UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

Construction d'arbres de discrimination pour expliquer les niveaux de contamination radioactive des végétaux

THESE

pour obtenir le grade de

DOCTEUR de l'UNIVERSITE MONTPELLIER II

Formation doctorale : Biostatistique
Ecole doctorale : Information, Structures, Systèmes

présentée et soutenue publiquement
par

Bénédicte BRIAND

le 25 avril 2008 devant le jury composé de :

Denis ALLARD	*Président du jury*
Avner BAR-HEN	*Rapporteur*
Mireille BATTON-HUBERT	*Rapporteur*
Philippe BESSE	*Examinateur*
Gilles DUCHARME	*Directeur*
Catherine MERCAT-ROMMENS	*Co-encadrant*
Ricco RAKOTOMALALA	*Examinateur*

Institut de Radioprotection et de Sûreté Nucléaire
Laboratoire d'Etudes Radioécologiques en Milieux Continental et Marin
IRSN

Remerciements

En tout premier lieu, je tiens à remercier Catherine Mercat-Rommens, co-encadrante de cette thèse. Merci pour ta disponibilité, ton dynamisme et ton efficacité. Malgré les différentes embûches déposées sur mon chemin, tu as toujours été là pour moi et tu m'as appris à ne jamais baisser les bras et à persévérer. J'ai énormément appris à tes côtés tant sur le plan professionnel que personnel. Travailler avec toi a été un vrai plaisir. Pour résumer, je dirai que tu es la co-encadrante idéale ! Je garderai d'excellents souvenirs de notre collaboration, en particulier ce fameux Nantes-Marseille via Toulouse…

Je remercie Gilles Ducharme qui a dirigé cette thèse pour la confiance qu'il m'a témoignée et ses conseils. Il m'a encouragé à faire cette thèse et aujourd'hui je lui suis très reconnaissante.

Je tiens à remercier l'ensemble des membres du jury d'avoir accepté d'évaluer ce travail. En particulier, je remercie Avner Bar-Hen, Philippe Besse et Ricco Rakotomalala pour avoir pris le temps de me rencontrer et de discuter de mes travaux.

Je pense également à toutes les personnes qui de près ou de loin ont contribué à cette thèse. Je remercie Eric Chojnacki et Laurent Garcia-Sanchez pour leurs conseils et leur disponibilité. Je vous suis reconnaissante d'avoir relu cette thèse.

Je remercie très chaleureusement Philippe Renaud pour son accueil au sein du Laboratoire d'Etudes Radioécologiques en milieux Continental et Marin. Ces années passées au laboratoire ont été très agréables et instructives, j'en garderai un excellent souvenir. Il est difficile de résumer tous les bons moments passés ensemble en quelques lignes mais je vais essayer d'être brève. En premier lieu, je souhaiterai remercier les amis que je me suis faits au cours de ces trois années et demi :

- Vanessa, merci d'avoir « sticsé» pour moi ! Merci pour ta bonne humeur, ton aide, ton soutien et tes conseils en agronomie. J'ai vraiment apprécié de travailler avec toi. J'ai également trouvé une amie et confidente. Je te remercie sincèrement pour tout ce que tu m'as apporté.

- Sabrina, merci pour tout, ta disponibilité et surtout ton efficacité, tu es la MEILLEURE assistante de direction ! Merci pour tous ces échanges, de la Nouvelle Star à nos problèmes personnels. Je garde un très bon souvenir de cette soirée aux Vannades : la situation n'est pas bloquée (n'est ce pas Franck ?). Je garde aussi d'excellents souvenirs de notre initiation au surf, j'ai encore mal rien que d'en parler !

- Lionel, merci pour ton franc parlé ! Merci pour tous ces bons moments partagés au ski, au foot, au UNO, et les fameuses parties de shikumi, je crois que je n'ai jamais dépassé celui de bronze…

- Je pense bien sûr à Laetitia et Franck de l'équipe Onectra. Laetitia, merci pour ta joie de vivre, ta gentillesse et ton soutien. Franck, merci pour toutes ces activités sportives ! Le volley aux Vannades, les sorties skis et bien sûr le FOOT ! Tu as été un super coach, j'espère que l'on gagnera plus d'un match cette année !

Je remercie également très chaleureusement les autres membres du LERCM et d'Onectra, pour leurs conseils, leur soutien, pour avoir partagé de très bons moments en salle café ou en mission : Christelle, Benoît (alias chouchou), David Mourier, Pascal (le fameux informateur…), Olivier (merci pour ces billards aux journées des thèses), Alain (je garde un très bon souvenir de la fameuse blague de la chaise mouillée), Laurent, Damien, Gilles Gontier, Gilles Salaun, Sylvie, Fred, Vincent et David Claval.

J'ai également une pensée pour Françoise qui m'a encadré durant mes six premiers mois au LERCM.

Un grand merci à mes collègues docteurs et amis, Christel, Yohann et Christelle pour leur aide, leur soutien et surtout pour tous ces instants partagés à Montpellier… ou à Nice, merci Christel !

Je pense également à tous mes amis. En particulier, mes deux poulettes Johanna et Audrey qui ont toujours été là pour moi. Vous avez suivi mon parcours de très près (« Alors ça avance cette thèse ? »). Merci les filles pour votre soutien et pour tous ces bons moments passés ensemble et à venir. Je pense également à ma chère cousine Mag, pour tous nos délires…

Un grand merci à ma famille et ma belle famille, plus particulièrement à mes parents. Papa, Maman, merci pour la confiance que vous m'avez toujours accordée. Vous m'avez permis de réaliser mes études dans les meilleures conditions. Vous avez toujours cru en moi et vous avez toujours été présents dans les moments difficiles. Pour tout ça, je vous dédie cette thèse.
Louise, Lydie et Jessy merci d'être là. Lydie, merci pour tout, ta compréhension, ton écoute, tes conseils, et puis tes blagues aussi… Tu es la sister de choc, je sais que je pourrai toujours compter sur toi. Merci à Jessy pour tous ces instants de détente. Je pense en particulier à cette fameuse soirée à la Rosière après un bon repas, n'est-ce pas Seb ? Louise, je suis très fière d'être ta marraine, j'espère te transmettre ma passion du travail.

Mes derniers remerciements vont à la personne la plus importante à mes yeux avec qui j'ai hâte de partager ma vie : Sébastien. Merci pour ton amour, ton soutien et tes coups de gueules…

Valorisation du travail

1. Revues scientifiques à comité de lecture :

Articles acceptés

- Briand B., Durand V. and Mercat-Rommens C. Identifying the relationships between agronomic and radioecological variables using a crop model applied to lettuce. Journal of Agronomy, *sous presse*.
- Durand V., Mercat-Rommens C., Curmi P., Benoit M. and Briand B., (2007). Modelling regional impacts of radioactive pollution on permanent grassland. Journal of Agronomy, 6 (1), pp 11-20.
- Mercat-Rommens C., Métivier J.M., Briand B. and Durand V. (2007). How spatial analysis can help in predicting the level of radioactive contamination of cereals. In: geoENV VI – Geostatistics for Environmental Applications, chapter 6, *à paraître*.
- Mercat-Rommens C., Roussel-Debet S., Briand B., Durand V., Besson B. et Renaud P., (2007). La sensibilité radioécologique des territoires : vers un outil opérationnel. Radioprotection Vol. 43, n°3, pp 177-295.

Articles soumis

- Briand B., Mercat-Rommens C. and Ducharme G. A similarity measure for classification trees. Soumis dans Computational Statistics & Data Analysis.

2. Communications à des congrès ou colloques :

Communications orales

- Briand B., Mercat-Rommens C. et Ducharme G., (2008). Une méthodologie statistique pour expliquer les différents niveaux de contamination radioactive des végétaux. 10èmes journées Européennes Agro-industrie et Méthodes statistiques, 23-25 janvier 2008, Louvain-la-Neuve, Belgique.
- Briand B., Mercat-Rommens C. et Ducharme G., (2007). Méthode de stabilisation par rééchantillonnage dans les nœuds pour construire des arbres de classification. XIVe Rencontre de la Société francophone de classification, 5-7 septembre 2007, Paris, France.

- Briand B., Mercat-Rommens C. and Ducharme G., (2007). Using stabilized classification trees in the field of radioecology. IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction, August 30th to September 1st, 2007, Aveiro, Portugal.
- Briand B., Mercat-Rommens C. and Ducharme G., (2007). Using classification trees techniques like sensitivity analysis in the field radioecology. Fifth International Conference on Sensitivity Analysis of Model Output, 18-22 june 2007, Budapest, Hungary.
- Briand B., Mercat-Rommens C. et Ducharme G., (2007). Utilisation d'arbres de classification stabilisés pour expliquer la contamination radioactive des végétaux. 39èmes Journées de Statistique, 11-15 juin 2007, Angers, France.
- Mercat-Rommens C., Métivier J.M., Briand B. and Durand V. (2006). How geostatistics can help in predicting the level of radioactive contamination of cereals. 6th European Conference of Geostatistics for Environmental Applications, 25-27 october 2006, Rhodes, Greece.

Posters

- Briand B., Mercat-Rommens C. et Ducharme G., (2006). Apports de la biostatistique à la radioécologie de terrain. 38ièmes Journées de Statistique, 29 mai au 2 juin 2006, Clamart, France.
- Briand B. and Mercat-Rommens C., (2006). Difficulties and lessons of environmental data processing to fit modelling parameters. SETAC Europe 16th Annual Meeting, 7-11 may 2006, The Hague, Nederland.

Table des matières

Liste des figures

Liste des tableaux

INTRODUCTION

Les conséquences pour l'homme et l'environnement d'une pollution radioactive chronique ou accidentelle dépendent de l'importance et de la nature de celle-ci, mais également du territoire qui la reçoit. La sensibilité radioécologique d'un territoire est définie comme l'intensité de sa réponse globale à une pollution radioactive. Pour un rejet déterminé, plus cette réponse sera élevée plus le territoire sera considéré comme sensible. La connaissance de la sensibilité radioécologique d'un territoire nécessite donc l'étude des facteurs qui déterminent cette sensibilité et qui sont susceptibles de la modifier, c'est-à-dire d'augmenter ou de diminuer les conséquences d'une pollution. Dans cet objectif, l'Institut de Radioprotection et de Sûreté Nucléaire (IRSN) a développé le projet de recherche SENSIB, acronyme pour SENSIBilité radioécologique (Mercat-Rommens et Renaud, 2005 ; Mercat-Rommens et al., 2007). Les deux principaux objectifs de ce projet sont : l'identification des spécificités des territoires français influant fortement sur le devenir d'un contaminant radioactif dans l'environnement et le développement d'outils et de méthodes de traitement de la connaissance afin de caractériser différents territoires vis-à-vis d'une pollution.

Ce travail de thèse se place dans l'un des volets du projet SENSIB orienté sur l'étude des milieux agricoles du territoire métropolitain français en contexte post-accidentel. L'objectif est de proposer une méthode permettant l'identification des facteurs conduisant à différents niveaux de contamination radioactive des végétaux. L'enjeu d'un tel outil est particulièrement important en contexte post-accidentel pour la connaissance des facteurs qui vont conduire à dépasser des niveaux de contamination maximum admissibles. La connaissance sur les caractéristiques des territoires pourra alors être utilisée, de façon anticipée par rapport aux situations accidentelles, pour émettre des recommandations en matière de gestion des territoires contaminés et hiérarchiser la prise de décision.

Pour répondre à cette problématique, la méthodologie que nous proposons est basée sur l'utilisation d'un modèle radioécologique de transfert des radionucléides dans l'environnement et d'une méthode de discrimination par arbre. Le modèle radioécologique sert à la génération d'échantillons artificiels de données relatifs à des contaminations radioactives de productions agricoles. La simulation des données a nécessité la recherche des distributions de valeurs et des éventuelles relations entre les différentes variables d'entrées du modèle étudié. Une fois les échantillons de données générés, une méthode statistique est appliquée permettant l'identification des valeurs spécifiques des entrées du

modèle radioécologique pour lesquelles le niveau de contamination radioactive des productions agricoles dépasse (ou ne dépasse pas) des valeurs limites. Les méthodes de construction d'arbre de discrimination[1] ont alors été identifiées comme des outils particulièrement intéressants, de par leur pouvoir explicatif important, pour répondre à cette problématique environnementale de pollution. De plus, la simplicité de représentation et d'interprétation des arbres en font des outils facilement utilisables par les radioécologistes et les décideurs. Dans un premier temps, la méthode CART (Breiman et al., 1984) a été utilisée. Son application a révélé l'un des principaux inconvénients de cette méthode : son instabilité (Breiman, 1996a). Le modèle de prédiction est assez sensible à de légères altérations de l'échantillon qui a permis sa construction. L'instabilité se remarque au niveau des prévisions et également au niveau de la structure de l'arbre (taille et divisions associées aux nœuds intermédiaires). Pour corriger cette instabilité, différentes approches ont été développées. Par exemple, pour améliorer la stabilité des prévisions, certains auteurs comme Breiman (1996b, 2001) ont proposé des méthodes basées sur l'agrégation d'arbres : le bagging (bootstrap aggregating) ou les forêts aléatoires dites random forests. Le principal désavantage de ces méthodes est la perte de la structure de l'arbre unique et donc de l'ensemble des règles de décision qui en découlent. D'autres auteurs ont préféré privilégier la structure de l'arbre : Dannegger (2000) et Ruey-Hsia (2001) ont développé des algorithmes permettant de stabiliser la procédure de sélection des divisions associées aux nœuds intermédiaires. Ces dernières méthodes sont particulièrement intéressantes dans notre contexte d'application. Elles permettent de réaliser le compromis souhaité : conserver l'aspect visuel de l'arbre (outil facilement exploitable par les décideurs) et stabiliser sa structure (indispensable dans un contexte post-accidentel car le décideur nécessite de disposer de règles de décision robustes).

Dans cette thèse, nous nous sommes donc attachés à développer une méthode de construction d'arbre de discrimination permettant de tenir compte de l'instabilité des arbres. La méthode proposée peut être vue comme une amélioration de la technique de Dannegger (2000), nous l'appelons la méthode REN[2]. En effet, nous avons identifié une faiblesse dans les travaux de Dannegger (2000). Ce dernier a développé un algorithme de stabilisation par rééchantillonnage bootstrap dans les nœuds et l'a appliqué à la méthode CART. Cependant, son approche présente un inconvénient : l'élagage, défini par Breiman et al., (1984) ne semble pas être applicable aux arbres construits selon son algorithme (dû

[1] En anglais *classification tree*, mais il peut se traduire en français par arbre de discrimination ou de classement. Les méthodes de classement consistent à attribuer une classe préexistante à un individu statistique. Par opposition, la classification (clustering, en anglais) correspond à la recherche de classes par groupement d'individus statistiques présentant les mêmes caractéristiques.

[2] Acronyme pour REéchantillonnage bootstrap dans les Nœuds.

à la définition des divisions qui ne sont plus optimales selon le critère de Breiman et al.). De ce fait, nous proposons d'élaguer les arbres par une méthode plus adaptée, basée sur les travaux de Quinlan (1987).
La stabilisation de la structure de l'arbre par cette méthode, et les méthodes similaires, est le plus souvent évaluée qualitativement. Il paraît assez difficile de comparer rigoureusement diverses méthodes destinées à stabiliser les procédures de construction d'arbre. Nous proposons également, en complémentarité de la méthode REN, une mesure de similarité afin d'évaluer quantitativement la proximité de deux arbres de discrimination. Cette mesure peut également être utilisée pour déterminer la « dispersion » d'un ensemble d'arbres et identifier un arbre « central ». Dans le contexte de notre application, nous l'utilisons principalement pour étudier les performances de stabilisation de la méthode REN comparativement à la méthode CART.

Le scénario de contamination utilisé pour le développement de cette méthodologie est un scénario simplifié d'une contamination, par dépôts de radioactivité lors de temps sec, de différents légumes-feuilles (laitue, épinard, chou et poireau) à la suite d'un rejet radioactif accidentel de ^{90}Sr.

Organisation du document
Le document s'articule en deux grandes parties :

PARTIE 1

La première partie de ce document consiste en la présentation des outils et méthodes utilisés pour le développement de notre méthodologie. Après une brève présentation des principaux concepts de radioécologie utilisés, nous exposons de manière générale dans le chapitre 1 la méthodologie développée ainsi qu'un état de l'art. Dans le chapitre 2, nous présentons la démarche qui a conduit à la génération d'échantillons artificiels de données. Le choix du modèle radioécologique est présenté ainsi que les outils et méthodes utilisés pour renseigner les variables d'entrées de ce modèle. Le chapitre 3 présente de manière générale le principe de construction des arbres de décision et traite de la méthode CART (Classification And Regression Trees), méthode de référence pour la construction d'arbres de régression et de discrimination. Dans le chapitre 4, nous discutons de l'instabilité de cette méthode et définissons en quoi elle est gênante pour notre application radioécologique. Les techniques les plus connues, développées pour pallier à ce problème, sont présentées. En particulier, nous focalisons sur un algorithme permettant la

stabilisation de la structure d'un arbre de discrimination par rééchantillonnage bootstrap dans les nœuds. Nous proposons une méthode permettant de construire des arbres de discrimination par cet algorithme (méthode REN). Nous définissons également une mesure de similarité, permettant de quantifier la proximité de deux arbres de discrimination.

PARTIE 2

La deuxième partie du document consiste en la présentation et l'analyse des résultats obtenus suite à l'application de notre méthodologie. Dans le chapitre 5, la constitution des échantillons artificiels de données pour notre scénario de contamination est présentée. Dans le chapitre 6, nous nous intéressons au choix de la taille des échantillons et des variables explicatives pour la construction des arbres de discrimination. Le chapitre 7 présente une comparaison empirique entre la méthode CART et la méthode REN, pour un cas de contamination. Les performances de ces deux méthodes sont comparées ainsi que la structure des arbres de discrimination obtenus. Dans le dernier chapitre (8), nous présentons les arbres de discrimination obtenus pour chaque légume-feuille de notre scénario de contamination et nous procédons à l'analyse et à l'interprétation de ces arbres.

PARTIE 1 : MATERIELS ET METHODES

1 Généralités

1.1 Quelques notions de radioécologie

La radioécologie a pour objectif l'évaluation de l'impact de la radioactivité naturelle et artificielle sur l'environnement et sur la population (Foulquier et Bretheau, 1998 ; IRSN, 2004). Dans ce chapitre, nous proposons au lecteur quelques notions de radioécologie, en particulier sur le transfert des radionucléides aux végétaux, qui faciliteront la lecture et la compréhension de certaines parties de ce document.

1.1.1 Les radionucléides

Les radionucléides sont des atomes dont les noyaux sont radioactifs. Deux types de radionucléides sont distingués :

- Les radionucléides naturels. Ces radionucléides sont présents dans tous les milieux de l'environnement. Par exemple, le radon 222 dans l'air extérieur, le potassium 40 dans l'eau de mer ou encore l'uranium 238 dans les sols.
- Les radionucléides artificiels. Ils sont d'origine humaine et proviennent d'activités civiles et militaires. La principale source de radioactivité artificielle dans l'environnement est due aux essais atmosphériques d'armes nucléaires. Elle est également due aux accidents d'installations nucléaires (ex : accident de Tchernobyl), aux rejets contrôlés de l'industrie nucléaire, aux utilisations médicales (radiothérapie par exemple),…

Chaque radionucléide est caractérisé par une période radioactive ou de demi-vie. Elle représente le temps nécessaire pour que la moitié des atomes radioactifs initialement présents se soient désintégrés. Au bout d'une période, l'activité est donc diminuée de moitié.

1.1.2 Les dépôts de radionucléides

La contamination de l'écosystème terrestre est en général le résultat de dépôts atmosphériques. En fonction de la météo et de la topographie, les radionucléides présents dans l'air peuvent se déposer de façon plus ou moins importante. Par temps sec, les radionucléides se déposent sous l'effet du vent et des turbulences qu'il engendre (dépôt sec). Par temps de pluie, les gouttes d'eau entraînent les particules vers le sol. Sous les averses, les dépôts dits humides peuvent être 10 fois plus importants. A la rencontre d'un relief, les turbulences de l'air augmentent et les précipitations s'intensifient, les dépôts

sont plus forts qu'en plaine. De même, une végétation haute et couvrante intercepte efficacement les radionucléides (Renaud et al., 2007).

1.1.3 Le transfert des radionucléides aux végétaux

Suite à un dépôt de radionucléides, deux grandes voies de transfert aux végétaux coexistent (Cf. Figure 1.a) (Renaud et al., 1997a) :

- le transfert foliaire (1). Les radionucléides contenus dans l'air ou dans l'eau de pluie sont interceptés par les feuilles et les tiges du végétal au moment où s'effectuent les dépôts. Cette interception sera d'autant plus efficace que la culture est développée et couvre le sol. La migration de certains de ces éléments vers la partie consommable de la plante est appelée translocation. Ce phénomène est plus intense durant la maturation du végétal (formation des grains par exemple).
- le transfert racinaire (3). Ce transfert à lieu lorsque les radionucléides ont pénétré suffisamment dans le sol pour qu'ils puissent y être puisés par la plante (2). Le labour est le plus souvent le facteur d'incorporation et d'homogénéisation des contaminants dans l'horizon racinaire.

L'activité massique de la production végétale à sa récolte peut aussi résulter de remise en suspension, par le vent ou lors de travaux agricoles, d'éléments déposés sur le sol, ou encore de l'éclaboussement des feuilles, lors des pluies, par des particules de sol contaminées.

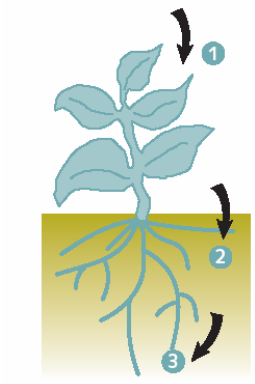


Figure 1.a : Transfert des radionucléides aux végétaux : les radionucléides contaminent directement les feuilles des végétaux (1), ils migrent ensuite dans le sol (2) où interviennent les transferts vers les racines (3) (IRSN, 2004)

Certaines espèces agronomiques présentent des similitudes vis-à-vis des modes de transfert de radioactivité et peuvent être assimilées, elles appartiennent alors à une unité radioécologique. Pour les produits maraîchers (étudiés dans le présent travail), trois grands groupes sont distingués (Calmon et Mourlon, 2003) :

- les légumes-feuilles : cultures maraîchères dont les feuilles sont consommées (ex : salade, épinard,…),
- les légumes-fruits : cultures maraîchères dont les fruits sont consommés (ex : tomate, petit pois,…),
- les légumes-racines : cultures maraîchères dont les racines ou les tubercules sont consommés et assimilés (ex : carottes, pomme de terre,…).

1.1.4 Le risque lié aux activités radiologiques et nucléaires

En cas d'accident impliquant le rejet de substances radioactives dans l'environnement, deux phases d'intervention sont généralement distinguées : une phase dite d'urgence nécessitant une réponse rapide et organisée dans le cadre des plans d'intervention (évacuation, mise à l'abri,…) et une phase différée de traitement des conséquences de l'événement, dite post-accidentelle. Il est donc important de disposer de façon anticipée d'outils méthodologiques utilisables pour ces phases d'urgence et de gestion post-accidentelle afin d'évaluer les conséquences sur les nombreux domaines impactés (agriculture, eau, gestion des déchets,…). La méthodologie que nous proposons a pour but d'expliquer et de prédire les niveaux de contamination radioactive dans certaines productions agricoles et donc d'apporter des éléments d'information utiles pour la gestion des territoires contaminés. Cette méthodologie est présentée de manière générale dans le paragraphe suivant.

1.2 Présentation générale de la méthodologie

1.2.1 Présentation de la méthodologie utilisée

La méthodologie développée est basée sur l'utilisation d'un modèle radioécologique de transfert et une méthode de discrimination par arbre (Cf. Figure 1.b). Le présent travail ne vise pas à valider le modèle radioécologique utilisé mais à montrer l'efficacité des arbres de discrimination.

La première étape de cette méthodologie consiste en la constitution d'échantillons de données. Dans un premier temps, nous avons cherché à utiliser des mesures de radioactivité dans le milieu agricole français. En effet, suite aux essais atmosphériques d'armes nucléaires et à l'accident de Tchernobyl, de nombreuses mesures de radioactivité sont disponibles (Cf. paragraphe 2.2). Cependant, les caractéristiques de prélèvements associées à ces mesures sont souvent peu nombreuses et imprécises. Ce manque d'information a donc conduit à recourir à un modèle radioécologique de transfert des radionucléides dans l'environnement pour générer des échantillons artificiels de données. Les variables d'entrées du modèle radioécologique sont identifiées comme les différentes

caractéristiques (agronomiques, radioécologiques,…) susceptibles de modifier les conséquences d'une pollution. En général ces variables sont de type quantitatives (dépôt de radioactivité le jour de l'accident, temps de croissance du végétal étudié,…).

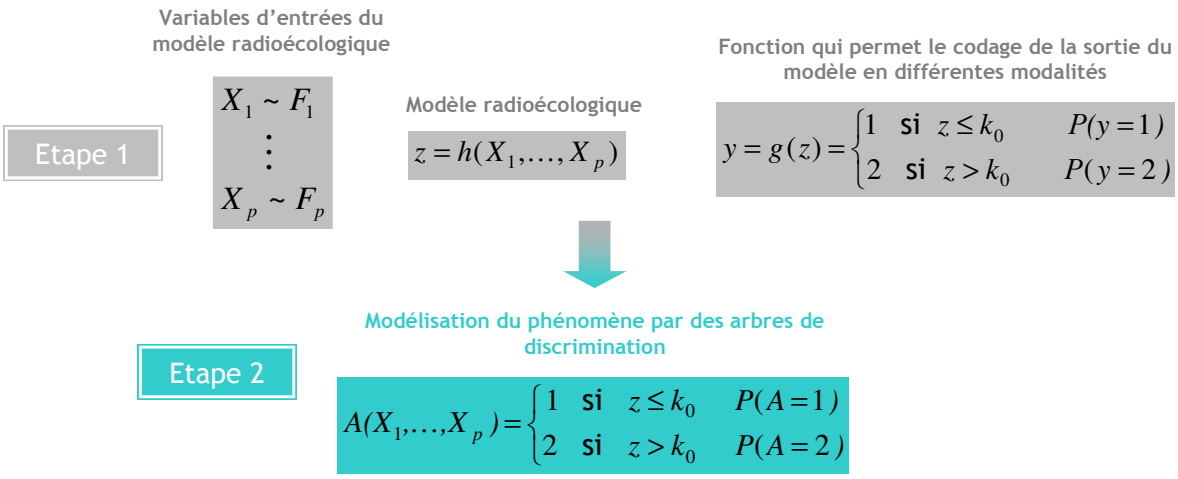


Figure 1.b : Présentation générale de la méthodologie (cas d'un codage de la sortie du modèle en deux modalités)

Pour chacune des entrées du modèle radioécologique, un travail de caractérisation est effectué afin de déterminer leur distribution de probabilité et de mettre en évidence d'éventuelles relations entres ces variables. La variable de sortie du modèle radioécologique est une variable quantitative, elle représente la contamination radioactive du végétal étudié. Pour répondre aux objectifs de l'étude, cette variable est codée selon des limites fixées par la radioprotection. Par exemple, sur la figure 1.b, la sortie du modèle radioécologique (z) a été codée en deux modalités (modalité 1, modalité 2) selon une valeur seuil (k_0). Les probabilités d'appartenance à chaque classe sont alors définies.

La deuxième étape de la méthodologie consiste à construire des arbres de discrimination à partir des échantillons de données créés lors de l'étape précédente. Les arbres de discrimination ont été identifiés comme un axe de recherche appliquée particulièrement intéressant pour développer des modèles de classement dans le cadre du projet SENSIB en raison de :

- leur simplicité de représentation et d'interprétation qui en font des outils facilement utilisables par les non-spécialistes,
- leur pouvoir explicatif important : les règles de décision fournies par les différentes branches de l'arbre permettent de mettre en évidence les variables explicatives et/ou les associations de variables explicatives qui conduisent aux différentes modalités de la variable à expliquer.

Une fois l'arbre de discrimination construit, nous exploitons à la fois son côté explicatif et son côté prédictif. L'idée est de pouvoir répondre à certaines questions qui se poseraient en contexte post-accidentel : quels sont les facteurs (environnementaux, anthropiques,…) qui conduisent à de forts niveaux de contamination radioactive des végétaux ? Peut-on identifier a priori les végétaux qui seraient fortement contaminés à la récolte?, etc. On pourrait alors envisager d'utiliser les règles les plus fiables (extraites de l'arbre de discrimination), afin de recommander des actions dans un contexte post-accidentel.

1.2.2 Etat de l'art

Dans la littérature, la méthodologie définie dans le paragraphe précédent peut être considérée comme une méthode particulière d'analyse de sensibilité (Mishra et al., 2003 ; Mokhtari et al., 2006). D'une manière générale, les méthodes d'analyse de sensibilité consistent à analyser et à quantifier l'influence des variables d'entrées d'un modèle sur la (les) sortie(s) de ce modèle. Il existe différents types d'analyse de sensibilité, Saltelli et al. (2000) proposent une classification de ces méthodes selon leurs capacités :

- Les méthodes de screening (Morris, 1991)

Ces méthodes sont généralement basées sur une configuration « un-à-la-fois » (one-at-time, OAT) : l'impact du changement d'une variable d'entrée est évalué (en fonction de son influence sur la variabilité de la sortie du modèle) alors que les autres entrées restent constantes. Ces méthodes fournissent une mesure de la sensibilité de type qualitative, elles permettent un classement par ordre d'importance des entrées, mais cette importance ne peut être quantifiée.

- Les méthodes locales (Turányi, 1990 ; Turányi and Rabitz, 2000)

Comme pour les méthodes de screening, ces techniques se concentrent sur l'impact local des entrées sur la (les) sortie(s) du modèle. Elles sont généralement basées sur le calcul de dérivées partielles de la fonction de sortie du modèle, par rapport aux variables d'entrées. Contrairement aux méthodes de screening, elles fournissent une mesure de la sensibilité de type quantitative. Depuis quelques années, il a été démontré (Cukier et al., 1973) que lorsque le modèle est non linéaire et que de nombreuses variables d'entrées sont affectées par des incertitudes, il est préférable de recourir à une méthode d'analyse de sensibilité globale.

- Les méthodes globales (Sobol, 1993 ; Cukier et al., 1973, 1975, 1978 ; Schaibly et Shuler, 1973)

Ces méthodes consistent à étudier la variabilité de la sortie d'un modèle due à la variabilité de ses entrées, elles peuvent être définies à l'aide de ces deux propriétés :

- Pour chaque entrée du modèle, le domaine de variation ainsi que la forme de la fonction de distribution de probabilité sont pris en compte pour estimer la mesure de sensibilité.
- Les estimations de sensibilité de chaque entrée du modèle sont évaluées alors que toutes les autres entrées varient en même temps.

Ces méthodes sont basées sur des techniques de simulation permettant de multiples évaluations du modèle. Par exemple, la méthode Monte Carlo (Robert et Casella, 1999) avec échantillonnage aléatoire simple consiste à sélectionner aléatoirement les valeurs des variables d'entrées du modèle selon leur distribution de probabilité et à évaluer la sortie de ce modèle. Les résultats de ces multiples évaluations sont alors utilisés pour étudier l'influence de la variabilité des entrées sur la variabilité de la sortie du modèle. Parmi les différentes méthodes globales existantes, les plus connues et utilisées sont celles basées sur l'étude de la variance. Elles consistent à estimer quelle part de variance de la sortie du modèle est due à la variance d'une entrée, ou d'un groupe d'entrées. L'estimation de ces indices peut alors se faire par différentes méthodes, les plus utilisées sont les méthodes de Sobol (Sobol, 1993) et FAST (Cukier et al., 1973, 1975, 1978 ; Schaibly et Shuler, 1973).

Plus récemment, une méthodologie originale d'analyse de sensibilité globale est apparue, utilisant les techniques d'arbre de discrimination (Mishra et al., 2003) et de régression (Mokhtari et al., 2006). Les arbres de discrimination permettent de traiter une variable de sortie qualitative, ce qui n'est pas souvent le cas avec les méthodes d'analyse de sensibilité habituelles plus restreintes aux sorties quantitatives. Cette méthodologie, qui peut être complémentaire aux méthodes classiques, permet de déterminer les combinaisons de plages de valeurs spécifiques des entrées du modèle qui engendrent des sorties particulières. Elle permet ainsi de mieux comprendre le phénomène modélisé en soulignant certaines relations entres les variables d'entrées et la variable de sortie du modèle étudié (par exemple, identification des variables d'entrées qui conduisent à des valeurs extrêmes de la sortie (Mishra et al., 2003)).

La première application des techniques d'arbres de discrimination en tant que méthode d'analyse de sensibilité apparaît dans le domaine du nucléaire (Mishra et al., 2003). L'objectif de cette étude était d'évaluer si une mine géologique (Yucca Mountain) située dans l'Etat du Nevada, Etats-Unis, pourrait être utilisée comme un site de stockage de déchets hautement radioactifs (combustibles usagés provenant de réacteurs nucléaires situés sur le territoire des Etats-Unis et autres déchets hautement radioactifs sous forme

de verre vitrifié). Pendant les vingt dernières années, le département de l'énergie des Etats-Unis (US Department of Energy, DOE), a effectué de nombreux travaux et enquêtes pour évaluer ce site en tant que site de stockage. Pour évaluer à long terme (milliers d'années), les performances de ce site potentiel (capacités à isoler les déchets radioactifs), un modèle a été développé (Total System Performance Assessment, TSPA). En prenant en compte de nombreux processus (physiques, radioécologiques,…), ce modèle permet de prédire, à long terme, le débit de dose pour une personne vivant à 20 km à l'aval du site. L'application des techniques d'arbres de discrimination, en tant qu'analyse de sensibilité globale, a alors permis d'identifier les variables d'entrées clés responsables des valeurs de dose extrêmes (faibles, fortes) et ceci pour différents scénarii d'application (dose à 70 000 et à 100 000 ans).

Une autre application récente de cette méthodologie, cette fois-ci basée sur les techniques d'arbres de régression, a été publiée dans *Journal of Food Protection* (Mokhtari et al., 2006). Le modèle étudié, *E. coli* O157 :H7, est un modèle de prédiction des risques alimentaires développé aux Etats-Unis, par le département de l'agriculture (US Department of Agriculture). Il permet d'estimer, tout au long du processus de transformation de la viande en steak haché la présence de la bactérie *E. coli*. Ce modèle est composé de divers modules (abattage, préparation,…) où chaque module peut avoir des entrées qui sont des sorties du module précédent. Le modèle est assez complexe, il est caractérisé par des non linéarités, des interactions entre les variables d'entrées, des points de saturation[3], des variables d'entrées quantitatives et qualitatives. Il est alors difficile d'appliquer une méthode d'analyse de sensibilité permettant de considérer l'ensemble de ces caractéristiques. Les méthodes de construction d'arbres de discrimination et de régression ont alors été identifiées comme des techniques prometteuses permettant de tenir compte des différentes particularités du modèle. Les sorties des modules « abattage » et « préparation » du modèle *E. coli* O157 :H7, étant quantitatives, les arbres de régression ont été utilisés comme méthode d'analyse de sensibilité globale. Ils ont permis d'identifier les variables et les associations de variables conduisant à des valeurs particulières de la sortie de chaque module considéré.

Contrairement aux études décrites ci-dessus, notre travail permet de tenir compte de l'instabilité des arbres de discrimination. La méthode de construction d'arbre de discrimination proposée est basée sur un algorithme de stabilisation des nœuds intermédiaires (Briand et al., 2007).

[3] Valeur d'une variable d'entrée à partir de laquelle cette dernière ne produit plus d'effet sur la sortie du modèle.

2 Les données utilisées

2.1 Introduction

Pour construire des arbres de discrimination, des données relatives à des contaminations radioactives sur le territoire français sont nécessaires. Dans un premier temps, nous avons cherché à utiliser des résultats de mesures effectuées dans l'environnement français. Une synthèse de ces données disponibles est présentée dans le paragraphe 2.2. Aboutissant au constat de l'impossibilité de traiter ces résultats de mesures, nous proposons dans le paragraphe 2.3 de générer des échantillons artificiels de données à partir d'un modèle radioécologique de transfert des radionucléides dans l'environnement. Après avoir exposé le scénario de contamination étudié, le choix du modèle radioécologique utilisé est présenté ainsi que les outils et méthodes qui ont été nécessaires pour renseigner les variables d'entrées de ce modèle.

2.2 Synthèse des données de mesures disponibles

2.2.1 Les essais atmosphériques d'armes nucléaires

Depuis 1945, plus de 2400 essais nucléaires, dont 543 essais atmosphériques, ont été réalisés par les Etats-Unis, la Russie, la Grande-Bretagne, la France et la Chine. Les sites des essais étaient répartis sur l'ensemble du globe, mais la plupart des explosions ont eu lieu dans l'hémisphère Nord. A partir de 1961, les tirs souterrains ont progressivement remplacé les explosions aériennes.

Après chaque explosion, les particules radioactives sont libérées dans l'atmosphère à une altitude qui dépend des conditions du tir. Elles y séjournent de quelques heures à quelques mois avant de retomber. Compte tenu de la répartition des sites, du nombre d'essais et de la variété de puissance des tirs, les dépôts ont affecté la planète toute entière. Dans l'hémisphère Nord, qui a reçu 75 % des retombées radioactives, les grandes circulations d'air ont concentré les dépôts dans les régions tempérées, notamment dans la bande comprise entre le 40° et le 50° de latitude, dans laquelle se trouve la France (Vray et Renaud, 2004).

La France a été l'un des premiers pays à mettre en place un réseau complet de surveillance de la radioactivité dans l'environnement. Dès 1961, les retombées radioactives consécutives aux essais atmosphériques ont été mesurées par différents organismes : le Service Central de Protection contre les Rayonnements Ionisants (SCPRI) et le Sous-comité Interministériels de protection Sanitaire devenu par la suite CEA/DPS[4] – organismes regroupés aujourd'hui au sein de l'IRSN. Ces mesures de radioactivité ont donné lieu à des

[4] Commissariat à l'Energie Atomique / Département de Protection Sanitaire.

rapports publics (mensuels pour le SCPRI et trimestriels pour le DPS). Ces rapports fournissent des résultats de mesures de différents radionucléides dans l'environnement et plus particulièrement dans la chaîne alimentaire (un exemple d'un extrait d'un rapport CEA/DPS est donné en annexe A). De 1961 à 1980, les réseaux du SCPRI et du CEA (Cf. Figure 2.a) fournissent près de 50 000 résultats de mesures.

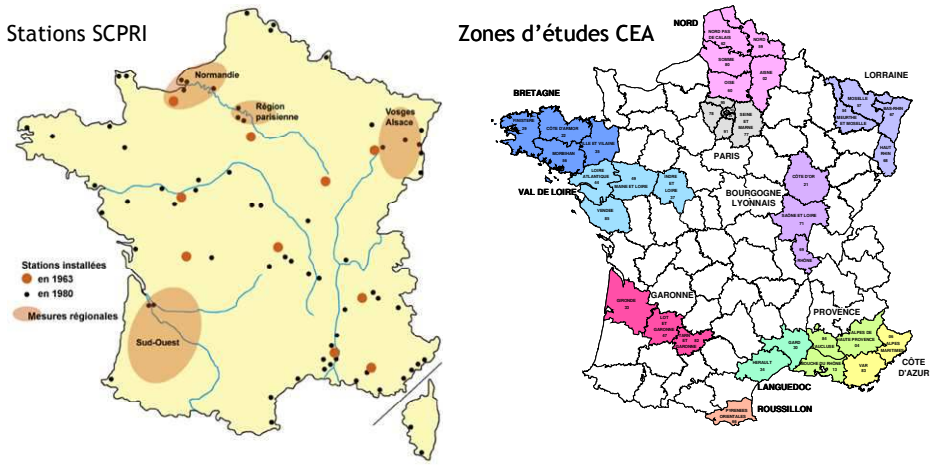


Figure 2.a : Régions et stations ayant donné lieu à des prélèvements (Renaud et al., 2003 ; Vray, 2002)

2.2.2 L'accident de Tchernobyl

Le 26 avril 1986, le plus grave accident survenu dans l'industrie nucléaire civile se produit sur le réacteur 4 de la centrale de Tchernobyl, en territoire ukrainien de l'Union soviétique. Cet accident provoque d'importants rejets radioactifs dans l'atmosphère pendant plusieurs jours. Le vent a d'abord emporté les substances radioactives rejetées le 26 avril vers le Nord-Ouest. Parvenu au-dessus des pays baltes et scandinaves, ce premier panache a été rabattu vers l'Est puis le Sud, ramenant les polluants radioactifs vers l'Europe centrale et balkanique. Les rejets du 27 avril ont été entraînés vers l'Ouest. Ceux de la nuit (0 - 6 heures) ont abordé la France par l'Est et ceux de l'après-midi par le Sud. Les rejets émis par la centrale à partir du 28 avril ont été emportés vers l'Est et le Sud : Russie, Méditerranée orientale et Europe centrale (Renaud et al., 2007).

Les rapports mensuels publiés par le SCPRI en avril et mai 1986 mettent à disposition de nombreuses mesures de radioactivité sur divers produits de la chaîne alimentaire (légumes, fruits, céréales,…).

Ainsi, suite aux essais atmosphériques d'armes nucléaires et à l'accident de Tchernobyl, de nombreuses mesures de radioactivité, effectuées dans l'environnement français, sont à disposition. Nous avons donc envisagé de travailler sur des échantillons de données relatifs au milieu agricole. A titre d'exemple, la figure 2.b présente, sur la période 1961-1980, les résultats de mesures disponibles en ^{90}Sr pour le légume-feuille poireau en distinguant les différentes régions de prélèvement (données issues des bulletins trimestriels CEA/DPS).

