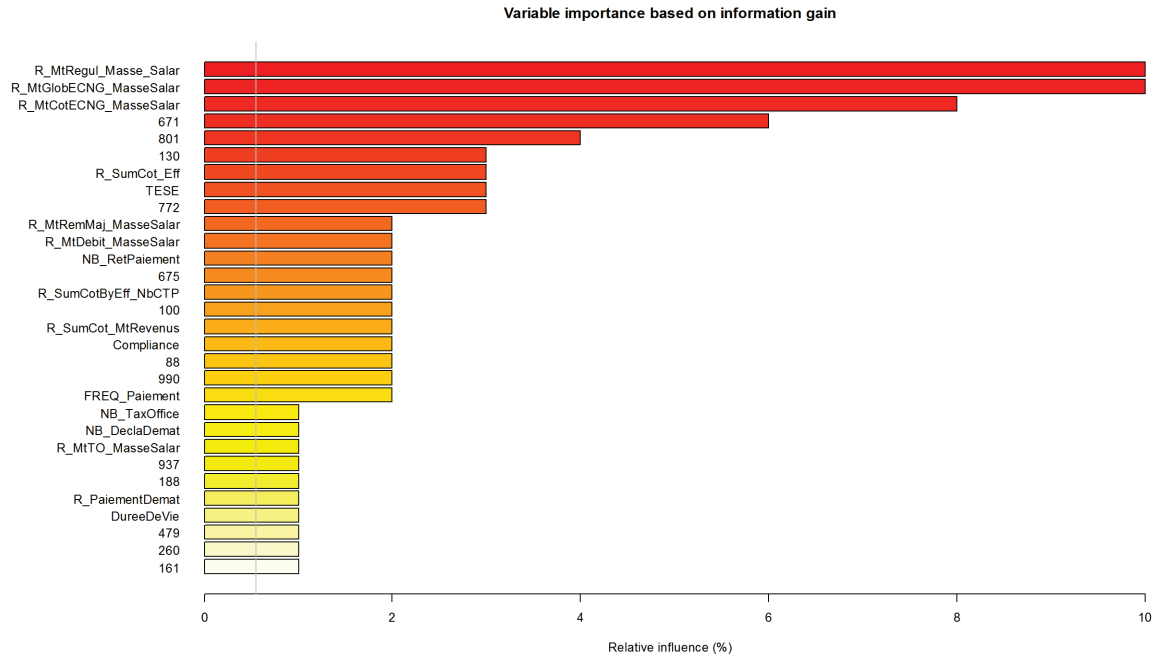

FIGURE 5.11 – Importance globale des variables pour l'estimation des montants de redressement.

L'importance globale des variables est beaucoup plus explicite dans le cas de la régression. Les 10 premières variables expliquent plus de 50% du lien avec les montants de redressement et les 30 premières en expliquent plus de 90%. Les variables désignées par un numéro correspondent aux codes-type de personnel (les catégories de cotisation) nommées explicitement ainsi. Un double effet est notable ici : la présence d'anomalies de cotisation (régularisations, écarts de cotisation) et l'expression explicite de leur localisation dans les catégories de cotisation. Par exemple, la variable "671", dite *"réduction des cotisations patronales sur les bas salaires"* ou réduction Fillon, est la principale mesure de réduction de cotisations (allègement de charges) et demeure une des principales sources d'irrégularités observées. Elle est, en particulier, associée à la variable "801" qui est une *"régularisation sur la réduction Fillon"*. La somme de leurs influences est supérieure à 10% de celle de toutes les variables et on note que la *réduction des cotisations patronales sur les bas salaires* aboutit à des irrégularités dont les régularisations conduisent elles-mêmes à de nouvelles irrégularités. Les trois variables les plus influentes sont, elles, toutes liées à des régularisations de cotisations ou des écarts entre cotisations reçues et attendues. Pour une définition des catégories de cotisation, nous renvoyons le lecteur vers le lien URSSAF qui suit : https ://fichierdirect.declaration.urssaf.fr/TablesReference.htm. La définition des autres variables est disponibles dans l'annexe du manuscrit.

De manière générale, l'importance globale des variables permet d'expliquer la quasi-totalité des montants de redressement par une situation liée soit à une anomalie de cotisation (régularisations, retards, taux de cotisation, conformité, dettes de cotisation, remises sur majorations,...), soit à des catégories spécifiques (mesures de réduction, "Cas général" de cotisation,...).

`Importance partielle`

vii) L'importance partielle est une mesure d'importance locale qui agit à la fois sur les variables les plus influentes et sur une caractéristique de la variable à expliquer. Elle répond à la question suivante : *quelles sont les variables explicatives les plus influentes, lorsqu'on assigne à la variable à expliquer une caractéristique spécifique ?*
Nous sommes précisément intéressés par les niveaux d'irrégularité (le rapport entre le montant redressé et la masse salariale de l'entreprise contrôlée) positifs et suffisamment importants. L'analyse effectuée dans le quatrième chapitre nous a permis de déterminer un seuil en dessous duquel la détection des irrégularités devient beaucoup plus problématique. Nous cherchons ici toutes les variables influentes, lorsque le niveau d'irrégularités est supérieur à 1% de la masse salariale de la période de contrôle considérée (2010-2011).

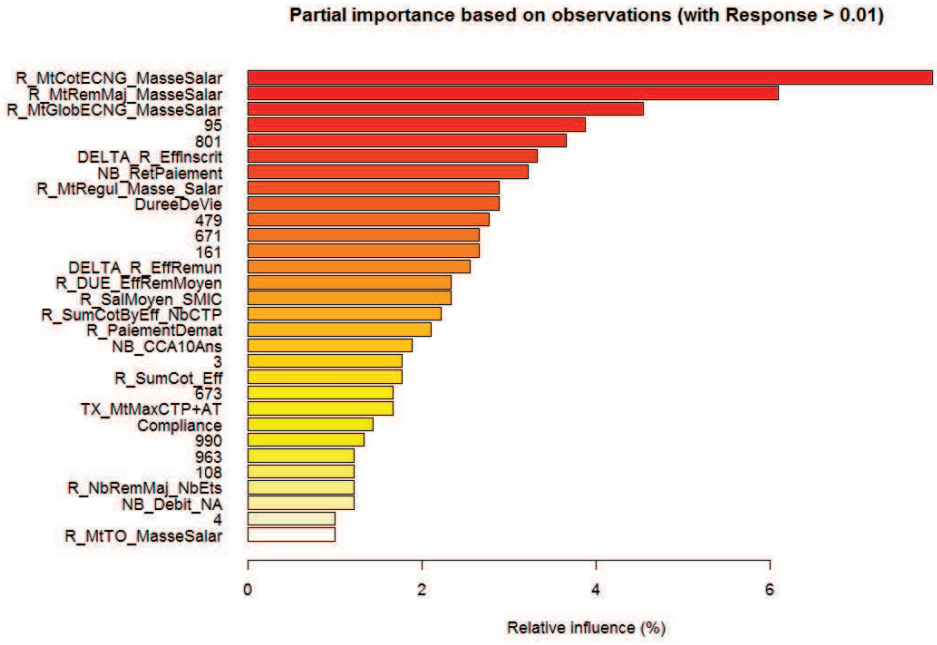

FIGURE 5.12 – Importance partielle des variables pour les montants de redressement supérieurs à 1% de la masse salariale.

Les 30 premières variables identifient 75% de tous les niveaux d'irrégularité supérieurs à 1%. Les montants de redressement supérieurs à 1% de la masse salariale concernent moins de 30% des entreprises redressées. Ce sont cependant ces cas qui nous intéressent car les variables identifiées interviennent dans toutes les situations. Seule leur répartition change : pour les montants de redressement inférieurs à 1% de la masse salariale, les régularisations de cotisation et les écarts de cotisations expliquent, à eux seuls, 25% des cas. Ceci procure un nouveau développement à notre analyse : les anomalies de cotisation, enregistrées par l'URSSAF à travers sa politique de recouvrement, déterminent une partie des contrôles qui seront effectués. Comme elles caractérisent, à la fois, l'absence et la présence d'irrégularités, la capacité à détecter ces dernières en est directement altérée. Bien que les anomalies soient décisives, il est nécessaire de les associer à plusieurs critères et variables supplémentaires afin de distinguer plus spécifiquement les irrégularités.