Pour tenter d'expliquer les causes des différents niveaux de contamination observés, une évaluation de la contribution des différentes variables explicatives est alors nécessaire. Intuitivement, la première variable susceptible d'influer les niveaux de contamination est l'ampleur du dépôt de radioactivité. Les facteurs agronomiques associés au végétal étudié peuvent aussi être considérés comme des variables explicatives (temps de croissance, rendement, stade de développement le jour de la mesure, …), de même pour les caractéristiques du sol (nature du sol, masse volumique,…). Cependant, lors des prélèvements effectués par les deux organismes cités précédemment, ces différents facteurs n'étaient pas systématiquement et précisément renseignés conjointement aux mesures de radioactivité. Par exemple, pour les données issues des rapports CEA/DPS, les caractéristiques de prélèvements associées aux mesures étaient : la nature de l'échantillon prélevé (lait, poireau, salade,…), le radionucléide mesuré (^{90}Sr, ^{137}Cs, ^{131}I,..), le mois et l'année du prélèvement (la date de mesure exacte dans le mois n'est pas connue) et le lieu (seule la région est identifiée, Cf. Figure 2.a, carte de droite). De plus, les valeurs de dépôt de radioactivité ne sont pas connues avec précision, elles sont le plus souvent estimées à partir de mesures effectuées dans l'air (Renaud et al., 2003). Ainsi, l'application et/ou le développement d'une méthodologie statistique à partir de ces données n'est pas envisageable à cause du manque d'information et de précision sur les variables explicatives potentielles.

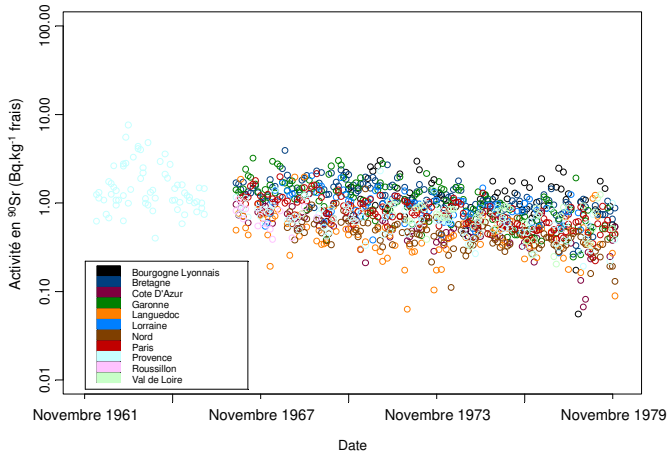


Figure 2.b : Représentation graphique de différents résultats de mesures du ^{90}Sr dans le légume-feuille poireau pour la période 1961-1980 (bulletins trimestriels CEA/DPS)

Cette impossibilité de traiter les données de mesures a également été mise en évidence lors de travaux portant sur l'estimation de trois principaux paramètres radioécologiques de transfert (Briand et al., 2006 ; Briand et Mercat-Rommens, 2006a,b). Cette étude traitant de la sensibilité de la contamination végétale aux rejets chroniques[5] est résumée dans le paragraphe suivant et présentée en détail dans l'annexe B.

2.2.3 L'impossibilité de traiter les données de mesures : exemple de l'estimation de paramètres d'une équation de transfert à partir de résultats de mesures

L'objectif initial de ce travail était d'obtenir des estimations régionales des principaux paramètres de transfert d'une équation d'un modèle radioécologique et, par le biais de ces valeurs, des renseignements (si possibles régionalisés) sur ces paramètres : intervalle de variation et/ou distribution. Nous envisagions d'utiliser certaines de ces informations pour réaliser la première étape de notre méthodologie (Cf. Figure 1.b).

Ces estimations ont été réalisées à partir de différents jeux de données disponibles, en particulier, à partir de mesures de divers radionucléides effectuées en France sur des légumes-feuilles sur la période 1960-1980 relative aux retombées des essais atmosphériques d'armes nucléaires. L'équation de transfert de la contamination radioactive vers les végétaux a été ajustée à partir de ces divers jeux de données. Dans l'ensemble, les nombreux cas étudiés ont mis en évidence une mauvaise qualité de l'ajustement. Ainsi, même si la plupart des valeurs estimées pour les trois paramètres de cette équation sont cohérentes avec la littérature (Cf. Annexe B), elles ne peuvent pas être utilisées pour proposer des informations régionalisées.

Dans ce paragraphe, nous nous contentons d'exposer deux des cas obtenus et de discuter des potentielles causes de ces mauvais ajustements. Nous avons choisi de présenter des résultats relatifs à l'exemple précédent (Figure 2.b). Il s'agit de la confrontation des valeurs prédites et des valeurs observées pour l'activité du ^{90}Sr dans les poireaux en région Nord et en région Bourgogne Lyonnais (Figure 2.c). Ces deux cas sont assez représentatifs de l'ensemble des résultats obtenus. D'autres exemples peuvent également être visualisés en annexe B. Le plus souvent, les prédictions d'activité par l'équation de la contamination des végétaux (dont les paramètres ont été ajustés) conduisent à de forts écarts avec les résultats de mesures. D'une manière générale, les faibles valeurs d'activité sont surestimées, tandis que les plus fortes sont sous-estimées.

[5] Contamination continue dans le temps.

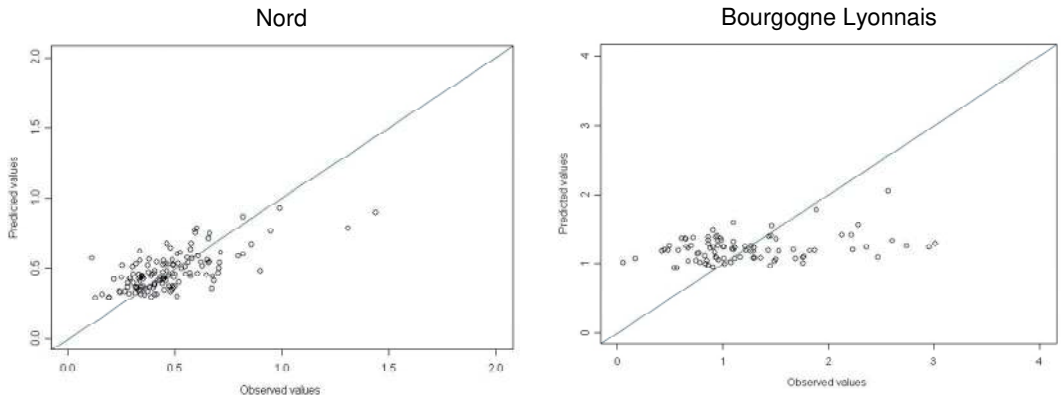


Figure 2.c : Valeurs prédites en fonction des valeurs observées (en Bq.kg^{-1} frais) pour les triplets (poireau, ^{90}Sr, Nord) et (poireau, ^{90}Sr, Bourgogne Lyonnais)

Pour certains cas, la présence de « points isolés » conduit à s'interroger sur la validité de certaines mesures. En effet, il est possible que certains de ces points soient aberrants.

La mauvaise qualité de prédiction obtenue semble être liée à la fois à l'équation du modèle radioécologique et aux données utilisées.
Le fait que l'équation suppose un dépôt constant sur le pas de temps de l'étude, qui dans notre cas était le mois, n'est peut-être pas vérifié. En effet, le dépôt est lissé sur le mois et ces variations sont négligées. Pour les prendre en compte, il faudrait réduire le pas de temps de l'étude en travaillant par exemple au pas de temps hebdomadaire, mais les données ne sont pas disponibles à cette échelle de temps.
Le manque de précision sur les données contribue probablement aussi à ces problèmes d'ajustements :

- Certaines entrées de l'équation sont fixées (temps de croissance ou masse volumique du sol), à défaut de mesures disponibles pour chaque observation de la variable réponse (activité dans le végétal).
- Les séries de données disponibles sur les activités massiques dans les végétaux sont imprécises en termes de date et de lieu. Le fait que les dates de prélèvement aient été affectées au 28 de chaque mois (Cf. Annexe B) contribue vraisemblablement à expliquer les écarts constatés entre valeurs prédites et observées.
- Certaines mesures semblent aussi très élevées : il est possible que les végétaux aient mal été lavés (présence de terre), les radionucléides présents dans la terre influençant alors fortement la concentration finale.
- L'activité dans les végétaux n'a peut-être pas été mesurée lorsqu'ils étaient à maturité.

Pour pouvoir utiliser des données de matrices environnementales afin de renseigner un modèle compartimental de transfert, il apparaît indispensable de connaître le plus précisément possible le terme source (ici l'activité déposée) et les caractéristiques de prélèvement associées à la mesure de l'activité massique. Les enseignements qui peuvent être tirés de cette analyse sont donc des recommandations, pour l'acquisition de données, en vue de l'estimation de paramètres d'un modèle radioécologique. Les améliorations suivantes peuvent être proposées :

1. Effectuer des mesures hebdomadaires (ou journalières) du dépôt, ou se donner les moyens de reconstituer les valeurs du dépôt à cette échelle de temps. Dans le cas du pas de temps mensuel, le dépôt sec est supposé partout identique à celui mesuré en région parisienne. Cette hypothèse n'est plus valide lorsqu'il s'agit d'un pas de temps hebdomadaire (ou journalier), car les masses d'air n'arrivent pas toutes la même semaine (ou le même jour) dans les différentes régions françaises. Afin de travailler par région, il faudrait alors disposer de mesures régionales hebdomadaires (ou journalières) du dépôt ou de méthode d'estimation de ces valeurs. Les travaux en cours dans le cadre de la thèse de L. Bourcier sur l'approche événementielle du dépôt (volet atmosphérique du projet SENSIB) ont pour objectif d'améliorer la connaissance des dépôts à l'échelle des événements pluvieux (Masson et al., 2005).

2. Renseigner lors du prélèvement certaines caractéristiques précises du végétal (stade de développement), du sol (nature du sol, masse volumique) et la localisation temporelle et spatiale de l'échantillon. A l'heure actuelle, les prélèvements réalisés au LERCM dans le cadre des études de terrain et stockés dans la base de données SYLVESTRE remplissent le plus souvent ces conditions. Les caractéristiques suivantes sont notamment référencées : date, lieu, coordonnées GPS, masse volumique et certaines caractéristiques du sol (pour environ un tiers des échantillons de légumes-feuilles). Depuis quelques années, pour certains échantillons, le stade de développement du végétal est aussi précisé en commentaire (plant, maturité, stade monté).

2.3 La simulation des données

2.3.1 La génération d'échantillons artificiels de données

L'impossibilité de traiter les données de mesures, nous a conduit à utiliser un modèle radioécologique de transfert des radionucléides dans l'environnement pour générer des échantillons artificiels de données. Pour cela, nous allons utiliser la méthode d'échantillonnage classiquement employée en analyse d'incertitude de modèle ou en

analyse de sensibilité globale (Cf. 1.2.2) : l'échantillonnage aléatoire simple. Elle permet de générer aléatoirement des valeurs de variables d'entrées à partir de leur loi de probabilité. Ainsi, pour simuler n valeurs de la variable de sortie du modèle, la procédure suivante est utilisée :

- choix des distributions de probabilité pour les variables d'entrées du modèle,
- recherche d'éventuelles relations entres les variables d'entrées,
- génération aléatoire de n valeurs des variables d'entrées (selon les distributions qui les caractérisent et en prenant en compte les éventuelles relations entre les variables),
- évaluation des n valeurs de sortie du modèle,
- codage de la sortie du modèle en différentes modalités.

Ainsi, via cette approche, des échantillons de données sont générés, constitués de p variables explicatives ($X_1,...,X_p$) et d'une variable à expliquer Y qualitative.

2.3.2 Le scénario de contamination

Le scénario de contamination considéré est un cas simplifié. Nous préférons consacrer le présent travail au développement de la méthodologie. Une fois testée sur ce cas simplifié, nous pourrons envisager de l'appliquer à un scénario de contamination plus complexe.

Nous avons choisi de travailler sur la contamination par voie sèche de légumes-feuilles à la suite d'un rejet radioactif accidentel de strontium 90 dans l'atmosphère. Le choix de cette catégorie de végétaux est particulièrement pertinent dans le contexte post-accidentel car c'est souvent la plus rapidement exposée aux retombées de polluants atmosphériques. La contamination de quatre principaux légumes-feuilles (en terme de consommation en France (Combris et al., 1995)) appartenant à cette catégorie est étudiée : l'épinard, le poireau, le chou ainsi qu'une variété de salade, la laitue. Le champ d'étude est restreint au transfert des contaminants par voie foliaire car ce mode de transfert est prépondérant par rapport à la voie racinaire durant la première année suivant l'accident. La principale limite de notre scénario de contamination est la nature du dépôt. En effet, le transfert foliaire des contaminants radioactifs s'effectue par voie sèche ou humide, selon les conditions météorologiques le jour de l'accident. Cependant, il est peu fréquent qu'un dépôt s'effectue exclusivement par voie sèche ou par voie humide (Renaud et al., 1997a). De plus, lors d'un accident, de nombreuses particules radioactives sont libérées. Nous avons choisi de limiter le scénario de contamination à un radionucléide : le strontium 90, qui fait partie des radionucléides potentiellement émis en cas d'accident.

2.3.3 Le choix du modèle radioécologique : le code de calcul ASTRAL

Afin d'évaluer les transferts des radionucléides dans les différents compartiments de l'écosystème, des codes de calcul opérationnels ont été développés par la communauté des radioécologistes. Il s'agit par exemple en France des codes FOCON (Rommens et al., 1999) et ASTRAL (Mourlon et Calmon, 2002), au Royaume-Uni de PC-CREAM (Simmonds et al., 1995) et FARMLAND (Brown et Simmonds, 1995), en Allemagne d'ECOSYS-87 (Müller et Pröhl, 1993) ou de codes développés à l'échelle européenne comme RODOS (Bartzis et al., 2000) ou COSYMA (Jones et al., 1995). Ces modèles mathématiques sont utilisés dans un but prédictif dans le cas d'accident grave et de pollutions chroniques mais aussi pour effectuer des simulations car les expérimentations sont souvent assez coûteuses ou même impossibles. Par exemple, considérons une étude dont l'objectif serait d'analyser l'impact d'une contamination radioactive sur un territoire (comme le territoire métropolitain français). Il paraît impossible de procéder, de manière expérimentale, à une telle contamination. De ce fait, les codes de calculs sont utilisés afin d'évaluer l'impact de ce type de scénario de contamination.

La simulation des données de cette étude s'appuie sur le code de calcul ASTRAL (Assistance Technique en Radioprotection post-AccidenteLle) développé par l'IRSN. Il permet d'évaluer le transfert des radionucléides dans la chaîne alimentaire (écosystème agricole et forestier) et l'évaluation des doses dues à l'exposition interne et externe suite à une émission atmosphérique accidentelle (Mourlon et Calmon, 2002). La donnée d'entrée principale est la valeur de dépôt sur l'ensemble du sol et des végétaux (Cf. Figure 2.d). Le module agricole est composé de trois sous-modules permettant de calculer les concentrations dans les productions végétales, les productions animales et les produits alimentaires qui en sont dérivés. Dans le cadre de notre analyse, le sous module productions végétales et les équations associées nous intéressent plus particulièrement. L'équation correspondant en partie au scénario de contamination défini au paragraphe précédent est celle des cultures maraîchères. Cette équation permet de calculer, pour un radionucléide r et un produit végétal v, l'activité dans le végétal (à la récolte) due au transfert foliaire :

$$C_{v,fol,r} = D_r\left(K_r.FTds_{v,r}.Dils_{Dat_D,H_p} + (1-K_r).FTdh_{v,r,H_p}.Dilh_{Dat_D,H_p}\right).e^{-(\lambda_b+\lambda_r)\Delta} \qquad (1)$$

où :

$C_{v,fol,r}$ (Bq.kg^{-1}) : activité du végétal (à la récolte) résultant du transfert foliaire,

D_r (Bq.m^{-2}) : dépôt de radioactivité le jour de l'accident,

K_r (sd) : proportion du dépôt sec dans le dépôt total,

$FTds_{v,r}$, $FTdh_{v,r,Hp}$ (m^2.kg^{-1} frais) : facteurs de transfert (respectivement par temps sec et par temps de pluie),

$Dils_{DatD,Hp}$, $Dilh_{DatD,Hp}$ (sd) : facteurs de protection du fait d'éventuelles serres (respectivement par temps sec et par temps de pluie). Ces paramètres permettent de tenir compte de la protection apportée par les serres utilisées pour les cultures maraîchères,

Dat_D (j) : date du dépôt,

H_p (mm) : hauteur de précipitation,

λ_b (j^{-1}) : constante de décroissance biomécanique du radionucléide pour le végétal. L'évolution de l'activité durant les jours suivants le dépôt est modélisée par une exponentielle décroissante caractérisée par λ_b. Cette décroissance rend compte des phénomènes de lessivage des feuilles par la pluie et de la croissance biologique du végétal,

λ_r (j^{-1}) : constante de décroissance physique du radionucléide,

Δ (j) : délai entre le dépôt et la récolte du végétal. Il peut aussi s'exprimer comme la différence entre le temps de croissance du végétal (T_c) et la date de l'accident (t).

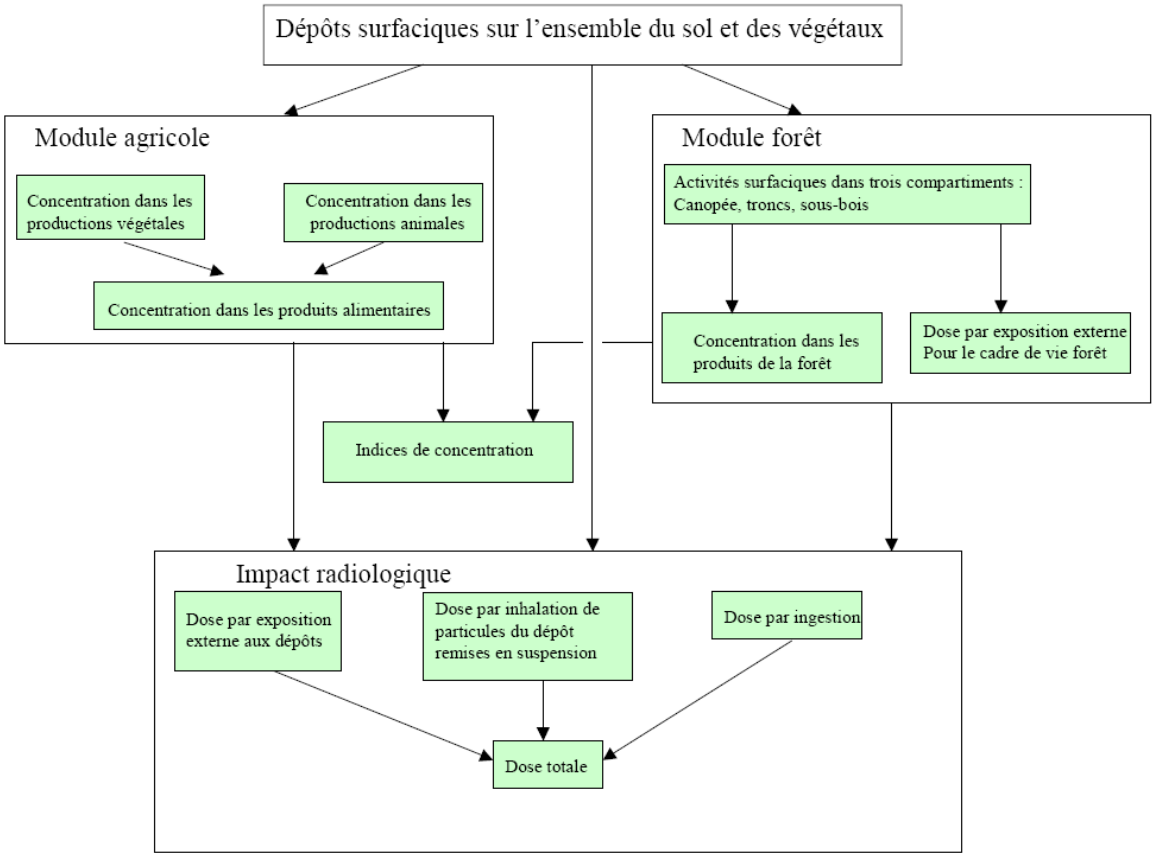


Figure 2.d : Organisation du code de calcul ASTRAL

La contamination est essentiellement déterminée par le facteur de transfert direct. Ce paramètre couvre à la fois les phénomènes d'interception, de translocation lorsque la partie du végétal consommée n'est pas celle qui est exposée à la contamination (cas de la

pomme de terre par exemple) et de remise en suspension. Les facteurs de transfert directs par dépôt sec et dépôt humide se décomposent de la façon suivante :

$$FTds_{v,r} = \frac{Rcs_{v,r}.Tlf_{v,r}}{Rdt_v} \qquad FTdh_{v,r,H_p} = \frac{Rch_{v,r}.Tlf_{v,r}}{Rdt_v}$$

où :

$Rcs_{v,r}$, $Rch_{v,r,Hp}$ (sd) : rapports de captation respectivement par temps sec et par temps de pluie. Le rapport de captation (par temps sec ou par temps de pluie) correspond à la fraction de l'activité totale déposée (par temps sec ou par temps de pluie) qui est interceptée par la partie aérienne des végétaux. Il désigne la fraction du dépôt exprimé en $Bq.m^{-2}$, qui est interceptée par la masse foliaire des végétaux se trouvant à la surface du sol. Il va donc dépendre du développement du végétal le jour de l'accident. Il varie entre 0 et 1 et s'exprime par un rapport d'activité sans dimension.

$Tlf_{v,r}$ (sd) : facteur de translocation. Les radionucléides déposés sur les parties foliaires peuvent être absorbés par la plante. La translocation correspond à la migration de ces radionucléides depuis le site d'absorption vers tous les organes comestibles de la plante : grain, fruit, racine, tubercule. Tout comme le rapport de captation, la valeur de ce paramètre sans dimension est comprise entre 0 et 1.

Rdt_v (kg frais.m^{-2}) : rendement cultural du végétal à la récolte.

Cette équation ne reflète pas totalement le scénario de contamination défini au paragraphe 2.3.2 car le dépôt total est un dépôt sec ($K_r = 1$) et pour les légumes-feuilles les phénomènes de translocation sont inexistants ($Tlf_{v,r}$ est affecté à la valeur 1). Ainsi l'équation (1) devient :

$$C_{v,fol,r} = \frac{D_r Rcs_{v,r} e^{-(\lambda_b + \lambda_r)\Delta}}{Rdt_v} \qquad (2)$$

Au total, les variables explicatives sont au nombre de 5 : le dépôt, le rapport de captation, la constante de décroissance biomécanique, le délai dépôt-récolte et le rendement cultural (la constante de décroissance physique du radionucléide n'est pas considérée comme une variable explicative du fait de sa valeur fixe pour chaque radionucléide).

Afin de mieux appréhender le rôle des variables intervenant dans cette équation, une illustration simplifiée d'un cas de contamination accidentelle sur une production végétale est présentée à la figure 2.e. Cette production végétale est caractérisée par un temps de croissance T_c et l'accident à lieu à la date t sur ce temps de croissance. Le végétal est alors contaminé par un dépôt de radioactivité et la fraction de l'activité totale qu'il intercepte dépend de son stade de développement. Puis, le végétal continue sa croissance, l'évolution de sa contamination durant le délai dépôt-récolte est modélisée par une

exponentielle décroissante ($e^{-(\lambda_b+\lambda_r)\Delta}$). A la récolte, lorsque le végétal est prêt à être consommé (il est caractérisé par un certain rendement), la mesure de son activité est effectuée.

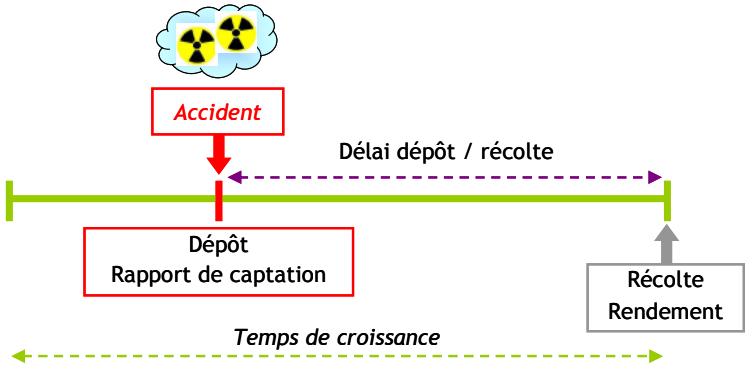


Figure 2.e : Schématisation simplifiée du rôle des principales variables intervenant dans l'équation (2)

Ainsi, cette représentation permet une meilleure compréhension du fonctionnement de l'équation et des relations existantes entre certaines variables intervenant explicitement et implicitement dans l'équation (par exemple la relation liant le rapport de captation au développement du végétal). Pour prendre en compte ce type de relation dans la génération des données et caractériser au mieux les distributions de probabilité de chacune des variables explicatives, un travail de recherche conséquent a été effectué. Il s'est principalement appuyé sur des références bibliographiques, des contacts avec divers organismes agricoles et l'utilisation d'un modèle agronomique de culture développé par l'INRA d'Avignon.

2.3.4 Outils et méthodes utilisés pour renseigner les entrées du modèle

2.3.4.1 Recherches bibliographiques et contact avec des organismes agricoles

Les distributions de certaines variables radioécologiques ont été déterminées à partir de références bibliographiques. Il s'agit de la constante de décroissance biomécanique et du dépôt de radioactivité. En ce qui concerne les variables agronomiques (le temps de croissance et le rendement pour certains légumes-feuilles), les recherches bibliographiques ont été accompagnées de prise de contact avec des personnes spécialisées relevant d'organismes agricoles (Centre Technique Interprofessionnel des Fruits et des Légumes (CTIFL), Institut National de la Recherche Agronomique (INRA), et Union Nationale Interprofessionnelle des Légumes Transformés (UNILET)).

Notre principale référence bibliographique est un rapport technique rédigé dans le cadre de travaux menés par le Groupe Radioécologie Nord-Cotentin[6] (GRNC). Ce rapport présente une étude de la variabilité et de l'incertitude de différents paramètres intervenant dans les modèles de transfert utilisés par le GRNC (GRNC, 2001).

2.3.4.2 Le modèle de culture STICS

Le modèle STICS (Simulateur mulTIdisciplinaire pour les Cultures Standard), développé par l'INRA, est un modèle de croissance des cultures à pas de temps journalier (Brisson et al., 1998 ; Brisson et al., 2003). Ses variables d'entrées sont relatives au climat, au sol et au système de culture. Ses variables de sorties sont relatives à la production (quantité et qualité), à l'environnement et à l'évolution des caractéristiques du sol sous l'effet de la culture (Cf. Figure 2.f). STICS a été conçu comme un outil de recherche, intégrant toutes les connaissances agronomiques, et un outil de simulation opérationnel en conditions agricoles. Son principal objectif est de simuler les conséquences des variations du milieu et du système de culture sur la production d'une parcelle agricole au cours d'une année. La culture est appréhendée globalement par sa biomasse aérienne et sa teneur en azote, son indice foliaire ainsi que le nombre et la biomasse des organes récoltés. L'indice foliaire calculé par le modèle STICS respecte la définition agronomique de l'indice de surface foliaire, qui ne prend en compte que la surface cumulée des feuilles vertes par unité de surface de sol. Le sol est assimilé à une succession de couches horizontales, chacune de ces couches étant caractérisée par sa réserve en eau, en azote minéral et en azote organique.

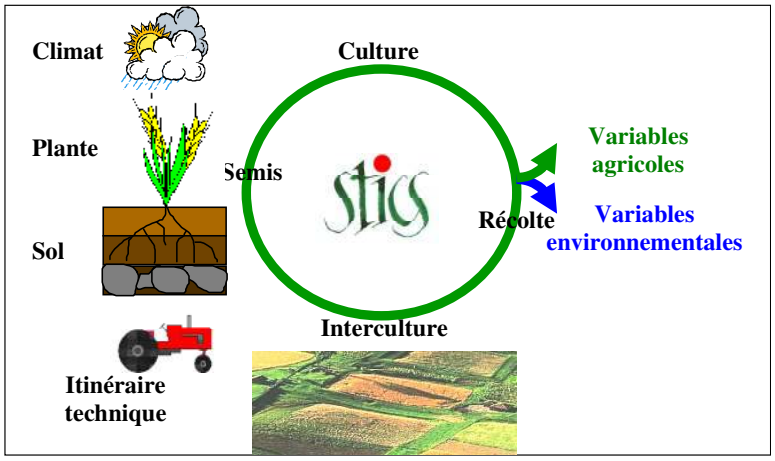


Figure 2.f : Entrées et sorties du modèle STICS (INRA)

[6] Le GRNC a été créé dans un contexte de controverses suite à la publication d'une étude épidémiologique sur l'incidence des leucémies dans le canton de Beaumont-La Hague. Il a été constitué pour faire une évaluation des risques, c'est-à-dire calculer le risque de leucémie radio-induit pour le comparer aux résultats de l'étude épidémiologique et tenter de sortir de la controverse.
(http://www.irsn.org/index.php?position=radioecologie_nord_cotentin)

Les interactions entre le sol et la culture sont assurées par les racines, celles-ci étant définies par une distribution de densité racinaire dans le profil de sol. Le modèle simule le bilan de carbone, le bilan d'eau et le bilan d'azote du système et permet de calculer à la fois des variables agricoles (rendement, consommations d'intrants) et des variables environnementales (pertes d'eau et de nitrates) dans diverses situations agricoles.

Dans le cadre du projet SENSIB, ce modèle a déjà été utilisé afin d'étudier l'effet de la régionalisation des variables *rapport de captation* et *rendement cultural* sur la contamination de productions agricoles. La caractérisation de ces variables (dans le sens reconstruction de l'évolution du rapport de captation en fonction du développement du végétal et détermination de valeurs de rendement cultural à la récolte) a été effectuée pour divers types de cultures : le blé d'hiver (Mercat-Rommens et al., 2006), les prairies permanentes (Durand et al., 2007b), la vigne (Levain et al., 2006) et la pomme de terre (Larue et al., 2007). Dans chacun des cas, différentes simulations de croissance de cultures ont été réalisées afin de mettre en évidence l'influence du climat, du sol et de l'itinéraire technique (Durand et al., 2007a). Par exemple, pour le cas de la vigne, 5 simulations ont été effectuées. Chaque simulation est basée sur le vignoble d'une commune représentative (Levain et al., 2006). Par exemple, l'évolution du rapport de captation par temps sec a pu être reconstituée en fonction du cycle végétatif de chaque vignoble étudié (Cf. Figure 2.g).

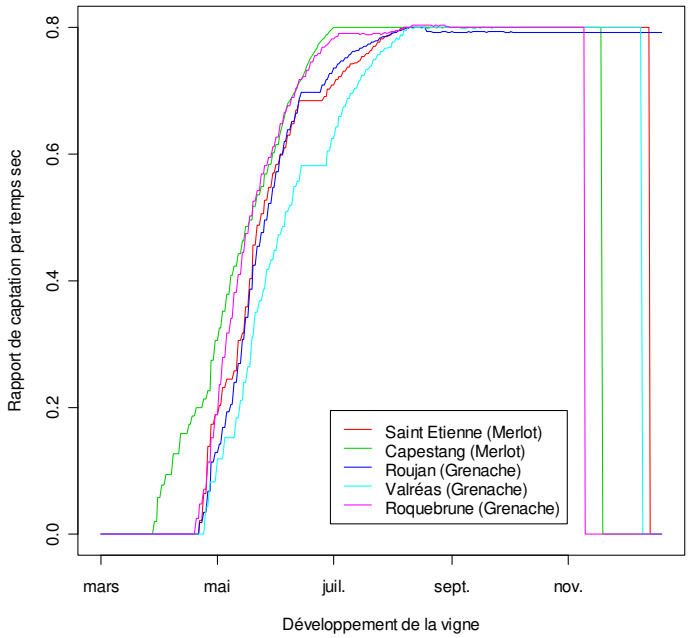


Figure 2.g : Evolution du rapport de captation par temps sec pour les 5 vignobles considérés (Levain et al., 2006)

Concernant les légumes-feuilles, le modèle STICS est uniquement paramétré pour une variété de salade : la laitue. Ce modèle a donc été utilisé pour caractériser le rapport de captation et le rendement d'une production de laitue et certains des résultats obtenus ont été extrapolés aux autres légumes-feuilles étudiés (Cf. 5.1.1).

2.3.4.2.1 Etude du rapport de captation par temps sec

La particularité du modèle STICS est de fournir, à un pas de temps journalier (de la plantation à la récolte), l'évolution de certaines variables agronomiques durant la croissance d'une culture. Notre objectif est donc de trouver une relation entre le rapport de captation et une sortie du modèle STICS afin de reconstruire l'évolution du rapport de captation en fonction du développement du végétal.

Une modélisation du rapport de captation par dépôt sec d'un végétal herbacé ou à l'état herbacé a été proposée par Chamberlain (1970). Cette expression exprime le rapport de captation en fonction du rendement de biomasse (kg.m^{-2}) de ce végétal :

$$Rcs = 1 - e^{-\mu \times Rdt}$$

où

μ (m^2.kg^{-1} frais ou sec) : constante d'interception ou coefficient de captation,

Rdt (kg frais ou sec.m^{-2}) : rendement de biomasse aérienne.

Un des objectifs du programme RESSAC (Réhabilitation des Sols et des Surfaces en cas d'Accident) initié par l'IRSN (Maubert et al., 1991) était notamment de généraliser cette équation à d'autres cultures. Les résultats des expérimentations menées en France ont conduit à l'élaboration de la formule de Chamberlain modifiée :

$$Rcs = S \times (1 - e^{-\mu \times Rdt})$$

où :

S (sd) : Coefficient de saturation. C'est la fraction maximale de dépôt captée par une culture donnée quelle que soit sa biomasse, soit $S = Rc_{max}$. Pour une culture donnée, S dépend de la géométrie du couvert et de la densité de plantation.

Le rendement cultural est une sortie du modèle agronomique STICS, qui s'exprime en fonction du développement de la culture considérée. Cependant, les valeurs des deux coefficients S et μ ont été estimées à partir de peu de valeurs et relatives à plusieurs types de plantes (salades, choux et betteraves) (programme RESSAC, Maubert et al., 1991). Ainsi, le choix de cette équation pour estimer le rapport de captation nous paraît peu pertinent.

Le rapport de captation par temps sec peut également être modélisé par l'indice foliaire (*LAI*) comme il a été démontré dans les études précédentes (Delboe et Mercat-Rommens, 2005 ; Durand et Mercat-Rommens, 2006 ; Levain et al., 2006 ; Larue et al., 2007). Cet indice se définit comme la surface susceptible de capter les aérosols, c'est-à-dire la surface de feuille par unité de surface de sol. On considère donc l'ensemble des surfaces des feuilles, cumulées entre les étages de végétation, par unité de surface du sol (m^2 feuilles/m^2 sol), indépendamment de leur activité photosynthétique. Cette définition diffère donc de celle agronomique qui considère uniquement la surface des feuilles vertes. Un code de calcul allemand, ECOSYS-87, qui a été testé suite à l'accident de Tchernobyl dans le cadre de programmes internationaux (Müller et Pröhl, 1993), propose un mode de calcul du dépôt sur la plante et du dépôt total, dont le rapport équivaut à la captation, qui utilise l'indice foliaire. Des études de comparaison entre les prédictions d'ECOSYS-87 et des mesures effectuées dans différents sites européens ont donné des résultats satisfaisants (Renaud et al., 1997b).
La variable choisie pour caractériser le stade de la plante est l'indice foliaire. L'équation donnée par le code de calcul est la suivante :

$$Rcs = \frac{LAI_i / LAI_{i\max}}{(LAI_i / LAI_{i\max}) + (Vg_s / Vg_{i\max})} \qquad (3)$$

où :
Vg_i : Vitesse de dépôt sur la plante i ($m.s^{-1}$),
Vg_{imax} : Vitesse de dépôt maximale sur la plante i ($m.s^{-1}$),
LAI_i : Leaf Area Index ou indice foliaire de la plante i,
LAI_{imax} : Indice foliaire maximal de la plante i,
Vg_s : Vitesse de dépôt sur le sol ($m.s^{-1}$), constante pour tout type de plante.

La vitesse de dépôt dépend de l'indice foliaire au moment du dépôt et au maximum de développement foliaire. Néanmoins, en l'absence d'expression reliant Vg_{imax} et LAI_{imax}, ce paramètre a été considéré comme constant. Müller et Pröhl (1993) préconisent notamment pour les légumes-feuilles et pour un dépôt de radionucléides sous forme d'aérosols, d'utiliser la valeur de 2 $mm.s^{-1}$ pour Vg_{imax}. De même, la vitesse de dépôt sur le sol Vgs est considérée comme constante pour tout type de plante et prend la valeur de 0,5 $mm.s^{-1}$.

Cependant, pour la culture de laitue, le *LAI* n'est pas une sortie du modèle STICS. Effectivement, dans le cas de la salade, compte tenu de sa structure en forme de pomme, l'indice foliaire, ordinairement retenu pour l'interception du rayonnement est remplacé

dans la modélisation de STICS par le taux de couverture du sol. Le calcul direct du taux de couverture du sol (*TAUXCOUV*) est une alternative simple au calcul du *LAI* et intervient comme variable d'état dans les calculs de l'interception du rayonnement et des besoins en eau. Ainsi dans le cas de la salade et dans le modèle STICS - Feuille, la variable *LAI* s'exprime en pourcentage de recouvrement et correspond au taux de couverture. La relation entre ces deux grandeurs peut être déduite des travaux menés par De Tourdonnet sur la maîtrise de la qualité et de la pollution nitrique en production de laitues sous abri plastique (De Tourdonnet, 1998). Il a construit, paramétré et validé sur des données indépendantes, un modèle de fonctionnement de l'agrosystème, intégrant une hétérogénéité du milieu. Il ressort que l'évolution de l'efficience d'absorption d'une culture de laitue est généralement modélisée à partir de l'évolution de son *LAI* selon la relation :

$$\varepsilon a = \varepsilon a_{\max} \times (1 - e^{-k \times LAI})$$

où

εa : efficience d'absorption du rayonnement,

$\varepsilon a_{\max}$: efficience d'absorption maximale du rayonnement,

k : coefficient d'extinction du rayonnement dans le couvert végétal (paramètre caractéristique de l'espèce),

LAI : indice foliaire.

Cette relation est fondée sur la loi d'extinction du rayonnement de Beer et sa validité repose sur l'hypothèse d'une distribution aléatoire des surfaces foliaires dans l'espace. Le coefficient k est un paramètre renseigné dans le fichier plante du modèle STICS, dans la fonction « interception du rayonnement ». Dans le cas de la salade, ce paramètre prend la valeur de 0,6 et est considéré comme constant. D'après De Tourdonnet (1998), la valeur de ce paramètre avoisine la valeur de 0,68 ± 0,02.

En outre, De Tourdonnet a exprimé l'efficience d'absorption en fonction du taux de couverture selon les relations suivantes :

$$\varepsilon a = \begin{cases} f \times TAUXCOUV & \text{Si } TAUXCOUV < 77\% \\ TAUXCOUV \times (1 - R) & \text{Si } TAUXCOUV \geq 77\% \end{cases}$$

Avec $f = 1{,}196$ (valeur ajustée) et $R = 0{,}08$ correspond à la réflectance du couvert.

Ces deux relations permettent de relier le taux de recouvrement, $TAUXCOUV$, à l'indice foliaire, LAI, et d'exprimer le LAI en fonction du taux de recouvrement selon les équations suivantes :

$$LAI = \begin{cases} -(1/k)\log\left(1 - \dfrac{f \times TAUXCOUV^2}{\varepsilon a_{\max}}\right) & \text{Si } TAUXCOUV < 77\% \\ -(1/k)\log\left(1 - \dfrac{TAUXCOUV(1-R)}{\varepsilon a_{\max}}\right) & \text{Si } TAUXCOUV \geq 77\% \end{cases} \quad (4)$$

Avec εa_{max}, efficience maximale d'absorption du rayonnement, fixée à 0,92.

D'après (3) et (4), le rapport de captation peut s'exprimer en fonction du taux de recouvrement :

$$Rcs = \begin{cases} \dfrac{\dfrac{-\log\left(1 - \dfrac{f \times TAUXCOUV^2}{\varepsilon a_{\max}}\right)}{k \times LAI_{\max}}}{\left(\dfrac{-\log\left(1 - \dfrac{f \times TAUXCOUV^2}{\varepsilon a_{\max}}\right)}{k \times LAI_{\max}}\right) + (Vgs/Vgi_{\max})} & \text{Si } TAUXCOUV < 77\% \\ \dfrac{\dfrac{-\log\left(1 - \dfrac{TAUXCOUV \times (1-R)}{\varepsilon a_{\max}}\right)}{k \times LAI_{\max}}}{\left(\dfrac{-\log\left(1 - \dfrac{TAUXCOUV \times (1-R)}{\varepsilon a_{\max}}\right)}{k \times LAI_{\max}}\right) + (Vgs/Vgi_{\max})} & \text{Si } TAUXCOUV \geq 77\% \end{cases} \quad (5)$$

La sortie $TAUXCOUV$ du modèle STICS pouvant s'exprimer en fonction du développement de la laitue, l'équation (5) permettra dans la suite de reconstituer l'évolution du rapport de captation sec en fonction du développement de la laitue.