`Dépendances partielles et croisement de données`

Nous illustrons ce cas ainsi que la différence entre la sélection de contrôles (à effectuer) par un croisement de données et celle déterminée par un modèle. Nous choisissons, pour cela, la variable la plus influente de l'importance partielle. Il s'agit du *montant de la différence entre cotisations attendues et reçues* (relativement à la masse salariale).
a- Dans un premier temps, nous déterminons sa dépendance partielle avec les montants de redressement puis observons le pourcentage d'entreprises (sélectionnées par cette méthode) en situation d'irrégularité. La dépendance partielle est calculée au second ordre, soit en considérant qu'il existe une autre variable, inconnue, plus importante que celle que l'on analyse.

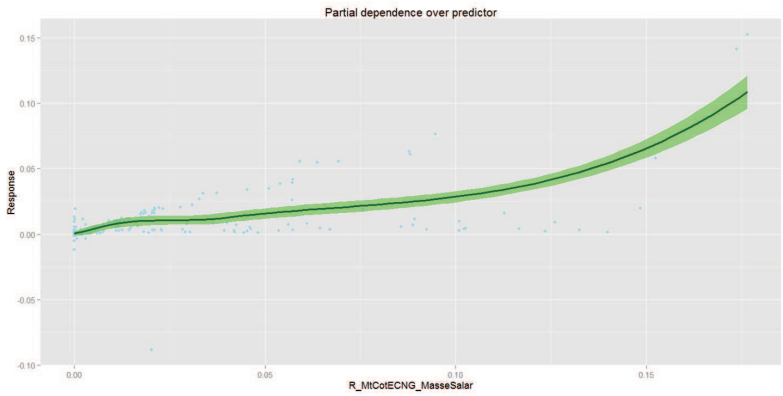

FIGURE 5.13 – Dépendance partielle de la différence entre cotisations attendues et reçues, et du montant de redressement.

Le graphique illustre l'effet des écarts de cotisations sur les montants de redressements. Plus les écarts sont importants, plus les montants de redressement le sont également. La dépendance partielle identifie 10% des entreprises contrôlées dans une situation dans laquelle les cotisations reçues sont inférieures à celles attendues. Parmi elles, 92% présentent un montant de redressement positif.
b- Comparons maintenant avec un croisement de données dans lequel on considérerait toutes les situations dans lesquelles cette différence entre cotisations attendues et reçues serait positive. 84% des entreprises sont dans cette situation et seule la moitié d'entre elles présente un montant de redressement positif. En augmentant le seuil, par exemple lorsque la différence est supérieure à 2% de la masse salariale, on diminue, naturellement, le nombre d'entreprises sélectionnées mais la répartition des montants de redressement ne change pas.

La dépendance partielle fournit avant tout un cadre prédictif préalable au cadre explicatif. Comme l'ensemble des outils d'analyse proposés (à l'exception de la mesure d'importance globale des variable), elle s'applique aussi bien à l'échantillon d'apprentissage qu'à n'importe quel échantillon de test. La relation entre écarts de cotisation et montants de redressement semble naturelle, à posteriori, mais elle ne l'est pas lorsqu'on croise les données en observant tous les contrôles. De manière encore plus décisive, la relation entre les montants de redressement et les anomalies de cotisation procède de plusieurs effets spécifiques qui interagissent. A travers les différentes mesures de dépendance, nous pouvons mesurer les interactions les plus importantes. Entre les écarts de cotisation et les régularisations, la dépendance est d'environ 0.3. C'est la corrélation la plus importante entre les variables explicatives influentes.
c- Nous souhaitons améliorer le croisement de données grâce à cette nouvelle information. En sélectionnant toutes les entreprises dont le montant des régularisations est positif et toutes celles dont les écarts de cotisation le sont également, environ deux tiers d'entre elles sont concernées. Et 60% de cette proportion présente alors un montant de redressement positif.
d- En exploitant uniquement la dépendance partielle, moins d'entreprises sont sélectionnées (15%), mais 93% d'entre elles présente un montant de redressement positif.

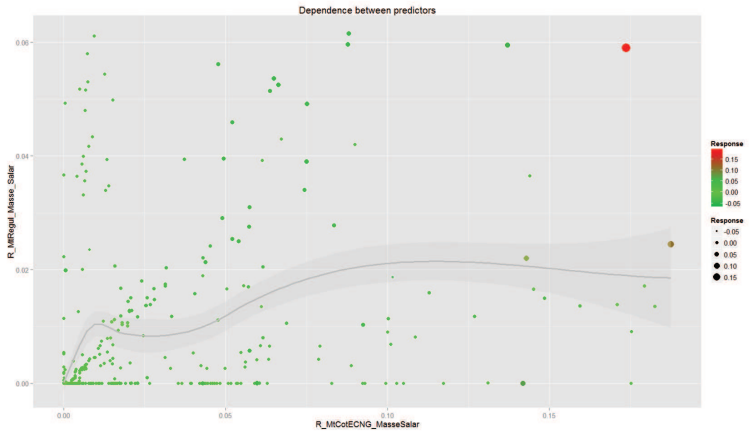

FIGURE 5.14 – Dépendance partielle de la différence entre cotisations attendues et reçues et du montant des régularisations, avec effet sur les montants de redressement.

Le graphique ci-dessus illustre la dépendance entre variables explicatives et l'effet sur les montants de redressement. Chaque point est un montant de redressement correspondant à la jonction entre un écart de cotisation, en abscisses, et une régularisation (les valeurs sont toutes relative à la masse salariale), en ordonnées. Plus un point est gros, plus le montant de redressement est important. La courbe, en ligne continue, illustre la dépendance entre les deux variables. L'augmentation des écarts de cotisations suffit, seule, à entraîner des niveau d'irrégularités positifs (et plus encore pour les seules régularisations). La dépendance entre les deux variables à un effet sur l'intensité du niveau d'irrégularités (tandis que la fréquence baisse) jusqu'à un certain niveau (environ 4% de montants régularisés et 10% d'écarts de cotisation, relativement à la masse salariale), puis se réduit brutalement. Une représentation en trois dimensions permet de mieux détailler cet effet.

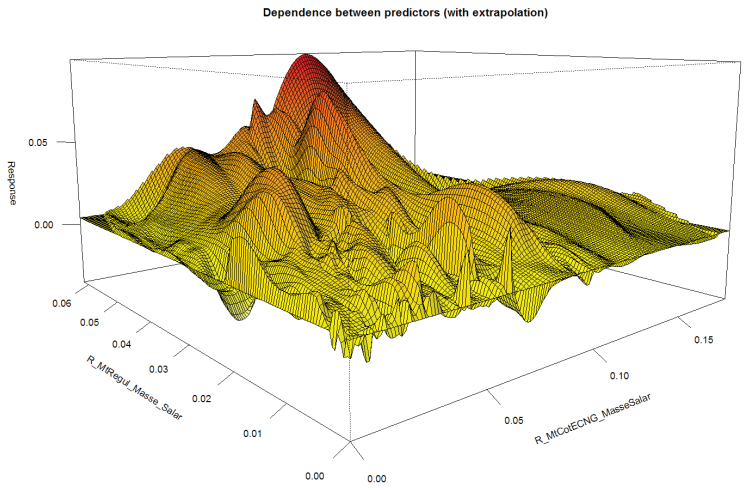

FIGURE 5.15 – Dépendance partielle entre les écarts de cotisation et le montant des régularisations, et effet sur le niveau d'irrégularités.

Le passage en trois dimensions permet de mieux situer la co-influence des variables sur le niveau d'irrégularités. Les écarts de cotisations et les régularisations ont un effet conjoint sur le niveau d'irrégularités jusqu'à une certaine limite. Au delà, leur co-influence sur un niveau d'irrégularités élevé devient caduque. De la même manière, lorsque les variables prennent toutes les deux de faibles valeurs (moins de 2% de régularisations et moins de 5% d'écarts de cotisation), les irrégularités demeurent à des niveaux faibles.