2.3.4.2.2 Etude du rendement cultural

La variable rendement de biomasse aérienne correspond à la masse du végétal en kg frais par m^2 de sol au moment de la récolte et plus particulièrement, à la date de récolte commerciale. Cette variable respecte la définition du « rendement agronomique » dans la mesure où la partie comestible de la laitue correspond à la biomasse aérienne.

Contrairement au rapport de captation présenté précédemment, cette variable est une sortie directe du modèle STICS.

3 Les arbres de décision et la méthode CART

3.1 Introduction

Les méthodes d'arbres de décision[7] ont vu le jour dans les années 1960, avec l'algorithme AID (*Automatic Interaction Detection*) développé par Morgan et Sonquist (1963). L'objectif de cette méthode était la prédiction d'une variable quantitative Y par la construction d'un arbre binaire de régression. Morgan et Messenger (1972) ont proposé d'étendre cet algorithme au cas où Y est une variable discrète avec la méthode THAID (*Theta Automatic Interaction Detection*). Au début des années 1980, Kass (1980) proposa une amélioration de ces deux algorithmes avec la méthode CHAID (*CHi-squared AID*). Proposée en 1984 par Breiman et al., la méthode CART (*Classification And Regression Trees*) présente de nombreuses nouveautés dans la construction des arbres de régression et de discrimination. Elle reste encore aujourd'hui l'une des références principales dans le domaine des méthodes de segmentation avec les méthodes CHAID et C4.5 (Quinlan, 1993).

Dans un premier temps, à partir d'un exemple illustratif issu de la monographie de Breiman et al. (1984), nous présentons la structure d'un arbre de discrimination ainsi que le vocabulaire associé aux arbres. Puis, nous exposons de manière générale le principe de construction des arbres de discrimination. Enfin, nous présentons de manière plus approfondie la méthode CART.

3.2 Description d'un arbre de discrimination

La construction d'un arbre de décision est basée sur la division successive d'un échantillon de données. Les données observées sur les sujets de cet échantillon, noté E, sont constituées d'une variable à expliquer Y quantitative ou qualitative à J modalités et de p variables explicatives $X_1,\ldots,X_p$ observées conjointement sur les n individus de E. Par exemple, la variable à expliquer peut être la présence ou l'absence d'une maladie sur des sujets et les variables explicatives, les caractéristiques associées aux individus malades ou non malades : l'âge, le poids,… L'objectif des arbres est de prédire la variable Y à partir des différentes variables explicatives. Selon la nature de la variable à expliquer (quantitative ou qualitative), on parlera d'arbre de régression ou de discrimination. L'échantillon E introduit précédemment qui sert pour la construction de l'arbre est appelé échantillon d'apprentissage. Lorsque l'erreur associée à l'arbre est estimée à partir de l'échantillon d'apprentissage[8] elle est généralement trop optimiste (l'échantillon sert de

[7] Connues aussi sous le nom de méthodes de segmentation ou de partitionnement récursif.
[8] On l'appelle le plus souvent erreur apparente ou erreur estimée par resubstitution.

base à la construction de l'arbre). Une estimation précise des performances réelles de l'arbre (erreur de généralisation) s'effectue le plus souvent à partir d'un second échantillon appelé test (échantillon de mesures sur les mêmes variables que l'échantillon d'apprentissage et n'ayant pas participé à la construction de l'arbre) ou d'une méthode de validation croisée (Cf. 3.4 pour la méthode CART).

Afin de mieux appréhender la structure d'un arbre de discrimination, nous proposons d'en expliquer les principes à partir d'un exemple tiré de la monographie de Breiman et al. (1984) sur la méthode CART (Cf. Figure 3.a). L'objectif de cette étude conduite au centre médical de San Diego (Université de Californie, Etats-Unis) était de développer une méthode permettant d'identifier de façon anticipée les patients présentant un risque élevé de décéder à la suite d'une crise cardiaque. Ainsi, lorsqu'un patient victime d'une crise cardiaque se présentait à ce centre médical, un certain nombre de variables relatives à son état de santé étaient mesurées durant les 24 premières heures suivant son admission (au total 19 variables comme la pression sanguine, l'âge,…).

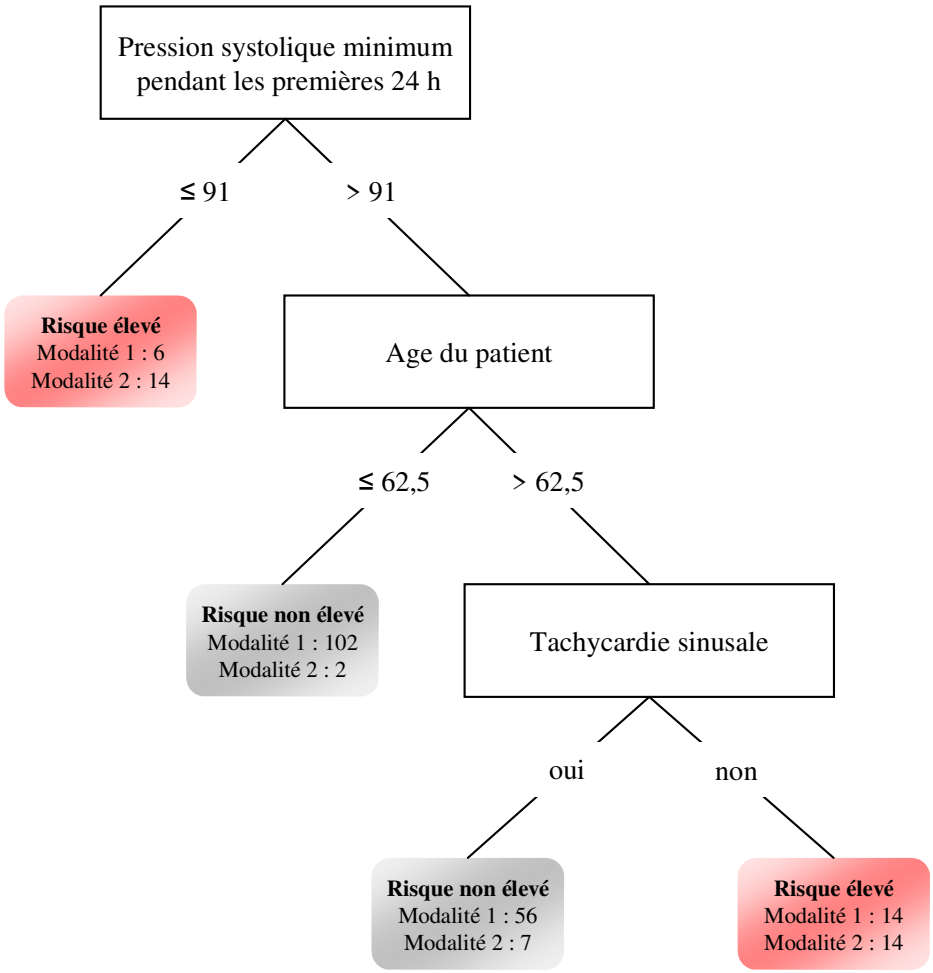


Figure 3.a : Exemple d'arbre de discrimination (Breiman et al., 1984)

Un échantillon d'apprentissage composé de 215 patients a été observé. Parmi ces patients, 37 n'ont pas survécu durant les 30 jours suivant leur admission (ils sont appelés dans la suite individus à risque élevé et sont affectés à la modalité 2) et 178 ont survécu (ces survivants sont des individus ne présentant pas un risque élevé de décéder, ils sont affectés à la modalité 1). Afin de prédire les patients présentant un risque élevé de décéder durant les 30 jours suivant une crise cardiaque (modalité 2) à partir des données mesurées pendant les 24 premières heures d'admission, un arbre de discrimination a été construit par la méthode CART (Cf. Figure 3.a). Cet arbre est caractérisé par des questions portant sur certaines variables explicatives (parmi les variables mesurées durant les 24 premières heures d'admission du patient). La première question porte sur la pression systolique[9] minimum durant les 24 premières heures. Si, pour un patient, cette variable est inférieure ou égale à 91 alors il présente un risque élevé de décès à la suite de son admission. Dans le cas contraire, si le patient présente une valeur supérieure à 91, on ne peut pas conclure directement et une nouvelle question est posée : quel est l'âge du patient ? De la même manière que précédemment, selon son âge on pourra conclure directement ou avoir besoin d'une information supplémentaire qui permettra d'aboutir à la réponse. Ainsi, la succession des questions sur certaines des variables explicatives conduit à une réponse particulière qui est l'appartenance du sujet à l'une des modalités de la variable à expliquer (risque élevé ou risque non élevé de décéder durant les 30 jours suivant une crise cardiaque).

Cet exemple met en avant le côté très attractif des arbres : ce sont des outils explicatifs et prédictifs. Parmi les 19 variables mesurées durant les 24 premières heures d'admission, seulement 3 sont finalement nécessaires pour pronostiquer l'état de santé des patients. La première division est très instructive, elle permet de déterminer une valeur particulière de la pression systolique minimum du patient qui conduit à l'identification d'un groupe de patients présentant un risque élevé de décéder. L'interprétation de ces règles, basées sur les variables explicatives, conduit à une meilleure compréhension du phénomène étudié. De plus, pour un nouvel individu dont on connaît les valeurs pour ces trois variables, les règles de décision permettent de prédire son appartenance à l'un des groupes définis par la variable Y à expliquer.

Dans le vocabulaire statistique lié aux arbres, chacune des questions décrites précédemment est appelée nœud. Un nœud est caractérisé par une division sur l'une des variables explicatives. Le premier nœud de l'arbre est appelé racine ou nœud racine, il contient l'ensemble de l'échantillon E. La division associée à ce nœud permet de créer deux sous-ensembles de l'échantillon, appelés nœuds enfants (ou nœuds descendants).

[9] Pression artérielle, il s'agit de la pression du sang dans les artères.

L'ensemble des descendants d'un nœud est appelé une branche. Contrairement aux nœuds intermédiaires qui fournissent des descendants, les nœuds non divisés sont appelés nœuds terminaux ou feuilles, ils proposent une prédiction de la variable Y (l'une des modalités de Y si cette dernière est qualitative ou une valeur prédite si Y est une variable quantitative). La complexité d'un arbre A, notée $\left|\tilde{A}\right|$ représente son nombre de feuilles. Elaguer un arbre à un nœud consiste à supprimer la branche issue de ce nœud, le nœud devenant alors une feuille. Un sous-arbre de A est un arbre obtenu à partir de A après avoir élagué une ou plusieurs branches.

Suite à la description de cet arbre de discrimination, plusieurs questions se posent quant à sa construction :

- Pour chacun des nœuds intermédiaires de l'arbre, quel critère utilise-t-on pour sélectionner la meilleure division ?
- Quand décide-t-on qu'un nœud est terminal ?
- Lorsqu'un nœud est terminal, comment lui affecte-t-on une modalité de Y ?

La suite du chapitre permet de répondre à ces questions en présentant le principe de construction d'un arbre de discrimination et la méthode CART.

3.3 Principe de construction d'un arbre de discrimination

Les arbres de discrimination sont construits par partitionnement récursif binaire (cas de la méthode CART, Cf. exemple précédent) ou n-aire, comme avec la méthode CHAID (Kass ,1980). Pour la plupart des méthodes, la division qui est sélectionnée est optimale au sens d'un indicateur de qualité de la partition engendrée. De nombreux critères pour la segmentation d'un nœud ont été développés. Parmi ces critères, certains sont fondés sur la théorie de l'information, comme le gain d'information proposé par Quinlan (1986) qui repose sur la notion d'entropie (Shannon et Weaver, 1949). D'autres critères sont basés sur la distance entre les distributions de probabilités comme le critère du Khi-2 (Kaas, 1980), l'indice de Gini, et le critère de twoing préféré lorsque le nombre de modalité de Y est important (Breiman et al., 1984). Le Khi-2 proposé par la méthode CHAID (Kaas, 1980) avantage les variables explicatives qualitatives ayant un nombre élevé de modalités. L'indice de Gini favorise l'isolement de la modalité majoritaire de Y dans un nœud homogène. Tandis que le critère de twoing et l'entropie tendent à produire des nœuds enfants de taille plus équilibré (Breiman, 1996c).

Une fois le critère de segmentation d'un nœud choisi, il faut procéder à la construction de l'arbre de décision. Il est important de déterminer la bonne taille de l'arbre car celle-ci aura un effet sur les performances de prédiction de l'arbre. Dans la littérature, deux types de techniques sont proposées pour définir cette taille :

- Le pré-élagage qui consiste, lors de la phase de construction de l'arbre, à arrêter la procédure de division par la fixation d'une règle. C'est par exemple, le cas de la méthode CHAID (Kass, 1980) où la règle d'arrêt de division d'un nœud est basée sur la valeur de la p-value du test du Khi-2.
- Le post-élagage, qui comporte deux étapes : la construction d'un arbre le plus homogène possible et la réduction de cet arbre (élagage) par suppression des branches les moins informatives au sens d'un critère. C'est le cas de la méthode CART par exemple (Cf. 3.4.3).

La première technique est souvent critiquée, la règle d'arrêt conduit généralement à des arbres trop détaillés ou à l'inverse pas assez détaillés (Celeux et Nakache, 1994). Lorsque l'arbre est très grand, il dépend de l'échantillon d'apprentissage et certaines divisions, en particulier dans les niveaux inférieurs de l'arbre, ne sont pas pertinentes. En revanche, des arbres peu complexes ne font pas apparaître toute l'information contenue dans l'échantillon d'apprentissage et l'évaluation du taux d'erreur est généralement plus élevée.
La construction d'un arbre en deux temps (construction de l'arbre maximal et élagage de cet arbre) est apparue avec la méthode CART (Breiman et al., 1984) et l'introduction de l'élagage de coût-complexité minimal (Cf. 3.4.3.2). Les différentes méthodes d'élagage développées par la suite, consistent le plus souvent en une évaluation de l'erreur de prédiction lors de la transformation de certaines branches en feuilles. Selon la méthode, l'estimation de l'erreur peut se faire à partir de l'échantillon d'apprentissage ou à partir d'un nouvel échantillon mesuré sur les mêmes variables mais qui n'a pas participé à la construction de l'arbre, appelé échantillon d'élagage ou de validation. Une comparaison empirique de six méthodes d'élagage les plus connues a conclu au fait qu'il n'y avait pas de différence globale entre les méthodes qui utilisent un échantillon supplémentaire lors de l'élagage et celles qui se restreignent à l'utilisation de l'échantillon d'apprentissage (Esposito et al., 1997).

Selon Breiman et al. (1984), l'arbre final serait plus sensible à la procédure d'élagage choisie qu'au critère de division, du moment qu'il produit des partitions homogènes au sens de la variable à expliquer.

3.4 La méthode CART

Ce paragraphe est consacré à la description de la méthode CART, en particulier à la construction d'un arbre de discrimination. C'est une méthode non paramétrique où les

variables explicatives peuvent être quantitatives ou qualitatives. Comme il a été vu dans l'exemple précédent (paragraphe 3.2), les arbres sont obtenus par partitionnement récursif binaire : les nœuds parents sont toujours divisés en deux nœuds enfants et ce processus est réitéré en considérant les nœuds descendants obtenus comme des nœuds parents.

Nous allons décrire les différentes étapes dans la construction d'un arbre et présenter la notion d'importance des variables. Nous aborderons, dans ce chapitre, uniquement la construction d'un arbre de discrimination par la méthode CART. Pour plus de détails concernant la construction d'un arbre de régression par cette méthode, le lecteur peut se référer à l'annexe C.

3.4.1 Notations

Afin de définir les prochaines notions, les notations suivantes sont nécessaires :

- n : taille de l'échantillon d'apprentissage E,
- n_j : nombre d'observations de E appartenant à la modalité j de Y, ($j = 1,\ldots,J$),
- n_{lj} : nombre d'observations de E appartenant à la modalité j de Y et affectées à la modalité l,
- π_j : probabilité d'appartenance à la modalité j, définie par la fréquence des modalités dans l'échantillon :

$$\pi_j = \frac{n_j}{n}$$

- $n(t)$: nombre d'observations appartenant au nœud t,
- $n_j(t)$: nombre d'observations appartenant au nœud t qui ont pour modalité j, ($j = 1,\ldots,J$),
- $P(j,t)$: probabilité qu'une observation appartienne au nœud t et soit de modalité j, estimée par :

$$p(j,t) = \pi_j \frac{n_j(t)}{n_j} = \frac{n_j(t)}{n}$$

- $P(t)$: probabilité d'appartenir au nœud t, estimée par :

$$p(t) = \sum_{j=1}^{J} p(j,t) = \sum_{j=1}^{J} \pi_j \frac{n_j(t)}{n_j} = \frac{n(t)}{n}$$

- $P(j|t)$: probabilité que l'observation soit de modalité j sachant qu'elle appartient au nœud t, estimée par :

$$p(j|t) = \frac{p(j,t)}{p(t)} = \frac{n_j(t)}{n(t)}$$

- $\gamma(j|l)$: coût de mauvais classement entraîné par l'affectation d'une observation à la modalité j alors qu'elle appartient à la modalité l, supposé unitaire : $\gamma(j|l) = 1$ pour tout $j \neq l$, avec $\gamma(j|j) = 0$.

3.4.2 Le critère de division d'un nœud

Pour chaque nœud intermédiaire, la division choisie est celle permettant d'obtenir les nœuds enfants les plus homogènes possibles relativement à la variable à expliquer Y. Afin de mesurer le mélange des modalités de Y dans un nœud, une fonction d'hétérogénéité, notée i, est définie. Elle est fonction des probabilités $p(j|t)$, pour $j = 1,\ldots,J$ et vérifie les propriétés suivantes :

i) Elle atteint son maximum pour l'équiprobabilité :

$$\arg\max_{p(\cdot|t)} \; i\big(p(1|t),\ldots,p(J|t)\big) = \left(\frac{1}{J},\ldots,\frac{1}{J}\right)$$

ii) Elle est nulle lorsque le nœud est homogène (il contient des observations d'une seule modalité de la variable Y) :

$i\big(p(1|t),\ldots,p(J|t)\big) = 0$ si $p(l|t) = 1$ et $p(r|t) = 0$ pour tout $l \neq r$, $l = 1,\ldots,J$; $r = 1,\ldots,J$

iii) Elle est symétrique en $p(1|t),\ p(2|t),\ldots,p(J|t)$.

Soit δ une division qui scinde le nœud t en deux nœuds enfants t_g et t_d. La proportion d'observations de t dirigées vers t_g (resp. t_d), notée p_g (resp. p_d) s'exprime par :

$$p_g = \frac{p(t_g)}{p(t)} \quad \left(\text{resp.} \quad p_d = \frac{p(t_d)}{p(t)}\right)$$

La réduction de l'hétérogénéité entraînée par la division δ au nœud t, s'exprime par :

$$\Delta i(\delta,t) = i(t) - p_g i(t_g) - p_d i(t_d) \qquad (6)$$

Il s'agit de la différence entre l'hétérogénéité au nœud parent et la somme pondérée des hétérogénéités dans les nœuds enfants issus de t. La division recherchée est celle rendant les nœuds t_g et t_d les plus homogènes au sens de la variable Y. Par conséquent, la meilleure division $\delta*$ est celle maximisant la réduction de l'hétérogénéité :

$$\Delta i(\delta^*,t) = \arg\max_{\delta \in D}\{\Delta i(\delta,t)\} \qquad (7)$$

où D est l'ensemble des divisions admissibles[10] du nœud t.

Pour la présente étude, nous avons choisi d'utiliser comme fonction d'hétérogénéité le critère d'entropie, vérifiant les propriétés précédentes et défini par :

$$i(t) = -\sum_{j=1}^{J} p(j|t) \log p(j|t) \qquad (8)$$

En pratique, c'est le plus souvent le logiciel statistique utilisé qui nous impose le choix de ce critère. Sous le logiciel SPAD, par exemple, seul l'indice de Gini peut être employé. Les logiciels Splus et R et leur librairie *rpart* (Atkinson et Therneau, 2000) proposent le choix de l'indice de Gini ou de l'entropie, afin de construire un arbre de discrimination.

3.4.3 Les différentes étapes dans la construction d'un arbre

La construction d'un arbre de discrimination par la méthode CART repose sur l'application successive de trois étapes décrites dans les paragraphes suivants.

3.4.3.1 Construction de l'arbre maximal

La première étape de l'algorithme CART consiste à construire un arbre très détaillé appelé arbre maximal et noté A_{max}. L'échantillon d'apprentissage est divisé successivement de manière à construire l'arbre le plus grand possible. Un nœud est déclaré feuille lorsque celui-ci est homogène (il ne peut donc plus être divisé) ou lorsque le nombre d'observations présentes dans le nœud est inférieur à un effectif fixé (généralement de 1 à 5). Il est alors affecté au groupe pour lequel $\sum_j \gamma(l|j)\, p(j|t)$ est minimal. Le coût de mauvais classement étant supposé unitaire, la feuille est affectée à la modalité de Y la plus fréquente. L'arbre maximal obtenu est très dépendant de l'échantillon d'apprentissage, en particulier ses niveaux inférieurs. En effet, les dernières divisions engendrent des nœuds de taille très faible.

3.4.3.2 Etape d'élagage

Afin de définir le principe de l'élagage de l'arbre maximal, quelques définitions et notations sont nécessaires :

- Le coût de mauvais classement associé au nœud t d'un arbre A (estimé par resubstitution) s'exprime par :

[10] Une division est dite admissible si les deux nœuds enfants issus de cette division ne sont pas vides.

$$C(t) = p(t)c(t)$$

où $c(t) = \min_l \sum_{j=1}^{J} \gamma(l|j)\, p(j|t)$ représente le coût de mauvais classement d'une observation au nœud t,

- Le coût de mauvais classement associé à l'arbre A (estimé par resubstitution) s'exprime par :

$$C(A) = \sum_{t \in \tilde{T}} C(t) = \sum_{j=1}^{J} \sum_{l=1}^{J} \pi_j \gamma(l|j) \frac{n_{lj}}{n_j}$$

- Afin de pénaliser la complexité d'un arbre, Breiman et al. (1984) proposent de définir la mesure de coût-complexité suivante :

$$C_\alpha(A) = C(A) + \alpha \left|\tilde{A}\right|$$

où α est un réel positif. C'est une combinaison linéaire entre le coût de mauvais classement de l'arbre et sa complexité. Dans le cas d'un nœud t, le coût-complexité est calculé en considérant le nœud comme une feuille :

$$C_\alpha(t) = C(t) + \alpha$$

La méthode d'élagage proposée par Breiman et al. (1984) consiste en la construction d'une série de sous-arbres emboîtés entre l'arbre maximal et sa racine, en cherchant à chaque étape l'arbre le moins complexe minimisant la mesure de coût-complexité.

Considérons l'arbre A et A^t la branche issue du nœud t et ayant t pour racine. Pour $\alpha = 0$, l'inégalité $C_\alpha(t) \geq C_\alpha(A^t)$ est toujours vraie (pour l'échantillon d'apprentissage, le coût de mauvais classement décroît ou reste stable après une division). Pour une certaine valeur de α, cette inégalité n'est plus vérifiée et les mesures de coût-complexité deviennent égales. En résolvant l'équation $C_\alpha(t) = C_\alpha(A^t)$, on obtient alors :

$$\alpha = \frac{C(t) - C(A^t)}{\left|\tilde{A}^t\right| - 1} \qquad (9)$$

Cette valeur critique de α est également connue sous le nom de paramètre de complexité ou critère d'élagage (Gueguen et Nakache, 1988). Comme les deux mesures de coût-complexité sont égales, le nœud t est privilégié à la branche A^t car il est beaucoup moins complexe.

Construction de la séquence d'arbres

Le point de départ de l'algorithme d'élagage est l'arbre maximal A_{max}. Pour chacun des nœuds intermédiaires de cet arbre, le calcul du critère d'élagage défini en (9) est effectué. L'arbre A_{max} est alors élagué au(x) nœud(s) pour le(s)quel(s) ce critère est minimal (cet élagage conduit à supprimer la (les) branche(s) la (les) moins informative(s) de l'arbre). Le sous-arbre obtenu est noté A_1 . Cette procédure est réitérée sur l'arbre A_1, afin d'obtenir un nouveau sous-arbre A_2, etc. De cette manière, une séquence d'arbres S est construite entre l'arbre maximal et l'arbre réduit à une feuille. D'après Breiman et al. (1984), chaque arbre de la séquence est, parmi tous les sous-arbres ayant le même nombre de nœuds terminaux, celui présentant le coût de mauvais classement le plus faible (estimé par resubstitution).

3.4.3.3 Sélection de l'arbre final

La sélection de l'arbre final est basée sur les performances de prédiction des arbres obtenus à l'étape précédente. Afin d'estimer le coût de mauvais classement, deux méthodes peuvent être utilisées :

- La méthode de validation croisée. Souvent préconisée lorsque l'échantillon d'apprentissage est de petite taille, cette méthode permet d'estimer le coût de mauvais classement de chaque arbre de la séquence uniquement à partir de l'échantillon d'apprentissage. Pour plus de détails sur cette technique le lecteur peut se référer à l'Annexe D.
- La méthode de l'échantillon de validation. C'est un échantillon mesuré sur les mêmes variables que l'échantillon d'apprentissage et qui n'a pas participé à la construction de l'arbre. Lorsque l'échantillon d'apprentissage est suffisamment grand, il peut être divisé pour réserver une partie de cet échantillon à la sélection de l'arbre optimal. Soit E^{valid} un tel échantillon. Le coût de mauvais classement associé à l'arbre A est alors déterminé par :

$$C^{valid}(A) = \sum_{j=1}^{J} \sum_{l=1}^{J} \pi_j \gamma(l|j) \frac{n_{lj}^{valid}}{n_j^{valid}}$$

  Les coûts de mauvais classement étant supposés unitaires et les probabilités *a priori* égales aux fréquences des groupes dans l'échantillon de validation, $C^{valid}(A)$ représente la proportion d'observations de l'échantillon de validation mal classées par l'arbre A.

Afin de sélectionner l'arbre optimal parmi cette séquence S, deux règles de sélection peuvent être utilisées (Breiman et al., 1984). La première règle (règle 0-SE) consiste à

choisir l'arbre A^{opt1} présentant le coût de mauvais classement minimal parmi tous les sous-arbres de la séquence S :

$$C^*(A^{opt1}) = Min\left\{C^*(A_r); A_r \in S\right\}$$

où $C^*(A)$ représente le coût de mauvais classement de l'arbre A estimé par un échantillon de validation ou par la méthode de validation croisée.

Cette sélection va dépendre des échantillons utilisés pour estimer le coût de mauvais classement (choix de l'échantillon de validation ou choix aléatoire des V sous-ensembles pour la méthode de validation croisée décrite en annexe D), ce qui va générer une instabilité dans le choix de l'arbre optimal. Afin de réduire cette instabilité, Breiman et al. (1984) proposent d'utiliser la règle de l'écart-type (règle 1-SE). Elle consiste à sélectionner l'arbre A^{opt2} présentant le plus petit nombre de nœuds terminaux et vérifiant :

$$C^*(A^{opt2}) \leq C^*(A^{opt1}) + \sigma(C^*(A^{opt1}))$$

où $\sigma(C^*(A^{opt1}))$ représente l'écart-type de $C^*(A^{opt1})$.

3.4.4 Les divisions de substitution et l'importance des variables

Une division de substitution peut être vue comme une division de remplacement de la division optimale δ. Une division de substitution reproduit le plus précisément possible le comportement de δ au nœud t (elle essaie de diriger les observations dans les nœuds enfants de la même manière que le fait la division optimale). Cette division ne doit pas être confondue avec la notion de division concurrente qui réalise la deuxième plus grande valeur dans la réduction de l'hétérogénéité définie en (6). Contrairement à la division optimale, la division concurrente ne dirigera pas les observations de t de la même manière que δ. En pratique, les divisions de substitution sont utilisées pour pallier le problème des données manquantes, déterminer l'importance des variables dans la construction de l'arbre ou détecter des variables masquées (Breiman et al., 1984 ; Ghattas, 1999).

Soient t un nœud issu d'un arbre A et δ la division optimale de t, basée sur la variable X_k, qui scinde t en deux nœuds enfants t_g et t_d. Notons $D_{k'}$ $(1 \leq k' \leq p)$ l'ensemble des divisions associées à la variable $X_{k'}$ (i.e. $X_{k'} \leq d$) et $\overline{D}_{k'}$ l'ensemble des divisions complémentaires (i.e. $X_{k'} > d$). Soit δ' appartenant à $D_{k'} \cup \overline{D}_{k'}$ qui divise le nœud t en t_g' et t_d'. Alors la probabilité $P(t_g \cap t_g')$ (resp. $P(t_d \cap t_d')$) qu'une observation appartienne aux nœuds t_g et t_g' (resp. t_d et t_d'), est estimée par :

$$p(t_g \cap t_g') = \sum_{j=1}^{J} \pi_j \frac{n_j(t_g \cap t_g')}{n_j} \quad \left(\text{resp.} \quad p(t_d \cap t_d') = \sum_{j=1}^{J} \pi_j \frac{n_j(t_d \cap t_d')}{n_j} \right)$$

où $n_j(t_g \cap t_g')$ (resp. $n_j(t_d \cap t_d')$) est le nombre d'observations appartenant à la modalité j commun au nœud t_g et t_g' (resp. t_d et t_d').
La probabilité pour que δ et δ' conduisent une observation du nœud t vers la gauche (resp. vers la droite) est estimée par :

$$p_g(\delta,\delta') = \frac{p(t_g \cap t_g')}{p(t)} \quad \left(\text{resp.} \quad p_d(\delta,\delta') = \frac{p(t_d \cap t_d')}{p(t)}\right)$$

L'estimation de la probabilité que la division δ' reproduise correctement δ est donnée par :

$$p(\delta,\delta') = p_g(\delta,\delta') + p_d(\delta,\delta')$$

Une division $\tilde{\delta}$ appartenant à $D_{k'} \cup \overline{D}_{k'}$ basée sur la variable X_k' est une division de substitution de δ si :

$$p(\delta,\tilde{\delta}) = \max\left\{p(\delta,\delta');\delta' \in D_{k'} \cup \overline{D}_{k'}\right\}$$

Souvent, l'arbre final ne fait pas intervenir toutes les variables explicatives (comme dans l'exemple illustratif présenté dans le paragraphe 3.2 ou seulement 3 des 19 variables sont utilisées pour prédire les deux modalités de Y). Si une variable X_s apparaissant dans l'arbre est supprimée de l'échantillon d'apprentissage et qu'un nouvel arbre est construit, une autre variable explicative, n'intervenant pas dans l'arbre précédent, peut prendre la place de X_s à chacun des nœuds. La qualité de prédiction de l'arbre obtenu peut être tout aussi satisfaisante que le premier arbre. Par conséquent, Breiman et al. (1984) proposent d'établir une hiérarchie des variables explicatives permettant de mettre en évidence les variables les plus importantes dans la construction de l'arbre. C'est à partir des divisions de substitution que l'importance de chacune des variables est évaluée. En reprenant les notations précédentes, l'importance de la variable X_k est définie par :

$$I(X_k) = \sum_{t \in A} \Delta i(\tilde{\delta},t) \qquad (10)$$

Il s'agit de la somme sur tous les nœuds de l'arbre de la réduction de l'hétérogénéité entraînée par la division de substitution sur la variable X_k.
En pratique, la mesure de l'importance des variables (10) est normalisée :

$$\tilde{I}(X_k) = 100\,\frac{I(X_k)}{\max\limits_{1 \le k \le p} I(X_k)}$$

où la variable la plus importante prend la valeur 100 et les autres prennent leurs valeurs dans l'intervalle $[0, 100[$.

4 Instabilité des arbres de discrimination et méthodes de stabilisation

4.1 Introduction

Une méthode statistique de prédiction est dite instable si une légère perturbation dans l'échantillon d'apprentissage conduit à une grande modification dans la prédiction finale (Breiman, 1996a). Les méthodes de construction d'arbre de décision ont été identifiées comme des méthodes instables (Breiman, 1996a ; Rakotomalala, 2005 ; Ruey-Hsia, 2001). Dans le paragraphe 4.2, nous illustrons l'instabilité des arbres de discrimination à partir de l'exemple de contamination radioactive de la laitue. Nous définissons en quoi cette instabilité va être gênante pour notre application radioécologique. Le paragraphe 4.3 décrit les principales méthodes développées pour améliorer les performances de prédiction des arbres de discrimination : bagging, random forests et boosting. Dans le paragraphe 4.4, nous présentons deux algorithmes développés afin de stabiliser la structure d'un arbre de discrimination. Nous décrivons plus particulièrement la proposition de Dannegger (2000) permettant de stabiliser les nœuds intermédiaires d'un arbre de discrimination à partir de rééchantillonnage bootstrap. Nous proposons une méthodologie, dans le paragraphe 4.4.3, afin de construire des arbres de discrimination par cet algorithme de sélection des divisions.

4.2 Illustration de l'instabilité

L'instabilité des arbres de discrimination a été illustrée par de nombreux auteurs (Ghattas, 2000 ; Dannegger, 2000 ; Ruey-Hsia, 2001). Déjà, Breiman et al. (1984) y faisaient référence dans leur monographie. Pour illustrer cette notion, nous allons construire différents arbres de discrimination par la méthode CART, à partir du scénario de contamination défini au paragraphe 2.3.2 et appliqué au cas de la laitue[11]. Nous générons 10 échantillons d'apprentissage de taille 1000 et construisons 10 arbres de discrimination[12] par validation croisée ($V = 10$ et la sélection de l'arbre final se fait par la règle de l'écart-type définie au paragraphe 3.4.3.3). Le tableau 4.A présente, pour chaque arbre de discrimination (A_1,.., A_{10}) le nombre de feuilles obtenues ainsi que le taux de mauvais classement (%), estimé à partir d'un échantillon test (de taille 1000). Dans un premier temps, les taux de mauvais classement sont comparés. Selon l'arbre, les estimations sont

[11] Pour plus de détails concernant la création des échantillons artificiels de données, le lecteur peut se référer au chapitre 5.
[12] Le logiciel utilisé pour la construction des arbres de discrimination est le logiciel statistique R.

assez variables, allant de 4,3 % jusqu'à 8,1 %. Nous verrons dans le paragraphe 4.3 les principales méthodes développées pour stabiliser les résultats de prévisions.

Arbre	Nombre de feuilles	Taux de mauvais classement (%)
A_1	6	4,3
A_2	9	5,9
A_3	6	7
A_4	6	5,3
A_5	11	7,4
A_6	9	4,9
A_7	8	5,1
A_8	14	4,5
A_9	8	8,1
A_{10}	8	5,6

Tableau 4.A : Nombre de feuilles et taux de mauvais classement (estimé par un échantillon test) associés aux 10 arbres de discrimination construits selon la méthode CART

Le deuxième type d'instabilité, plus gênant dans notre contexte, se remarque au niveau de la structure des arbres. Le tableau 4.A permet d'identifier des variations dans le nombre de feuilles associées à chaque arbre de discrimination (de 6 à 14 nœuds terminaux). Une structure d'arbre synthétisant les informations relatives aux 7 premiers nœuds des 10 arbres de discrimination construits est présentée sur la figure 4.a. Pour chaque nœud intermédiaire, nous précisons le nombre de fois où l'une des variables explicatives est sélectionnée pour effectuer la segmentation, et, lorsque celle-ci est sélectionnée plusieurs fois, la valeur minimale et la valeur maximale de la division optimale. Le premier nœud est assez stable au niveau du choix de la variable de segmentation, 9 fois sur 10 il est basé sur le rapport de captation (R_c). A partir des niveaux inférieurs, il y a une certaine instabilité dans le choix de la variable et de la division associée. Par exemple, pour le nœud n°3, le dépôt est sélectionné 6 fois sur 10 et la division peut aller de 3099 $Bq.m^{-2}$ à presque 10000 $Bq.m^{-2}$. Ces variations sont dues aux faibles effectifs présents dans les niveaux inférieurs des arbres engendrant des partitions pas toujours identiques. Elles entraînent des modifications dans la structure de l'arbre et donc dans les règles de décision qui en découlent. Selon le contexte d'application, cette instabilité peut être assez gênante pour l'utilisateur. Même si deux règles différentes peuvent conduire à une même action (Breiman et al., 1984 ; Rakotomalala, 2005), le problème du choix de la règle se pose. L'un des objectifs de notre application radioécologique est l'identification des valeurs spécifiques des entrées du modèle radioécologique pour lesquelles le niveau de

contamination radioactive des végétaux dépasse (ou ne dépasse pas) des valeurs limites. Nous envisageons d'utiliser certaines règles extraites des arbres de discrimination afin de proposer des recommandations dans un contexte post-accidentel. Cette instabilité est donc un vrai problème pour les décideurs qui ont besoin de disposer de règles de décision robustes. Dans le paragraphe 4.4, nous allons décrire deux algorithmes développés afin de stabiliser la procédure de sélection des divisions. En particulier, nous nous intéressons à une méthode de rééchantillonnage bootstrap dans les nœuds développée par Dannegger (2000).

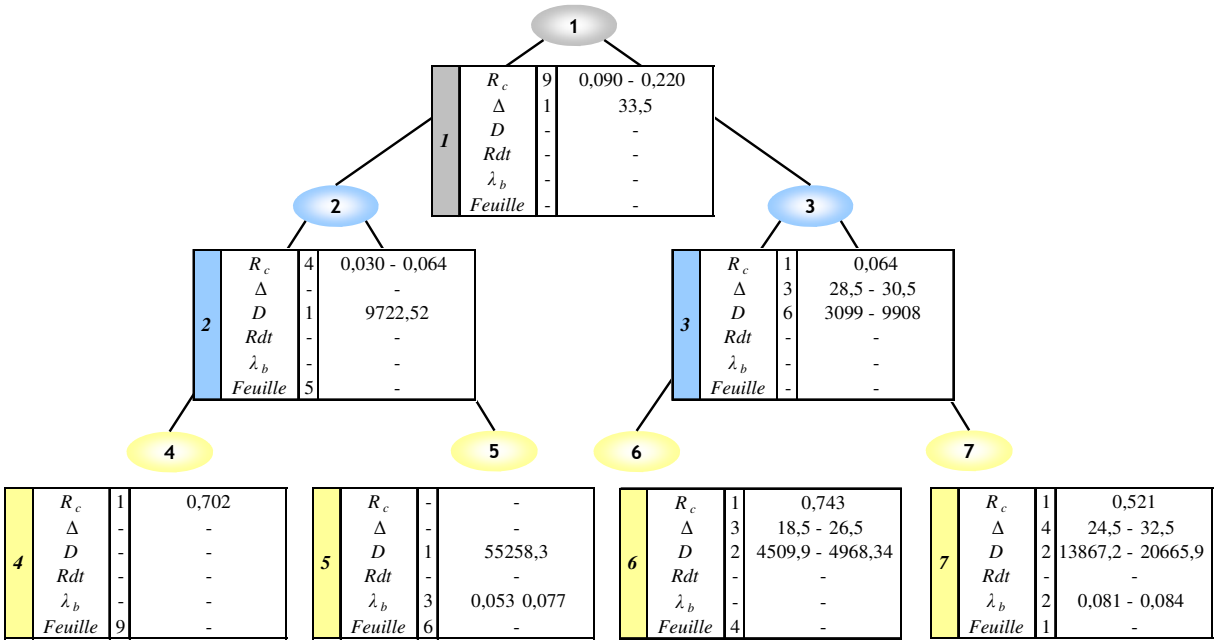


Figure 4.a : Synthèse des 7 premiers nœuds obtenus des 10 arbres de discrimination construits par la méthode CART. Les variables explicatives sont au nombre de 5 : rapport de captation (R_c), délai dépôt-récolte (Δ), dépôt de radioactivité (D), rendement cultural (Rdt) et constante de décroissance biomécanique (λ_b).

4.3 Stabiliser les prédictions d'un arbre de discrimination

Dans ce paragraphe, nous décrivons les principales méthodes développées permettant d'améliorer les performances de prédiction des arbres de discrimination.

4.3.1 L'agrégation par bootstrap : bagging

La méthode bagging (***bootstrap aggregating***) (Breiman, 1996b), consiste en la construction de nombreuses versions d'un modèle de prédiction, à partir de l'échantillon d'apprentissage, et en leur agrégation afin d'obtenir un nouvel estimateur plus performant. Breiman (1996b) montre que la procédure d'agrégation améliore nettement les performances de prédiction lorsque l'estimateur est affecté d'une grande variance

(estimateur instable). La construction des différentes versions du prédicteur s'effectue à partir de rééchantillonnage bootstrap[13] de l'échantillon d'apprentissage E. Un échantillon bootstrap est obtenu par rééchantillonnage de l'échantillon d'apprentissage selon la fonction de distribution empirique, qui associe une probabilité $1/n$ à chaque observation (bootstrap non-paramétrique) ou, s'il est connu, selon un modèle paramétrique (bootstrap paramétrique, c'est le cas de notre application car les échantillons de données sont générés selon les lois de probabilités des différentes variables[14]).

Le principe du bagging est décrit dans l'algorithme suivant :

Pour b variant de 1 à B ***Faire***

Générer un échantillon bootstrap E_b

Construire un estimateur ($\hat{y}_b$) à l'aide de l'échantillon bootstrap

Fin pour

La prédiction d'un nouvel individu x est déterminée par :

$$\hat{y}_{bagg}(x) = \frac{1}{B}\sum_{i=1}^{B} \hat{y}_b(x) \qquad \textit{dans le cas de la régression}$$

$$\hat{y}_{bagg}(x) = \arg\max_j \#\{b; \hat{y}_b(x) = j\} \qquad \textit{dans le cas de la discrimination}$$

Une application du bagging à la méthode CART est présentée par Breiman (1994). Dans le cas de la discrimination, 7 jeux de données sont utilisés. Ils sont divisés aléatoirement de manière à construire un échantillon d'apprentissage et un échantillon test. Un arbre simple est construit par la méthode CART et $B = 50$ échantillons bootstrap sont utilisés pour construire l'estimateur agrégé. Cette procédure est répétée 100 fois et, à chaque étape, les taux de mauvais classement sont estimés sur l'échantillon test. De la même manière, 5 jeux de données sont étudiés pour le cas de la régression ($B = 25$ et la qualité de prédiction est évaluée par le calcul de l'erreur quadratique moyenne). Pour chaque jeu de données étudiés, les performances obtenues par le bagging sont toujours meilleures que celles de l'arbre simple. Il y a cependant deux cas pour lesquels le bagging ne donnera pas d'amélioration notable :

- lorsqu'il est appliqué à une méthode identifiée comme stable, par exemple à l'algorithme des k plus proches voisins (Breiman et al., 1996a),

[13] Les techniques du bootstrap sont basées sur des rééchantillonnages de l'échantillon initial pour améliorer la précision des estimations statistiques. Par exemple, ces techniques peuvent être utilisées pour l'estimation d'une médiane, d'un écart-type ou encore d'un intervalle de confiance (Efron and Tibshirani, 1993).

[14] Pour être plus précis, ce sont des simulations Monte Carlo. Néanmoins, nous utiliserons dans la terminologie bootstrap qui est plus générale.

- pour les jeux de données où le prédicteur est très proche de la réalité observée, les performances ne pourront qu'être très faiblement améliorées.

4.3.2 La méthode Random Forests

Développée par Breiman (2001), cette méthode consiste à améliorer les capacités prédictives d'un modèle CART, par l'agrégation de plusieurs arbres perturbés au moyen d'une double randomisation. Breiman (2001) démontre que les performances de prédiction d'une forêt aléatoire dépendent de deux paramètres : la qualité de prédiction de chaque arbre et leur indépendance[15]. Plus les arbres seront indépendants et auront une bonne capacité de prédiction individuelle, meilleure sera la qualité de prédiction de la forêt aléatoire.

Comme pour le bagging, la construction des différents arbres se fait à partir de rééchantillonnage bootstrap de l'échantillon d'apprentissage E. Les arbres ne sont cependant pas construits selon la méthode CART classique. Afin de les rendre plus indépendants et de préserver leur qualité de prédiction, Breiman (2001) propose deux types de perturbations :

- la meilleure division d'un nœud est choisie parmi q des p variables explicatives, où q est sélectionné aléatoirement (Random Forests–RI). Dans le même esprit, Dietterich (1999) avait proposé la méthode *randomization* afin de construire un arbre de discrimination où chaque division était sélectionnée aléatoirement parmi les K meilleures.
- une combinaison linéaire aléatoire des variables explicatives est utilisée pour effectuer la division d'un nœud (Random Forests–RC).

En pratique la méthode Random Forests–RI est la plus souvent utilisée (programmée dans la librairie *randomForest* de R). Le principe de cette méthode est décrit dans l'algorithme suivant :

Pour b variant de 1 à B ***Faire***
  Générer un échantillon bootstrap E_b
  Construire un arbre sur cet échantillon :
    Pour chaque nœud intermédiaire de l'arbre ***Faire***
    Tirage aléatoire de q variables explicatives ($q=1,\ldots,p$)
    Sélection de la variable de segmentation parmi les q variables

[15] Dans ce cas, le concept d'indépendance est une notion difficile à définir, elle peut être vue comme l'absence de liaison entre les arbres de la forêt. Pour plus de détails, la notion de corrélation entre les arbres d'une forêt aléatoire est définie par Breiman (2001).

Fin pour

L'estimateur obtenu est noté $\hat{y}_b$

Fin pour

La prédiction d'un nouvel individu x est déterminée par

$$\hat{y}_{rf}(x) = \frac{1}{B}\sum_{i=1}^{B} \hat{y}_b(x) \qquad \textit{dans le cas de la régression}$$

$$\hat{y}_{rf}(x) = \arg\max_j \#\{b; \hat{y}_b(x) = j\} \qquad \textit{dans le cas de la discrimination}$$

La sensibilité de l'algorithme (Random Forests-RI) au choix du nombre de variables explicatives q a été testée empiriquement par Breiman (2001) à partir de nombreux jeux de données. Lorsque la valeur de q augmente, la qualité de prédiction de chaque arbre croît rapidement avant d'atteindre un pallier pour un certain nombre de variables. La dépendance entre les arbres est une fonction croissante du nombre de variables. De ce fait, la qualité de prédiction de la forêt aléatoire atteint rapidement son maximum avant de se détériorer progressivement. L'effet de l'addition de bruit sur les performances de l'algorithme est également testé par Breiman (2001). Les résultats obtenus sont généralement peu altérés, contrairement à ceux obtenus par une méthode de boosting Adaboost (Cf. 4.3.3) assez sensible à l'ajout de bruit.

Le principal inconvénient de cette méthode est la perte de la structure de l'arbre unique et donc de la facilité d'interprétation. Conscient que cet outil explicatif est très utile dans certains domaines comme la médecine, Breiman (2001, 2002) propose en complémentarité de sa méthode plusieurs mesures permettant de quantifier l'importance des variables explicatives afin de mieux appréhender le phénomène modélisé.

4.3.3 Le boosting

Le boosting (Freund et Schapire, 1999) est un autre moyen d'améliorer les performances d'un estimateur. D'une manière générale, le principe de cette technique est identique à celui du bagging dans le sens où le résultat final consiste en une agrégation d'une famille de modèles. La principale différence réside dans la construction de chaque modèle. Nous nous restreignons ici à la présentation de l'algorithme de boosting de référence AdaBoost (pour Adaptative Boosting) proposé par Freund et Schapire (1995). L'objectif de cet algorithme est la prédiction d'une variable qualitative à deux modalités : $Y = \{-1,+1\}$.

Soient $\{(\mathbf{x}_1, y_1),...,(\mathbf{x}_n, y_n)\}$ un échantillon de données, $w_1(i) = \frac{1}{n},\ i = 1,...,n$ les poids initiaux associés à chaque individu i de l'échantillon et h le classifieur (modèle de classement qui

prédit l'appartenance à l'une des deux modalités, par exemple un arbre de discrimination). Le principe de l'algorithme AdaBoost est le suivant :

Pour t variant de 1 à T ***Faire***

Estimer le classifieur h_t à partir de l'échantillon pondéré par $w_t(i)$

Calculer l'erreur associée : $\varepsilon_t = \sum_{i=1}^{n} w_t(i) \begin{cases} 0 \text{ si } h_t(\mathrm{x}_i) = y_i \\ 1 \text{ si } h_t(\mathrm{x}_i) \neq y_i \end{cases}$

Calculer α_t : $\alpha_t = \frac{1}{2}\log\left(\frac{\varepsilon_t}{1-\varepsilon_t}\right)$

Mettre à jour les poids :

$$w_{t+1}(i) = \frac{w_t(i)}{Z_t} \begin{cases} e^{-\alpha_t} & \text{si } h_t(\mathrm{x}_i) = y_i \\ e^{\alpha_t} & \text{si } h_t(\mathrm{x}_i) \neq y_i \end{cases}$$

où Z_t est une valeur de normalisation faisant que la somme des poids soit égale à 1.

Fin pour

Le classement d'un nouvel individu x se fait en utilisant la règle :

$$H(x) = signe\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

L'une des idées principales de l'algorithme AdaBoost est d'affecter des poids $w_t(i)$ à chaque individu appartenant à l'échantillon d'apprentissage. A chaque itération t, un classifieur est calculé à partir de l'échantillon (pondéré par les poids) et sa qualité de prédiction est mesurée. Au départ, tous les poids sont égaux, puis à chaque itération, des poids plus élevés sont affectés aux individus mal classés afin de forcer le classifieur à se concentrer sur ces exemples difficiles. A chaque étape, α_t peut être vu comme l'importance accordée au classifieur h_t : au plus l'erreur associée à h_t est faible, au plus la valeur de α_t est élevée. Ainsi, à la fin de l'algorithme, tous les classifieurs construits sont pondérés par leurs poids respectifs α_t et le classement d'un nouvel individu peut être déduit par le signe de cette expression.

Suite à la publication de cet algorithme, de nombreuses versions ont été développées. Parmi elles, on peut citer une version aléatoire proposée par Breiman (1996d), une généralisation de l'algorithme lorsque la variable à expliquer est à plus de deux modalités (Freund et Shapire, 1996) et une adaptation au cas où la variable à expliquer est quantitative (Drucker, 1997). Des comparaisons empiriques montrent que les méthodes de

boosting appliquées aux cas de la discrimination donnent généralement de meilleures performances que le bagging (Freund et Schapire, 1996 ; Dietterich, 1999). Néanmoins en présence de données bruitées, les algorithmes de boosting sont beaucoup moins performants (Dietterich, 1999). Ce résultat a également été mis en évidence pour le cas de la régression par Gey et Poggi (2006).

4.4 Stabiliser la structure d'un arbre de discrimination

Contrairement aux approches décrites dans le paragraphe précédent qui permettent de stabiliser les performances, les méthodes présentées dans cette partie permettent de préserver la structure de l'arbre de discrimination, avec pour objectif l'obtention de règles de décision plus robustes. Dannegger (2000) et Ruey-Hsia (2001) ont montré que le comportement instable des méthodes de construction d'arbre de discrimination était causé par l'instabilité dans la sélection des divisions associées aux nœuds intermédiaires de l'arbre. Ces deux auteurs ont donc proposé des procédures permettant de stabiliser la division associée à un nœud intermédiaire. Les paragraphes 4.4.1 et 4.4.2 décrivent respectivement les algorithmes proposés par Ruey-Hsia (2001) et Dannegger (2000). Nous présentons dans le paragraphe 4.4.3 une méthodologie permettant de construire des arbres de discrimination par la technique développée par Dannegger (2000).

4.4.1 Construction d'un arbre stable par associations de divisions

La méthode proposée par Ruey-Hsia (2001), consiste à rechercher, pour chaque nœud intermédiaire, des associations de divisions ayant des propriétés particulières. Afin d'expliquer le principe de cette approche, les définitions suivantes sont nécessaires :

- Une modification de l'échantillon d'apprentissage peut correspondre à l'ajout, à la suppression ou à la modification de certaines observations.
- La « sensibilité de la division optimale[16] d^{opt} » correspond à la plus petite modification de l'échantillon d'apprentissage qui fait que d^{opt} n'est plus considérée comme la division optimale pour le nouvel échantillon.
- Pour un nœud intermédiaire t, « la sensibilité de la division optimale d^{opt} à une division admissible d^k » est définie comme étant la plus petite modification de l'échantillon d'apprentissage (en %) qui conduit à préférer d^k pour segmenter le nœud t.

[16] Division maximisant la réduction de l'hétérogénéité, Cf. chapitre 3, équation (7).

- Une division admissible est dite *presque aussi bonne*[17] que la division optimale si la sensibilité de d^{opt} à cette division est inférieure ou égale à p, un pourcentage donné de modification des données d'apprentissage.

Pour illustrer ces définitions, considérons l'exemple suivant issu de Ruey-Hsia (2001). Soient t un nœud intermédiaire issu d'un arbre de discrimination, d_1 la division optimale associée à t et d_2, d_3, d_4 3 divisions candidates au nœud t. Le calcul de la sensibilité de d_1 aux divisions d_2, d_3 et d_4 est le suivant :

c_2 = sensibilité de d_1 à d_2 = 1 %

c_3 = sensibilité de d_1 à d_3 = 2 %

c_4 = sensibilité de d_1 à d_4 = 5 %

Alors, la sensibilité de d_1 est de 1 % (c'est le plus faible pourcentage de modification qui conduit à ne plus considérer d_1 comme la division optimale au nœud t) et les divisions presque aussi bonnes que d_1 en fonction d'un changement de 3 % des données d'apprentissage sont d_1, d_2 et d_3.

La procédure de stabilisation proposée par Ruey-Hsia (2001) consiste à utiliser, pour la segmentation d'un nœud intermédiaire, des combinaisons de divisions ayant un même degré d'importance que la division optimale (notion de division *presque aussi bonne* que la division optimale). Pour déterminer la division associée au nœud t, le principe général de l'algorithme est le suivant :

(a) Déterminer la division optimale δ^* associée au nœud t : $\Delta i(\delta^*,t) = \arg\max_{\delta \in D}\{\Delta i(\delta,t)\}$.

(b) Identifier les divisions presque aussi bonnes que δ^* en fonction d'un pourcentage p fixé d'un changement dans l'échantillon d'apprentissage.

(c) Former des expressions logiques à partir des divisions obtenues en (b) : identification de toutes les associations possibles entre les variables explicatives et leur division associée, en utilisant l'opérateur logique « et ». Par exemple $2 \le X_1 \le 5$ ou encore $2 \le X_1$ et $X_2 = \{1\}$.

(d) Sélectionner les associations obtenues en (c) dont la réduction de l'hétérogénéité est supérieure à $\Delta i(\delta^*,t)$.

(e) A partir des associations sélectionnées en (d) former d'autres associations logiques, cette fois-ci en utilisant l'opérateur logique « ou ».

[17] Traduit littéralement de l'anglais *almost as good as*.

(f) Sélectionner la meilleure association parmi celles obtenues en (d) et (e) qui sera utilisée pour segmenter le nœud : celle dont la réduction de l'hétérogénéité est maximale et supérieure à $\Delta i(\delta^*, t)$.

Ruey-Hsia (2001) a appliqué l'algorithme de sélection décrit ci-dessus à la méthode CART. Les arbres obtenus par cet algorithme modifié et l'algorithme classique[18] ont été comparés. Les principaux résultats obtenus sont décrits ci-dessous.

Dans un premier temps, la comparaison a porté uniquement sur la structure des arbres. Pour cela, des générateurs de données artificielles ont été utilisés : les variables explicatives sont toujours booléennes et la variable à expliquer est une association logique entre les variables explicatives. Par exemple, pour le premier jeu de données, les variables explicatives sont au nombre de 5 (X_i, i=1,...,5) leurs valeurs sont générées dans $\{0,1\}$ et la variable à expliquer est définie comme suit :

$$Y = \begin{cases} 1 \text{ si } X_1 = 1 \text{ et } X_2 = 1 \\ 0 \text{ sinon} \end{cases}$$

De la même manière, six autres jeux de données ont été considérés. Pour déterminer les divisions de l'étape (b) le pourcentage de modification a été fixé à 1 %. Les résultats empiriques ont montré que la méthode proposée par Ruey-Hsia (2001) tend à produire des arbres moins complexes et plus concis, mais logiquement équivalents à ceux obtenus par la méthode CART. De ce fait, la qualité de prédiction n'a pas été comparée. Pour l'exemple précédent, l'arbre obtenu par la méthode CART comprend deux nœuds intermédiaires (segmentés respectivement par $X_2 = 1$ et $X_1 = 1$) et 3 feuilles, tandis que l'algorithme proposé par Ruey-Hsia (2001) génère un arbre de discrimination ayant qu'un seul nœud, la racine, où la division synthétise les informations précédentes : $X_1 = 1 \wedge X_2 = 1$.

Une comparaison empirique plus intéressante, proposée par Ruey-Hsia (2001), porte sur la sensibilité des deux algorithmes à de faibles modifications de l'échantillon d'apprentissage. Cet échantillon est obtenu de la même manière que dans les expériences précédentes, mais 20 % des données sont volontairement mal classées. Comme précédemment, les deux premiers arbres construits sont logiquement équivalents. Cependant, si 1,5 % des données d'apprentissage sont modifiées aléatoirement, la structure de l'arbre construit selon la méthode CART est altérée, tandis que l'autre arbre reste identique (construit selon l'algorithme proposé par Ruey-Hsia (2001)). Il en est de même pour une modification de 2,3 %. Ce n'est qu'à partir d'un changement de 2,5 % que de légères modifications sont observées dans la structure de l'arbre construit par l'algorithme modifié. De plus, la qualité de prédiction de cet arbre, reste toujours

[18] La mesure d'hétérogénéité utilisée par Ruey-Hsia (2001) est le critère de Gini (Breiman et al., 1984).

supérieure à celle de l'arbre obtenu par la méthode CART. Les nombreuses comparaisons empiriques effectuées par Ruey-Hsia (2001) montrent que cet algorithme est tout de même assez coûteux en temps de calcul, mais qu'il permet d'obtenir des arbres plus stables et plus concis que ceux obtenus par la méthode CART.

Cependant, l'existence d'expressions logiques pour segmenter chaque nœud intermédiaire de l'arbre peut vite devenir gênant pour le décideur, en particulier dans notre contexte d'application. En situation d'urgence, il est nécessaire de disposer de règles de décision simples, hiérarchiques pour prioriser la quête d'information. De nombreux connecteurs logiques dans une règle peuvent entraîner une complexification de celle-ci et de ce fait engendrer une confusion dans l'esprit du décideur ayant du mal à synthétiser des recommandations. Ainsi, pour notre application radioécologique, nous préférons utiliser une méthode permettant de stabiliser, à chaque nœud intermédiaire de l'arbre, le choix de la variable explicative et de sa division associée.

4.4.2 Construction d'un arbre stable par rééchantillonnage bootstrap dans les nœuds

L'algorithme proposé par Dannegger (2000) consiste à effectuer des rééchantillonnages bootstrap dans les nœuds intermédiaires afin d'obtenir des divisions plus stables. Pour chaque nœud t intermédiaire de l'arbre de taille $n(t)$, il s'agit de réaliser B rééchantillonnages bootstrap de taille $n_B(t)$ et de rechercher pour chaque échantillon bootstrap la division optimale. La variable associée à la division apparaissant le plus souvent en première position est alors sélectionnée pour segmenter le nœud et sa division est choisie en prenant la médiane des différentes réplications bootstrap.

Dannegger (2000) teste les capacités de sa méthode en l'appliquant à un jeu de simulation (de taille n=500) dans le contexte de la discrimination. Il construit 3 types de modèles : le premier est un arbre de discrimination obtenu par l'application de la méthode CART, le deuxième est un modèle agrégé (bagging appliqué à la méthode CART) et le troisième un arbre de discrimination construit par la méthode de rééchantillonnage dans les nœuds (B=100). Cette expérience est répétée 50 fois et les taux moyens de mauvais classement, pour l'échantillon d'apprentissage et un échantillon test, sont comparés. Les résultats du bagging sont les plus convaincants, cette méthode permet de réduire le taux moyen de mauvais classement estimé sur l'échantillon test de 20 % à moins de 4 %. La méthode de rééchantillonnage dans les nœuds se place en deuxième position, globalement, elle permet de réduire l'erreur de prédiction de 20 % à 6 % (en moyenne). Il est cependant assez dommage que l'auteur ne donne pas un exemple de la structure de l'un des arbres de discrimination obtenus. Notons aussi que dans certaines situations, cette méthode de

rééchantillonnage dans les nœuds peut engendrer une forte augmentation de l'erreur de prédiction. Lorsque deux divisions assez distantes (basées sur une même variable) sont en compétition, le fait de prendre la médiane de la distribution bootstrap conduit à choisir une division instable (Dannegger, 2000).

Comme pour l'algorithme proposé par Ruey-Hsia (2001), cette méthode est assez coûteuse en temps de calcul. Elle apparaît tout de même comme un bon compromis entre la méthode CART et les techniques décrites au paragraphe 4.3 car elle permet de stabiliser les résultats de prédiction tout en conservant la structure de l'arbre. Nous mettons cependant en avant deux points qui mériteraient d'être précisés par l'auteur :

- La procédure d'élagage utilisée. Il s'avère que l'élagage de coût-complexité minimum défini par Breiman et al. (1984), ne s'applique pas aux arbres construits selon cet algorithme car les divisions obtenues ne sont plus optimales au sens de la réduction de l'hétérogénéité (7). Nous y reviendrons dans le paragraphe 4.4.3.2.
- La taille des arbres obtenus. Les arbres peuvent être assez complexes et donc difficilement interprétables.

Enfin, cette méthode pourrait être enrichie par la possibilité de quantifier la stabilité des arbres de décision obtenus. Nous allons définir dans le paragraphe 4.4.3.3 une mesure de similarité afin de comparer la structure de deux arbres de discrimination.

4.4.3 Appropriation de la méthode de Dannegger (2000) : la méthode REN

Dans ce paragraphe, nous allons définir une méthodologie (appelée méthode REN) afin de construire un arbre de discrimination selon l'algorithme de sélection d'une division décrit par Dannegger (2000). Pour élaguer les arbres de discrimination, nous utiliserons la deuxième technique d'élagage (post-élagage) définie au paragraphe 3.3. La méthode REN se décompose donc en deux étapes, présentées dans les paragraphes suivants : la construction de l'arbre maximal (Cf. 4.4.3.1) et la méthode d'élagage choisie (Cf. 4.4.3.2).

4.4.3.1 La construction de l'arbre maximal

L'arbre maximal est construit à partir de l'échantillon d'apprentissage E selon l'algorithme suivant :

Pour chaque nœud t intermédiaire de taille n_t ***Faire***
    Pour b variant de 1 à B ***Faire***

Générer un échantillon bootstrap de taille $n_t(b)$

Pour chaque variable explicative X_k, $k=1,...,p$, rechercher la division optimale $\delta^k(b)$ selon (7)

Fin pour

Sélection de la variable X_d qui sera utilisée pour effectuer la division :

$$X_d = \arg\max_d \#\left\{\delta^d(b) = best\ split; b = 1,...,B; d = 1,...,p\right\}$$

Détermination de la division pour la variable X_d : $\breve{\delta}^d = \begin{cases} \dfrac{\delta^d_{B/2} + \delta^d_{B/2+1}}{2} & \text{si } B \text{ est pair} \\ \delta^d_{(B+1)/2} & \text{si } B \text{ est impair} \end{cases}$

Fin pour

Les critères d'arrêts de division d'un nœud sont les mêmes que ceux de la méthode CART (Cf. 3.4.3.1) ; lorsque le nœud est homogène ou lorsque l'effectif dans le nœud est inférieur à une valeur fixée.

4.4.3.2 La méthode d'élagage

Nous avons choisi d'utiliser une méthode d'élagage proposée par Quinlan (1987), appelée *reduced error pruning*, consistant à évaluer les performances des différentes branches de l'arbre maximal A_{max} à l'aide d'un échantillon de validation. Cette méthode nous a séduit par sa simplicité de réalisation et le fait que l'arbre obtenu par élagage est le plus petit arbre (issu de l'arbre maximal) dont la qualité de prédiction est maximale sur l'échantillon de validation. Une comparaison empirique de cette méthode avec d'autres méthodes d'élagage peut être trouvée dans (Quinlan, 1987 ; Mingers, 1989 ; Esposito et al., 1997).

Le point de départ de l'algorithme d'élagage est l'arbre A_{max}. Soient $\hat{T}(A^t)$ l'estimation du taux de mauvais classement de la branche A^t (branche issue du nœud t et ayant t pour racine) et $\hat{T}(t)$ l'estimation du taux de mauvais classement au nœud t si ce dernier est considéré comme une feuille. Pour chaque nœud intermédiaire t, la quantité suivante est calculée, à partir de l'échantillon de validation :

$$g(t) = \hat{T}(A^t) - \hat{T}(t) \qquad (11)$$

La valeur de (11) peut être :

(a) positive : la branche A^t est moins pertinente, si elle est conservée, elle engendre une augmentation du taux de mauvais classement,

(b) nulle : la branche A^t n'apporte pas plus d'information que le nœud t car le taux de mauvais classement n'évolue pas,

(c) négative : la branche A^t contient des feuilles pertinentes car sa suppression conduit à une détérioration des prévisions.

Les nœuds vérifiant (a) et (b) sont sélectionnés et la branche maximisant (11) est élaguée. Il y a cependant une contrainte : cette branche ne doit pas contenir de sous-branche présentant une erreur inférieure (car cela conduirait à supprimer une branche dont la racine est peu informative mais dont certaines sous-parties apportent des informations supplémentaires). Ce processus est alors réitéré jusqu'à ce que l'arbre contienne uniquement des nœuds présentant une valeur négative du critère (11).

L'un des inconvénients de cette méthode est la dépendance à l'échantillon de validation. Il est possible que certains cas rares présents dans l'échantillon d'apprentissage et modélisés par l'arbre soient absents de l'échantillon de validation (Quinlan, 1987). Ce qui peut entraîner la suppression de certaines branches de l'arbre caractéristiques de ces cas rares. Ce problème est particulièrement mis en évidence lorsque la taille de l'échantillon de validation est beaucoup plus petite que celle de l'échantillon d'apprentissage, mais devient moins gênant lorsque le nombre d'observations dans l'échantillon de validation augmente (Esposito et al., 1997).

Notons que la méthode de coût-complexité minimal proposée par Breiman et al., (1984), ne paraît pas être adaptée à l'élagage des arbres construits selon la méthode de rééchantillonnage bootstrap dans les nœuds. En effet, la procédure de construction de l'arbre maximal par l'algorithme de sélection des divisions proposé par Dannegger (2000) diffère de celle de la méthode CART. Pour chaque nœud intermédiaire, des échantillons bootstrap sont utilisés pour rechercher une division stable qui n'est alors plus optimale au sens de (7). Il n'est donc pas pertinent d'utiliser l'échantillon d'apprentissage pour calculer les valeurs du critère d'élagage (9). Contrairement à ce que l'on observe habituellement, l'évolution du taux de mauvais classement sur l'échantillon d'apprentissage ne sera pas forcement une fonction strictement décroissante du nombre de feuilles.

4.4.3.3 Une mesure de similarité pour comparer la structure de deux arbres de discrimination

Pour pouvoir quantifier la stabilité des arbres de discrimination construits selon la méthode REN, nous définissons dans ce paragraphe une mesure de similarité.

Soient A_1 et A_2 des arbres de discrimination construits sur les mêmes variables explicatives (quantitatives) et à expliquer, et à partir de deux méthodes différentes ou de deux échantillons d'apprentissage distincts. Nous supposons dans un premier temps, que le

domaine de valeurs associées à chaque variable est fini et que les deux arbres ont la même architecture (ensemble des nœuds intermédiaires et leurs positions respectives sans prendre en compte l'information apportée par la division à chaque nœud).

Notons t_0, t_1,…, t_T les nœuds intermédiaires d'un arbre, numérotés par ordre croissant du nœud racine et de la gauche vers la droite. Le nœud t de l'arbre A_1 est segmenté par $X_k \leq$ ou $> \delta_1$ tandis qu'avec l'arbre A_2 il est segmenté par $X_{k'} \leq$ ou $> \delta_2$.

Soit

$$S_t = \mathrm{I}\{k = k'\}\left(1 - \frac{|\delta_1 - \delta_2|}{range(X_k)}\right) \tag{12}$$

la dissimilarité entre les arbres A_1 et A_2 au nœud t, où $\mathrm{I}\{k = k'\}$ est l'indicatrice de l'évènement $X_k = X_{k'}$. Nous avons $S_t = 0$ quand la segmentation du nœud t est effectuée à partir de variables différentes ($X_k \neq X_{k'}$) ou, si les variables de segmentation coïncident, lorsque les divisions sont situées aux extrémités opposées de l'étendue de X_k. $S_t = 1$ si les divisions au nœud t sont identiques. Dans les autres cas $S_t \in]0, 1[$.

Considérons q_0, q_1,…, q_T des poids non négatifs sommant à 1, associés à chaque nœud intermédiaire de l'arbre. La mesure de similarité entre les arbres de discrimination A_1 et A_2 est définie par :

$$d(A_1, A_2) = 1 - \sum_{t=0}^{T} q_t S_t \tag{13}$$

Cette quantité est symétrique et satisfait $0 \leq d(A_1, A_2) \leq 1$ avec $d(A_1, A_2) = 0$ si les deux arbres sont identiques. A l'opposé, lorsque aucune variable de segmentation ne coïncide, $d(A_1, A_2) = 1$.

Quelques remarques concernant ce critère :

- Si l'étendue de la variable X_k est infinie, elle peut être remplacée par l'étendue observée de cette variable dans les échantillons (combinés). Une autre variante, serait de remplacer le ratio (12) par une autre mesure de dispersion relative, par exemple $|F_k(\delta_1) - F_k(\delta_2)|$, où F_k est la fonction de répartition cumulée (peut-être empirique) de X_k.

- Dans (13), les poids q_t sont à renseigner par l'utilisateur. Ils sont introduits pour pouvoir accorder plus d'importance à certaines parties de l'arbre qui seraient plus pertinentes pour la comparaison. Dans beaucoup d'applications pratiques, la partie

supérieure de l'arbre serait associée à des poids élevés, tandis que les nœuds intermédiaires proches des feuilles recevraient moins d'importance. Des poids constants ($q_t = 1/(T + 1)$) peuvent aussi être utilisés. Nous verrons que dans notre application ce choix n'est pas crucial.

- La mesure de similarité (13) peut être adaptée à des arbres ayant des architectures différentes. Chaque fois qu'une feuille est rencontrée dans l'un des arbres où un nœud intermédiaire est présent dans l'autre, une « branche fantôme » de la même structure est ajoutée afin de remplacer la feuille du premier arbre. Dans la branche fantôme, les divisions associées aux nœuds fantômes ne sont pas définies et engendrent, trivialement, toutes les valeurs de S_t à 0.

- La mesure de similarité ne prend pas en compte les feuilles de l'arbre. Dans notre application, nous sommes principalement intéressés par la hiérarchie des divisions, qui est traduite en règles de décision. Cependant, la méthode précédente peut être étendue aux feuilles en ajoutant dans $d(A_1, A_2)$ un terme mesurant la diversité des classes dans le nœud terminal.

Avant de terminer ce paragraphe, nous proposons quelques usages de notre mesure de similarité. Soit $\{A_1,\ldots, A_M\}$ un ensemble d'arbres de discrimination faisant intervenir les mêmes variables explicatives et à expliquer. La « dispersion » de cet ensemble peut être mesurée par la formule suivante :

$$\frac{1}{M(M-1)} \sum_{1 \le i < j \le M} d(A_i, A_j) \qquad (14)$$

qui est similaire à l'expression alternative de la variance pour un ensemble d'observations $\{X_1,.., X_n\}$ donné, par exemple, par Serfling (1980) :

$$n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \;=\; \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} (X_i - X_j)^2 \,.$$

Elle peut également être utilisée pour définir un arbre « central ». Le fait que la médiane puisse être définie comme l'élément de $\{X_1,.., X_n\}$ qui minimise sur a la quantité $\sum_{i=1}^{n}|X_i - a|$ conduit à définir l'arbre central comme un élément de $\{A_1,\ldots, A_M\}$ qui minimise sur A la quantité :

$$\sum_{i=1}^{M} d(A_i, A) \,.$$

PARTIE 2 : RESULTATS

5 La simulation des données

Ce chapitre présente les choix effectués pour caractériser les distributions des différentes variables explicatives définies dans la partie précédente, pour les quatre légumes-feuilles de notre scénario de contamination (paragraphe 5.1). Les résultats des simulations effectuées à l'aide du modèle de culture STICS sont également présentés. Le paragraphe 5.2 présente la valeur limite utilisée pour effectuer le codage de l'activité du ^{90}Sr dans les légumes-feuilles.

5.1 Les variables explicatives

5.1.1 Les variables dépendantes du végétal

5.1.1.1 Le temps de croissance

5.1.1.1.1 La laitue

Sur une surface totale estimée à 13 3000 hectares la laitue est cultivée toute l'année et à peu près partout en France : 40 % des laitues sont cultivées en hiver sous tunnel et 60 % en plein air aux autres saisons (Thicoïpé, 1997). En hiver, il s'agit majoritairement d'une culture de grands abris concentrée dans les régions à climat doux (4 100 hectares sont cultivés sous grands tunnels, principalement dans le Sud-Est). Au printemps, la culture de plein champ prend le pas et s'étend aux ceintures vertes maraîchères[19], particulièrement celles qui connaissent des étés tempérés (Thicoïpé, 1997).

D'après Thicoïpé (1997) et les différents organismes agricoles contactés (AGRIAL, INRA), le cycle de croissance de la laitue varie entre 30 et 90 jours, le cycle le plus fréquent se situant autour de 45 jours. Les travaux menés par le GRNC dans le cas de l'analyse d'incertitude atour des évaluations du risque (GRNC, 2001) proposent d'utiliser des distributions triangulaires pour les temps de croissance des légumes-feuilles. De ce fait, une loi triangulaire (min=30, mode=45, max=90) est utilisée pour générer des valeurs de temps de croissance de la laitue.

Les travaux effectués à l'aide du modèle de culture STICS nous permettent également d'obtenir les dates de récolte associées à chaque plantation simulée et donc de déduire les temps de croissance (Cf. Annexe E, qui présente les jeux de simulations utilisés afin de faire fonctionner le modèle de culture STICS ainsi que les résultats de ces simulations). Néanmoins, comme nous le précisons dans la partie résultats de l'annexe E, une différence d'une dizaine de jours s'établit entre les valeurs réelles des temps de croissance des laitues, de la plantation à la récolte, et les valeurs issues du logiciel STICS. Ces valeurs ne

[19] Espace agricole et forestier situé entre des agglomérations.

peuvent donc pas être utilisées pour caractériser précisément la distribution du temps de croissance de la laitue.

5.1.1.1.2 L'épinard

Afin d'obtenir des informations sur la culture de l'épinard, l'Union Nationale Interprofessionnelle des Légumes Transformés (UNILET) a été contactée. En France, 6500 hectares (ha) sont consacrés à la production de l'épinard : 5000 ha pour l'industrie agro-alimentaire (surgelés et conserves) et 1500 ha pour la consommation en frais. Pour le premier cas, seulement deux régions sont concernées : le Nord-Picardie et la Bretagne, pour le second, on en trouve dans toutes les ceintures vertes des villes. Globalement, les cycles de croissance de l'épinard varient entre 35 et 75 jours, le cycle le plus fréquent étant de 50 jours. Les cycles les plus longs correspondent aux plantations qui débutent au printemps. Ainsi, d'après ces informations et les formes de distribution de probabilité suggérées par le GRNC (2001), une loi triangulaire (min=35, mode=50, max=75) est utilisée pour caractériser le temps de croissance de l'épinard.

5.1.1.1.3 Le poireau

Le poireau est l'une des rares espèces légumières, comme pour la laitue, que l'on peut cultiver en toute latitude et en toute période de l'année (Le Bohec et al., 1993). Ainsi, il est cultivé dans de nombreuses régions françaises (PACA, Bretagne, Ile-de-France,…). Les temps de croissance ont été renseignés à partir d'un guide pratique sur la culture du poireau édité par le CTIFL (Le Bohec et al., 1993). Grâce à un calendrier de production par zones nous avons pu déterminer les cycles de croissance du poireau qui varient entre 90 et 270 jours, le cycle le plus fréquent étant de 180 jours. Par conséquent, une loi triangulaire (min=90, mode=180, max=270) est utilisée pour générer des valeurs de temps de croissance associées au poireau.

5.1.1.1.4 Le chou

Il existe différentes espèces de choux : les choux raves, les choux chinois, les choux fleurs, les choux pommés, les choux de Bruxelles, …. Les choux dont les feuilles sont consommées sont dénommés choux pommés. De nombreuses variétés sont distinguées : le chou vert, le chou rouge, le chou blanc… Leur culture s'établit principalement en Bretagne et dans le nord de la France. Les cycles de culture varient de 60 à 160 jours, le cycle le plus fréquent étant de 80 jours (CTIFL, communication personnelle). Les cycles les plus courts correspondent aux plantations effectuées au printemps tandis que les cycles les plus longs correspondent aux plantations de plein été (juillet à août) et de septembre. Comme pour les trois autres légumes-feuilles présentés précédemment, nous proposons une loi

triangulaire (min=60, mode=80, max=160) pour caractériser le temps de croissance du chou pommé.

5.1.1.2 Le rendement

5.1.1.2.1 La laitue

Pour étudier le rendement cultural de la laitue, le modèle de culture STICS a été utilisé (Cf. 2.3.4.2). L'annexe E présente en détail les jeux de simulations employés ainsi que les résultats préliminaires qui seront utilisés par la suite pour étudier le rendement cultural et le rapport de captation de la laitue.

Les résultats obtenus à la suite des différentes simulations (une pour chaque date de plantation) fournissent l'évolution journalière de la biomasse produite (en $t.ha^{-1}$ de matière sèche) depuis le stade plantule jusqu'à la récolte. Dans le contexte de notre étude, nous nous intéressons uniquement aux valeurs de rendement à maturité commerciale car l'évaluation de la contamination des végétaux est intéressante surtout lorsqu'ils sont prêts à être consommés. La plupart des modèles radioécologiques de transfert de la radioactivité vers les végétaux expriment la contamination radioactive d'un végétal en becquerel par kg frais de végétal. Un changement d'unité a donc été effectué pour convertir les valeurs de rendement (à maturité du végétal) obtenus par le modèle agronomique STICS de $t.ha^{-1}$ de matière sèche en $kg.m^{-2}$ de matière fraîche. Comme précédemment (Cf. Annexe B), c'est une valeur moyenne du rapport poids frais/poids sec pour les légumes-feuilles (18 $kg\ frais.kg^{-1}$ sec) qui a été utilisée pour effectuer ce changement d'unité (calculée à partir des données issues de la base de données SYLVESTRE). Les différentes valeurs obtenues de rendement (issues des régions de Rennes et d'Orange) sont synthétisées dans le tableau 5.A. Comme nous l'avons cité dans l'annexe E, la valeur moyenne est cohérente avec les rendements réels et avec la valeur de Thicoïpé (1997) qui propose un rendement moyen d'une laitue de 4,2 kg matière fraîche.m^{-2}.

L'utilisation du modèle de culture STICS a également permis de mettre en évidence une relation linéaire entre le rendement (à la récolte) et le temps de croissance de la laitue (Briand et al., 2008).

Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
2,25	3,46	4,33	4,90	5,91	9,23

Tableau 5.A : Résumés statistiques relatifs à la variable rendement ($kg\ frais.m^{-2}$) pour la production de laitue

Estimate	Std. Error	t value	Pr(>|t|)
0,12	0,00	40,80	<2E-16

Tableau 5.B : Résultats relatifs à l'estimation de la pente

Les paramètres de la droite de régression (pente et ordonnée à l'origine) ont été estimés et leur significativité a été testée. L'ordonnée à l'origine peut être considérée comme nulle et les résultats relatifs à l'estimation de la pente sont présentés dans le tableau 5.B (estimation de la pente, estimation de l'écart-type pour la valeur estimée, valeur de la statistique t de Student et probabilité critique). L'équation de la droite d'ajustement est :

$$\hat{R}dt = 0{,}117 \times T_c$$

où $\hat{R}dt$ représente la valeur estimée du rendement et T_c le temps de croissance de la laitue.

Après vérification de l'hypothèse de normalité des résidus, l'intervalle de prédiction à 95 % pour la droite de régression a été déterminé. L'ensemble de ces résultats est présenté graphiquement sur la figure 5.a. L'intervalle de prédiction pour la droite de régression nous permet de disposer d'un intervalle de valeurs probables de rendement pour chaque temps de croissance étudié. Ainsi, connaissant le temps de croissance du végétal, une valeur de la variable rendement peut être générée selon une loi uniforme dans l'intervalle de valeurs possibles associées.

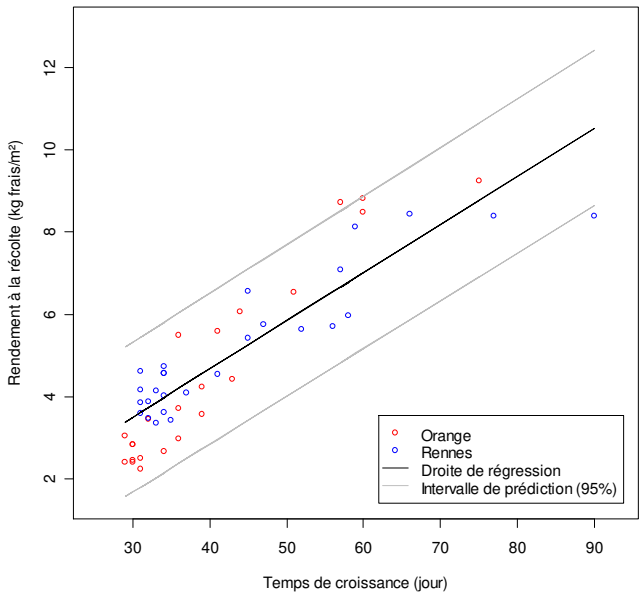


Figure 5.a : Relation entre le rendement et le temps de croissance des laitues

5.1.1.2.2 L'épinard

Les rendements du légume-feuille épinard restent, comme pour la laitue, beaucoup plus importants pour des semis de printemps et beaucoup plus faibles pour des semis d'été, correspondant à des temps de croissance plus faibles (UNILET, communication personnelle). Seulement trois valeurs de rendement associées à des temps de croissance ont été trouvées (UNILET, communication personnelle). Les données sont présentées sur la figure 5.b. L'idée est de reconstruire, comme pour le cas de la laitue, la relation liant le rendement au temps de croissance. Malgré le peu de points existants, une régression linéaire a été effectuée. Ainsi, afin de générer des valeurs de rendement tout en prenant en compte la relation entre le rendement et le temps de croissance, nous calculons directement les valeurs de rendement à partir de l'équation de la droite de régression (Cf. Figure 5.b) et des valeurs simulées de temps de croissance. Nous sommes néanmoins conscients, vu le faible nombre de données disponibles (temps de croissance, rendement), que cette technique n'est pas très fiable. C'est cependant le seul moyen trouvé pour prendre en compte cette relation.