L'analyse des anomalies de cotisation permet d'expliquer la présence de montants de redressement positifs. Plus il y a d'interactions (dans une certaine limite) entre elles, plus les montants redressés sont importants. Une seconde explication provient des catégories de cotisation. Nous ne l'avons pas illustré, hormis par l'importance partielle. Certaines catégories spécifiques contribuent massivement aux irrégularités. C'est le cas des mesures de réduction de cotisations. Ces deux situations, les anomalies et certaines catégories de cotisation, fournissent un cadre explicatif général car elles interviennent également lorsqu'il n'y a pas d'irrégularités. Ce dernier point permet de comprendre les raisons pour lesquelles la détection des situations d'irrégularité est délicate. Un modèle linéaire ou un croisement de données arrivent à mettre en évidence certaines des variables explicatives identifiées, mais ne permettent pas de les combiner avec de bonnes capacités prédictives. Dans le cas des forêts uniformément aléatoires, le modèle prédictif exploite les interactions et leurs effets et les combine. A cet égard, on peut comparer la seule détection d'irrégularités (classification) à l'estimation des montants de redressement (régression). Dans le second cas, la continuité de la variable à expliquer permet de déterminer avec plus de précision l'influence des variables, en particulier celles liées aux catégories de cotisation. La classification (selon l'existence ou non d'une irrégularité) agit comme un instrument prédictif alors que la régression (par l'estimation des montants de redressement) fournit un cadre explicatif plus explicite du phénomène. De manière plus générale, les variables disponibles ne présentent pas de caractère explicatif global du niveau d'irrégularités et une hypothèse probable est le caractère majoritairement involontaire des ces dernières.

5.7 Expérimentation opérationnelle

L'expérimentation opérationnelle est la phase de validation *sur le terrain* des résultats de l'algorithme. Ici, 167 entreprises (sur 500 proposées) ont pu été contrôlées par l'URSSAF sur la base des recommandations de l'algorithme. *Le critère de sélection est la probabilité d'existence d'une irrégularité (> 0.5), déterminée par le modèle, dans la déclaration de cotisation de chaque entreprise.* Nous illustrons ci-dessous la répartition des effectifs des entreprises ainsi que celle de leur masse salariale. Dans le cas des forêts uniformément aléatoires, la répartition est celle des 500 entreprises recommandées.

Effectif	Répartition de la masse salariale en IdF	Répartition des entreprises d'Île-de-France	Répartition des entreprises dans les contrôles de l'URSSAF	Répartition des entreprises dans le modèle
[0-9]	37%	84%	73%	36%
[10-49]	14%	13%	22%	49%
[50-149]	10%	2%	4%	12%
≥ 150	39%	1%	1%	3%

TABLE 5.5 – Répartition, en 2011, de la masse salariale totale et des entreprises d'Île-de-France, par tranche d'effectif.

Les contrôles considérés sont les contrôles comptables d'assiette (CCA). Notons qu'en 2011, 36% des entreprises contrôlées par l'URSSAF n'ont pas d'effectif renseigné dans les bases de données. La masse salariale en Île-de-France est majoritairement répartie entre les micro-entreprises (moins de 10 salariés), et les entreprises de plus de 150 salariés, dont les entreprises de taille intermédiaire (ETI), grandes et très grandes entreprises (GE et TGE). Les effectifs, eux, sont massivement issus des micro-entreprises qui emploient près de 85% des salariés de la région.

- La répartition des contrôles de l'URSSAF montre une adéquation limitée avec la répartition des entreprises par tranche d'effectif. Les forêts uniformément aléatoires s'en détachent nettement, principalement à cause de l'algorithme de traitement générique qui effectue une première sélection de l'échantillon d'apprentissage. La seconde raison est due à la nature de l'évaluation : le modèle considère d'abord les entreprises pour lesquelles la présence d'irrégularité est détectable avec le plus de probabilité.
- L'évaluation par le modèle est constituée pour moitié d'entreprises entre 10 et 49 salariés et, au moment où sont effectivement choisies les 167 entreprises à contrôler, aucune entreprise de plus de 250 salariés n'y est présente.

Nous n'avons pas eu accès aux résultats détaillés de chaque entreprise contrôlée mais seulement aux résultats globaux. Toutefois, avant leur réalisation, nous pouvons fournir une estimation de la précision minimale attendue (le rapport entre le nombre d'irrégularités effectivement trouvées et le nombre d'irrégularités estimé par le modèle). Pour cela, le modèle est recalculé sur un échantillon d'apprentissage compatible avec le nombre d'entreprises qui seront contrôlées. La précision est estimée grâce à la règle de décision $\bar{g}^{(B)}_{\mathcal{P},oob}$ et le corollaire de la section 5.5.2 permet d'en calculer la marge d'erreur.

Relativement aux résultats de laboratoire de la section précédente, nous modifions uniquement le nombres d'arbres, B, fixé ici à 500.

`Résultats:` *erreurs de référence pour l'expérimentation opérationnelle*

```
Out-of-bag (OOB) evaluation
OOB estimate of error rate: 24.5%

OOB confusion matrix:
          Reference
Prediction   0   1 class.error
         0 191  56      0.2267
         1  42 111      0.2745

Theorical (Breiman) bounds
Prediction error (expected to be lower than): 37.36%
Upper bound of prediction error: 49.51%
Average correlation between trees: 0.0612
Strength (margin): 0.3752
Standard deviation of strength: 0.264
```

La précision estimée, $\widehat{P}_r$, vaut $1 - 0.2745 = 72.55\%$ et reste proche de celle obtenue sur l'échantillon total d'apprentissage de 4069 entreprises. Il n'est pas nécessaire d'estimer la précision plusieurs fois pour obtenir une moyenne : la *marge d'erreur*, ϵ, suffit à obtenir une borne inférieure. Précisons que la modélisation ci-dessus n'est pas celle qui a fourni les recommandations, ces dernières ayant été choisies à partir de la modélisation présentée dans la section 5.6. Appliquer le modèle avec moins d'exemples d'apprentissage, pour estimer les erreurs, a plusieurs avantages. Le premier est lié à la capacité de généralisation des forêts aléatoires qui augmente avec le nombre d'observations de l'échantillon d'apprentissage. Avec 10 fois moins d'exemples, la modélisation ci-dessus est moins performante que les précédentes. Les erreurs estimées y sont donc pessimistes et permettent de s'adapter à une situation dans laquelle l'échantillon de test est petit.
- Pour estimer au mieux la *précision*, nous avons considéré dans l'apprentissage le même nombre d'irrégularités que le nombre de recommandations de contrôle réalisées par les inspecteurs. En considérant la précision comme une erreur de test au sein de l'expérimentation, nous pouvons rappeler le corollaire de la section 5.5.2 pour évaluer la marge d'erreur commise sur l'estimation de la précision. L'espérance de la précision est le meilleur estimateur que l'on puisse obtenir et on a :

$$\mathbf{P}\left\{\left|\widehat{P}_r - \mathbf{E}\left\{P_r\right\}\right| > \epsilon\right\} \leq 2\mathsf{e}^{-2n\epsilon^2},$$

où $\widehat{P}_r$ est la *précision* estimée.
$n = 167$. On pose $\epsilon = 10\%$, la marge d'erreur que l'on ne souhaite pas dépasser. La probabilité pour que la différence entre la précision estimée et la précision effective soit supérieure à ϵ est alors inférieure ou égale à $2\mathsf{e}^{-2n\epsilon^2} = 7.08\%$. En d'autres termes, une fois les 167 contrôles effectués, la précision obtenue est, avec une probabilité de 92.92%, supérieure à 62.55%.