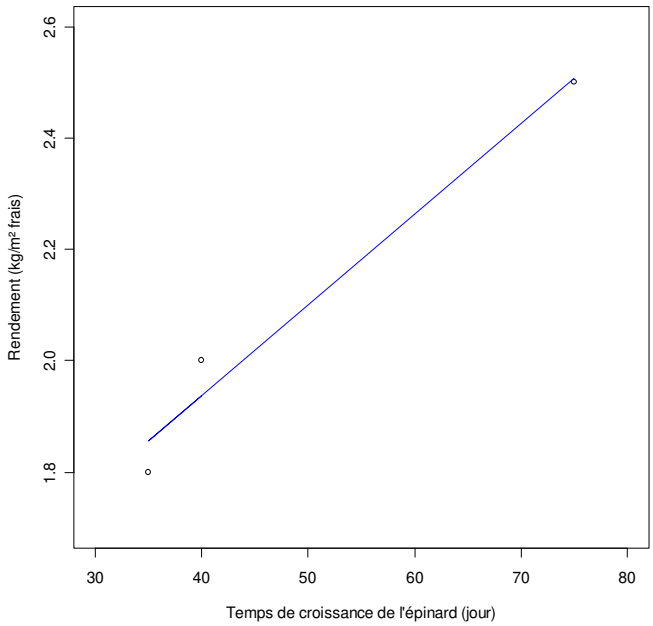


Figure 5.b : Rendement de l'épinard en fonction de son temps de croissance

5.1.1.2.3 Le poireau

Comme il a été énoncé précédemment, le poireau est présent en toutes régions, latitudes et altitudes françaises. Néanmoins, quatre bassins de production se détachent du reste : grand Ouest avec 30 % de la production française, Nord et région parisienne avec 15 % de la production française, grand Sud-Est avec 15 % de l'ensemble et Sud-Ouest avec 8 % du total. L'ensemble des autres départements représente environ 1/3 de la production

nationale (Le Bohec et al., 1993). Selon les données proposées par le Larousse Agricole (1981 ; 2002) et les organismes agricoles contactés, le rendement du poireau varie de 1 à 5 $kg.m^{-2}$ selon le type de culture et l'époque. En l'absence d'informations et de données sur la relation liant le rendement au temps de croissance, nous proposons de générer les valeurs de rendement selon une loi uniforme dans l'intervalle [1, 5 ($kg.m^{-2}$)].

5.1.1.2.4 Le chou

Contrairement à d'autres légumes, les choux pommés sont tous plantés en même temps, il n'y a pas d'étalement des plantations. Les producteurs jouent avec les variétés de choux à cycles plus ou moins précoces et plus ou moins tardifs. Les plantations sont généralement de 36 000 à 48 000 plants/ha, voire jusqu'à 50 000 plants/ha. La récolte n'a pas forcement lieu au stade final de maturité (risque d'éclatement) mais en cours de croissance. Généralement, ce sont des choux d'en moyenne 2 kg. Mais selon les cycles et le marché (colis de 6, en vrac…), le poids varie de 1,2 – 1,5 kg à 2 – 2,5 kg, les choux pouvant faire jusqu'à 4 kg (CTIFL, communication personnelle). Comme pour le poireau, nous n'avons pas d'information supplémentaire sur cette variable, nous proposons donc d'utiliser une loi uniforme et déduisons des informations précédentes l'intervalle de variation [4, 12 ($kg.m^{-2}$)].

5.1.1.3 Le rapport de captation par temps sec

5.1.1.3.1 La laitue

D'après l'équation (5) présentée dans le chapitre 2, le rapport de captation par temps sec est fonction de la variable taux de recouvrement. Cette variable pouvant s'exprimer comme une fonction du temps, le rapport de captation devient lui aussi une fonction du temps. Durant toute la phase de croissance du végétal, le rapport de captation prend alors différentes valeurs ; il évolue ainsi, depuis la sortie hors du sol (0) jusqu'à la maturité commerciale du végétal (palier maximal proche de 1). Pour les différents jeux de données étudiés (Cf. Annexe E), l'évolution du rapport de captation en fonction du développement de la laitue a été reconstituée à partir de la sortie taux de recouvrement du modèle STICS et de l'équation (5). Des résumés statistiques de la série d'observations obtenue sont présentés dans le tableau 5.C. Le rapport de captation par temps sec varie entre 0 lorsque la laitue vient d'être plantée (les deux premières feuilles du plant de laitue conduisent à un taux de recouvrement quasi nul) et 0,8 lorsqu'elle atteint son recouvrement maximal.

Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
0,00	0,02	0,70	0,48	0,80	0,80

Tableau 5.C : Résumés statistiques relatifs à la variable rapport de captation par temps sec

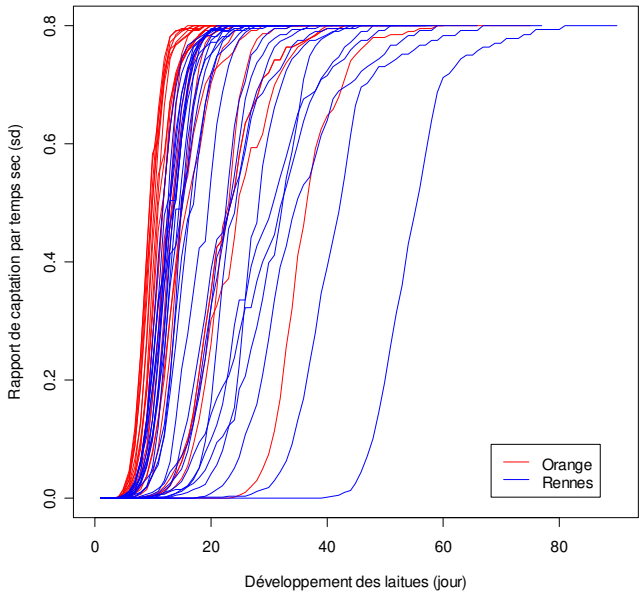


Figure 5.c : Evolution du rapport de captation par temps sec en fonction du développement des laitues

La valeur moyenne (0,48) est cohérente avec celle de 0,5 proposée pour les légumes-feuilles dans le code de calcul FOCON (Rommens et al., 1999) et légèrement supérieure à la valeur de 0,3 proposée dans le code de calcul FARMLAND (Brown et Simmonds, 1995).

La représentation du rapport de captation par temps sec en fonction du développement de la laitue, pour l'ensemble des jeux de données étudiés, permet de mettre en évidence une relation de type sigmoïde entre ces deux variables (Figure 5.c). Les courbes présentent la même allure sigmoïde que les courbes du taux de recouvrement données en fonction du développement des laitues (Cf. Annexe E) mais avec des pentes beaucoup plus importantes. Le rapport de captation par temps sec est maximal et prend la valeur de 0,8 lorsque le taux de recouvrement du sol atteint lui aussi son maximum, qu'il soit égal à 70 % ou à 95 %. Par ailleurs, les courbes sont très similaires pour les deux régions étudiées, le climat ne semble pas avoir une influence sur le rapport de captation par temps sec.

A partir des résultats obtenus, nous allons reconstruire une partie de la variabilité du rapport de captation en fonction du développement de la laitue et ceci pour les différents temps de croissance considérés (de 30 à 90 jours, Cf. 5.1.1.1.1). Ainsi, pour chaque jeu de simulation, une régression non linéaire entre les variables rapport de captation et développement de la laitue (jour) a été effectuée. Les données observées ont alors été approximées par une fonction sigmoïde d'équation :

$$\hat{R}cs = \frac{0{,}8}{1 + e^{-b(D_v + a)}} \qquad (15)$$

où $\hat{R}cs$ représente la valeur estimée du rapport de captation et D_v, le développement de la laitue (en jours). Les paramètres a et b ont été estimés par la fonction *nls* du logiciel statistique R. Notons que le coefficient b contrôle la pente de la courbe tandis que a représente l'abscisse du point d'inflexion (en ce point la tangente « traverse » la courbe).
Les résultats des estimations sont présentés en annexe F et illustrés par la figure 5.d, pour un temps de croissance de 31 jours. Pour cet exemple, les pentes des courbes semblent plus élevées pour la région d'Orange jusqu'à un facteur 1,6 par rapport à celles de la région de Rennes.

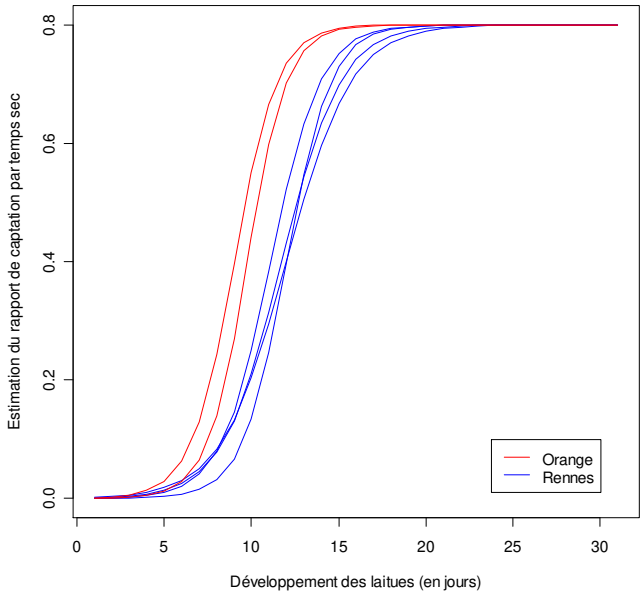


Figure 5.d : Estimation du rapport de captation pour les laitues ayant un temps de croissance de 31 jours

Si l'on peut disposer, pour chaque temps de croissance étudié (30 à 90 jours), de différentes valeurs pour les coefficients a et b, alors une partie de la variabilité du rapport de captation par temps sec peut être reconstruite à l'aide de l'équation (15). Ainsi, les valeurs estimées des coefficients a et b ont été représentées en fonction des temps de croissance obtenus pour les deux régions. Deux relations ont alors été mises en évidence :

- une relation linéaire entre le coefficient a et le temps de croissance de la laitue : au plus le temps de croissance est élevé, au plus le décalage entre l'origine du repère et l'abscisse du point d'inflexion est grand,
- une relation linéaire (par transformation logarithmique des deux variables) entre le coefficient b et le temps de croissance de la laitue : les valeurs estimées du paramètre b ont tendance à diminuer lorsque le temps de croissance augmente.

Les résultats relatifs aux estimations des paramètres de ces deux droites de régression sont présentés en annexe G. L'hypothèse de normalité des résidus étant vérifiée dans les deux cas, les intervalles de prédiction à 95 % ont été déterminés. L'ensemble de ces résultats est présenté graphiquement sur la figure 5.e. Par ces deux relations, des intervalles de valeurs probables des coefficients a et b de l'équation (15) peuvent être déterminés en fonction des différents temps de croissance étudiés (de 30 à 90 jours). Par combinaison de ces valeurs, un encadrement de valeurs probables pour le rapport de captation par temps sec peut alors être proposé pour chaque temps de croissance (Briand et al., 2008). Ainsi, connaissant le temps de croissance du végétal (par exemple 40 jours) et la date (t) de l'accident (par exemple, l'accident a lieu au 20ème jour de croissance du végétal), une valeur de rapport de captation[20] peut être générée selon une loi uniforme dans l'intervalle de valeurs possibles associées (Cf. Figure 5.f).

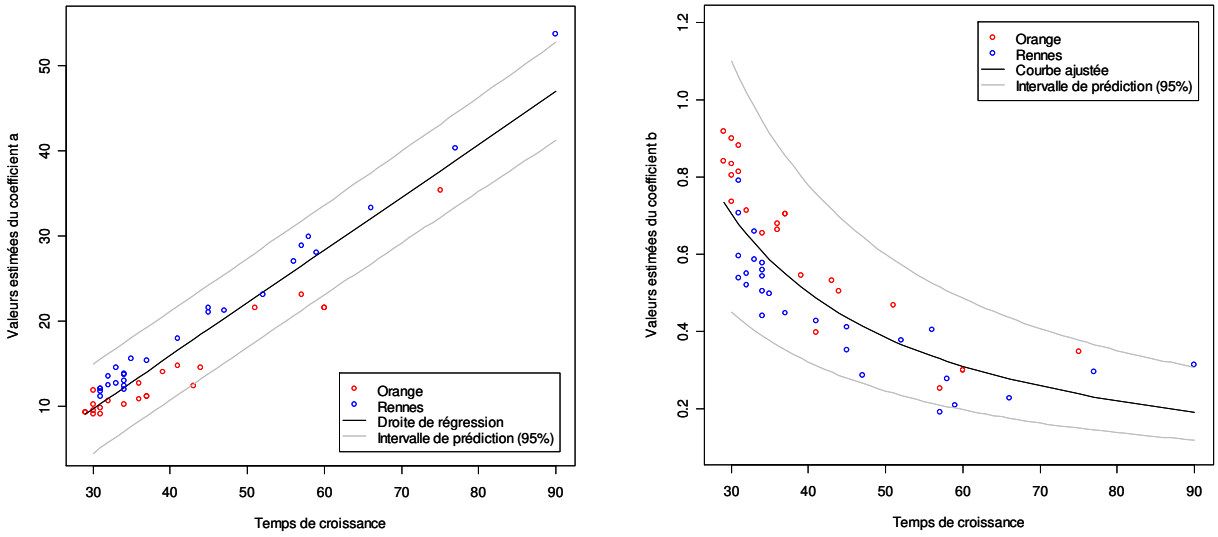


Figure 5.e : Relations entre les coefficients a et b et le temps de croissance des laitues

[20] Dans la suite du document, nous utiliserons la notation R_c pour faire référence à la variable rapport de captation dont les valeurs ont été générées selon la méthode décrite dans ce paragraphe.

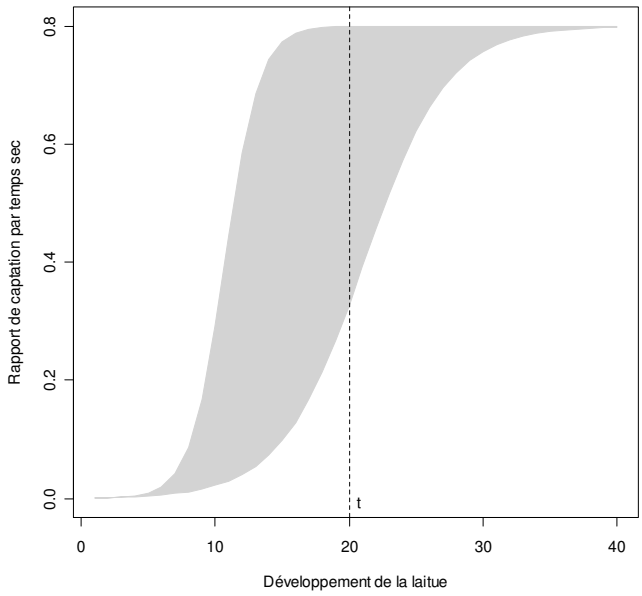


Figure 5.f : Variabilité du rapport de captation par temps sec pour une laitue ayant un temps de croissance de 40 jours

5.1.1.3.2 Les autres légumes-feuilles étudiés

Le modèle de culture STICS est paramétré pour un seul légume-feuille (la laitue). A défaut d'information disponible sur les valeurs de rapport de captation spécifiques pour les légumes-feuilles épinard, poireau et chou, nous utilisons les valeurs de rapport de captation par temps sec obtenus pour le cas de la laitue. En effet, même si l'espèce est différente, ces végétaux présentent des similitudes vis-à-vis des modes de transfert de radioactivité. Ainsi, les résultats présentés en figure 5.e sont utilisés pour renseigner les valeurs de rapport de captation par temps sec associées à ces autres légumes. Lorsque les temps de croissance de ces végétaux sont plus élevés que ceux de la laitue, les valeurs des coefficients a et b sont alors extrapolées à l'aide des relations présentées sur la figure 5.e. C'est le cas du poireau par exemple, qui peut avoir un cycle de croissance de 270 jours.

5.1.1.4 La constante de décroissance biomécanique

Dans l'équation (2) la constante de décroissance biomécanique est associée à l'exponentielle décroissante et modélise principalement la diminution de l'activité durant la croissance du végétal. Cette quantité va donc dépendre du végétal étudié (dans un même intervalle de temps, une salade et un poireau ne vont pas prendre le même poids) et du développement de ce dernier au moment du dépôt. Cependant, l'absence de données et de modèles permettant de prendre en compte ces relations nous conduisent à utiliser une loi triangulaire ayant les caractéristiques suivantes : min=0,03; mode=0,046;

max=0,14 (d'après les travaux effectués par le GRNC (2001) et des résultats d'expérimentations menées par l'IPSN[21] de 1988 à 1993 (Renaud et al., 1999)).

5.1.2 Les variables indépendantes du végétal

5.1.2.1 La date de l'accident

Lors d'un accident, le végétal peut être contaminé à n'importe quel stade de son développement. Ainsi, la date de l'accident est aléatoire sur le temps de croissance du végétal. La date de l'accident (notée t) est donc générée selon une loi uniforme dans l'intervalle $[1, T_c]$ (1 représente le premier jour de la plantation du végétal et T_c le nombre de jour nécessaire pour qu'il soit à maturité).

5.1.2.2 Le dépôt sec

Les valeurs de dépôt vont dépendre du radionucléide étudié. D'après le scénario simplifié de contamination défini au paragraphe 2.3.2, seul le strontium 90 est étudié. Il reste à déterminer l'intervalle de variation du dépôt. Pour déterminer la borne supérieure, nous nous sommes basés sur les valeurs proposées par les critères de zonage définies par les lois biélorusses de 1991 suite à l'accident de Tchernobyl (Bataille et Croüail, 2005). Le niveau de contamination de 111 kBq.m^{-2} est identifié comme la valeur charnière entre une zone de relogement « consécutif »[22] et une zone de relogement obligatoire et immédiat. Ainsi, nous choisissons pour borne supérieure de l'intervalle un dépôt de 100 kBq.m^{-2}. La borne inférieure est choisie de telle sorte qu'en dessous de cette valeur, la concentration dans le végétal ($C_{v,fol,r}$, Cf. équation 2) n'atteint jamais un seuil fixé à 10 Bq.kg^{-1}, jugé négligeable. La détermination de cette borne s'effectue par la résolution de l'équation (2) en affectant aux différentes variables d'entrées (R_c, Rdt, λ_b et Δ) leurs valeurs les plus pénalisantes. Pour chaque légume-feuille étudié, la valeur obtenue se situe autour de 20 Bq.m^{-2}, nous définissons donc ce dépôt comme borne inférieure de l'intervalle. Ainsi, les valeurs de dépôt (en Bq.m²) sont générées dans l'intervalle $[20,\ 100\,000]$ selon une loi uniforme.

5.2 La variable à expliquer

La variable à expliquer (ou variable réponse) est la sortie de l'équation (2), elle représente l'activité massique du ^{90}Sr dans la production végétale. Nous avons choisi de la coder selon

[21] Institut de Protection et de Sûreté Nucléaire devenu aujourd'hui IRSN.
[22] Les personnes vivant dans les zones où les contaminations en ^{90}Sr étaient supérieures à 111 kBq.m^{-2} étaient relogées de façon prioritaire, venaient ensuite les populations vivant dans les zones de relogement consécutif.

les valeurs limites proposées par la Commission du Codex Alimentarius[23] (CAC/GL 5, 1989). Ces valeurs ont été adoptées lors de la 18ème session de la Commission du Codex Alimentarius à Genève en 1989. Les niveaux proposés sont des limites indicatives sur les concentrations des denrées commercialisées (radionucléides dans les aliments), elles sont synthétisées dans le tableau 5.D. Dans le cas du strontium, la valeur limite est de 100 Bq.kg^{-1}, nous proposons donc une discrétisation en deux modalités :

modalité 1 : $C_{v,fol,r} \leq 100$ Bq.kg^{-1}, modalité 2 : $C_{v,fol,r} > 100$ Bq.kg^{-1}.

Radionucléides dans les aliments	Limites en Bq/kg
^{239}Pu, ^{241}Am	10
^{90}Sr	100
^{134}Cs, ^{137}Cs, ^{131}I	1000

Tableau 5.D : Limites indicatives sur les concentrations des denrées commercialisées (CAC/GL 5, 1989)

[23] Codex Alimentarius Commission (CAC).

6 Choix de la taille des échantillons et des variables explicatives pour la construction des arbres de discrimination

6.1 Paramètres relatifs à la construction des arbres de discrimination

En lien avec notre domaine d'application, nous ne voulons pas obtenir des arbres trop détaillés car ils seraient difficilement interprétables et utilisables en contexte post-accidentel. C'est pourquoi nous avons choisi de limiter l'arbre maximal[24] A_{max} à 6 niveaux (le premier niveau représente le nœud racine, le deuxième niveau les deux nœuds enfants issus du nœud racine,…, le dernier niveau correspond aux feuilles de l'arbre). Pour les arbres construits à partir de la méthode REN, la valeur de B (nombre de rééchantillonnages bootstrap) est fixée arbitrairement à 100.

En ce qui concerne l'élagage, les arbres construits selon la méthode CART sont élagués à partir de la technique décrite dans le paragraphe 3.4.3.2. N'étant pas concerné par le problème de la taille de l'échantillon d'apprentissage (les données étant générées artificiellement), un échantillon de validation est systématiquement utilisé pour estimer l'erreur associée aux arbres de la séquence d'élagage. La sélection de l'arbre optimal s'effectue par la règle de l'écart-type (Cf. 3.4.3.3). Les arbres construits selon la méthode REN sont élagués par la méthode de l'erreur réduite (*reduced error pruning*) proposée par Quinlan (1987) (Cf. 4.4.3.2).

6.2 Choix de la taille des échantillons et des variables explicatives

Afin de construire un arbre de discrimination, nous utilisons trois échantillons : un échantillon d'apprentissage pour construire l'arbre maximal, un échantillon de validation pour élaguer cet arbre et un échantillon test pour estimer les performances de prédiction de l'arbre obtenu. Le choix de la taille des échantillons est important car elle doit être suffisamment élevée pour garantir une bonne couverture de l'espace des variables explicatives et une fiabilité des résultats. Nous utilisons la méthode CART comme méthode de référence ainsi que les données relatives au scénario de contamination (laitue, ^{90}Sr).

6.2.1 La taille des échantillons d'apprentissage et de validation

Les arbres de décision sont connus pour leur instabilité, il est donc assez difficile de comparer des arbres construits à partir d'échantillons de différentes tailles. En effet, le nombre de feuilles et les prédictions vont être variables d'un arbre à l'autre, même pour

[24] Construit par la méthode CART où la méthode REN.

deux arbres construits à partir de deux échantillons de même taille. Ghattas (2000) l'illustre empiriquement dans ces travaux : à partir de 10 échantillons bootstrap d'un échantillon de base ($N = 822$), il construit 10 arbres de discrimination par validation croisée. Il obtient alors des arbres qui sont différents à partir du nœud racine et de taille très variable. On peut cependant penser qu'à partir d'une certaine taille d'échantillon, les premières divisions seront basées sur les mêmes variables et les arbres se distingueront uniquement par leurs niveaux inférieurs. Afin d'étudier l'influence de la taille des échantillons (d'apprentissage et de validation) sur la structure des arbres de discrimination, nous utilisons le concept de l'importance des variables explicatives défini par Breiman et al. (1984) et étudié par Ghattas (1999).

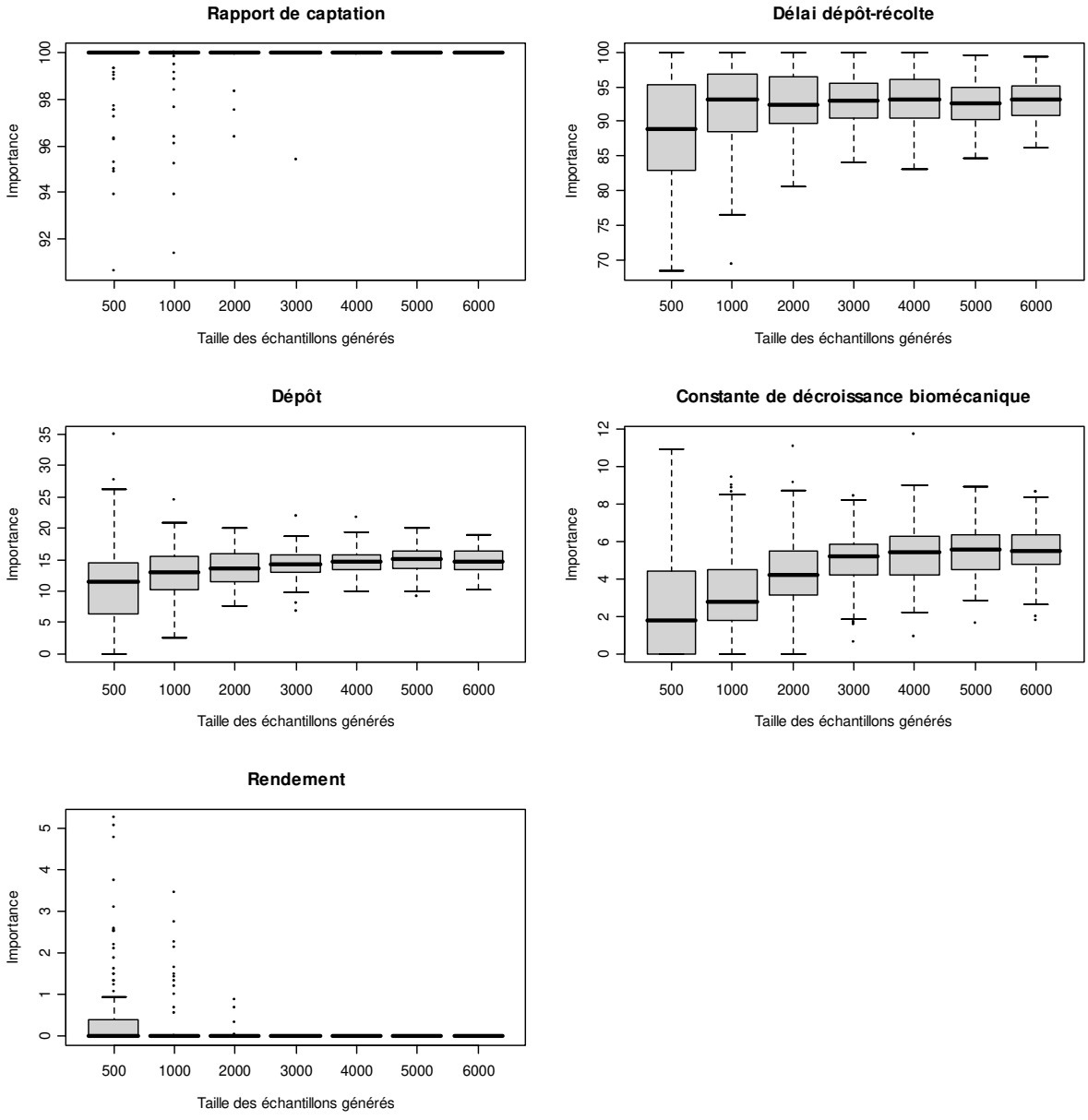


Figure 6.a: Importance des 5 variables explicatives en fonction de la taille des échantillons

L'importance des variables permet de quantifier et de hiérarchiser le pouvoir explicatif de chaque variable (Cf. 3.4.4). L'idée développée ici est d'étudier le comportement du concept d'importance des variables pour des arbres construits à partir d'échantillons de différentes tailles. L'objectif est de mettre en évidence, à partir d'une certaine taille d'échantillon, des comportements communs dans la hiérarchie des variables explicatives et donc une stabilité globale dans l'importance des variables. Pour que les résultats obtenus ne soient pas dépendants des échantillons utilisés, 100 couples d'échantillons sont générés (composés d'un échantillon d'apprentissage et d'un échantillon de validation) de tailles respectives 500, 1000, 2000, 3000, 4000, 5000 et 6000 observations. L'importance de chaque variable explicative est alors déterminée pour les différents arbres de discrimination construits à partir des échantillons simulés. Les résultats sont présentés sur la figure 6.a. Pour chacune des variables, l'importance est calculée en fonction de la taille des échantillons générés et est représentée sous forme de box-plot. Lorsque la taille des échantillons atteint 4000 observations, la hiérarchie des variables explicatives est identique, quel que soit l'arbre de discrimination : la variable rapport de captation est la plus importante (importance =100) suivi du délai entre le dépôt et la récolte, du dépôt de radioactivité, de la décroissance biomécanique et du rendement qui a une importance nulle. A la vue de ces résultats et afin d'obtenir une bonne couverture de l'espace des variables explicatives, nous choisissons, dans la suite du travail, de générer des échantillons d'apprentissage et de validation de taille 5000.

6.2.2 Le choix des variables explicatives pour la construction des arbres

La figure 6.b représente l'importance des 5 variables explicatives pour des arbres de discrimination construits à partir d'échantillons de taille 5000. Cette représentation permet de mettre en évidence deux groupes de variables : le premier groupe composé des variables ayant un fort pouvoir discriminant (rapport de captation et délai dépôt-récolte) et un deuxième groupe composé des variables moins importantes dans le processus de discrimination (dépôt, constante de décroissance biomécanique et rendement cultural). Les deux dernières variables ont une importance négligeable. De plus, elles sont peu utilisables en contexte post-accidentel car difficilement mesurables dans l'environnement (en particulier la constante de décroissance biomécanique). Nous décidons donc de ne pas les faire participer à la construction des arbres de discrimination. Notons que le dépôt paraît lui aussi peu important mais nous choisissons tout de même de le conserver car cette variable est utile pour classer les accidents en fonction de l'ampleur des rejets dans l'environnement.

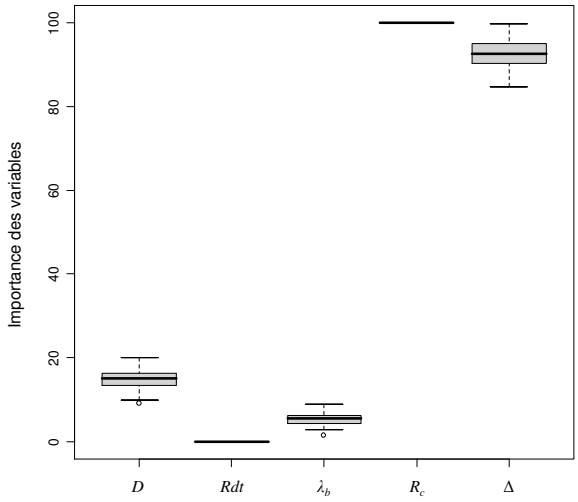


Figure 6.b : Importance des 5 variables explicatives (échantillons de taille 5000)
D : dépôt de radioactivité, Rdt : rendement cultural, λ_b : constante de décroissance biomécanique, R_c : rapport de captation et Δ : délai dépôt-récolte.

Pour juger de la qualité du modèle retenu et s'assurer que la suppression des variables λ_b et Rdt ne conduit pas à une augmentation considérable du taux de mauvais classement, des comparaisons sur la base de différents échantillons sont effectuées. Dix couples d'échantillons sont générés (apprentissage, validation) et 10 arbres de discrimination sont construits en utilisant les 5 variables explicatives décrites précédemment. Puis, 10 nouveaux arbres sont construits à partir des mêmes échantillons mais en considérant cette fois-ci uniquement les variables les plus discriminantes (rapport de captation, délai dépôt-récolte et dépôt). Les comparaisons portent sur la complexité des arbres et sur le taux de mauvais classement estimé par un échantillon test (de taille 5000). Les résultats sont présentés dans le tableau 6.A. Les arbres construits à partir des échantillons à 5 variables présentent des pourcentages de mauvais classement assez faibles, de l'ordre de 4,2 %. Si l'on s'intéresse aux arbres construits à partir des mêmes échantillons mais restreints aux 3 variables les plus discriminantes, les arbres obtenus sont généralement moins complexes, ce qui présente un avantage par rapport aux précédents car ils seront plus facilement interprétables et utilisables dans un contexte post-accidentel. Ne pas faire participer les variables λ_b et Rdt à la construction des arbres de discrimination, conduit cependant à une légère augmentation du taux de mauvais classement (de l'ordre de 1 % dans le cas de cette analyse).

Arbres 3 variables	
Nombre de feuilles	% mauvais classement
8	5,56
10	5,52
8	5,52
13	5,04
9	5,42
11	5,42
13	5,32
10	5,72
8	5,44
7	5,52

Arbres 5 variables	
Nombre de feuilles	% mauvais classement
15	4,04
13	4,22
12	4,5
12	4,36
11	4,6
13	4,28
14	4,14
13	4,52
14	3,94
15	4,06

Tableau 6.A : Comparaison des arbres construits à partir d'échantillons à 3 ou à 5 variables explicatives

6.2.3 La taille de l'échantillon test

Dans les paragraphes précédents, nous avons défini la taille des échantillons d'apprentissage et de validation ainsi que les variables explicatives qui seront utilisés dans la suite de ce travail pour construire les arbres de discrimination. Un autre point important qui va être traité dans ce paragraphe est le choix de la taille de l'échantillon test. Cet échantillon qui ne participe pas à la construction de l'arbre permet d'estimer les performances de prédiction de cet arbre. Nous allons donc tester différentes tailles d'échantillons test et retenir la taille pour laquelle les résultats sont les plus fidèles à la réalité observée (dans notre cas ce sont les valeurs issues de la simulation). Pour que les résultats ne soient pas dépendants de l'arbre choisi, quatre arbres de discrimination sont construits (à partir d'échantillons d'apprentissage et de validation de taille 5000 et des trois variables explicatives les plus importantes). De plus, comme dans le paragraphe précédent, pour ne pas avoir de dépendance avec l'échantillon test utilisé, 100 échantillons test sont générés de tailles respectives 500, 1000,…, 6000 observations. Les résultats sont présentés sur la figure 6.c. Pour chaque arbre de discrimination, le pourcentage de mauvais classement est calculé pour les échantillons test de différentes tailles et est représenté sous forme de box-plot. Les résultats obtenus sont assez semblables aux précédents, l'estimation du taux de mauvais classement est assez variable lorsque les échantillons test sont de petites tailles (500 et 1000 observations). A partir de 4000 observations, quel que soit l'arbre de discrimination, les taux de mauvais classement se stabilisent. Ainsi, nous proposons d'utiliser des échantillons test de taille 5000 afin d'estimer fidèlement les performances de prédiction des différents arbres de discrimination qui seront construits dans la suite de ce travail.

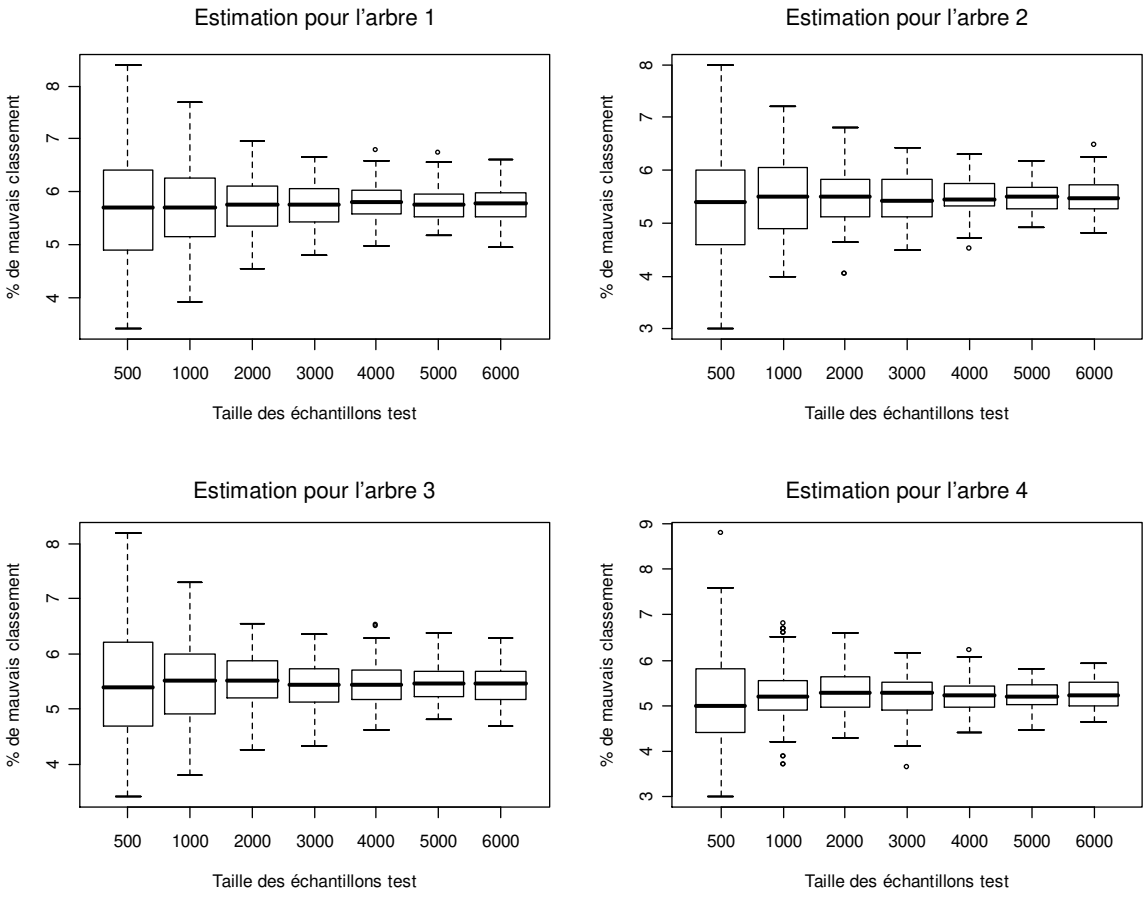


Figure 6.c : Estimation du pourcentage de mauvais classement par les 4 arbres de discrimination en fonction de la taille des échantillons test

Ces choix ont été effectués en utilisant la méthode CART comme méthode de référence et à partir de jeux de données simulés relatifs au cas de contamination (laitue, ^{90}Sr). Pour les arbres de discrimination construits selon la méthode REN, nous utiliserons les mêmes tailles d'échantillons, le fait d'avoir un échantillon de validation de même taille que l'échantillon d'apprentissage permettra d'éviter les problèmes de sur-élagage discuté en 4.4.3.2. De plus, compte tenu des similitudes vis-à-vis des modes de transfert de radioactivité des autres légumes-feuilles étudiés, nous utiliseront les mêmes critères pour construire les arbres associés au cas de contamination (épinard, ^{90}Sr), (poireau, ^{90}Sr) et (chou, ^{90}Sr).

7 Comparaisons empiriques de deux méthodes de construction d'arbre de discrimination : la méthode CART et la méthode REN

Dans ce chapitre, nous allons effectuer de nombreuses comparaisons empiriques entre la méthode CART et la méthode REN. L'objectif est de mettre en évidence la stabilité des arbres de discrimination construits selon la méthode REN. Dans un premier temps, de nombreux arbres de discrimination sont construits par les deux méthodes. Les performances et la structure de ces arbres sont comparées. Nous utilisons la mesure de similarité définie au paragraphe 4.4.3.3. Dans un deuxième temps, nous comparons la sensibilité des deux méthodes (CART et REN) à de légères modifications de l'échantillon d'apprentissage. L'ensemble de ces analyses est effectué à partir du scénario de contamination (laitue, ^{90}Sr).

7.1 Comparaisons de la méthode CART et de la méthode REN

Afin de comparer les arbres obtenus par les deux méthodes, nous générons $k = 30$ couples d'échantillons (apprentissage, validation) selon les critères définis au chapitre 6. Chaque couple est appelé échantillon d'entraînement et il sert à construire deux arbres de discrimination : le premier par la méthode CART et le deuxième par la méthode REN. Au total, nous disposons donc de 60 arbres de discrimination.

7.1.1 Comparaison des performances

Pour comparer les performances des deux méthodes, nous générons arbitrairement 50 échantillons test. Chaque échantillon d'entraînement est utilisé pour construire deux arbres de discrimination et pour chacun de ces arbres, les taux de mauvais classement sont estimés sur les 50 échantillons test. Au total nous effectuons donc 3000 estimations (Cf. figure 7.a).

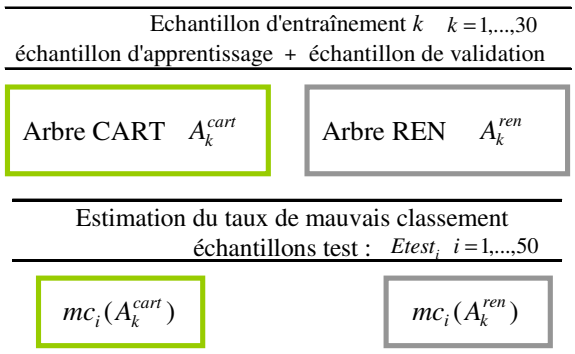


Figure 7.a : Notations utilisées pour faire référence aux arbres de discrimination construits par les méthodes CART et REN

Pour chaque paire d'arbres obtenue (A_k^{cart}, A_k^{ren}), k = 1,…,30, nous calculons le nombre de fois où $mc_i(A_k^{cart}) > mc_i(A_k^{ren})$, i = 1,…,50. Si cette valeur est nulle cela signifie que les taux de mauvais classement calculés à partir du modèle CART sont toujours inférieurs à ceux calculés par le modèle REN (pour les 50 échantillons test). Lorsque la valeur est maximale (égale à 50), la méthode REN est toujours plus performante que la méthode CART. Les résultats de ces comparaisons sont présentés sur la figure 7.b. Pour chaque valeur de k, nous avons représenté, en ordonnée, le nombre de fois où $mc_i(A_k^{cart}) > mc_i(A_k^{ren})$, i = 1,…,50. D'une manière générale, la méthode REN paraît plus performante que la méthode CART. Pour 21 couples d'arbres de discrimination (chaque couple correspond à une valeur de k), les prédictions par la méthode CART sont les plus souvent de moins bonne qualité. En particulier, pour les couples 11 et 30, l'erreur de classement calculée par les arbres de discrimination construits selon la méthode REN est toujours inférieure à celle calculée à partir des modèles CART quel que soit l'échantillon test utilisé. Par contre, pour les couples 1, 2, 3, 4, 5, 17, 22, 28 et 29, la méthode CART donne de meilleurs résultats. Pour analyser plus précisément les performances de ces méthodes et avoir une idée des valeurs associées aux taux de mauvais classement, certains cas particuliers sont étudiés. Nous comparons, pour $k = 11$ et $k = 30$ (cas où la méthode REN est toujours la plus performante), les taux de mauvais classement estimés par ces deux méthodes sur les 50 échantillons test. Les résultats sont présentés sous forme de box-plot en figure 7.c.

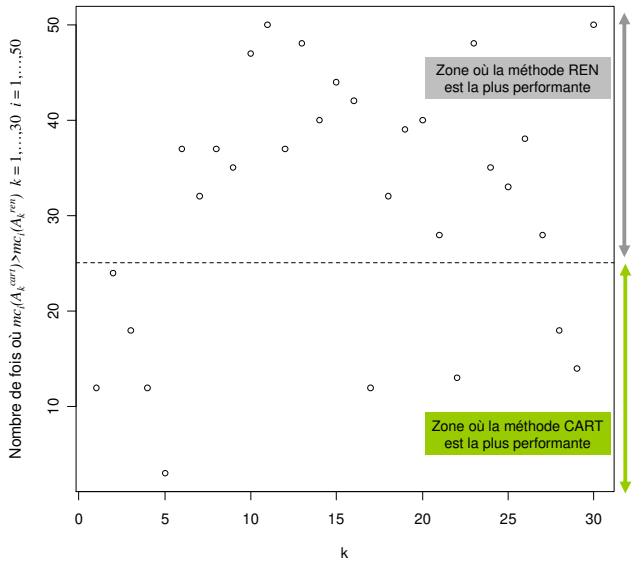


Figure 7.b : Performances comparées de la méthode CART et de la méthode REN pour les 30 couples d'arbres de discrimination construits

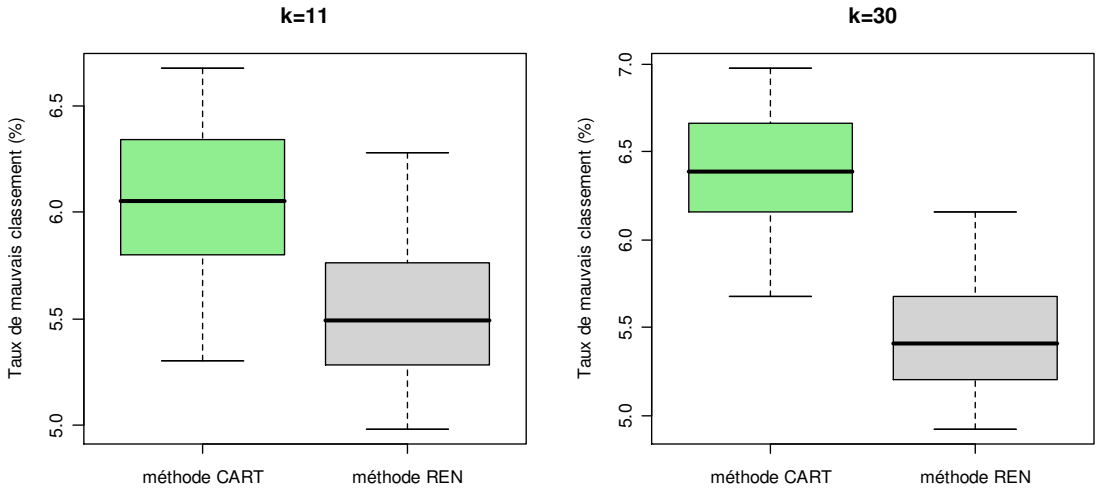


Figure 7.c : Comparaison des taux de mauvais classement (%) estimés sur les 50 échantillons test par les deux couples d'arbres de discrimination : $(A_{11}^{cart}, A_{11}^{ren})$ et $(A_{30}^{cart}, A_{30}^{ren})$

Même si les deux arbres construits selon la méthode REN sont plus performants, la différence entre les taux de mauvais classement est assez faible. Pour les arbres construits selon la méthode CART, la médiane est de 6,05 % pour $k = 11$ et 6,39 % lorsque $k = 30$. Les médianes sont plus stables pour les arbres construits selon la méthode REN, de l'ordre de 5,5 %.

De la même manière, nous nous intéressons aux valeurs de k pour lesquelles la méthode CART a tendance à donner de meilleurs résultats que la méthode REN. Pour ces 9 couples d'arbres de discrimination, les représentations graphiques peuvent être visualisées sur la figure 7.d. Pour ces cas, les écarts observés sont encore plus faibles que précédemment, les médianes des estimations du taux de mauvais classement se situant toujours autour de la valeur de 5,5 % pour la méthode REN. Finalement, l'ensemble de ces comparaisons empiriques entre les deux méthodes de construction d'arbre de discrimination montre que les performances de ces deux méthodes restent très proches.

Nous avons également comparé ces erreurs de classement avec celles obtenues par les méthodes présentées dans le paragraphe 4.3 du chapitre 4 : random forests et les principes du bagging et du boosting appliqués à la méthode CART. Plutôt que de construire 30 modèles à partir de chaque méthode, nous avons choisi de restreindre la comparaison à deux cas particuliers. Le premier ($k = 5$) correspond au cas où la méthode CART semble être la plus performante et le deuxième ($k = 11$) celui où l'erreur de classement calculée par la méthode REN est toujours inférieure à celle de la méthode CART.

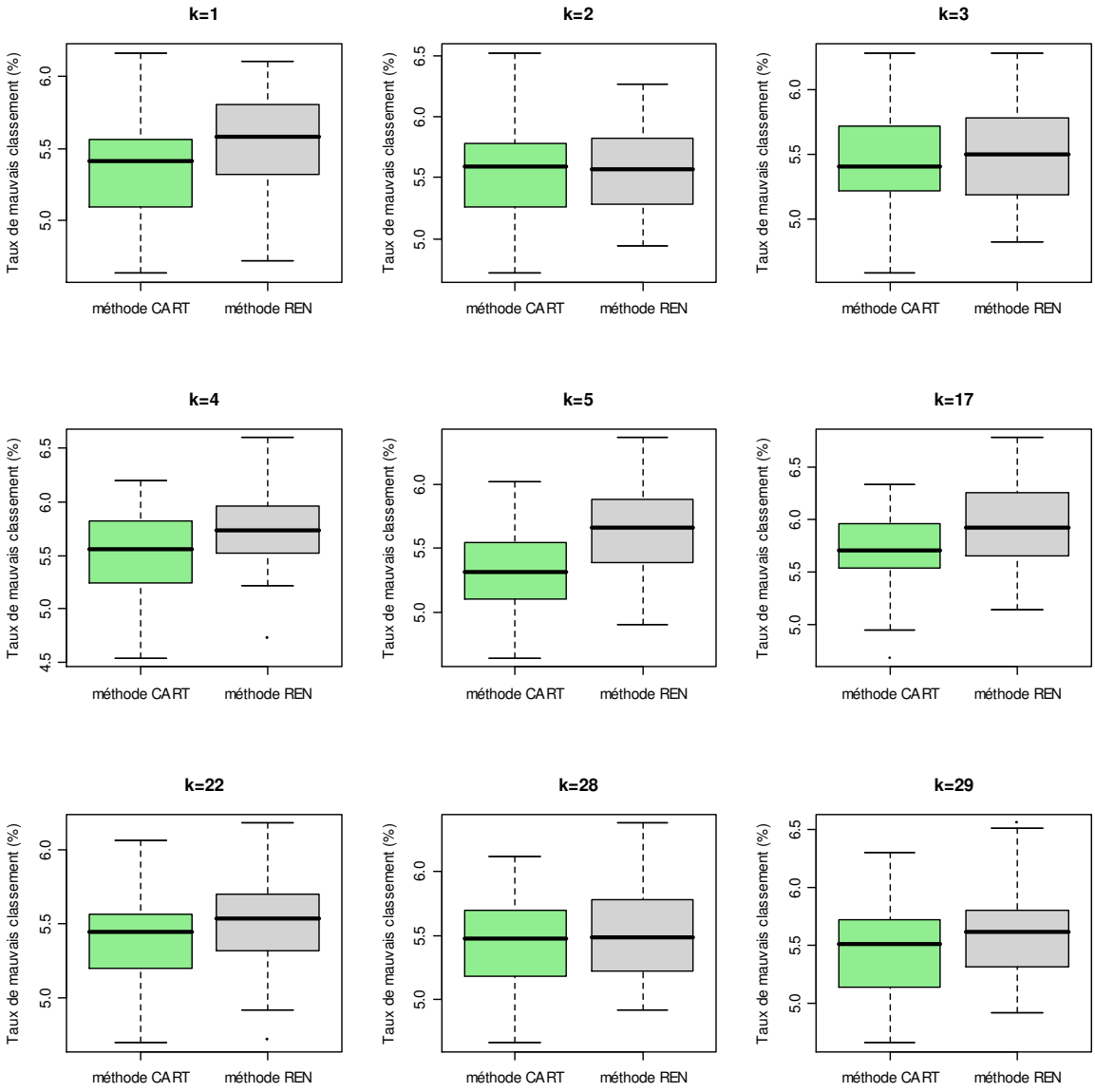


Figure 7.d : Comparaison des taux de mauvais classement (%) estimés sur les 50 échantillons test par les 9 couples d'arbres de discrimination

Pour chacune des trois méthodes (random forests, bagging et boosting (algorithme Adaboost)), des expérimentions préalables ont permis de fixer les paramètres suivants :

- bagging, random forests : le nombre d'échantillons bootstrap est fixé à $B = 100$,
- boosting : le nombre d'itérations est fixé à $T = 100$.

Pour la méthode random forests, Breiman (2002) conseille de choisir, pour valeur de q (Cf. 4.3.2), la racine carrée du nombre de variables explicatives. Dans notre cas $\sqrt{3} = 1{,}73$, nous proposons donc de construire deux modèles : le premier avec $q = 1$ (noté RF-1) et le deuxième avec $q = 2$, noté RF-2. Une fois l'ensemble des modèles construits, les prédictions sont effectuées sur les 50 échantillons test. Les résultats obtenus (taux de mauvais classement) sont présentés sous forme de box-plot sur la figure 7.e pour chacune

des valeurs de k. Comme l'on pouvait s'y attendre, lorsque l'une de ces trois méthodes est utilisée, les qualités de prédictions sont meilleures. Pour la méthode de boosting et la méthode random forests, les estimations du taux de mauvais classement sont comparables. De plus, pour la méthode random forests, il n'y a pas de différence entre choisir aléatoirement une variable explicative pour segmenter chaque nœud ($q = 1$) où choisir parmi 2 tirées aléatoirement. Les taux de mauvais classement estimés par le principe du bagging sont légèrement plus élevés, que ceux estimés par les deux autres méthodes, mais restent tout de même dans un intervalle de valeurs très étroites.

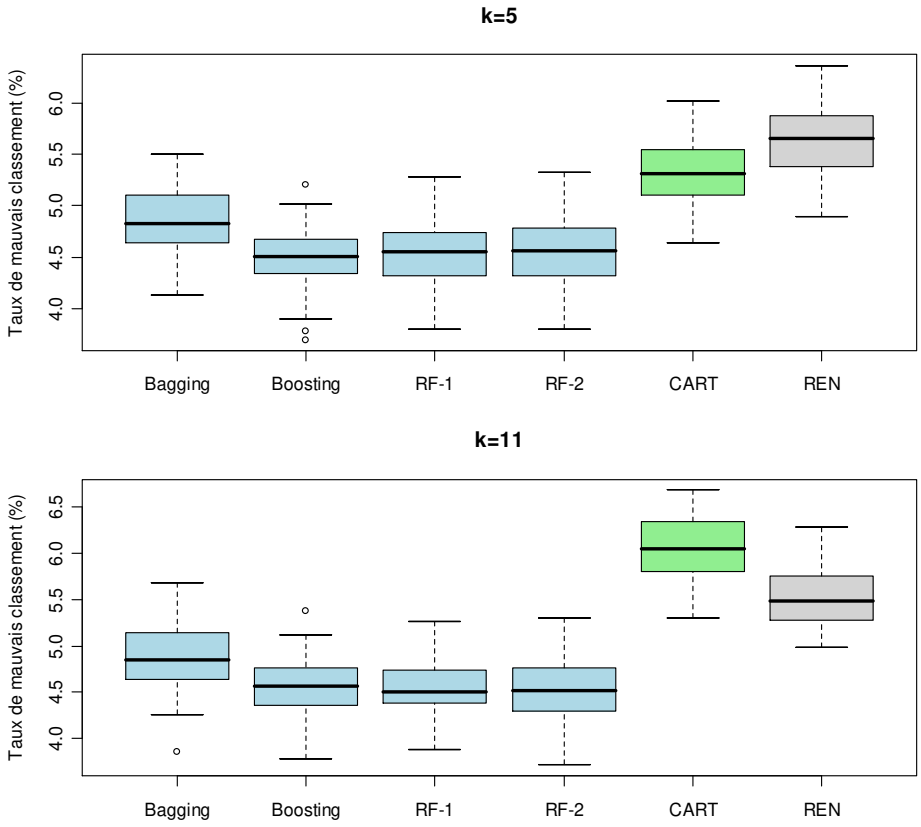


Figure 7.e : Comparaison des taux de mauvais classement (%) estimés sur les 50 échantillons test par les 5 méthodes, pour les deux cas particuliers $k = 5$ et $k = 11$

7.1.2 Comparaison de la structure des arbres de discrimination

7.1.2.1 Le nombre de feuilles et les divisions

Dans un premier temps, la taille des arbres de discrimination est examinée. La figure 7.f présente une comparaison de la complexité des arbres : les 30 points de coordonnées $\left(\left|\tilde{A}_k^{cart}\right|,\left|\tilde{A}_k^{ren}\right|\right)$ sont présentés sur le graphe de gauche tandis que le graphe de droite établit une comparaison, sous forme de box-plot, du nombre de feuilles des arbres de discrimination construits par la méthode CART et la méthode REN.

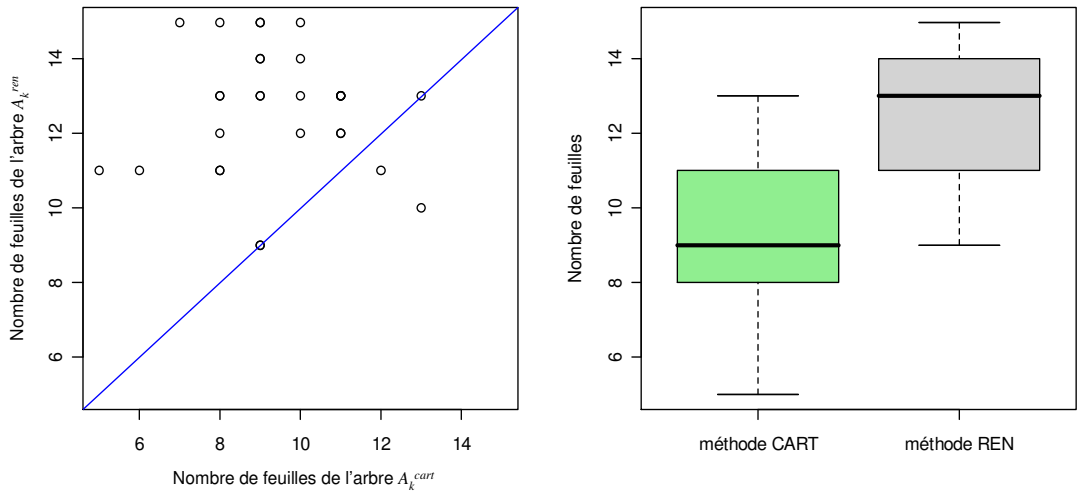


Figure 7.f : Comparaison du nombre de feuilles des arbres de discrimination obtenus par la méthode CART et la méthode REN

Il semble que les arbres de discrimination construits selon la méthode CART aient une taille plus variable que celle des arbres obtenus par la méthode REN. Néanmoins, la méthode CART fournit des arbres de discrimination de complexité plus faible, la médiane se situant à 9 feuilles. Les arbres construits selon la méthode REN ont tendance à avoir un plus grand nombre de feuilles, la valeur médiane étant de 13 nœuds terminaux.

Dans un deuxième temps, nous étudions la stabilité des nœuds intermédiaires (choix de la variable de segmentation et division associée) des différents arbres de discrimination. Le tableau 7.A constitue une synthèse des variables[25] utilisées pour diviser les 15 premiers nœuds des arbres de discrimination construits par les méthodes CART et REN. Chaque nœud est identifié par un numéro (par ordre croissant du nœud racine et de la gauche vers la droite) et sa position dans l'arbre. Pour le nœud racine, que ce soit pour les arbres construits selon la méthode CART ou la méthode REN, le rapport de captation est la variable qui est toujours utilisée pour effectuer la segmentation. Pour le troisième nœud (enfant droit issu du nœud racine), lorsque la méthode CART est utilisée, les variables de division ne sont pas toujours les mêmes : l'une est basée sur le rapport de captation, treize sont basées sur le dépôt et seize sur la variable délai dépôt-récolte. Tandis que pour la méthode REN, les divisions sont toujours effectuées à partir du délai dépôt-récolte. En analysant les variables utilisées pour effectuer la segmentation des nœuds descendants, la stabilité de la méthode REN, dans le choix de la variable de division, apparaît clairement.

[25] Nous rappelons que les variables explicatives sont au nombre de trois : dépôt de radioactivité (D), rapport de captation (R_c) et délai dépôt-récolte (Δ).

Méthode CART				Nœud	Méthode REN			
D	R_c	Δ	*Feuille*		D	R_c	Δ	*Feuille*
0	30	0	0	*1 - racine*	0	30	0	0
0	16	0	14	*2 - g*	0	30	0	0
13	1	16	0	*3 - d*	0	0	30	0
0	0	0	30	*4 - gg*	0	0	0	30
10	0	6	14	*5 - gd*	30	0	0	0
17	0	13	0	*6 - dg*	30	0	0	0
17	0	4	9	*7 - dd*	30	0	0	0
0	0	0	30	*8 - ggg*	0	0	0	30
0	0	0	30	*9 - ggd*	0	0	0	30
5	0	0	25	*10 - gdg*	0	0	0	30
0	0	10	20	*11 - gdd*	0	0	30	0
24	0	6	0	*12 - dgg*	30	0	0	0
3	0	1	26	*13 - dgd*	0	0	0	30
4	1	1	24	*14 - ddg*	13	0	0	17
3	2	9	16	*15 - ddd*	0	0	17	13

Tableau 7.A : Synthèse des variables utilisées pour segmenter les 15 premiers nœuds des arbres de discrimination (D : dépôt de radioactivité, R_c : rapport de captation et Δ : délai dépôt-récolte)

Intéressons nous maintenant aux valeurs de division associées à ces variables. Le nœud racine et le nœud n°2 (enfant gauche issu de la racine) étant toujours segmentés par la variable rapport de captation, nous avons choisi d'étudier ces deux cas plus particulièrement. La figure 7.g présente deux graphes (un pour chaque nœud) qui permettent d'effectuer la comparaison entre les valeurs de division proposées par la méthode CART et celles proposées par la méthode REN. Comme précédemment, ces résultats permettent d'illustrer la stabilité de la méthode REN. Lorsque les arbres de discrimination sont construits par la méthode CART, les divisions associées au nœud racine varient de 0,103 à 0,229. Avec la méthode REN, l'écart entre les deux divisions extrêmes est considérablement réduit : la division minimale est de 0,139 et la division maximale de 0,155.

Les autres nœuds pourraient être comparés de la même manière, mais cela devient plus difficile car les divisions ne sont pas toujours basées sur les mêmes variables (en particulier pour la méthode CART). Il serait cependant intéressant de pouvoir comparer globalement deux arbres de discrimination. Cela pourrait renseigner sur la proximité de leur structure et donc sur la stabilité de leurs règles de décision. Ces constats nous ont amené à définir une mesure (Cf. 4.4.3.3) permettant de quantifier la similarité de deux arbres de discrimination. L'application de cette mesure est présentée dans le paragraphe suivant.

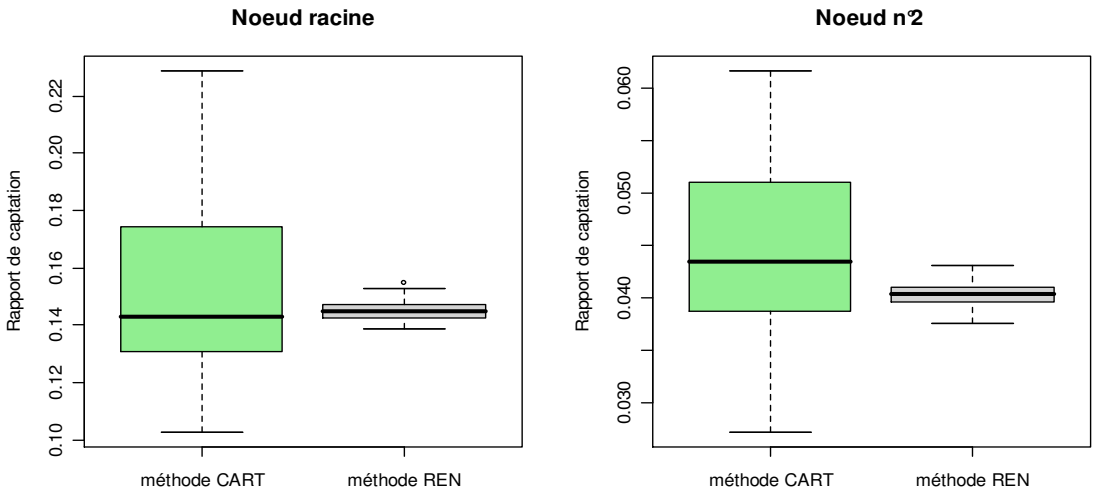


Figure 7.g : Comparaison des valeurs de division basées sur le rapport de captation pour les deux premiers nœuds des arbres de discrimination

7.1.2.2 Utilisation de la mesure de similarité

Dans ce paragraphe, nous utilisons la mesure de similarité afin d'étudier les performances (en terme de stabilisation) de la méthode REN et de la méthode CART. Ces comparaisons sont effectuées à partir de deux types de pondérations :

(a) $q_t = 1/(T+1)$, $t = 1,\ldots,T$ (poids constants),

(b) q_t proportionnel à $1/2^{(L_t-1)}$ (sommant à 1) où L_t représente le niveau de l'arbre (1 pour le nœud racine, 2 pour le deuxième niveau,…).

Dans un premier temps, la mesure de similarité (13) a été calculée, avec les deux types de pondérations (a) et (b), pour les 30 couples d'arbres de discrimination. Les résultats sont présentés graphiquement sur la figure 7.h. On retrouve en abscisse, pour chaque couple (A_k^{cart}, A_k^{ren}) $k = 1,\ldots,30$, les valeurs de la mesure en utilisant la pondération (a) et en ordonnée les mêmes calculs en utilisant la pondération (b). Chaque point est identifié par sa valeur de k. La linéarité (dans la représentation des données) montre que l'impact du choix de la pondération n'est pas très grand. Lorsque les poids (b) sont utilisés, la valeur de la mesure est toujours plus faible quel que soit le couple d'arbres de discrimination. Cette pondération accorde beaucoup plus d'importance à la partie haute de l'arbre. Comme nous l'avons illustré précédemment par le tableau 7.A, les arbres de discrimination construits par la méthode CART et REN utilisent la même variable de segmentation pour diviser les deux premiers nœuds intermédiaires (racine et son nœud descendant gauche). Ainsi, le fait d'accorder beaucoup plus de poids aux premiers niveaux conduit à des arbres de discrimination plus proches au sens de la mesure de similarité.

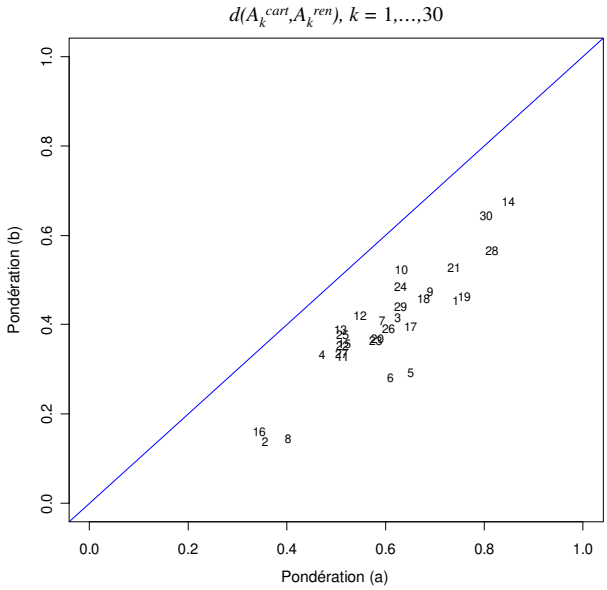


Figure 7.h : Mesure de similarité pour les 30 couples d'arbres de discrimination en utilisant les deux types de pondérations (a) et (b)

Pour l'ensemble des couples comparés, la mesure de similarité varie de 0,345 à 0,848 lorsque la pondération (a) est utilisée et de 0,141 à 0,678 lorsque la pondération (b) est utilisée. Pour illustrer les valeurs obtenues par notre mesure, nous avons choisi de représenter dans le tableau 7.B les valeurs de division associées aux nœuds des deux couples d'arbres de discrimination (A_2^{cart}, A_2^{ren}) et $(A_{14}^{cart}, A_{14}^{ren})$ qui ont, respectivement, la mesure la plus faible et la plus élevée. Seuls les 5 premiers niveaux des 4 arbres ont été représentés (ils sont distingués dans le tableau par des droites horizontales en pointillés) car le sixième niveau ne fait intervenir que les feuilles de certains arbres. Les valeurs entre parenthèses permettent d'identifier la variable utilisée pour effectuer la segmentation (Δ : délai, D : dépôt et R_c : rapport de captation).

Pour le couple (A_2^{cart}, A_2^{ren}), l'architecture des trois premiers niveaux des deux arbres de discrimination est identique (position des nœuds intermédiaires) et les variables de segmentation sont toujours les mêmes. Les divisions associées sont assez semblables dans les deux premiers niveaux et plus variables dans le troisième. Les autres niveaux se distinguent par des variables de segmentation différentes et de nouvelles branches pour l'arbre A_2^{ren}. Les arbres relatifs au couple $(A_{14}^{cart}, A_{14}^{ren})$ se distinguent dès leur deuxième niveau.

	k=2		k=14	
	CART	REN	CART	REN
1 rac	(R_c) 0,17	(R_c) 0,14	(R_c) 0,14	(R_c) 0,15
2 g	(R_c) 0,04	(R_c) 0,04	f	(R_c) 0,04
3 d	(Δ) 28,50	(Δ) 27,50	(D) 4593,15	(Δ) 27,50
4 gg	f	f	-	f
5 gd	(D) 50234,00	(D) 36604,66	-	(D) 38419,11
6 dg	(D) 4807,90	(D) 7066,01	(Δ) 18,50	(D) 6788,71
7 dd	(D) 16207,90	(D) 20704,36	(Δ) 29,50	(D) 21025,09
8 ggg	-	-	-	-
9 ggd	-	-	-	-
10 gdg	f	f	-	f
11 gdd	(Δ) 38,50	(Δ) 37,50	-	(Δ) 36,50
12 dgg	(Δ) 15,50	(D) 1697,91	(D) 2025,35	(D) 1601,21
13 dgd	f	f	f	f
14 ddg	f	(D) 7835,85	f	(D) 8025,56
15 ddd	f	f	(D) 26890,40	f
16 gggg	-	-	-	-
⋮	⋮		⋮	
22 gddg	(R_c) 0,08	(R_c) 0,08	-	(R_c) 0,08
23 gddd	f	f	-	f
24 dggg	f	f	f	f
25 dggd	f	(Δ) 16,25	f	f
26 dgdg	-	-	-	-
27 dgdd	-	-	-	-
28 ddgg	-	f	-	f
29 ddgd	-	(Δ) 33,50	-	(R_c) 0,34
30 dddg	-	-	(R_c) 0,41	-
31 dddd	-	-	f	-

Tableau 7.B : Comparaison des divisions associées aux nœuds intermédiaires relatifs aux couples d'arbres de discrimination (A_2^{cart}, A_2^{ren}) et $(A_{14}^{cart}, A_{14}^{ren})$

Pour l'arbre A_{14}^{cart} le nœud n°2 est une feuille tandis que pour A_{14}^{ren} c'est la racine d'une branche. Les branches opposées (issues du nœud n°3) se distinguent par leurs variables de segmentation et leurs divisions associées. Ainsi, ces illustrations montrent que notre mesure de similarité capture bien les différences d'architecture.

En fonction de la valeur de k, il y donc une certaine variabilité entre les arbres construits par la méthode CART et la méthode REN. Quelquefois, les arbres de discrimination sont assez proches et dans d'autres cas, leur structure diffère. Pour identifier la cause de cette variabilité et, en particulier, montrer que les arbres de discrimination construits par la méthode REN sont plus stables, de nouvelles comparaisons sont effectuées. La mesure de similarité est appliquée à tous les arbres de discrimination construits d'une part par la méthode CART et d'autre part par la méthode REN.

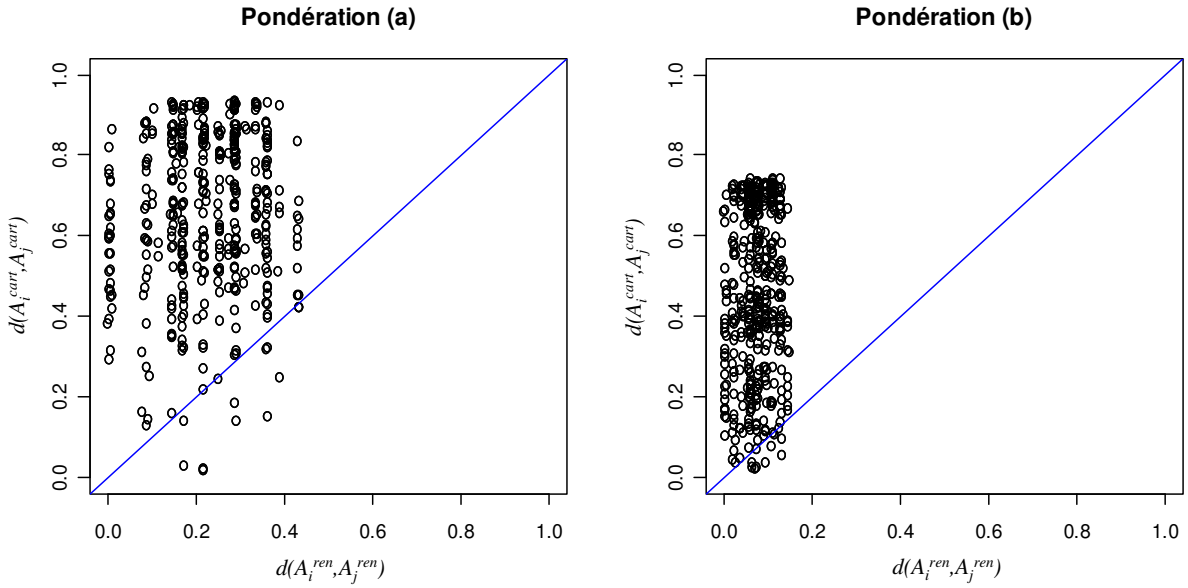


Figure 7.i : Comparaison des valeurs de la mesure de similarité pour tous les couples d'arbres de discrimination construits par les deux méthodes

Au total, $(30\times29)/2=435$ comparaisons des arbres deux à deux sont effectuées pour chaque méthode. Les résultats sont présentés sur la figure 7.i en fonction de la pondération utilisée. Les 435 points présentés sur ces graphes ont pour coordonnées $\left(d(A_i^{ren},A_j^{ren}),d(A_i^{cart},A_j^{cart})\right)$. Les deux graphes ont, de manière générale, la même structure. Elle est cependant beaucoup plus étalée lorsque la pondération (a) est utilisée.

Ces représentations permettent d'illustrer la stabilité des arbres de discrimination construits selon la méthode REN. Lorsque la pondération (b) est utilisée la variabilité de la mesure de similarité, calculée pour les arbres de discrimination construits par la méthode REN, est considérablement réduite. Le tableau 7.C propose une synthèse des valeurs minimales et maximales obtenues en fonction de la pondération et de la méthode utilisées. D'après le tableau 7.A, les arbres de discrimination construits par la méthode REN se distinguent à partir des nœuds intermédiaires du niveau 4 (nœuds n°14 et 15).

	Mesure de similarité			
	Méthode CART		Méthode REN	
	min	*max*	*min*	*max*
Pondération (a)	0,019	0,935	0,002	0,433
Pondération (b)	0,022	0,743	0,002	0,150

Tableau 7.C : Synthèse des valeurs minimales et maximales de la mesure de similarité en fonction du type de pondération et de la méthode utilisés

Avant ces nœuds, les variables de segmentation sont toujours les mêmes quelle que soit la position considérée dans l'arbre. Les divisions associées à ces variables sont toujours très proches comme l'illustre la comparaison des deux premiers nœuds des arbres en figure 7.g. De ce fait, lorsque moins d'importance est attribuée aux niveaux inférieurs, les arbres de discrimination construits par la méthode REN sont presque tous identiques selon notre mesure. La comparaison des arbres de discrimination construits par la méthode CART montre une grande variabilité dans les valeurs de la mesure de similarité (Cf. Tableau 7.C), quels que soient les poids utilisés. La pondération (b) permet tout de même de réduire légèrement l'écart entre les valeurs minimales et maximales. Néanmoins, elle ne permet pas d'affirmer que les premiers niveaux des arbres CART sont quasi semblables, contrairement aux arbres de discrimination construits par la méthode REN.

Nous avons également calculé la « dispersion » (14) relative aux deux ensembles d'arbres de discrimination $\left\{A_1^{cart},...,A_{30}^{cart}\right\}$ et $\left\{A_1^{ren},...,A_{30}^{ren}\right\}$ (Cf. Tableau 7.D). Comme l'on pouvait s'y attendre, elle est beaucoup plus faible pour l'ensemble d'arbres de discrimination construits par la méthode REN, en particulier lorsque la pondération (b) est utilisée, ce qui confirme les résultats précédents.

	"Dispersion"	
	$\{A_1^{cart},\ldots,A_{30}^{cart}\}$	$\{A_1^{ren},\ldots,A_{30}^{ren}\}$
Pondération (a)	0,332	0,113
Pondération (b)	0,230	0,038

Tableau 7.D : Calcul la « dispersion » des ensembles $\{A_1^{cart},\ldots,A_{30}^{cart}\}$ et $\{A_1^{ren},\ldots,A_{30}^{ren}\}$

7.2 Effet de la modification aléatoire de l'échantillon d'apprentissage

Dans ce paragraphe, nous nous inspirons des comparaisons empiriques effectuées par Ruey-Hsia (2001) afin de tester la stabilité de sa méthode. Nous générons un couple d'échantillons composé d'un échantillon d'apprentissage et d'un échantillon de validation. A partir de ce couple, nous construisons deux arbres de discrimination : le premier par la méthode CART et le deuxième par la méthode REN. Nous procédons ensuite à une modification aléatoire[26] d'une proportion $p_m = 5\ \%$ de l'échantillon d'apprentissage et comparons les deux nouveaux arbres obtenus. La valeur de p_m est ensuite fixée à $10\ \%$.

[26] Une proportion p_m d'individus de l'échantillon d'apprentissage est sélectionnée aléatoirement et nous procédons à la génération de nouvelles données.

			p_m = 5 %		p_m = 10 %	
	CART	REN	CART	REN	CART	REN
1 rac	(R_c) 0,18	(R_c) 0,14	(R_c) 0,14	(R_c) 0,14	(R_c) 0,17	(R_c) 0,14
2 g	(R_c) 0,05	(R_c) 0,04	*f*	(R_c) 0,04	(R_c) 0,05	(R_c) 0,04
3 d	(Δ) 24,50	(Δ) 27,50	(Δ) 26,50	(Δ) 27,50	(Δ) 23,50	(Δ) 28,50
4 gg	*f*	*f*	-	*f*	*f*	*f*
5 gd	(D) 39452,40	(D) 37478,49	-	(D) 36375,69	(D) 39452,40	(D) 38893,60
6 dg	(D) 5337,59	(D) 6454,70	(D) 5337,59	(D) 6164,65	(D) 5337,59	(D) 8111,45
7 dd	(D) 14909,20	(D) 19879,93	(D) 29335,80	(D) 21594,08	(D) 9921,42	(D) 21106,33
8 ggg	-	-	-	-	-	-
9 ggd	-	-	-	-	-	-
10 gdg	*f*	*f*	-	*f*	*f*	*f*
11 gdd	(Δ) 44,50	(Δ) 37,50	-	(Δ) 36,50	(Δ) 41,50	(Δ) 37,50
12 dgg	(D) 1557,60	(D) 1665,10	(Δ) 18,50	(D) 1656,88	(D) 1683,35	(D) 1851,34
13 dgd	*f*	*f*	*f*	*f*	*f*	*f*
14 ddg	(D) 6714,95	(D) 8013,03	(D) 6714,95	(D) 8114,19	(D) 6543,66	(D) 8398,50
15 ddd	*f*	*f*	*f*	(Δ) 34,50	*f*	(Δ) 35,50
16 gggg	-	-	-	-	-	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮
22 gddg	(Δ) 31,50	(Δ) 31,50	-	(R_c) 0,07	(Δ) 29,50	(R_c) 0,08
23 gddd	*f*	*f*	-	*f*	*f*	*f*
24 dggg	*f*	*f*	(D) 751,08	*f*	(Δ) 8,50	*f*
25 dggd	(Δ) 17,50	(Δ) 16,50	*f*	(Δ) 15,50	(Δ) 17,50	(Δ) 17,50
26 dgdg	-	-	-	-	-	-
27 dgdd	-	-	-	-	-	-
28 ddgg	*f*	*f*	*f*	*f*	*f*	*f*
29 ddgd	(Δ) 33,50	(R_c) 0,38	(Δ) 32,50	(Δ) 33,50	(Δ) 30,50	(R_c) 0,34
30 dddg	-	-	-	*f*	-	*f*
31 dddd	-	-	-	(D) 50381,97	-	(R_c) 0,30
44 gddgg	*f*	*f*	-	*f*	*f*	*f*
45 gddgd	*f*	*f*	-	*f*	*f*	*f*
48 dgggg	-	-	*f*	-	*f*	-
49 dgggd	-	-	*f*	-	*f*	-
50 dggdg	*f*	*f*	-	*f*	*f*	*f*
51 dggdd	*f*	*f*	-	*f*	*f*	*f*
58 ddgdg	*f*	*f*	*f*	*f*	*f*	*f*
59 ddgdd	*f*	*f*	*f*	*f*	*f*	*f*
62 ddddg	-	-	-	*f*	-	*f*
63 ddddd	-	-	-	*f*	-	*f*

Tableau 7.E : Comparaison des arbres de discrimination, construits par les méthodes CART et REN, selon le pourcentage de modification de l'échantillon d'apprentissage

Pour comparer les différents arbres de discrimination obtenus, nous avons choisi de représenter dans le tableau 7.E les valeurs de division associées aux nœuds des trois couples d'arbres de discrimination construits par les deux méthodes. Tous les niveaux des arbres sont représentés (du premier au sixième). Les deux premiers arbres construits à partir de l'échantillon de base ont exactement la même architecture (position des nœuds et des feuilles). Les variables de segmentation sont également identiques, excepté pour le nœud n°29. Lorsque 5 % des données de l'échantillon d'apprentissage sont modifiées aléatoirement (250 observations car l'échantillon d'apprentissage est de taille 5000), l'arbre construit par la méthode CART n'a plus la même architecture (le nœud n°2 devient

une feuille). Pour certains nœuds intermédiaires, les valeurs de division sont modifiées, par exemple pour le nœud racine, la division basée sur le rapport de captation passe de 0,14 à 0,18 ou alors elles ne sont plus basées sur les mêmes variables (nœud n°12 par exemple). Lorsque 10 % des données d'apprentissage sont modifiées aléatoirement (250 observations de l'échantillon précédent qui a déjà été modifié sont sélectionnées parmi les 4750 initiales de l'échantillon d'apprentissage, soit au total 500 observations de l'échantillon d'apprentissage sont modifiées) l'arbre de discrimination obtenu par la méthode CART est assez similaire à celui de référence, il semble que cette modification de l'échantillon d'apprentissage influe peu sur l'arbre obtenu. Les deux arbres obtenus par la méthode REN ont quasiment la même architecture que celui de référence, avec l'ajout d'une branche aux nœuds n°15 et 31. Leurs deux premiers niveaux sont pratiquement identiques. D'une manière générale les autres divisions associées aux nœuds intermédiaires sont assez semblables.

Les taux de mauvais classement associés à ces arbres sont également estimés à partir d'un échantillon test, ils sont présentés (en %) dans le tableau 7.F. Pour la méthode CART, l'erreur de classement est variable (elle se dégrade) en fonction du nombre d'observations modifiées dans l'échantillon d'apprentissage. Comme nous l'avions mis en évidence précédemment, la méthode est sensible à de légères altérations de l'échantillon d'apprentissage. Par contre, pour les arbres de discrimination construits selon la méthode REN, les modifications aléatoires effectuées dans l'échantillon d'apprentissage n'ont pas de répercussion sur la qualité du modèle. Le taux de mauvais classement reste identique pour une modification aléatoire de 5 % et diminue faiblement lorsque qu'une modification de 10 % est effectuée dans l'échantillon.