`Résultats obtenus`

Nous reportons ci-dessous les résultats obtenus lors de l'expérimentation opérationnelle des forêts uniformément aléatoires. Nous rappelons les dénominations utilisées par l'URSSAF : la *fréquence des redressements positifs* est équivalente à la *précision*. Elle est définie par :

$$fréquence\ des\ redressements\ positifs = P_r = \frac{nombre\ de\ redressements\ positifs}{nombre\ de\ contrôles}.$$

La *fréquence des redressements positifs* correspond au *taux de détection des irrégularités aux cotisations sociales*. L'URSSAF reporte généralement la *fréquence de redressement*.
Notons N_{R^+}, le nombre de redressements positifs,
N_{R^-}, le nombre de redressements négatifs,
et N_C, le nombre de contrôles réalisés.
On a :

$$fréquence\ de\ redressement = \frac{N_{R^+} + N_{R^-}}{N_C},$$

où les redressements négatifs désignent les contrôles ayant entraîné un remboursement de cotisations aux entreprises, correspondant à un montant des redressements négatifs, noté R^-. Notons également R^+, le montant des redressements positifs. Nous mesurons le *rendement* d'un modèle par son *montant comptable moyen*, par contrôle, défini par :

$$montant\ comptable\ moyen = \bar{R} = \frac{\sum_{i=1}^{N_C} R_i^+ - \sum_{i=1}^{N_C} R_i^-}{N_C}.$$

Les résultats obtenus sont résumés dans le tableau suivant :

Nombre de contrôles réalisés	Fréquence des redressements positifs	Montant des redressements positifs	Montant des redressements négatifs	Montant comptable moyen par contrôle
167	69%	1 118 993 euros	230 735 euros	5319 euros

TABLE 5.6 – Forêts uniformément aléatoires et résultats de l'expérimentation opérationnelle au sein de l'URSSAF d'Île-de-France, en 2012, pour des entreprises de moins de 250 salariés.

i) Le *montant net redressé* est de 888 258 euros pour 167 contrôles réalisés. Il correspond à la différence entre le montant des redressements positifs et le montant des redressements négatifs. Le montant net redressé est la somme qui figure au crédit des recettes supplémentaires de la Sécurité sociale, à condition qu'il soit effectivement récupéré. Le montant net redressé est assimilable à un *enjeu financier*.

ii) La *fréquence des redressements positifs* (la *précision*) est de 69%, soit 7 points au-dessus de la précision minimale estimée. Elle est, surtout, plus importante (+14 points) que celle réalisée par l'URSSAF sur toute la période 2006-2012, ce qui demeure la justification principale d'un modèle. Le *montant comptable moyen* (le *rendement*) est de 5319 euros (correspondant à un montant net moyen par contrôle).

`Contributions aux résultats de l'URSSAF`

Une question essentielle consiste à déterminer dans quelle mesure la généralisation d'un algorithme constitue une amélioration significative de la politique de contrôle menée. Nous décrivons quelques caractéristiques des résultats des contrôles comptables d'assiette (CCA) pour l'année 2012.

i) Environ 20 000 contrôles de ce type ont été réalisés. Le montant total net (comptable) redressé de ces contrôles CCA est d'environ 265 millions d'euros (303 millions d'euros en ne comptant que les redressements positifs). Précisons que, ni les chiffres du travail illégal, ni ceux des autres formes de contrôle ne sont intégrés ici.

ii) Comme nous l'avons déjà indiqué, les résultats financiers des contrôles de l'URSSAF reposent dans leur grande majorité sur les contrôles réalisés sur les grandes et très grandes entreprises. Pour les autres entreprises, il existe une difficulté intrinsèque à la mise en place d'une politique plus efficiente d'abord due aux données. Le traitement demandé pour une exploitation plus décisive est très important car elles présentent la plupart du temps les mêmes caractéristiques, que l'entreprise soit en situation d'irrégularité ou non. La seconde difficulté, corollaire du premier point, est le caractère généralement involontaire des irrégularités.

iii) Deux caractéristiques résument les résultats financiers des contrôles de l'URSSAF sur la période 2006-2011 :
- moins de 200 entreprises (essentiellement des grandes et très grandes entreprises, ainsi que celles de secteurs d'activité très spécifiques) génèrent 70% du montant total net redressé par l'URSSAF et 10% des entreprises en génèrent 90% ;
- la fréquence des redressements positifs est, sur toute la période 2006-2012, inférieure à 56%.

Bien que les montants redressés soient importants, ils indiquent une asymétrie importante entre les entreprises. Nous illustrons cette situation en décomposant les résultats des contrôles de l'URSSAF selon la taille des entreprises contrôlées.

Type d'entreprise	Nombre de contrôles	Fréquence des redressements positifs	Montant total des redressements positifs
GE/TGE/Autres	326	> 90 %	203 000 000 euros
TPE/PME	19646	< 56 %	100 000 000 euros

TABLE 5.7 – Répartition des résultats des contrôles comptables d'assiette (CCA) de l'URSSAF d'Île-de-France, en 2012.

En 2012, moins de 2% des entreprises ont généré, environ, 67% des montants redressés en faveur de l'URSSAF. La progression des résultats relativement aux années précédentes est essentiellement due à la nature exceptionnelle, par leur valeur, de quelques redressements. Deux points de vue permettent de mieux situer les modèles probabilistes dans la

détection des irrégularités :
- le premier consiste à observer le montant total des redressements sans en mesurer la répartition. Dans cette situation, ce montant apparaît important et le principal problème est celui de sa récupération effective.
- Le second point de vue généralise le précédent en ne considérant la capacité à détecter les irrégularités que du point de vue des données. Ces dernières sont issues des résultats accumulés grâce aux contrôles précédents et le modèle en dégage une synthèse qui peut être déclinée dans n'importe quelle politique de contrôle.

Nous faisons le lien avec ce second point de vue en examinant le *score d'importance* de la détection, noté S_D, pour l'expérimentation opérationnelle. Il donne un point de vue cohérent avec les enjeux financiers, la fréquence des redressements positifs et le montant comptable moyen. Rappelons sa définition.
Notons M_C, la masse salariale des entreprises contrôlées sur la base des recommandations d'un modèle,
M_R, la masse salariale des entreprises redressées et N_R, leur nombre.
On a :

$$S_D = \frac{\left(P_r - \frac{1-P_r}{P_r}\right)\left(1 + \frac{M_C}{M_R N_R}\right)\sum_{i=1}^{N_R} R_i}{M_R} \times 100,$$

avec $R_i = R_i^+ - R_i^-$, le montant net redressé pour la i-ème entreprise.

Comme nous n'avons à disposition que les résultats globaux de l'année 2012, nous remplaçons ici la masse salariale par la somme des montants contrôlés et posons $M_C = M_R$. On peut noter que dans ce cas, le taux de redressement est aussi pris en compte. On obtient un estimateur (avec un biais négatif) de S_D, noté $\tilde{S}_D$ et sa valeur est donnée par :

$$\begin{aligned}\tilde{S}_D &= \frac{\left(0.69 - \frac{1-0.69}{0.69}\right)\left(1 + \frac{1}{167}\right) \times 888258}{94643842} \times 100 \\ &= 0.2272.\end{aligned}$$

Le score d'importance peut être interprété de la manière qui suit. Supposons que le montant total des irrégularités relativement à la masse salariale de toutes les entreprises d'Île-de-France ait été estimé avec une marge d'erreur faible. Alors le score d'importance représente, en valeur relative, l'estimation moyenne du montant total net des redressements qu'il est possible de détecter en généralisant le modèle utilisé. Deux conditions sont nécessaires : le modèle doit pouvoir être généralisable (soit, être étendu à toutes les entreprises) et sa précision ne doit pas s'effondrer avec l'augmentation du nombre de contrôles. Par exemple, si le montant total des irrégularités relativement à la masse salariale est de 2% en Île-de-France, le score d'importance indique que 11.36% (0.2272/2) de ce montant pourrait être, au minimum, redressé sur la base des recommandations du modèle sans la nécessité de contrôler toutes les entreprises (plus clairement, le nombre d'entreprises serait connu). Cela semble faible, mais rappelons que la masse salariale annuelle en Île-de-France dépasse 195 Mds d'euros en 2011 et que les contrôles portent sur deux à trois années de cotisations. Un modèle dont le score d'importance est négatif détecterait moins de la moitié de toutes les irrégularités s'il était généralisé. Un modèle,

dont le score d'importance est supérieur à 1, ne peut pas être généralisé mais reste très adapté à des situations spécifiques. Cela est le cas des contrôles réalisés par l'URSSAF sur les grandes et très grandes entreprises (leur score est supérieur à 1). En comptabilisant l'ensemble des entreprises, le score devient négatif à cause de la précision limitée de la capacité de détection. Malgré l'importance des sommes récupérées, le score d'importance indique qu'une très large part des irrégularités n'est pas détectée.

5.8 Evaluation complète : ensemble des entreprises d'Île-de-France

L'évaluation complète consiste à d'abord attribuer, à chaque entreprise enregistrée à l'URSSAF d'Île-de-France et pour l'année 2013, une probabilité d'observation d'une irrégularité, définie à partir du modèle proposé et de l'échantillon d'apprentissage des entreprises contrôlées pour le compte de la période 2010-2011. Lorsque cette probabilité est supérieure à 0.5, la déclaration de cotisations de l'entreprise présente une ou plusieurs irrégularités pour l'année 2010 et/ou 2011, avec la probabilité estimée. Dans les forêts uniformément aléatoires, une irrégularité n'existe qu'avec une probabilité explicite.

Indépendamment de la détection d'irrégularités, le modèle estime également, pour chaque entreprise, un montant de redressement associé à un intervalle de confiance.

Une fois l'évaluation effectuée, les recommandations du modèle, pour chaque entreprise, sont un ensemble d'informations dont les plus importantes sont la probabilité d'observation d'une irrégularité et, si cette probabilité est supérieure à 0.5, le montant de redressement estimé associé à son intervalle de confiance.

Nous présentons les résultats de l'évaluation réalisée sur les 311 904 entreprises enregistrées en Île-de-France en 2011. 17 387 d'entre elles sont automatiquement exclues par l'algorithme de traitement générique, lequel assiste le moteur de détection. Cette évaluation correspond à une industrialisation du modèle.

`MATICS : Modèle(s) Aléatoire(s) pour le Traitement des Irrégularités aux Cotisations Sociales`

Le moteur de détection, nommé MATICS, est constitué de trois algorithmes principaux qui permettent d'évaluer n'importe quelle entreprise d'Île-de-France de façon transparente pour l'utilisateur. Un premier algorithme pilote l'ensemble des informations : du traitement des bases de données à l'exportation des résultats de l'évaluation. Le second est l'algorithme de traitement générique qui optimise les données en vue de leur apprentissage et de l'évaluation des nouvelles par les forêts uniformément aléatoires. Les deux particularités du moteur de détection sont l'utilisation d'un logiciel libre pour l'ensemble des opérations et le traitement de toutes les entreprises d'Île-de-France sur une station de travail. Aucune donnée nominative n'est utilisée par l'algorithme.

`Garanties`

Evaluer l'ensemble des entreprises permet d'entrevoir les sommes qui peuvent manquer aux recettes de la Sécurité sociale. *A la différence de l'inférence statistique classique, les montants évalués par un algorithme d'apprentissage statistique correspondent à des sommes dont le modèle fournit la démarche explicite conduisant à leur récupération. Parmi les critiques souvent attribuées à des modèles, figurent leur manque de précision, les hypothèses contraignantes qu'ils nécessitent ou la difficulté à les mettre en oeuvre de manière pratique. Ici, ces contraintes sont levées. En particulier, une manière de s'en assurer consiste à remarquer que les résultats peuvent être suivis en temps réel.*

A mesure que les contrôles sont réalisés sur les recommandations du modèle, leurs résultats peuvent être immédiatement comparés (par exemple tous les 10 ou 20 contrôles), à la fois, à l'évaluation et aux résultats de la politique de contrôle courante, et conduire à écarter le modèle s'il ne répond pas aux garanties annoncées. Dans l'expérimentation opérationnelle, les résultats obtenus ont été meilleurs que les objectifs annoncés, dont le premier et le plus important est le suivant : *la fréquence des redressements positifs (la précision) minimale estimée par le modèle est plus petite ou égale à la fréquence des redressements positifs effectivement mesurée sur la base de ses recommandations. De plus, elle est strictement supérieure à celle réalisée par l'URSSAF sur l'ensemble de ses contrôles.*

`Evaluation`

Pour évaluer les entreprises, nous procédons en plusieurs étapes afin d'en faciliter la lisibilité.
i) Les entreprises déjà contrôlées en 2010 ou 2011 (environ 60 000 entreprises de SIREN unique, tous contrôles confondus) ne peuvent plus l'être en 2013. Nous les enlevons donc de l'évaluation. De plus, on ne s'intéresse pas à la fraude (travail dissimulé).
ii) La période évaluée par le modèle ne concerne que la présence d'irrégularités pour les déclarations de cotisations des années 2010 et 2011 à cause de l'indisponibilité des données pour l'année 2012 au moment où l'évaluation a été réalisée. En conséquence, l'évaluation proposée minore d'environ 30% le montant total net des redressements estimé. Cette minoration correspond au taux de faux-positifs maximal du modèle, ce qui permet de fournir simplement une évaluation minimale.
iii) Les (très) grandes entreprises sont presque systématiquement contrôlées par l'URSSAF. Nous enlevons donc, après évaluation, toutes les entreprises dont les effectifs sont supérieurs à 250 salariés. Notons que les grandes entreprises suggérées par le modèle peuvent être différentes de celles contrôlées par l'URSSAF.
iv) Dans la pratique, les ressources de l'URSSAF ne permettent pas de contrôler un nombre d'entreprises beaucoup plus important que celui qui est mesuré à ce jour et il faut s'adapter à cette contrainte.

A) Dans un premier temps, nous fournissons les *résultats bruts* de l'évaluation et seuls les points *i)* et *ii)* sont pris en compte. Les forêts uniformément aléatoires incrémentales permettent une plus grande adaptation à la problématique posée et nous les avons donc utilisées en remplacement du modèle standard ; ce fut la seule optimisation réalisée pour l'évaluation. Tous les montants sont arrondis.

Nombre de recommandations de contrôle	Précision minimale	Masse salariale totale cumulée des entreprises recommandées	Intervalle de confiance du montant net des redressements estimé relativement à la masse salariale
26 409	70.79%	49.8 Mds euros	[0.966%, 2.585%]

TABLE 5.8 – Evaluation des irrégularités aux cotisations sociales, par une forêt uniformément aléatoire, pour l'ensemble des entreprises d'Île-de-France, en 2013.

Montant net des redressements estimé	Montant minimal des redressements estimé	Montant maximal des redressements estimé	Montant comptable moyen par contrôle
557 000 000 euros	254 000 000 euros	861 000 000 euros	21 106 euros

TABLE 5.9 – Evaluation des irrégularités aux cotisations sociales, par une forêt uniformément aléatoire, pour l'ensemble des entreprises d'Île-de-France, en 2013.

Le tableau ci-dessus correspond à une généralisation des forêts uniformément aléatoires à l'ensemble des entreprises d'Île-de-France pour la détection des irrégularités et l'évaluation du montant total net (comptable) des redressements espérés, pour les déclarations de cotisations des années 2010 et 2011.
Si le modèle avait été généralisé, les contrôles auraient été réalisés en 2013 et auraient concerné les déclarations de cotisations des années 2010 à 2012.
Le nombre de recommandations de contrôle est d'environ 36% inférieur à l'ensemble des contrôles réalisés par l'URSSAF d'Île-de-France en 2012 et le montant net des redressements (constitué de la somme des montants nets de redressement, estimés pour chaque entreprise) est supérieur de 65% au montant total net redressé cette même année (337 millions d'euros).
La précision minimale est équivalente à la fréquence des redressements positifs minimale qui serait réalisée, avec une probabilité supérieure à 92%.
La masse salariale des entreprises recommandées est d'environ 50 Mds d'euros et le score d'importance de la détection est estimé à 0.3302.
Le niveau d'irrégularité global, soit le rapport entre le montant total net des redressements estimé et la masse salariale totale cumulée (de 2010 à 2011) des entreprises recommandées, est d'environ 1.12%.
Le niveau moyen d'irrégularité estimé, soit la moyenne du rapport entre le montant net redressé et la masse salariale d'une entreprise recommandée, est, avec un niveau de confiance de 99%, compris entre 0.966% et 2.585%.