	Taux de mauvais classement (%)	
	méthode CART	méthode REN
	5,36	5,14
$p_m = 5\ \%$	5,64	5,14
$p_m = 10\ \%$	5,82	5,12

Tableau 7.F : Comparaison du taux de mauvais classement (%) associé aux différents arbres de discrimination selon le pourcentage de modification de l'échantillon d'apprentissage

7.3 Conclusion

Les résultats empiriques obtenus dans ce chapitre permettent de confirmer la stabilité de la méthode REN dans la construction des arbres de discrimination. La méthode REN permet de réduire la variabilité observée dans le choix de la variable de segmentation et de sa

division associée. De ce fait, à partir de l'arbre de discrimination obtenu des règles de décision plus stables pourront être proposées. Ces résultats reposent en partie sur la mesure de similarité définie dans le paragraphe 4.4.3.3. Cette mesure permet de comparer très facilement la structure de deux arbres de discrimination et offre à l'utilisateur la possibilité de pondérer les niveaux de l'arbre et donc d'accorder plus d'importance aux parties qui lui paraissent les plus importantes.

En ce qui concerne les performances de prédiction, la méthode REN est globalement meilleure que la méthode CART. Néanmoins, les comparaisons empiriques ont montré que la différence entre les erreurs de classement issues de ces deux méthodes était assez faible, les taux de mauvais classement variant dans un intervalle de valeurs très réduit.

8 Analyse et interprétation des arbres de discrimination obtenus

Dans ce chapitre, nous procédons à l'analyse et à l'interprétation, pour chaque légume-feuille étudié (laitue, chou, épinard et poireau), de l'arbre de discrimination construit par la méthode REN. En particulier, nous interprétons les différents chemins des arbres conduisant aux deux niveaux de contamination radioactive définis en 5.2.

8.1 Le cas de la laitue

8.1.1 Exploration de l'arbre de discrimination

Dans le chapitre précédent, nous avons construit 30 arbres de discrimination par la méthode REN. Nous utilisons la mesure de similarité (13) définie précédemment afin de proposer un arbre « central » plutôt que de générer arbitrairement un couple d'échantillons d'apprentissage et de validation. L'approche décrite en 4.4.3.3 est appliquée à l'ensemble $\{A_1^{ren},...,A_{30}^{ren}\}$. L'arbre de discrimination obtenu est présenté sur la figure 8.a. Pour chaque nœud intermédiaire, le nombre entre parenthèses représente le vote associé au choix de la variable de segmentation. Par exemple, pour le nœud racine, sur les 100 échantillons bootstrap générés, la division optimale est toujours basée sur le rapport de captation. Pour chaque nœud intermédiaire de l'arbre, le vote associé au choix de la variable de division est généralement assez élevé, excepté pour deux nœuds proches des feuilles de l'arbre qui présentent des valeurs inférieures à 50.

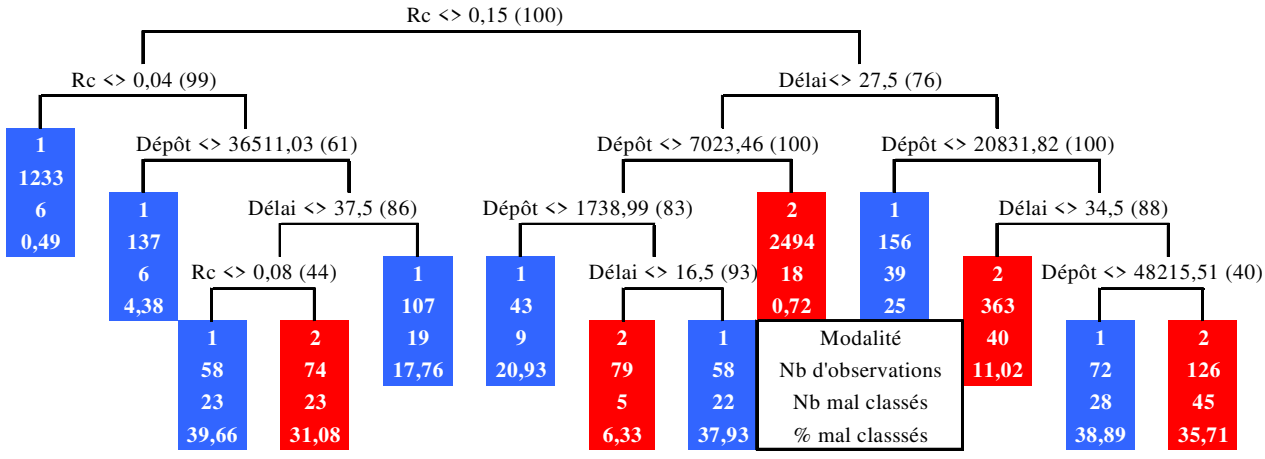


Figure 8.a : Arbre de discrimination central relatif au légume-feuille laitue obtenu par la méthode REN, les taux de mauvais classement dans les feuilles sont estimés par un échantillon test

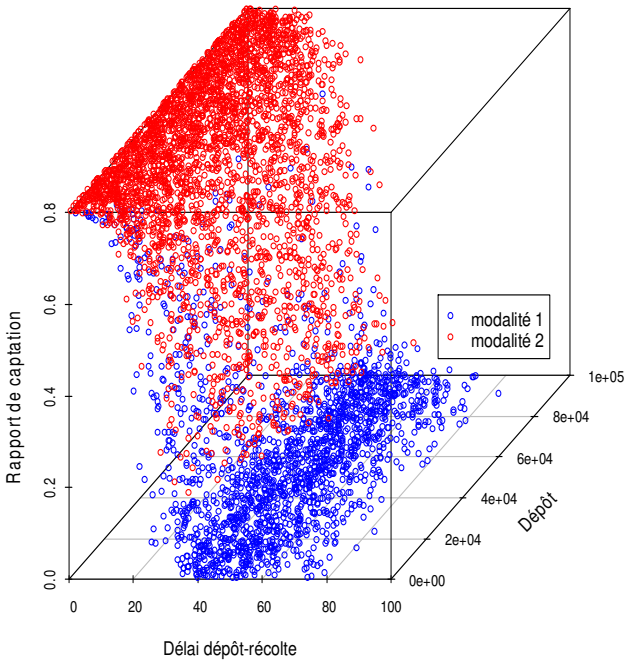


Figure 8.b : Représentation, en trois dimensions, de l'échantillon test utilisé pour estimer les performances de l'arbre de discrimination central

Le taux de mauvais classement associé à cet arbre est estimé par un échantillon test (représenté en trois dimensions sur la figure 8.b), il est de 5,66 %. Il est également estimé dans les feuilles de l'arbre (Cf. Figure 8.a) : pour chaque nœud terminal, nous indiquons le nombre total d'observations dans le nœud, le nombre de mal classés ainsi que le pourcentage de mauvais classement associé. Pour mesurer plus précisément la qualité de prédiction de notre modèle, une matrice de confusion est également construite sur l'échantillon test (Cf. Tableau 8.A). Elle consiste à confronter les valeurs prédites par l'arbre de discrimination aux valeurs observées dans l'échantillon. On retrouve dans la diagonale principale de cette matrice le nombre d'observations correctement prédites pour chacune des modalités. Les éléments hors diagonale permettent de détecter les différentes erreurs de classement. Il semble que le modèle ait légèrement plus de mal à classer certaines observations appartenant à la modalité 2 : 152 observations sont affectées à la modalité 1 alors qu'elles appartiennent à la modalité 2 et 131 sont affectées à la modalité 2 alors qu'elles appartiennent en réalité à la modalité 1. Dans ce qui suit, nous choisissons d'examiner les feuilles présentant les plus forts taux de mauvais classement. Pour plus de facilité dans les notations, nous synthétisons, dans le tableau 8.B la position de chaque feuille dans l'arbre (mêmes notations que celles utilisées dans le chapitre précédent pour caractériser les nœuds d'un arbre) ainsi que les taux de mauvais classement associés.

		Prédit		Total
		modalité 1	modalité 2	
Observé	modalité 1	1712	131	1843
	modalité 2	152	3005	3157
	Total	1864	3136	5000

Tableau 8.A : Matrice de confusion sur l'échantillon test

Nous sélectionnons les feuilles 14, 44, 45, 51, 62 et 63 qui présentent des taux de mauvais classement supérieur ou égal à 25 %. Au total, elles contribuent à plus de 50 % des erreurs de classement (180 sur 283 observations). Pour chacune de ces feuilles, nous sélectionnons les valeurs quantitatives de la variable à expliquer (activité massique du ^{90}Sr, en $Bq.kg^{-1}$) correspondant aux observations incorrectement classées. Une représentation graphique de ces valeurs est proposée sur la figure 8.c. Pour les feuilles affectées à la modalité 1, plus de la moitié des observations mal classées présentent des valeurs de la variable à expliquer comprises dans l'intervalle]100, 200] $Bq.kg^{-1}$, proche de la valeur seuil qui a été utilisée (100 $Bq.kg^{-1}$) pour effectuer le codage de l'échantillon en deux modalités. Les autres valeurs sont plus élevées, en particulier pour certaines observations appartenant à la feuille n°62. Pour les feuilles n°45 et 63, les valeurs d'activités de ^{90}Sr ont tendance à se répartir de manière plus uniforme, dans l'intervalle [0, 100] $Bq.kg^{-1}$. Dans ce qui suit, nous allons explorer l'arbre de discrimination, en particulier examiner les différentes variables explicatives apparaissant dans l'arbre.

Position feuille		Taux de mauvais classement (%)	Modalité
4	gg	0,49	1
10	gdg	4,38	1
13	dgd	0,72	2
14	ddg	25,00	1
23	gddd	17,76	1
24	dggg	20,93	1
30	dddg	11,02	2
44	gddgg	39,66	1
45	gddgd	31,08	2
50	dggdg	6,33	2
51	dggdd	37,93	1
62	ddddg	38,89	1
63	ddddd	35,71	2

Tableau 8.B : Position de chaque feuille dans l'arbre de discrimination relatif au légume-feuille laitue et taux de mauvais classement associé

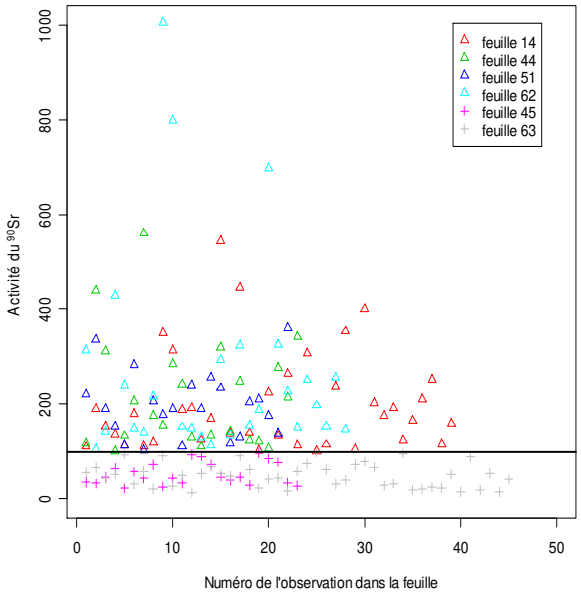


Figure 8.c : Représentation des observations mal classées dans les feuilles 14, 44, 51, 62, 45 et 63 de l'arbre de discrimination relatif au légume-feuille laitue

La représentation graphique de l'échantillon test (Cf. Figure 8.b) permet de mettre en évidence deux variables qui semblent importantes pour discriminer les observations appartenant aux deux modalités. Il s'agit du rapport de captation (R_c) et du délai dépôt-récolte (Δ). Les travaux effectués dans le paragraphe 6.2.2 du chapitre 6 sur l'importance des variables avaient déjà permis leur identification. Il n'est donc pas étonnant de retrouver ces deux variables dans les deux premiers niveaux de l'arbre. Pour illustrer le pouvoir discriminant du rapport de captation et du délai dépôt-récolte, nous représentons sur la figure 8.d les partitions engendrées, sur le jeu de données test, par les divisions associées au nœud racine et au nœud n°3. La première division, basée sur le rapport de captation, permet de créer deux sous-ensembles, dont un particulièrement homogène du point de vue de la variable à expliquer. Il est constitué de 1609 observations dont 105 seulement appartiennent à la modalité 2. La division associée au nœud n°3, basée sur le délai dépôt-récolte, engendre également deux sous-ensembles, dont un permettant d'isoler de nombreuses observations appartenant à la modalité 2. Dans le deuxième sous-ensemble engendré par le nœud n°3 ($R_c > 0{,}15$ et $\Delta \geq 28$), il semble plus difficile de distinguer des groupes d'observations ayant des modalités communes. Nous détaillons donc plus particulièrement la partition du jeu de données test produite par la branche droite de l'arbre issue du nœud racine (Cf. Figure 8.e). Chaque feuille est identifiée par son numéro (Cf. Tableau 8.B) et correspond à un sous-ensemble rectangulaire délimité par des droites de séparation (divisions associées aux nœuds intermédiaires de l'arbre).

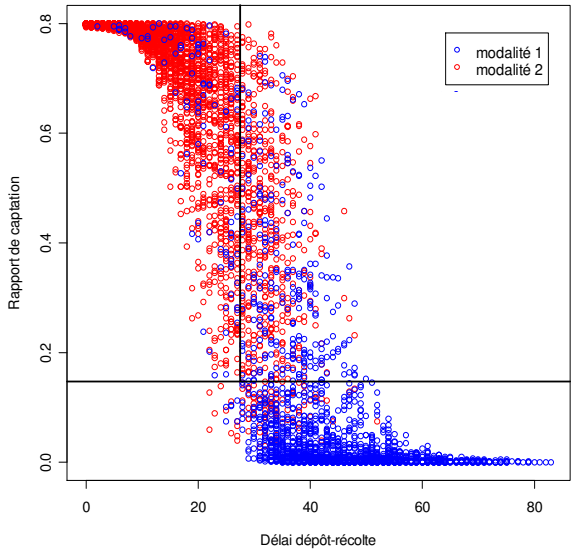


Figure 8.d : Partition engendrée par le nœud racine et le nœud n°3 de l'arbre de discrimination relatif au légume-feuille laitue (données relatives à l'échantillon test)

La représentation de la figure 8.e permet de mieux appréhender les règles de décision dans cette partie de l'arbre. Comme nous l'avions mis en évidence précédemment, la feuille n°13 est très homogène, elle contient seulement 18 observations appartenant à la modalité 1. Les autres sous-ensembles sont de tailles plus faibles, en particulier les feuilles n°24 et n°51. Dans certains cas, il semble difficile d'isoler les observations appartenant à l'une des deux modalités. Par exemple, pour les feuilles n°62 et n°63, aucune structure n'apparaît clairement dans les données.

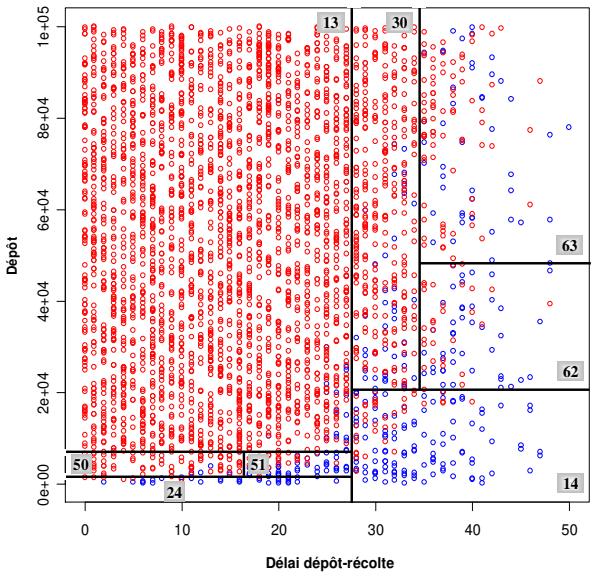


Figure 8.e : Partition engendrée par la branche droite issue du nœud racine de l'arbre de discrimination relatif au légume-feuille laitue (données relatives à l'échantillon test)

8.1.2 Interprétation des chemins les plus pertinents

8.1.2.1 Sélection des chemins

Etant donné le contexte de notre application, nous ne pouvons pas sélectionner tous les chemins car l'erreur de classement associée à certains est très élevée. Nous sélectionnons donc les règles de décision dont les taux de mauvais classement, estimés sur l'échantillon test, sont les plus faibles. Cela nous conduit à retenir les 5 chemins présentés dans le tableau 8.C. Pour chacun d'entre eux, les caractéristiques suivantes sont renseignées : la feuille associée (colorée selon son appartenance à l'une des deux modalités), et pour le jeu de données test : l'effectif de l'échantillon test (%) vérifiant le chemin ainsi que le taux de mauvais classement (%). Dans un premier temps, comme il a été fait précédemment pour les feuilles de plus mauvaise qualité, nous étudions la répartition des observations mal classées pour chacun des chemins sélectionnés. Pour chaque feuille, les valeurs de la variable à expliquer relatives aux observations mal classées sont sélectionnées et présentées sous forme de box-plot en figure 8.f. Pour les deux feuilles affectées à la modalité 1 (n°4 et n°10), les valeurs d'activités du ^{90}Sr sont relativement faibles. De ce fait, même si une observation est incorrectement classée, l'activité du ^{90}Sr associée à la production de laitue, reste assez proche de la valeur seuil utilisée pour le codage de l'échantillon (100 $Bq.kg^{-1}$), la valeur maximale sur l'échantillon test étant de 160 $Bq.kg^{-1}$. Pour les observations affectées à la modalité 2, qui appartiennent, en réalité, à la modalité 1, les activités sont également proches de la valeur de codage, en particulier pour la feuille n°50. Afin d'illustrer les erreurs de classement, les observations du jeu de données test dans chacune des feuilles de l'arbre sont représentées graphiquement (Cf. Figure 8.g pour les feuilles n°4 et n°10 et Annexe H pour les feuilles n°13, 30 et 50). Les mauvais classements effectués dans les feuilles n°4 et n°10 paraissent liés à de fortes valeurs de dépôt et de rapport de captation. En particulier, pour la feuille n°4, les observations incorrectement classées sont associées à des dépôts très élevés, variant de 69 580 à 97 330 $Bq.m^{-2}$.

Chemin	Feuille	Effectif (%)	Taux de mauvais classement (%)
$Rc \leq 0{,}15$ & $Rc \leq 0{,}04$	4	24,66	0,49
$Rc \leq 0{,}15$ & $Rc > 0{,}04$ & $D \leq 36511{,}03$	10	2,74	4,38
$Rc > 0{,}15$ & $\Delta \leq 27$ & $D > 7023{,}46$	13	49,88	0,72
$Rc > 0{,}15$ & $\Delta \geq 28$ & $D > 20831{,}82$ & $\Delta \leq 34$	30	7,26	11,02
$Rc > 0{,}15$ & $\Delta \leq 27$ & $D \leq 7023{,}46$ & $D > 1738{,}99$ & $\Delta \leq 16$	50	1,58	6,33

Tableau 8.C : Sélection des chemins dont les taux de mauvais classement sont les plus faibles (estimés sur l'échantillon test)

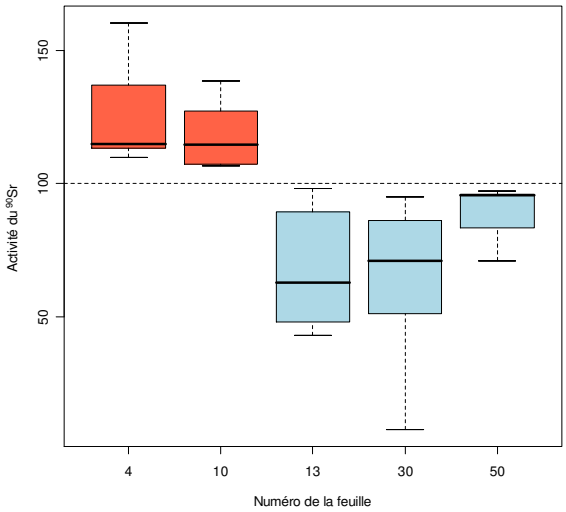


Figure 8.f : Etude des observations mal classées dans les feuilles n°4, 10, 13, 30 et 50 de l'arbre de discrimination relatif au légume-feuille laitue

Pour la feuille n°13, il semble que la variable délai dépôt-récolte soit en partie responsable des observations affectées à la modalité 1. Pour cette variable, les 18 observations présentent des valeurs très proches de la valeur de division associée au nœud n°3 (variant de 19 à 27 jours, avec une valeur médiane de 26 jours). La feuille n°50 est caractérisée par des observations présentant de fortes valeurs de rapport de captation (proches de la borne maximum de cette variable qui est de 0,8).

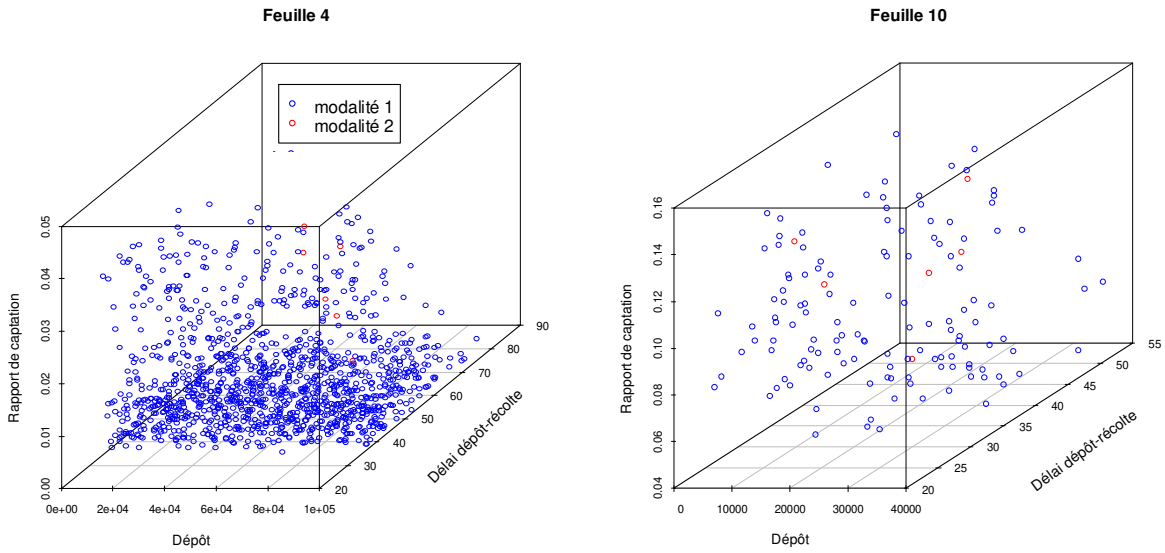


Figure 8.g : Représentation, en trois dimensions, des observations de l'échantillon test dans les feuilles n°4 et n°10 de l'arbre de discrimination relatif au légume-feuille laitue

Il semble que les erreurs de classement soient liées au délai entre le dépôt et la récolte de la production de laitue et au dépôt de radioactivité. En ce qui concerne la feuille n°30, il paraît plus difficile d'identifier les variables explicatives responsables des erreurs de classement.

Nous nous sommes également intéressés à la distribution de la variable à expliquer dans chacune de ces feuilles. Divers indicateurs statistiques ont été calculés, ils sont synthétisés dans le tableau 8.D. Parmi les deux chemins conduisant à la modalité 1, le premier (identifié par la feuille n°4) a tendance à prédire de faibles valeurs d'activités pour le radionucléide ^{90}Sr. Lorsque l'on s'intéresse aux chemins conduisant à la modalité 2, les plus fortes valeurs sont prédites par la feuille n°13 tandis que la feuille n°50 a tendance à prédire des valeurs d'activités plus faibles.

Feuille	Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
4	0,00	0,15	0,78	5,51	4,12	160,70
10	0,07	5,35	17,02	26,75	35,69	138,80
13	42,77	987,40	2269,00	3430,00	4762,00	25470,00
30	7,36	191,90	416,10	569,70	754,70	3914,00
50	71,17	177,70	296,90	356,80	444,70	1396,00

Tableau 8.D : Indicateurs statistiques des activités massiques du ^{90}Sr (Bq.kg^{-1}) pour chaque feuille de l'arbre sélectionnée

8.1.2.2 Interprétation des chemins

8.1.2.2.1 Les chemins conduisant à la modalité 1

Précédemment, deux chemins de l'arbre de discrimination conduisant à la modalité 1 de la variable à expliquer ont été identifiés. La première règle se résume par l'expression suivante :

$$Si \;\; R_c \leq 0{,}04 \;\; Alors \;\; \text{activité du } ^{90}\text{Sr} = \text{modalité 1} \qquad (r.1)$$

Le rapport de captation représente la fraction du dépôt radioactif qui est interceptée par la masse foliaire des végétaux se trouvant à la surface du sol. Comme nous l'avons vu dans le chapitre 5, c'est une variable qui évolue dans le temps en fonction du développement du végétal. De ce fait, une laitue présentant un faible rapport de captation est une laitue dont le système foliaire est encore peu développé, donc en début de croissance. La valeur de 0,04 est très faible, correspondant aux premiers jours suivant la plantation de la laitue. Cette première règle peut donc être interprétée de la manière suivante : quel que soit le dépôt radioactif, une laitue contaminée les premiers jours suivant sa plantation présente à la récolte une contamination radioactive inférieure ou égale à 100 Bq.kg^{-1}, et peut donc être commercialisée.

La deuxième règle, conduisant également à la modalité 1, a pour expression :

$$\textit{Si}\ R_c \leq 0{,}15\ \text{et}\ R_c > 0{,}04\ \text{et}\ D \leq 36511^{27}\ \textit{Alors}\ \text{activité du } {}^{90}\text{Sr} = \text{modalité 1} \qquad \text{(r.2)}$$

Contrairement à (r.1), cette règle fait intervenir le dépôt de radioactivité. Le rapport de captation, lui, est toujours compris entre deux valeurs relativement faibles. Cette règle nous permet d'avancer qu'une laitue au début de sa croissance et contaminée quelques jours après sa plantation par un dépôt inférieur à 36 511 Bq.m^{-2} peut également être commercialisée. D'après les résultats présentés dans le tableau 8.D, les valeurs d'activités du ^{90}Sr ont tendance à être légèrement plus élevées pour les observations vérifiant cette règle, que pour celles vérifiant la précédente (r.1) (Cf. Tableau 8.D).

8.1.2.2.2 Les chemins conduisant à la modalité 2

Nous allons maintenant nous intéresser aux chemins qui conduisent à la modalité 2 de la variable à expliquer. Toutes ces règles sont relatives à la branche droite de l'arbre issue du nœud racine, elles ont donc toutes en commun la première condition sur la variable rapport de captation : $R_c > 0{,}15$. Ainsi, toutes les interprétations effectuées dans la suite de ce paragraphe concerneront des cultures de laitue qui peuvent être en début de croissance (de quelques jours à quelques semaines après leurs plantations) ou prête à être récoltées. La première règle a pour expression :

$$\textit{Si}\ R_c > 0{,}15\ \text{et}\ \Delta \leq 27\ \text{et}\ D > 7023\ \textit{Alors}\ \text{activité du } {}^{90}\text{Sr} = \text{modalité 2} \qquad \text{(r.3)}$$

Si une culture de laitue est contaminée dans les 27 jours qui précèdent sa récolte et que le dépôt de radioactivité est supérieur à 7 023 Bq.m^{-2} alors la contamination radioactive de la production de laitue est de niveau 2 (> 100 Bq.kg^{-1}). Le deuxième chemin, conduisant à la feuille n°30, est le suivant :

$$\textit{Si}\ R_c > 0{,}15\ \text{et}\ \Delta \geq 28\ \text{et}\ D > 20831\ \text{et}\ \Delta \leq 34\ \textit{Alors}\ \text{activité du } {}^{90}\text{Sr} = \text{modalité 2} \qquad \text{(r.4)}$$

Dans ce cas, le dépôt de radioactivité et le délai dépôt-récolte sont plus élevés que précédemment : le dépôt doit être supérieur à 20 831 Bq.m^{-2} et le délai compris entre 27 et 34 jours. La dernière règle peut être synthétisée par l'expression suivante :

$$\textit{Si}\ R_c > 0{,}15\ \text{et}\ D \leq 7023\ \text{et}\ D > 1739\ \text{et}\ \Delta \leq 16\ \textit{Alors}\ \text{activité du } {}^{90}\text{Sr} = \text{modalité 2} \qquad \text{(r.5)}$$

Un dépôt inférieur à ceux proposés dans les deux règles précédentes combiné à un délai dépôt-récolte également plus faible (inférieur ou égal à 16 jours), engendre une contamination radioactive de niveau 2. D'après le tableau 8.D, les observations vérifiant

[27] Pour faciliter la lecture des règles, nous proposons pour l'extraction de cette règle et des suivantes d'arrondir les valeurs du dépôt de radioactivité.

cette règle ont tendance à présenter des valeurs d'activités massiques du ^{90}Sr plus faibles que celle vérifiant (r.3).

8.1.2.2.3 Synthèse des interprétations

L'extraction de ces chemins permet d'identifier les associations de variables qui sont responsables des différents niveaux de contamination radioactive de la laitue. Après interprétation de ces règles, il ressort que les chemins relatifs à la branche gauche de l'arbre issue de la racine conduisent à des contaminations radioactives de niveau 1 (≤ 100 $Bq.kg^{-1}$), tandis que les règles de décision extraites de la branche opposée (droite issue de la racine) permettent de discriminer les contaminations radioactives de niveau 2 (> 100 $Bq.kg^{-1}$). La variable qui permet de distinguer ces deux grands groupes d'observations est le rapport de captation, variable dépendante du développement de la laitue. Dans la branche droite de l'arbre, le délai entre le dépôt et la récolte du végétal est aussi une variable importante permettant de discriminer les observations de modalité 2. Cette variable est également liée à la croissance de la laitue : au plus le délai est élevé, au plus le végétal a été contaminé au début de sa croissance. De ce fait, le développement de la culture de laitue le jour du dépôt de radioactivité semble être la principale cause du niveau de contamination de la production à la récolte.

8.2 Les autres légumes-feuilles étudiés

Dans ce paragraphe, nous nous intéressons aux arbres de discrimination associés aux autres légumes-feuilles étudiés dans notre scénario de contamination (Cf. 2.3.2) : le chou, l'épinard, et le poireau. Chaque arbre est construit à partir d'un couple d'échantillons générés arbitrairement. Pour faciliter la lecture de la suite de ce paragraphe, nous utilisons les notations A_{laitue}^{REN}, A_{chou}^{REN}, $A_{épinard}^{REN}$ et $A_{poireau}^{REN}$ pour faire référence à chaque arbre de discrimination, construit par la méthode REN, pour les légumes-feuilles laitue, chou, épinard et poireau. Nous avons choisi de présenter les arbres de discrimination obtenus pour l'épinard et le chou en Annexe I, car l'organisation de ces deux arbres est assez similaire à celle de l'arbre de discrimination obtenu pour le cas de la laitue.

	Mesure de similarité		
	$d\left(A_{laitue}^{REN}, A_{épinard}^{REN}\right)$	$d\left(A_{laitue}^{REN}, A_{chou}^{REN}\right)$	$d\left(A_{épinard}^{REN}, A_{chou}^{REN}\right)$
Pondération (a)	0,43	0,77	0,41
Pondération (b)	0,64	0,88	0,58

Tableau 8.E : Calcul de la mesure de similarité pour les trois couples d'arbres de discrimination $(A_{laitue}^{REN}, A_{épinard}^{REN})$, $(A_{laitue}^{REN}, A_{chou}^{REN})$ et $(A_{épinard}^{REN}, A_{chou}^{REN})$

	Mesure de similarité		
	$d\left(A^{REN}_{poireau}, A^{REN}_{épinard}\right)$	$d\left(A^{REN}_{poireau}, A^{REN}_{laitue}\right)$	$d\left(A^{REN}_{poireau}, A^{REN}_{chou}\right)$
Pondération (a)	0,13	0,06	0,06
Pondération (b)	0,08	0,05	0,05

Tableau 8.F : Calcul de la mesure de similarité pour les trois couples d'arbres de discrimination $(A^{REN}_{poireau}, A^{REN}_{épinard})$, $(A^{REN}_{poireau}, A^{REN}_{laitue})$ et $(A^{REN}_{poireau}, A^{REN}_{chou})$

La mesure de similarité (13) est calculée pour ces trois couples d'arbres, en utilisant les deux types de pondérations définies dans le chapitre précédent (Cf. Tableau 8.E). L'arbre de discrimination le plus semblable à l'arbre A^{REN}_{laitue} est A^{REN}_{chou}, leurs deux architectures sont quasi analogues. Elles se différencient uniquement par une branche supplémentaire dans l'arbre A^{REN}_{laitue} (nœud n°25, position *dggd*). Les variables de segmentation sont toutes identiques, excepté pour le nœud n°31 (position *dddd*). L'arbre de discrimination $A^{REN}_{épinard}$ est moins similaire à l'arbre A^{REN}_{laitue}. Ceci est sans doute lié à la variable délai dépôt-récolte n'apparaissant pas dans l'arbre $A^{REN}_{épinard}$, et remplacée par la variable rapport de captation.

Nous avons choisi de présenter, dans ce paragraphe, l'arbre obtenu pour le légume-feuille poireau, car sa structure se distingue des 3 autres arbres de discrimination (Cf. Tableau 8.F). Cet arbre est présenté sur la figure 8.h en utilisant les mêmes notations que dans le paragraphe précédent. Le taux de mauvais classement, estimé par un échantillon test est de 7,8 %. Cet arbre de discrimination se différencie des autres dès ses premiers niveaux, comme l'illustre les calculs de notre mesure de similarité, lorsque la pondération (b) est utilisée (Cf. Tableau 8.F). Contrairement aux arbres précédents, la segmentation du nœud racine est basée sur la variable délai dépôt-récolte. La partition de l'échantillon test engendré par la division associée à cette variable est présentée sur la figure 8.i, elle permet de créer deux sous-ensembles très homogènes relativement à la variable à expliquer. Les deux feuilles les plus intéressantes (dont les taux de mauvais classement sont les plus faibles) sont situées dans les deux branches opposées de l'arbre, il s'agit des feuilles n°3 (position *d*) et n°9 (position *ggd*). La première règle extraire de l'arbre est la suivante :

$$\textit{Si}\ \ \Delta \geq 61\ \ \textit{Alors}\ \ \text{activité du } {}^{90}\text{Sr} = \text{modalité 1} \qquad \text{(r.6)}$$

Quelle que soit la valeur du dépôt de radioactivité, il suffit que le délai entre le dépôt et la récolte de la production de poireau soit supérieur ou égal à 61 jours, pour que la contamination radioactive de la culture de poireau soit de niveau 1.

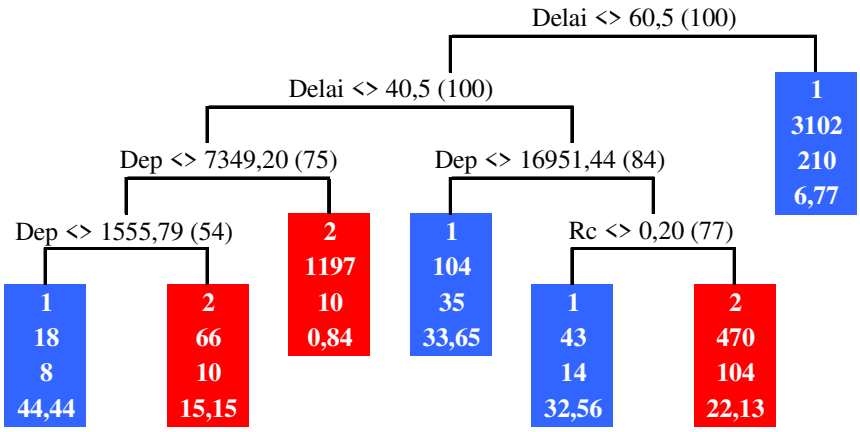


Figure 8.h : Arbre de discrimination construit par la méthode REN pour le légume-feuille poireau, les taux de mauvais classement dans les feuilles sont estimés par un échantillon test

Le deuxième chemin conduisant à la modalité 2 est synthétisé par l'expression suivante :

$$Si \ \Delta \leq 40 \text{ et } D > 7349 \ Alors \text{ activité du } {}^{90}Sr = \text{modalité } 2 \qquad (r.7)$$

Pour que l'activité du ^{90}Sr, dans la production de poireau, soit supérieure à 100 $Bq.kg^{-1}$, le dépôt de radioactivité doit être supérieur à 7 349 $Bq.m^{-2}$ et le délai inférieur ou égal à 40 jours. On retrouve donc le même type de règle que dans les arbres précédents, le rapport de captation étant remplacé ici par le délai dépôt-récolte. Si le nombre de jour entre le dépôt de radioactivité et la récolte du végétal est élevé, cela signifie que le végétal est en début de croissance et donc qu'il présente un faible rapport de captation. A l'inverse, si le délai entre le dépôt et la récolte du végétal est faible cela signifie qu'il est contaminé quelques jours avant sa récolte, il présente donc un fort rapport de captation. Ainsi, contrairement aux trois légumes-feuilles étudiés précédemment, la partie gauche de l'arbre de discrimination conduit principalement aux observations de modalité 2 et la partie droite à celles de modalité 1, comme l'illustre la figure 8.i.

Par conséquent, compte tenu des similitudes entre les arbres associés aux différents légumes-feuilles, nous proposons la construction d'un arbre de discrimination pour l'ensemble des légumes-feuilles étudiés. Les résultats sont présentés dans le paragraphe suivant.

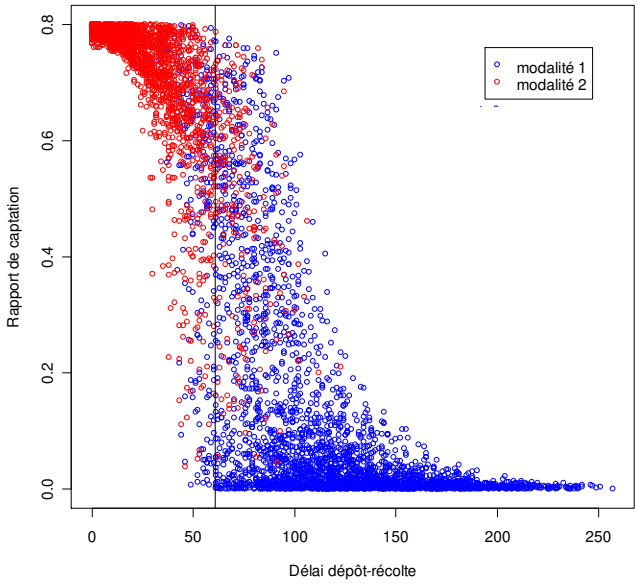


Figure 8.i : Partition du jeu de données test engendrée par la division basée sur la variable délai dépôt-récolte (racine de l'arbre de discrimination de l'arbre relatif au légume-feuille poireau)

8.3 Un arbre de discrimination pour les légumes-feuilles

Afin de construire l'arbre de discrimination relatif au 4 légumes-feuilles, nous générons trois échantillons : un échantillon d'apprentissage, un échantillon de validation pour construire l'arbre selon la méthode REN et un échantillon test afin d'estimer les performances de cet arbre. Une variable explicative supplémentaire est prise en compte, notée X_{lf}. Elle est de type qualitative et comprend 4 modalités, chacune correspondant à l'un des 4 légumes-feuilles étudiés. Pour générer un échantillon de taille n, nous tirons aléatoirement à chaque étape i, $i = 1,\ldots, n$, une modalité de la variable X_{lf} et nous générons les valeurs des variables correspondant au légume-feuille considéré. L'arbre de discrimination obtenu est présenté sur la figure 8.j, le taux de mauvais classement (%) estimé par l'échantillon test est de 6,9 %. La variable légume-feuille n'apparaît pas dans l'arbre et l'on retrouve le rapport de captation comme variable de segmentation du nœud racine. Ainsi, comme pour les trois légumes-feuilles laitue, chou et épinard, la partie gauche de l'arbre conduit principalement aux observations de modalité 1, tandis que celles de modalité 2 sont essentiellement dirigées du côté droit. La position des feuilles ayant les plus faibles taux de mauvais classement reste inchangée : n°4 et n°13. Les règles de décision associées à ces feuilles sont synthétisées dans les expressions ci-dessous.

$$\textit{Si}\ R_c \leq 0{,}03\ \textit{Alors}\ \text{activité du } {}^{90}\text{Sr} = \text{modalité 1} \qquad \text{(r.8)}$$

$$\textit{Si}\ R_c > 0{,}17\ \text{et}\ \Delta \leq 35\ \text{et}\ D > 8841\ \textit{Alors}\ \text{activité du } {}^{90}\text{Sr} = \text{modalité 2} \qquad \text{(r.9)}$$

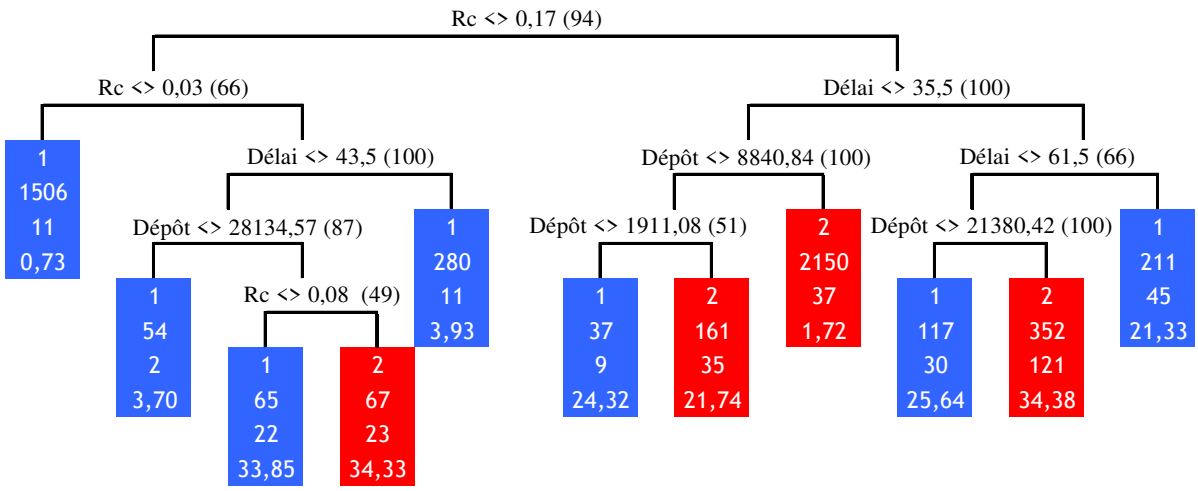


Figure 8.j : Arbre de discrimination obtenu par la méthode REN pour les 4 légumes-feuilles étudiés, les taux de mauvais classement dans les feuilles sont estimés par un échantillon test

D'autres chemins, dans la partie gauche de l'arbre, conduisant à la modalité 1 de la variable à expliquer sont également intéressants (vérifiés par moins d'observations du jeu de données test que les règles précédentes) :

$$\textit{Si}\ R_c \leq 0{,}03\ \text{et}\ \Delta \leq 43\ \text{et}\ D \leq 28135\ \textit{Alors}\ \text{activité du}\ {}^{90}\text{Sr} = \text{modalité 1} \qquad \text{(r.10)}$$

$$\textit{Si}\ R_c \leq 0{,}03\ \text{et}\ \Delta > 43\ \textit{Alors}\ \text{activité du}\ {}^{90}\text{Sr} = \text{modalité 1} \qquad \text{(r.11)}$$

Nous retrouvons donc le même type de règle que dans les arbres précédents (en particulier les règles (r.1) et (r.3)), mais cette fois-ci, les règles sont valables quel que soit le légume-feuille étudié (laitue, épinard, poireau ou chou). Les valeurs d'activités du ^{90}Sr correspondant aux observations de l'échantillon test et appartenant à la feuille n°4 sont présentées sur le graphe gauche de la figure 8.k, en fonction du rapport de captation. L'appartenance à l'une des 4 modalités de la variable légume-feuille est également précisée. Dans cette feuille, les plus fortes valeurs d'activités sont relatives à l'épinard, responsable des 11 observations incorrectement classées pour lesquelles les activités du ^{90}Sr varient de 101,7 à 187,5 Bq.kg^{-1}. Les légumes-feuilles présentant les plus faibles valeurs étant le chou et le poireau (Cf. Annexe J, où une synthèse des différentes valeurs d'activités du ^{90}Sr pour les feuilles n°4 et n°13, en fonction du légume-feuille étudié, est proposée).