B) La généralisation du modèle n'est cependant pas envisageable d'une manière simple. Nous prenons alors en compte les contraintes énoncées dans les points *iii)* et *iv)*. L'objectif est, essentiellement, de donner un aspect opérationnel aux contrôles qui ne peuvent, habituellement, pas être systématisés.

Nombre de recommandations de contrôle	Précision minimale	Masse salariale totale cumulée des entreprises recommandées	Intervalle de confiance du montant net des redressements estimé relativement à la masse salariale
26 042	70.79%	27.3 Mds euros	[0.448%, 1.603%]

TABLE 5.10 – Evaluation des irrégularités aux cotisations sociales, par une forêt uniformément aléatoire, pour l'ensemble des entreprises de moins de 250 salariés d'Île-de-France, en 2013.

Montant net moyen des redressements estimé	Montant net minimal des redressements estimé	Montant net maximal des redressements estimé	Montant comptable moyen par contrôle
363 000 000 euros	125 000 000 euros	456 000 000 euros	13 942 euros

TABLE 5.11 – Evaluation des irrégularités aux cotisations sociales, par une forêt uniformément aléatoire, pour l'ensemble des entreprises de moins de 250 salariés d'Île-de-France, en 2013.

Les seules recommandations de contrôle du modèle pour les très petites, petites et moyennes entreprises génèrent environ deux tiers du montant de l'ensemble des redressements estimés pour toutes les entreprises. Plusieurs éléments peuvent être notés :
- le temps de contrôle est limité (2 jours, en moyenne, par contrôle) ;
- le modèle est complémentaire de la politique de contrôle de l'URSSAF, les montants redressés s'ajoutent à ceux des contrôles réalisés sur les entreprises de plus grande taille ;
- les sommes redressées sont récupérables beaucoup plus facilement.

Une dernière question est cruciale : les redressements issus des recommandations de contrôle peuvent-ils constituer des recettes supplémentaires pour la Sécurité sociale ?
Les forêts uniformément aléatoires, et les modèles issus de l'apprentissage statistique et de l'apprentissage automatique, n'apportent pas de réponse à cette question. A la place, elles garantissent la position suivante : si le niveau des irrégularités (et de la fraude) est celui estimé par les autorités les plus compétentes sur ce sujet, alors, avec le concours des inspecteurs du contrôle, des équipes statistiques de l'URSSAF et avec un ordinateur, il est possible d'en récupérer la plus grande partie, sans autres moyens que ceux disponibles à ce jour.

Le travail réalisé à l'URSSAF d'Île-de-France a eu pour objectif de proposer un modèle visant à améliorer la détection des irrégularités aux cotisations sociales. Les forêts uniformément aléatoires en sont la mise en oeuvre théorique et pratique. Après le départ de l'ex-équipe dirigeante de l'URSSAF, à la fin 2012, le modèle a été abandonné.

Annexe A

Nous présentons ci-dessous l'ensemble des variables définies pour la détection des irrégularités aux cotisations sociales. Les variables sont issues de deux sources :
- la situation de l'entreprise auprès de l'URSSAF à travers la politique de recouvrement de cette dernière ;
- l'ensemble des *catégories de cotisation* susceptibles d'être déclarées par une entreprise. Leur dénomination officielle est *code-type de personnel* (CTP).

L'identifiant unique de chaque entreprise est son SIREN et il ne fait, naturellement, pas partie des variables.

`Définition des variables`

DureeDeVie : nombre d'années écoulées depuis la date d'assujettissement de l'entreprise auprès de l'URSSAF.

R_SumCotByEff_NbCTP : pour chaque entreprise, on calcule la somme de ses cotisations que l'on divise par sa masse salariale. Le résultat est lui-même divisé par le produit du nombre de salariés et de CTP. Le résultat final est équivalent, pour chaque entreprise, au taux de répartition des cotisations par salarié et par catégorie de cotisation.

TESE (Titre Emploi Service Entreprise) : dispositif d'allègement des formalités liées à l'emploi. La variable vaut 1 si l'entreprise en bénéficie. 0 sinon.
www.letese.urssaf.fr/tesewebinfo/cms/index.html

DELTA_R_EffInscrit : différence entre deux ratios ; le premier est le rapport entre l'effectif inscrit moyen et celui inscrit en début de période. Le second est le rapport entre l'effectif inscrit moyen et celui inscrit en fin de période.

DELTA_R_EffRemun : identique à la variable précédente. L'effectif inscrit est remplacé par l'effectif rémunéré.

R_SumCot_Eff : ratio entre le montant total des cotisations de l'entreprise et le nombre de salariés. Equivalent au taux de cotisation moyen par salarié et par entreprise.

FREQ_Paiement : inverse du nombre de paiements de cotisation (mensuel ou trimestriel) par an.

NB_DeclaDemat : nombre de déclarations de cotisation effectuées par voie dématérialisée.

NB_Debit_NA : nombre d'origines inconnues d'écarts de cotisation en faveur de l'URSSAF (débits). Dans cette situation, les débits n'ont pas d'origine identifiée.

R_MtDebit_MasseSalar : ratio entre le montant du débit et la masse salariale.

R_SalMoyen_SMIC : niveau du salaire moyen brut relativement au SMIC (valeur de référence calculée sur la période 2006-2009).

R_NbCTPExo_NbCTP : ratio entre le nombre de catégories de cotisation donnant lieu à une réduction de cotisations et le nombre total de catégories de cotisation de l'entreprise.

R_DUE_EffRemMoyen : ratio entre le nombre de déclarations uniques d'embauche et l'effectif rémunéré moyen.

R_NbRemMaj_NbEts : ratio entre le nombre de remises sur majoration de cotisation et le nombre d'établissements (identifiés par le SIRET) de l'entreprise (identifiée par son SIREN)

R_MtRemMaj_MasseSalar : ratio entre le montant de remises sur majoration de cotisation et la masse salariale.

R_MtGlobECNG_MasseSalar : ratio entre le montant global des écarts négatifs de cotisation et la masse salariale. Ce ratio concerne tout l'historique connu.

R_MtCotECNG_MasseSalar : ratio entre le montant, en cotisations, des écarts négatifs de cotisation et la masse salariale. Ce ratio ne concerne que la période courante.

R_MtPen_MasseSalar : ratio entre le montant des pénalités et la masse salariale.

NB_RetPaiement : nombre de retards de paiement de cotisations.

NB_TaxOffice : nombre de taxations d'office.

R_MtTO_MasseSalar : ratio entre le montant des taxations d'office et la masse salariale.

R_MtRegul_Masse_Salar : ratio entre le montant des régularisations et la masse salariale.

R_DelaisAcc_DelaisDem : ratio entre le nombre de délais de paiements accordés et le nombre de délais demandés. Il vaut 0 si aucun délai n'a été demandé.

R_PaiementDemat : ratio entre le nombre de paiements dématérialisés et le nombre de paiements total.

DernierCtrl : délai écoulé (en années) depuis le dernier contrôle de l'entreprise.

NB_CCA10Ans : nombre de contrôles comptables d'assiette (CCA) sur les 10 dernières années.

Compliance : ratio de conformité aux obligations liées à la déclaration de cotisations.

R_SumCot_MtRevenus : ratio entre le montant des cotisations versées à l'URSSAF et le montant des revenus versés au salariés. C'est le taux global de cotisation de l'entreprise.

R_NbCTP_NbETS : ratio entre le nombre total de CTP et le nombre d'établissements de l'entreprise.

TX_CotNetTheorique : taux de cotisation net théorique de l'entreprise (reconstitué par une heuristique).

TX_MtMaxCTP+AT : taux de cotisation de la catégorie la plus importante pour chaque entreprise.

R_AssPlaf_AssDeplaf : ratio entre l'assiette (la partie du salaire brut) plafonnée et celle déplafonnée.

`Codes-type de personnel (CTP)` : tous les montants sont transformés en ratios, relativement à la masse salariale de l'entreprise.

-9 à *-1* et *1000* à *1023* : Codes-type de personnel virtuels pour les besoins du modèle. Initialisés à 0.

0 à *999* : cf https ://fichierdirect.declaration.urssaf.fr/TablesReference.htm pour la liste complète. Nous en indiquons ci-dessous quelques uns parmi les plus importants :

3 : déduction salariale, heures supplémentaires.

4 : déduction patronale, heures supplémentaires. Maximum de 20 salariés.

5 : déduction patronale, heures supplémentaires. Plus de 20 salariés.

99 : aide aux 35 heures. Loi Aubry I.

100 : Régime Général. Cas général. Elle correspond à la cotisation la plus importante payée par une entreprise, sauf cas particuliers. Son taux a changé en 2014.

110 : plan emploi jeune. Cas particulier du code-type 100.

125 : activités économiques réduites. Cas particulier du code-type 100.

130 : concierges et employés d'immeubles. Cas général. Elle correspond également à une exception du code-type 100.

260 : CSG et CRDS. Régime Général. Elle admet des cas particuliers.

400 : Crédit Impôt Compétitivité Emploi (CICE). Introduit en 2013.

671 : réduction des cotisations patronales sur les bas salaires, dite réduction Fillon.

900 : versement transport.

990 : pénalités.

Bibliographie

ACOSS, 2012. *"Le contrôle des cotisants 2012"*. www.acoss.fr

Alquier, P., Biau, G., 2013. Sparse Single-index Model. *The Journal of Machine Learning Research* 14, 243-280.

Amit, Y., Geman, D., 1997. Shape Quantization and Recognition with Randomized Trees. *Neural Computation* 9, 1545-1588.

Balcan, M.-F., 2008. *New Theoretical Frameworks for Machine Learning*. PhD thesis, School of Computer Science, Carnegie Mellon University.

Bellemare, C., Fortin, B., Joubert, N., Marchand, S., 2012. *Effets de pairs et fraude sociale : une analyse économétrique sur données françaises* (No. 2012s-32). CIRANO.

Bertail, P., Politis, D.N., Romano, J.P., 1999. On Subsampling Estimators with Unknown Rate of Convergence. *Journal of the American Statistical Association* 94, 569-579.

Biau, G., 2012. Analysis of a Random Forests Model. *The Journal of Machine Learning Research* 13, 1063-1095.

Biau, G., Cérou, F., Guyader, A., 2010. On the Rate of Convergence of the Bagged Nearest Neighbor Estimate. *The Journal of Machine Learning Research* 11, 687-712.

Biau, G., Devroye, L., 2010. On the Layered Nearest Neighbour Estimate, the Bagged Nearest Neighbour Estimate and the Random Forest Method in Regression and Classification. *Journal of Multivariate Analysis* 101, 2499-2518.

Biau, G., Devroye, L., 2013. Cellular Tree Classifiers. *Electronic Journal of Statistics* 7.

Biau, G., Devroye, L., Lugosi, G., 2008. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research* 9, 2015-2033.

Blanchard, G., 2003. *Generalization Error Bounds for Aggregate Classifiers*, in : Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification, Lecture Notes in Statistics. Springer New York, pp. 357-367.

Boulesteix, A.-L., Janitza, S., Kruppa, J., König, I.R., 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery* 2, 493-507.

Bolton, R.J., Hand, D.J., 2002. Statistical Fraud Detection : A Review. *Statistical Science* 17, 235-255.

Bousquet, O., Boucheron, S., Lugosi, G., 2004. Introduction to Statistical Learning Theory. *Advanced Lectures on Machine Learning*, Springer Berlin Heidelberg, pp 169-207.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 1145-1159.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C., 1984. *Classification and Regression Trees*. New York : Chapman and Hall.

Breiman, L., 1996. Bagging predictors. *Machine learning* 24, 123-140.

Breiman, L., 1996. Heuristics of instability and stabilization in model selection. *The annals of statistics* 24, 2350-2383

Breiman, L., 1999. Pasting Small Votes for Classification in Large Databases and On-Line. *Machine Learning* 36, 85-103.

Breiman, L., 2001. Random forests. *Machine learning* 45, 5-32.

Breiman, L., 2001. Statistical Modeling : The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* 16, 199-231

Brown, G., Wyatt, J.L., Tino, P., 2005. Managing Diversity in Regression Ensembles. *The Journal of Machine Learning Research* 6, 1621-1650.

Buja, A., Stuetzle, W., 2006. Observations on bagging. *Statistica Sinica* 16(2), 323.

Bühlmann, P., Yu, B., 2002. Analyzing Bagging. *Annals of Statistics* 30, 927-961.

Burczynski, M.E., Peterson, R.L., Twine, N.C., Zuberek, K.A., Brodeur, B.J., Casciotti, L., Maganti, V., Reddy, P.S., Strahs, A., Immermann, F., Spinelli, W., Schwertschlag, U., Slager, A.M., Cotreau, M.M., Dorner, A.J., 2006. Molecular Classification of Crohns Disease and Ulcerative Colitis Patients Using Transcriptional Profiles in Peripheral Blood Mononuclear Cells. *Journal of Molecular Diagnostics* 8, 51-61.

Carslaw, D.C. and K. Ropkins, (2012) *openair — an R package for air quality data analysis*. Environmental Modelling and Software. Volume 27-28, 52-61

Chen, C., Liaw, A., Breiman, L., 2004. *Using Random Forest to Learn Imbalanced Data*, Statistics Technical Reports. Statistics Department, University of California, Berkeley, University of California at Berkeley, Berkeley, California.

Ciss, S., 2014. *randomUniformForest : random Uniform Forests for Classification and Regression*. R package version 1.0.6, http ://CRAN.R-project.org/package=randomUniformForest.

Conseil des prélèvements obligatoires, 2007. La fraude aux prélèvements obligatoires et son contrôle. *La Documentation française*.

Criminisi, A., Shotton, J., Konukoglu, E., 2012. Decision Forests : A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. *Foundations and Trends® in Computer Graphics and Vision* 7, 81-227.

Cutler, A., Zhao, G., 2001. PERT-perfect random tree ensembles. *Computing Science and Statistics* 33, 490-497.

Davis, J., Goadrich, M., 2006. The Relationship Between Precision-Recall and ROC Curves, in : *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06. ACM, New York, NY, USA, pp. 233-240.

Denil, M., Matheson, D., de Freitas, N., 2013. Consistency of Online Random Forests. *arXiv :1302.4853 [stat.ML]*.

Devroye, L., 1981. On the Almost Everywhere Convergence of Nonparametric Regression Function Estimates. *The Annals of Statistics* 9, 1310-1319.

Devroye, L., Györfi, L., Lugosi, G., 1996. *A probabilistic theory of pattern recognition*. New York : Springer.

Devroye, L.P., Wagner, T.J., 1980. Distribution-Free Consistency Results in Nonparametric Discrimination and Regression Function Estimation. *The Annals of Statistics*. 8, 231-239.

Diaz-Uriarte, R., 2007. GeneSrF and varSelRF : a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics* 8, 328.

Díaz-Uriarte, R., Andrés, S.A. de, 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 3.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. *Statistical Science*, 189-212

Dietterich, T.G., 2000. Ensemble Methods in Machine Learning, in : *Multiple Classifier Systems*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1 ?15.

Dietterich, T.G., 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees : Bagging, Boosting, and Randomization. *Machine Learning* 40, 139-157.

Dietterich, T.G., Domingos, P., Getoor, L., Muggleton, S., Tadepalli, P., 2008. Structured machine learning : the next ten years. *Machine Learning* 73, 3-23.

Dietterich, T. G., Kong, E. B., 1995. *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms*. Technical Report. Department of Computer Science, Oregon State University

Dionne, G., Giuliano, F., Picard, P., 2009. Optimal Auditing with Scoring : Theory and Application to Insurance Fraud. *Management Science* 55, 58-70.

Domingos, P., n.d. *A Unified Bias-Variance Decomposition and its Applications*. In Proceedings of the 17th International Conference on Machine Learning.

Domingos, P., 2012. A Few Useful Things to Know About Machine Learning. *Communications of the ACM* 55, 78-87.

Eddelbuettel, D., Francois, R., 2011. Rcpp : Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8), 1-18. URL http ://www.jstatsoft.org/v40/i08/.

Eddelbuettel, D., 2013. *Seamless R and C++ Integration with Rcpp*. Springer, New York.

Efron, B., 1979. Bootstrap Methods : Another Look at the Jackknife. *The Annals of Statistics* 7, 1-26.

Efron, B., 2003. Second Thoughts on the Bootstrap. *Statistical Science* 18, 135-140.

Efron, B., Tibshirani, R.J., 1994. *An Introduction to the Bootstrap*. CRC Press.

Efron, B., Tibshirani, R., 1997. Improvements on Cross-Validation : The 632+ Bootstrap Method. *Journal of the American Statistical Association* 92, 548 ?560.

Elisseeff, A., Evgeniou, T., Pontil, M., 2005. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, pp. 55-79.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 861-874.

Freund, Y., Schapire, R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55, 119-139.

Friedman, J.H., Hall, P., 2007. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, Special Issue on Nonparametric Statistics and Related Topics : In honor of M.L. Puri 137, 669-683.

Friedman, J.H., 2001. Greedy function approximation : A gradient boosting machine. *Annals of Statistics* 29, 1189-1232.

Friedman, J.H., 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis* 38, 367-378.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33, 1-22. http ://www.jstatsoft.org/v33/i01/.

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M. Bouchachia, A., 2013. A Survey on Concept Drift Adaptation, *ACM Computing Surveys* 46(4), p44.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. *Neural Computation* 4, 158.

Genuer, R., 2010. *Forêts aléatoires : aspects théoriques, sélection de variables et applications*. Thèse de doctorat. Université Paris-Sud.

Genuer, R., 2012. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 120.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. *Pattern Recognition Letters* 31, 2225-2236.

Genuer, R., Michel, V., Eger, E., Thirion, B., 2010. *Random Forests based feature selection for decoding fMRI data*. Presented at the Proceedings of the 19th COMPSTAT, pp. 1071-1078.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. *Machine Learning* 63, 3-42.

Gey, S., 2002. *Bornes de risque, détection de ruptures, Boosting : trois thèmes statistiques autour de CART*. Thèse de doctorat, Université Paris-Sud.

Gey, S., 2012. Risk bounds for CART classifiers under a margin condition. *Pattern Recognition*, 45(9), 3523-3534

Grömping, U., 2009. Variable Importance Assessment in Regression : Linear Regression versus Random Forest. *The American Statistician* 63, 308-319.

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *The Journal of Machine Learning Research* 3, 1157-1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46, 389-422.

Hand, D.J., 2006. Classifier Technology and the Illusion of Progress. *Statistical Science* 21, 1-14.

Hapfelmeier, A., Ulm, K., 2013. A new variable selection approach using Random Forests. *Computational Statistics and Data Analysis* 60, 50-69.

Hastie, T., Tibshirani, R., Friedman, J.J.H., 2001. *The elements of statistical learning*. New York : Springer.

Ho, T.K., 1998. The random subspace method for constructing decision forests.*IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 832-844.

Hoeffding, W., 1963. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* 58, 13.

James, G.M., 2003. Variance and bias for general loss functions. *Machine Learning* 51(2), pp. 115-135.

Joubert, N., 2009. Processus de détection et évaluation de la fraude sociale. *Revue Economique*. Septembre 2009, Numéro 5, Volume 60, 235-1256.

Kane, M.J., Emerson, J.W., 2011. *bigmemory : Manage massive matrices with shared memory and memory-mapped files*. R package version 4.2.11. http ://CRAN.R-project.org/package=bigmemory

Kleinberg, E.M., 2000. On the algorithmic implementation of stochastic discrimination. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 22, 473-490.

Kubat, M., Holte, R.C., Matwin, S., 1998. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning* 30, 195-215.

Larivière, B., Poel, D.V.D., 2004. Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques. *Expert Systems with Applications*, 29(2), 472-484.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. *R News* 2(3), 18–22.

Lin, Y., Jeon, Y., 2002. Random Forests and Adaptive Nearest Neighbors. *Journal of the American Statistical Association* 101-474.

Lounici, K., Pontil, M., Tsybakov, A.B., van de Geer, S., 2009. Taking Advantage of Sparsity in Multi-Task Learning. *arXiv :0903.1468 [math, stat]*.

MacKay, D. J., 2003. *Information theory, inference, and learning algorithms* (Vol. 7). Cambridge : Cambridge university press.

Massart, P., 2003. *Concentration inequalities and model selection*. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6-23, 2003.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2014. *e1071 : Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-3. http ://CRAN.R-project.org/package=e1071

Murthy, S. K., Salzberg, S. 1995. *Decision Tree Induction : How Effective Is the Greedy Heuristic ?*. In KDD, pp. 222-227.

Murthy, S.K., Salzberg, S.L., 1999. *Investigations of the Greedy Heuristic for Classification Tree Induction*.

Neville, J., Jensen, D., 2008. A bias/variance decomposition for models using collective inference. *Machine Learning* 73, 87-106.

Oza, N.C., 2005. *Online bagging and boosting*, in : 2005 IEEE International Conference on Systems, Man and Cybernetics, pp. 2340-2345 Vol. 3.

Prenger, R.J., Lemmond, T.D., Varshney, K.R., Chen, B.Y., Hanley, W.G., 2010. *Class-specific Error Bounds for Ensemble Classifiers*, in : Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 843-852.

Phua, C., Lee, V., Smith, K., Gayler, R., 2012. A Comprehensive Survey of Data Mining-based Fraud Detection Research. *Computers in Human Behavior* 28, 1002-1013.

Quinlan, J.R., 1986. Induction of decision trees. *Machine Learning* 1, 81-106.

Quinlan, J.R., 1993. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

R Core Team (2014). *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http ://www.R-project.org/.

Ramey, J.A., 2012. *datamicroarray : Collection of Data Sets for Classification*. R package version 0.2. https ://github.com/ramey/datamicroarray, http ://ramhiser.com

Revolution Analytics, 2013. *rmr2 : R and Hadoop Streaming Connector*. R package version 2.3.0.

Ridgeway, G., with contributions from others, 2013. *gbm : Generalized Boosted Regression Models*. R package version 2.1. http ://CRAN.R-project.org/package=gbm

Rokach, L., Maimon, O., 2005. Decision Trees, in : Maimon, O., Rokach, L. (Eds.), *Data Mining and Knowledge Discovery Handbook*. Springer US, pp. 165-192.

Rokach, L., Maimon, O., 2005. Top-down induction of decision trees classifiers - a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews* 35, 476-487.

Saffari, A., Leistner, C., Santner, J., Godec, M., Bischof, H., 2009. *On-line Random Forests*, in : 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1393-1400.

Seeger, M., Langford, J., Megiddo, N., 2001. *An improved predictive accuracy bound for averaging classifiers*. In Proceedings of the 18th International Conference on Machine Learning (No. EPFL-CONF-161321, pp. 290-297).

Scornet, E., Biau, G., Vert, J. P., 2014. Consistency of Random Forests. *arXiv* preprint arXiv :1405.2881.

Shannon, C.E., 1949. *The Mathematical Theory of Communication*. University of Illinois Press.

Simm, J., de Abril, I.M., 2013. *extraTrees : ExtraTrees method*. R package version 0.4-5. http ://CRAN.R-project.org/package=extraTrees

Song, L., Langfelder, P., Horvath, S., 2013. Random generalized linear model : a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics* 14, 5.

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

Steinwart, I., Hush, D. R., Scovel, C., 2005. A classification framework for anomaly detection. *Journal of Machine Learning Research*, pp. 211-232.

Stone, C. J., 1977. Consistent nonparametric regression. *The annals of statistics*, 595-620.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures : Illustrations, sources and a solution. *BMC Bioinformatics* 8, 25.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9, 307.

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning : Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323-348.

Tibshirani, R., 1996. *Bias, variance, and prediction error for classification rules.* Technical Report, Statistics Department, University of Toronto.

Vapnik, V.N., 1995. *The nature of statistical learning theory.* Springer-Verlag New York.

Viaene, S., Derrig, R.A., Baesens, B., Dedene, G., 2002. A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. *Journal of Risk and Insurance* 69, 373-421.

Wickham, H., 2011. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1), 1-29. http ://www.jstatsoft.org/v40/i01/.