Pour la feuille n°13, la représentation des valeurs d'activités du ^{90}Sr est effectuée sous forme de box-plot en fonction des 4 légumes-feuilles (Cf. graphe droit de la figure 8.k). Les valeurs de contamination les plus élevées sont relatives au poireau et, comme pour la feuille n°4, à l'épinard, les valeurs médianes respectives étant de 5150 Bq.kg^{-1} et de 5113

Bq.kg^{-1}. Les légumes-feuilles laitue et chou présentent les plus faibles valeurs (cf. Annexe J). Comme nous l'avons mis en évidence précédemment par notre mesure de similarité, les arbres de discrimination relatifs à ces deux légumes-feuilles sont relativement proches.

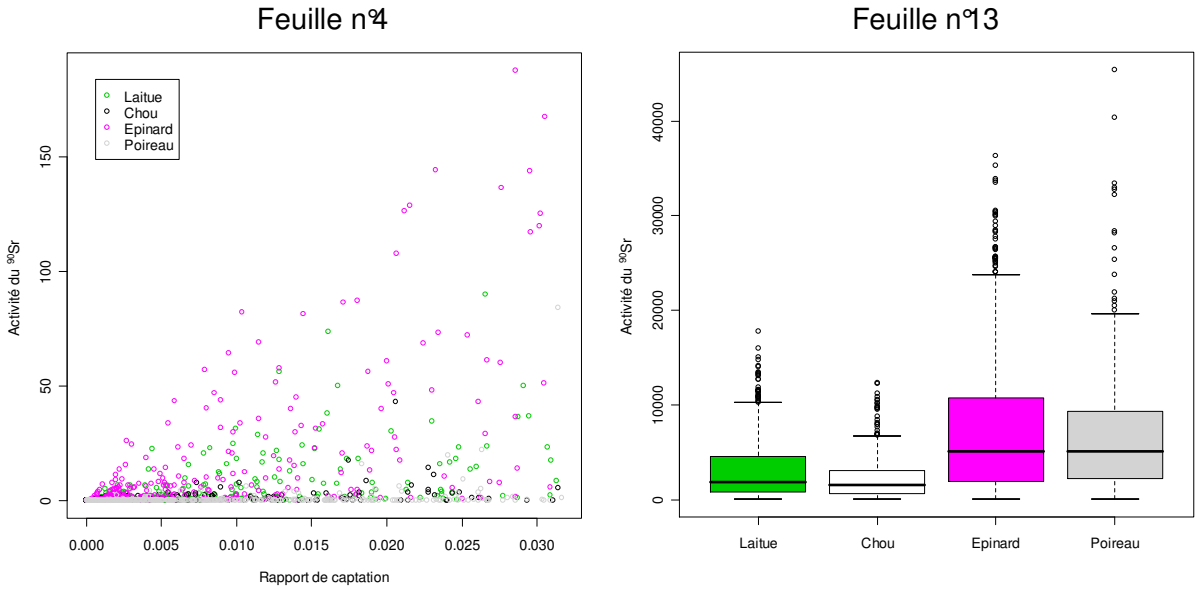


Figure 8.k : Représentation des données de l'échantillon test dans les feuilles n°4 et n°13 de l'arbre de discrimination relatif aux quatre légumes-feuilles

8.4 Conclusion

Pour chaque légume-feuille, l'interprétation des arbres de discrimination permet d'identifier les variables et les divisions qui discriminent les deux niveaux de contamination radioactive étudiés. Il apparaît que les deux variables les plus importantes dans la contamination des légumes-feuilles sont le rapport de captation et le délai entre le dépôt de radioactivité et la récolte du végétal, variables intimement liées au développement du végétal. Ces conclusions conduisent à proposer un arbre de discrimination pour les quatre légumes-feuilles étudiés, permettant de synthétiser l'ensemble des résultats obtenus en un seul arbre. Une précédente étude du projet SENSIB (Mercat-Rommens et al., 2006), avait également permis de mettre en évidence que le développement du végétal le jour de l'accident était une variable en grande partie responsable de la contamination à la récolte. Cette étude portant sur le blé d'hiver, avait identifié la date de floraison du blé comme une variable clé, responsable de la contamination des grains.

Chaque arbre de discrimination, étudié dans ce chapitre, présente deux chemins, conduisant respectivement à la modalité 1 et à la modalité 2 de la variable à expliquer, vérifiés par plus de 20 % des observations des jeux de données test et dont l'erreur de

classement est relativement faible (le taux de mauvais classement étant toujours inférieur à 2,5 %). De ce fait, nous pouvons envisager d'utiliser ces règles de décision dans un contexte post-accidentel afin de proposer un classement, selon les deux niveaux de contamination radioactive, de différents légumes-feuilles contaminés par un rejet accidentel de ^{90}Sr durant leur croissance.

CONCLUSION ET PERSPECTIVES

L'objectif de ce travail de thèse a été le développement d'une méthode permettant l'identification des facteurs (agronomiques, radioécologiques,…) responsables des niveaux de contamination radioactive des végétaux. La méthodologie développée, a permis de répondre à cette problématique environnementale de pollution. Appliquée à un scénario simplifié de contamination, cette méthodologie a mis en évidence deux variables très importantes dans la contamination des légumes-feuilles, à la suite d'un rejet accidentel : le rapport de captation, qui correspond à la fraction du dépôt radioactif interceptée par la masse foliaire des végétaux se trouvant à la surface du sol, et le délai entre le dépôt de radioactivité et la récolte du végétal. De ce fait, le stade de développement de la culture, au moment du dépôt de radioactivité, a été identifié comme une variable clé, en grande partie responsable du niveau de contamination des différents légumes-feuilles étudiés. Les arbres de discrimination construits nous permettent également de disposer de règles de décision, qui peuvent être utilisées dans un contexte post-accidentel afin de proposer des recommandations. En particulier, pour le scénario de contamination étudié, nous pouvons envisager un classement des légumes-feuilles (laitue, chou, épinard et poireau) en fonction de leur niveau de radioactivité (commercialisation ou non-commercialisation) à la suite d'un rejet accidentel de ^{90}Sr dans l'atmosphère. Par ces premiers résultats très encourageants, les arbres de discrimination ont été identifiés comme des outils prometteurs dans le domaine de la radioécologie pour leur côté décisionnel et explicatif. La structure de l'arbre est facilement exploitable et son interprétation est à la portée de non spécialistes du domaine. Leur utilisation est particulièrement intéressante dans un contexte post-accidentel pour faciliter la prise de décision.

Il est à noter que les interprétations obtenues sont liées à l'équation du modèle radioécologique utilisée. Une équation issue d'un autre code de calcul pourrait modéliser la contamination des végétaux d'une manière différente et faire intervenir d'autres variables. Le choix du modèle de culture STICS présente aussi des limites. Ce modèle a été utilisé pour caractériser le rapport de captation et le rendement d'une production de laitue. Les simulations ont été restreintes à deux grandes zones de production. A défaut de valeurs disponibles, les résultats relatifs au rapport de captation de la laitue ont été extrapolés pour les autres légumes-feuilles étudiés. De plus, toutes les relations entre les variables explicatives n'ont pas pu être prises en compte. La relation liant le rendement au temps de croissance, mise en évidence pour la laitue, n'a pu être considérée pour les deux autres légumes-feuilles du fait d'un manque d'information (excepté pour l'épinard,

mais la relation n'est basée que sur trois données). De même, la constante de décroissance biomécanique dépend du végétal étudié et de son développement au moment du dépôt de radioactivité mais ces relations n'ont pas pu être considérées du fait de l'absence de données et de modèle permettant de les prendre en compte.

L'un des principaux inconvénients des arbres de décision est leur instabilité, assez gênante pour les praticiens (Rakotomalala, 2005). Nous l'avons illustrée empiriquement : plusieurs échantillons issus d'un même modèle probabiliste ne conduisent pas à des arbres de même structure. D'un arbre à l'autre, les règles de décision extraites ne sont pas toujours identiques, même si le plus souvent, elles conduisent à la même action. Dans notre contexte d'application, le décideur a besoin de disposer de règles de décision robustes. C'est pourquoi nous avons proposé la méthode REN qui, par une procédure de stabilisation des nœuds intermédiaires, permet la construction d'un arbre de discrimination dont les règles de décision sont plus stables. Les différentes comparaisons empiriques entre cette méthode et la méthode CART ont montré que la méthode REN permet de réduire la variabilité associée aux arbres de discrimination. La mesure de similarité que nous avons introduite a permis de mettre en évidence la stabilité des arbres de discrimination construits selon cette méthode. En ce qui concerne les performances de prédiction, la méthode REN est de manière générale meilleure que la méthode CART. Toutefois, nous avons montré que la différence entre les erreurs de classement est relativement faible. Les méthodes basées sur l'agrégation d'arbres comme les principes du bagging, boosting ou les forêts aléatoires (random forests) restent meilleures en termes de prédiction mais ne permettent pas de conserver la structure de l'arbre.

Suite à ces travaux, différentes perspectives sont envisageables :

1. Sur le plan applicatif :

- La complexification du scénario de contamination. Dans un premier temps, les conditions du dépôt de radioactivité pourraient être généralisées. Nous avons choisi de travailler sur la contamination de végétaux par voie sèche mais la voie humide pourrait également être considérée. Dans ce cas, l'équation de la contamination des végétaux (2) étudiée dans ce document se décomposerait en deux parties : la contamination suite à un transfert par voie sèche et la contamination suite à un transfert par voie humide. Des travaux supplémentaires seraient alors nécessaires pour caractériser la variable rapport de captation par temps humide. Le scénario de contamination pourrait également être étendu à d'autres productions agricoles

comme les céréales, les légumes-racines et les fruits. Dans ce cas, une variable supplémentaire devrait être prise en compte : la translocation qui modélise la migration des radionucléides vers la partie consommable de la plante. Des travaux récents, effectués dans le cadre du projet SENSIB portent sur l'étude de la sensibilité radioécologique de la vigne (Levain et al., 2006) et permettraient d'étayer une telle étude. En effet, lors de ces travaux différentes simulations ont été effectuées à l'aide du modèle de culture STICS pour caractériser l'évolution du rapport de captation en fonction du développement de la culture ainsi que le rendement cultural. Ces résultats pourraient être analysés, comme pour ceux obtenus pour la laitue, en vue de la construction d'un arbre de discrimination relatif à la contamination d'un vignoble. D'autres radionucléides pourraient également être étudiés. L'étude s'est restreinte à un radionucléide, le strontium 90, qui fait partie des radionucléides potentiellement émis en cas d'accident. D'autres pourraient être traités, par exemple le césium 137 ou des radionucléides de période radioactive plus courte comme le zirconium 95 ou l'iode 131. Le transfert des radionucléides par voie racinaire pourrait également être pris en compte.

Cette étude s'est focalisée sur un scénario de contamination accidentelle (ponctuelle et de forte intensité) mais l'on pourrait également travailler sur des contaminations chroniques, de plus faible intensité mais continues dans le temps. Ce type de scénario nous conduirait alors à utiliser un modèle radioécologique de transfert des radionucléides suite à l'émission chronique de radionucléides, comme le code de calcul FOCON (Rommens et al., 1999).

- L'application de cette méthodologie à d'autres thématiques radioécologiques. Par exemple, il serait intéressant, après avoir travaillé sur la contamination de la chaîne alimentaire, d'étudier les facteurs conduisant à dépasser la valeur limite sur la dose relative à l'ingestion d'aliments contaminés (la valeur limite est fixée à 1mSv/an pour la population). Suite aux travaux effectués dans le cadre de notre scénario de contamination, nous avons travaillé sur la dose due à l'ingestion, au cours d'une année, des quatre légumes-feuilles plus ou moins contaminés. Les valeurs de dose obtenues étaient alors très faibles, ne dépassant quasiment jamais la valeur limite. De ce fait, il faudrait envisager de travailler sur l'ensemble de la chaîne alimentaire pour pouvoir identifier les facteurs qui conduisent à dépasser les niveaux admissibles.

- L'identification des légumes « sensibles » et des valeurs d'activités associées. Nous pouvons envisager d'exploiter les résultats de l'arbre de discrimination relatif aux

légumes-feuilles d'une manière originale pour identifier les légumes les plus « sensibles » dans certaines branches de l'arbre, c'est-à-dire les légumes présentant les plus fortes valeurs de contamination suite à un accident radioactif. Par exemple, pour l'arbre de discrimination relatif aux légumes-feuilles, l'épinard est le légume le plus sensible vérifiant la règle (R.8) (Cf. Figure 8.k, graphe gauche). Une telle exploitation permettrait de déterminer les caractéristiques des échantillons à mesurer pour garantir que l'ensemble d'une production est inférieure à un niveau de contamination fixé. L'identification des légumes sensibles et de leurs valeurs d'activité paraît alors particulièrement intéressante en contexte post-accidentel pour proposer des procédures de prélèvements sur le terrain.

2. Sur le plan méthodologique :

- L'amélioration de la méthode d'échantillonnage. En ce qui concerne la génération des échantillons artificiels de données, nous avons choisi une méthode facile à mettre en œuvre, l'échantillonnage aléatoire simple. Une autre méthode alternative pourrait être utilisée : l'échantillonnage stratifié ou par hypercubes latins (McKay, 1979). Cette méthode consiste à diviser le domaine de variation de chaque variable en n intervalles équiprobables et, pour chaque variable, à choisir une valeur dans chacun de ces intervalles. Les valeurs tirées sont alors associées de façon aléatoire. Contrairement à la méthode d'échantillonnage aléatoire simple, cette technique permettrait de mieux couvrir l'espace des variables explicatives.

- Une méthode d'élagage alternative. Nous avons utilisé une méthode d'élagage proposée par Quinlan (1987), appelée *reduced error pruning*. Cette méthode nécessite un deuxième échantillon dit de validation. Dans certaines applications, les données sont rares et la taille de l'échantillon ne permet pas le découpage de ce dernier en échantillon d'apprentissage et de validation. Une autre méthode d'élagage pourrait alors être utilisée. Parmi les méthodes existantes, deux peuvent être envisagées, nécessitant le seul usage de l'échantillon d'apprentissage. La première, appelée *pessimistic pruning* (Quinlan, 1987), consiste à estimer l'erreur de prédiction à partir des données de l'échantillon d'apprentissage. Conscient que le taux d'erreur obtenu est trop optimiste, Quinlan propose d'introduire dans l'estimation une correction de continuité basé sur la distribution binomiale, permettant d'obtenir un taux d'erreur plus réaliste. La deuxième méthode d'élagage, *error-based pruning* (Quinlan, 1993), peut être vue comme une amélioration de la précédente. La principale nouveauté réside dans le fait que l'on peut remplacer une branche par une portion de celle-ci.

Dans certaines situations, cette opération peut être préférée à l'élagage de la branche.

- L'utilisation de la mesure de similarité. Dans la deuxième partie de ce document, cette mesure a été calculée à partir de deux types de pondérations. D'autres poids pourraient être utilisés, les résultats pourraient alors être comparés aux précédents pour juger de l'impact de ces nouvelles pondérations. Nous avons restreint la comparaison des arbres par cette mesure aux méthodes CART et REN. D'autres méthodes pourraient être confrontées, du moment que les arbres construits ont la même architecture (dans le sens arbre binaire, n-aire,…). Il paraît plus difficile, par exemple, de comparer un arbre binaire obtenu par la méthode CART et un arbre n-aire obtenu par la méthode CHAID.

  La mesure de similarité pourrait également être adaptée au cas où les variables de division sont qualitatives, en modifiant l'expression de la dissimilarité (12) au nœud t. De plus, la mesure pourrait être enrichie par la prise en compte des feuilles de l'arbre. Il suffirait alors d'ajouter dans (13) un terme permettant la comparaison de deux feuilles issues d'arbres de discrimination distincts. En considérant cette nouvelle mesure, les conclusions obtenues pourraient être identiques ou au contraire s'éloigner des précédentes.

- La comparaison des performances de prévisions avec d'autres méthodes. Nous avons restreint la comparaison des performances de la méthode REN à la méthode CART et aux méthodes basées sur l'agrégation d'arbres de discrimination. D'autres méthodes pourraient être appliquées. Par exemple, une régression logistique pourrait être mise en œuvre. Nous pensons également à une méthode d'apprentissage supervisé plus récente, les SVM (Support Vector Machines) (Vapnik, 1995). Développée pour des problèmes de classement binaire, cette méthode est basée sur des fonctions noyaux permettant une séparation optimale des données. La comparaison des erreurs de prévision avec ces méthodes ou d'autres nous permettrait de mieux situer la méthode REN par rapport à ces diverses méthodes de classement.

REFERENCES

Atkinson E.J. and Therneau T.M., (2000). An Introduction to Recursive Partitioning Using the RPART Routines. Mayo Fondation, 33 p.

Bartzis J., Ehrhardt J., French S., Lochard J., Morrey M., Papamichail K.N., Sinkko K. and Sohier A., (2000). RODOS: decision support for nuclear emergencies. Decision Making: Recent Developments and Worldwide Applications, Kluwer Academic Publishers, pp 381-395.

Bataille C. et Croüail P., (2005). Analyse des dispositifs réglementaires concernant le contrôle et le suivi de la contamination des sols, des denrées alimentaires et des produits commerciaux en Biélorussie. CEPN, rapport n°291, 54 p.

Bernier J., Parent E. et Boreux J.J., (2000). Statistique pour l'environnement. Traitement bayésien de l'incertitude. Editions TEC & DOC, 363 p.

Breiman L., (1994). Bagging Predictors. Technical Report N°421, University of California, Department of Statistics.

Breiman L., (1996a). Heuristics of instability and stabilization in model selection. The Annals of Statistics, Vol 24, N°6, pp 2350-2383.

Breiman L., (1996b). Bagging predictors. Machine Learning, 24 (2), pp 123-140.

Breiman L., (1996c). Technical Note: Some Properties of Splitting Criteria. Machine Learning, 24(1), pp 41-47.

Breiman L., (1996d). Arcing classifiers. Technical report, University of California, Department of Statistics.

Breiman L., (2001). Random Forests. Machine Learning, 45(1), pp 5-32.

Breiman L., (2002). Manual-Setting Up, Using, And Understanding Random Forests V3.1, 29 p, (ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_V3.1.pdf).

Breiman L., Friedman J.H., Olshen R. and Stone C.J., (1984). Classification and Regression Trees, Chapman & Hall, 358 p.

Briand B., Durand V. and Mercat-Rommens C., (2008). Identifying the relationships between agronomic and radioecological variables using a crop model applied to lettuce. Journal of Agronomy sous presse.

Briand B. et Mercat-Rommens C., (2006a). Utilisation de séries de mesures environnementales pour étudier la sensibilité de la contamination végétale aux rejets chroniques – Projet SENSIB. Rapport IRSN/DEI/SESURE 06-07, 77 p.

Briand B. et Mercat-Rommens C., (2006b). Difficulties and lessons of environmental data processing to fit modeling parameters. SETAC Europe 16th Annual Meeting, 7-11 mai 2006, The Hague, Nederland.

Briand B., Mercat-Rommens C. et Ducharme G., (2006). Apports de la biostatistique à la radioécologie de terrain. 38ièmes Journées de Statistique, 29 mai au 2 juin 2006, Clamart.

Briand B., Mercat-Rommens C. and Ducharme G., (2007). Using classification trees techniques like sensitivity analysis in the field radioecology. Fifth International Conference on Sensitivity Analysis of Model Output, 18-22 june 2007, Budapest, Hungary.

Brisson N., Gary C., Justes E., Mary B., Roche R., Ripoche D., Zimmer D., Sierra J., Bertuzzi P., Burger P., Bussière F., Cabidoche Y. M., Cellier P., Debaeke P., Gaudillère J. P., Maraux F., Seguin F.B. and Sinoquet H., (2003). An overview of the crop model STICS. European Journal of Agronomy, 18, pp 309-332.

Brisson N., Mary B., Ripoche D., Jeuffroy M.H., Ruget F., Nicoullaud B., Gate P., Devienne, F., Antonioletti R., Dürr C., Richard G., Beaudoin N., Recous S., Tayot X., Plénet D., Cellier P., Machet J.M., Meynard J.M. and Delécolle R., (1998). STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parametrization applied to wheat and corn. Agronomie: agriculture and environment, 18, pp 311-346.

Brown J. and Simmonds J.R., (1995). FARMLAND, a dynamic model for the transfer of radionuclides through terrestrial foodchain. NRPB report n°R273.

Calmon P. et Mourlon C., (2003). Equations et paramètres du logiciel ASTRAL V2.1. Rapport IRSN DPRE/SERLAB 03-16, 113 p.

Celeux G. et Nakache J.P., (1994), Analyse discriminante sur variables qualitatives, Polytechnica, 270 p.

Chamberlain A.C., (1970). Interception and retention of radioactive aerosols by vegetation. Atmospheric Environment n°4, pp 57-78.

Codex Alimentarius Commission (1989). Guideline Levels for Radionuclides in Foods following accidental Nuclear Contamination for use in International Trade, CAC/GL 5.

Combris P., Bertail P., Boizot C. et Poupa JC., (1995). La consommation alimentaire en 1991 : distribution des quantités consommées à domicile. Observatoire des consommations alimentaires.

Cukier R.I., Fortuin C.M., Shuler K.E., Petschek A.G. and Schaibly J.H., (1973) Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I theory. Journal Chemical Physics, 59(8), pp 3873–3878.

Cukier R.I., Levine R.I. and. Shuler K.E., (1978). Nonlinear sensitivity analysis of multiparameter model systems. Journal Computational Physics, 26(1), pp 1–42.

Cukier R.I., Shuler K.E. and Schaibly J.H., (1975). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients - analysis of the approximations. Journal Chemical Physics, 63(3), pp 1140–1149.

Dannegger F., (2000). Tree stability diagnostics and some remedies for instability. Statistics in Medicine, 19, pp 475-491.

Delboe A. et Mercat-Rommens C., (2005). Étude régionalisée de l'impact d'une pollution radioactive accidentelle sur le blé d'hiver- Projet SENSIB. Rapport IRSN DEI/SESURE 05-15.

Département de Protection Sanitaire (CEA/IPSN/DPS), (1961-1980). Surveillance de la radioactivité de la chaîne alimentaire et de prélèvements divers. Bulletins trimestriels de mesures du Service d'Hygiène Atomique, Fontenay-aux-Roses.

De Tourdonnet S., (1998). Maîtrise de la qualité et de la pollution nitrique en production de laitue sous abri plastique : diagnostic et modélisation des effets des systèmes de cultures. Thèse Institut National Agronomique Paris-Grignon, Paris, 192p.

Dietterich T.G., (1999). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Tress: Bagging, Boosting and Randomization. Machine Learning, pp 1-22.

Drucker, H., (1997). Improving Regressors using Boosting Techniques. In: Kauffman, M. (Ed.), Proceedings of the 14th International Conference on Machine Learning, pp 107-115.

Duffa C., Antonelli C., Vray F., Marant M.J. et Salaun G., (2007). La base de données SYLVESTRE au service de la Qualité et de l'expertise en radioécologie à l'IRSN. Radioprotection, Vol. 42, n° 2, pp 219-225.

Durand V., Briand B. et Mercat-Rommens C., (2006). Simulations avec le modèle agronomique STICS pour l'étude de la culture de la laitue – Projet SENSIB. Rapport IRSN DEI/SESURE 06-48, 37 p.

Durand V. et Mercat-Rommens C., (2006). Etude régionalisée de l'impact d'une pollution radioactive accidentelle sur une prairie permanente – Projet SENSIB. Rapport IRSN DEI/SESURE 06-01, 46 p.

Durand V., Mercat-Rommens C., Briand B., Levain A. et Besson B, (2007a). Utilisation du logiciel STICS pour l'évaluation des conséquences d'une pollution radioactive accidentelle du milieu agricole. Séminaire STICS-INRA, 20 au 22 mars 2007, Reims.

Durand V., Mercat-Rommens C., Curmi P., Benoit M. and Briand B., (2007b). Modelling regional impacts of radioactive pollution on permanent grassland. Journal of Agronomy, 6 (1), pp 11-20.

Efron B. and Tibshirani R.J., (1993). An introduction to the bootstrap. New York: Chapman and Hall, 436 p.

Esposito F., Malerba D. and Semeraro G., (1997). A comparative analysis of methods for pruning decision trees. IEEE Trans. Pattern. Anal. Mach. Intell., 19(5), pp 476-491.

Foulquier L. et Bretheau F., (1998). Les installations nucléaires et l'environnement. EDP Sciences, Les Ulis (France). Collection IRSN, 171 p.

Freund Y. and Schapire R.E., (1995). A decision-theoric generalization of on-line learning and application to boosting. Proceedings of the Second European Conference of Computational Learning Theory.

Freund Y. and Schapire R.E., (1996). Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the Thirteenth International Conference.

Freund Y. and Schapire R.E., (1999). A short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence, 14(5), pp 771-780.

Gey S. and Poggi J.M., (2006). Boosting and instability for regression trees. Computational Statistics & Data Analysis 50, pp 533-550.

Ghattas B., (1999). Importance des variables dans les méthodes CART, Revue de Modulad, 24, pp 17-28.

Ghattas B., (2000). Agrégation d'arbre de classification. Revue de Statistique Appliquée, tome 48, n°2, pp 85-98.

Golub G.H. and Pereyra V., (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM J. Numerical Analysis, 10(2), pp 413-432.

Groupe Radioécologie Nord-Cotentin, Rapport Technique (2001). Identification et estimation des incertitudes associées aux évaluations de doses obtenues par le modèle mathématique du Groupe Radioécologique Nord-Cotentin. Rapport IPSN DPRE/SERNAT 01-13, 176 p.

Gueguen A. et Nakache J.P., (1988). Méthode de discrimination basée sur la construction d'un arbre de décision binaire. Revue de Statistique Appliquée, tome 36, n°1, pp 19-37.

IAEA, (1994). Handbook of parameter values for the prediction of radionuclides transfer in temperate environment. International Atomic Energy Agency: technical reports series n°364.

Institut de Radioprotection et de Sûreté Nucléaire, (2004). La radioécologie, Connaître et comprendre l'évolution des niveaux de radioactivité dans l'environnement. Les livrets de l'IRSN, 21 p.

Jones J.A., Mansfield P.A., Haywood S.M., Hasemann I., Steinhauer C., Ehrhardt J. and Fraude D., (1995). PC COSYMA (Version 2): an accident consequence assessment package for use on a PC. EUR report 16239.

Kass G., (1980). An exploratory technique for investigating large quantities of categorical data, Applied Statistics, 29(2), pp 119-127.

Larousse agricole (1981) – Direction de Jean-Michel Clément, 1207 p.

Larousse agricole (2002) - Direction de Marcel Mazoyer, Le monde paysan au XXIème siècle, 767 p.

Larue C., Durand V. et Mercat-Rommens C., (2007). Etude de la sensibilité d'une culture de pommes de terre à une pollution radioactive accidentelle. Rapport IRSN DEI/SESURE 07-62.

Le Bohec J., Erard P. et Leteinturier J., (1993). Le Poireau guide pratique. Ctifl (Centre technique interprofessionnel des fruits et légumes), 185 p.

Levain A., Mercat-Rommens C. et Roussel-Debet S., (2006). Etude de la sensibilité radioécologique de la vigne à une pollution radioactive accidentelle. Rapport IRSN DEI/SESURE 06-54, 49 p.

Masson O., Saey L., Paulat P. et Bois E., (2005). Etude évènementielle du lessivage de l'atmosphère. SFRP 2005, Nantes.

Maubert et al., (1991). Programme Réhabilitation des Sols et des Surfaces après un Accident. Acquis expérimentaux 1985-1990. Note RESSAC 04/91.

McKay M.D., Beckman R. and Conover W., (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2), pp 239–245.

Météo France (1996). Le climat de la France. CD-ROM.

Météo France (2005). Données climatiques journalières pour une année calendaire relatives à deux stations française : Orange et Rennes.

Mercat-Rommens C., Métivier J., Briand B. and Durand, V., 2006. How geostatistics can help in predicting the level of radioactive contamination of cereals. 6th European Conference of Geostatistics for Environmental Applications, 25- 27 October 2006, Rhodes, Grèce.

Mercat-Rommens C. and Renaud P., (2005). From radioecological sensitivity to risk management: the SENSIB Project. Second International Conference Radioactivity in the Environment, October 2005, Nice, France.

Mercat-Rommens C., Roussel-Debet S., Briand B., Durand V., Besson B. et Renaud P., (2007). La sensibilité radioécologique des territoires : vers un outil opérationnel. Radioprotection Vol. 43, n°3, pp 177-295.

Mingers J., (1989). An Empirical Comparison of Pruning Methods for Decision Tree Induction. Machine Learning, 4, pp 227-243.

Mishra S., Deeds N.E. and RamaRao B.S., (2003). Application of classification trees in the sensitivity analysis of probabilistic model results. Reliability Engineering and System Safety 79, pp 123-129.

Mokhtari A., Frey H.C and Jaykus L.A., (2006). Application of Classification and Regression Trees for Sensitivity Analysis of the Escherichia coli O157:H7 Food Safety Process Risk Model. International Association for Food Protection. Journal of Food Protection, Volume 69, Number 3, pp 609-618.

Morgan J.A. and Messenger R.C., (1972). A modal search technique for predictive nominal scale multivariate analysis, J. Amer. Statis. Ass. 67, pp 768-772.

Morgan J.A. and Sonquist J.N., (1963). Problems in the analysis of survey data, and a proposal. J. Amer. Stat. Assoc., 58(302), pp 415-434.

Morris M.D., (1991). Factorial sampling plans for preliminary computational experiments. Technometrics, 33(2), pp 161-174.

Mourlon C. and Calmon P., (2002). ASTRAL: a code for assessing situations after a nuclear accident. 12th annual meeting of SETAC Europe, 12-16 May 2002, Vienna, Austria.

Müller H. and Pröhl G., (1993). ECOSYS-87: a dynamic Model for assessing radiological consequences of nuclear accidents. Health Physics, vol. 64, n°3, pp 232-252.

Nisbet A.F., Woodman R.F.M. and Haylock R.G.E., (1999). Recommended soil-to-plant transfer factors for radiocesium and radiostrontium for use in arable systems. National Radiological Protection Board, NRPB-R304, Chilton, 50 p.

Quinault J.M., Cartier Y. et Bourdeau F., (1989). Guide d'évaluation de l'impact de rejets radioactifs atmosphériques. Electricité de France, Direction de l'équipement.

Quinlan J.R., (1986). Induction of decision trees. Machine Learning,1(1), pp 81-106.

Quinlan J.R., (1987). Simplifying decision trees. Int. J. Man-Machine Studies, 27, pp 221-234.

Quinlan J.R., (1993). C4.5. Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.

Rakotomalala R., (2005). Arbres de Décision. Revue de Modulad, 33, pp 163-187.

Renaud P., Champion D. et Brenot J., (2007). Les retombées radioactives de l'accident de Tchernobyl sur le territoire français – Conséquences environnementales et exposition des personnes. Collection sciences et techniques, Editions TEC & DOC, 190 p.

Renaud P., Louvat D. et Vray F., (2003). Les retombées en France des essais atmosphériques d'armes nucléaires. Production, fractionnement, dispersion atmosphérique et dépôt des produits de fission. Rapport IRSN DEI/SESURE 03-03, 27p.

Renaud P., Maubert H. et BERNIÉ J.C., (1997a). Prise en compte des paramètres contextuels en radioécologie post-accidentelle. Radioprotection, Vol. 32, n°2, pp 181-195.

Renaud P., Maubert H. et Duffa C., (1997b). Contamination des productions agricoles de base suite à une émission atmosphérique accidentelle. Deuxième partie : Le modèle ASTRAL Crise. Document SERE 97-018.

Renaud P., Réal J., Maubert H and Roussel-Debet S., (1999). Dynamic modelling of the cesium, strontium and ruthenium transfer to grass and vegetables. Health Physics, 76(5), pp 495-501.

Robert C. and Casella G., (1999). Monte Carlo Statistical Methods. Springer.

Rommens C. (1997). Etude bibliographique et choix des données par défaut pour les logiciels de calcul des impacts dosimétriques. Note technique SEGR/SAER/97 n°25, 29 p.

Rommens C., Morin A. et Merle-Szeremeta A., (1999). Le modèle FOCON d'évaluation de l'impact dosimétrique des rejets radioactifs atmosphériques des installations nucléaires en fonctionnement normal. Radioprotection – vol 34, n°2, pp 195-209.

Roussel-Debet S. et Duffa C., (2005). Radioactivité en ^{137}Cs dans l'environnement terrestre des sites électronucléaires d'Electricité de France - Interprétation des données acquises de 1989 à 2004. Rapport IRSN DEI/SESURE/LERCM 05-36, 70 p.

Roussel-Debet S., Masson O. et Salaun G., (2005). Radioactivité en ^{137}Cs de l'environnement terrestre français - Interprétation des données OPERA acquises de 1993 à 2004. Rapport IRSN DEI/SESURE/LERCM 05-10, 81 p.

Ruey Hsia L., (2001). Instability of decision tree classification algorithms. PhD Thesis, University of Illinois, Urbana-Champaign, 86 p.

Saltelli A., Chan K. and Scott M., (2000). Sensitivity Analysis, John Wiley & Sons publishers, Probability and Statistics series, 475 p.

Santucci P., (1995). Manuel d'utilisation du code ABRICOT, version 2.0. Rapport IPSN, DPEI/SERGD/LESTS 95-03.

Schaibly J.H. and Shuler K.E., (1973). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficeints. Part II Applications. Journal Chemical Physics, 59, pp 3879–3888.

Serfling R.J., (1980). Approximation theorems of mathematical statistics. John Wiley & Sons publishers, Probability and Mathematical Statistics series.

Service Central de protection contre les Rayonnements Ionisants (SCPRI), (1961-1978). Bulletins mensuels de résultats de mesures. République Française, Ministère de la Santé.

Shannon C. and Weaver W., (1949). The Mathematical Theory of Communication. University of Illinois Press, Urbana.

Simmonds J.R., Lawson G. and Mayall A., (1995). Methodology for assessing the radiological consequences of routine releases of radionuclides to the environment. Report EUR 15760,Radiation Protection 72 , National Radiological Protection Board, 350 p.

Sobol I.M., (1993). Sensitivity estimates for nonlinear mathematical models. Mathematical Modelling and Computational Experiments, 1(4), pp 407–414.

Thicoïpé J.P., (1997). Laitues. Centre Technique Interprofessionnel des Fruits et Légumes (CTIFL) - SERAIL, 281 p.

Turányi T., (1990). Sensitivity analysis of complex kinetic system, tools and applications. Journal of Mathematical Chemistry, 5(3), pp 203–248.

Turányi T. and Rabitz H., (2000). Local methods. In: Saltelli A., Chan K., Scott M., Sensitivity Analysis, John Wiley & Sons publishers, Probability and Statistics series, 2000, pp 81-99.

UNSCEAR, (1982). Rayonnements ionisants : sources et effets biologiques. Rapport à l'Assemblée Générale du Comité Scientifique des Nations Unies pour l'étude des effets des rayonnements ionisants.

UNSCEAR, (2000). Sources and effects of ionizing radiation. United Nations Scientific Committee on the Effects of Atomic Radiation. Report to the General Assembly.

Vapnik V., (1995). The Nature of Statistical Learning Theory. Springer Verlag, New York.

Vray F., (2002). Contamination des végétaux par dépôt atmosphérique chronique de ^{90}Sr : informations tirées de l'étude de la période 1961-1980. Rapport IRSN DPRE/SERNAT 02-29, 64 p.

Vray F. et Renaud P., (2004). Contamination de la chaîne alimentaire par les produits de fission émis lors des essais aériens d'armes nucléaires. Rapport IRSN DEI/SESURE 04-19, 42 p.

ANNEXES

ANNEXE A : Extrait d'un rapport CEA/DPS

RADIOACTIVITE DES LEGUMES ET FRUITS — Mois : MAI 1970

Réseau : D.P.S. — Prélèvements et Mesures effectués par : D.P.S.

LOCALITE DE REFERENCE OU REGION DE PRELEVEMENT		LANGUEDOC			PROVENCE		
NATURE DE L'ECHANTILLON		Carottes	Poireaux	Salades	Carottes	Poireaux	Salades
1		2	3	4	5	6	7
Cendres	g/kg. frais	8,0	11,0	11,0	7,7	11,4	12,1
Activité β totale *	pCi/kg. frais	2632	2523	3080	1476	2092	2601
^{40}K	pCi/kg. frais	1897	1861	2235	1524	2146	2587
^{90}Sr équivalent	pCi/kg. frais	72	64	102	18	51	185
	pCi/g. Ca	95	77	116	47	43	111
^{90}Sr vrai	pCi/kg. frais	13,9	24,7	19,6	7,5	17,9	17,5
	pCi/g. Ca	18,3	29,6	22,2	19,5	15,0	10,5
^{137}Cs	pCi/kg. frais	0	0	0	0	0	0
	pCi/g. K	0	0	0	0	0	0

LOCALITE DE REFERENCE OU REGION DE PRELEVEMENT		COTE D'AZUR			ROUSSILLON		
NATURE DE L'ECHANTILLON		Carottes	Poireaux	Salades	Salades		
8		9	10	11	12	13	14
Cendres	g/kg. frais	8,5	10,3	14,7	21,0		
Activité β totale *	pCi/kg. frais	2061	2184	3310	6741		
^{40}K	pCi/kg. frais	2167	1011	3049	4998		
^{90}Sr équivalent	pCi/kg. frais	22	70	116	255		
	pCi/g. Ca	41	70	60	144		
^{90}Sr vrai	pCi/kg. frais	9,7	14,7	15,5	36,0		
	pCi/g. Ca	18,5	14,8	8,0	20,3		
^{137}Cs	pCi/kg. frais	0	0	0	0		
	pCi/g. K	0	0	0	0		

* Etalonnage ^{40}K

0,0 = Elément recherché et non trouvé
* = Absence de mesure
< LD = Inférieur à la limite de détection.

ANNEXE B : Utilisation de séries de mesures environnementales pour étudier la sensibilité de la contamination végétale aux rejets chroniques

Dans cette annexe, nous présentons les travaux effectués pour procéder à l'estimation de trois principaux paramètres du transfert aux légumes-feuilles suite à un rejet chronique de radionucléides : le rapport de captation, le facteur de transfert racinaire et la constante de décroissance de la biodisponibilité du radionucléide dans le sol. Dans un premier temps, nous présentons l'équation de la contamination des végétaux étudiée ainsi que les différents jeux de données disponibles. Dans un deuxième temps, nous présentons les méthodes d'estimation utilisées et les principaux résultats obtenus.

1 Les différentes composantes de l'équation de la contamination des végétaux et les données associées

La modélisation de la contamination des végétaux repose sur une équation de calcul de l'activité massique ($Bq.kg^{-1}$ frais) d'un végétal suite à un dépôt continu sur un pas de temps donné. Selon Vray (2002), cette expression est voisine de celle utilisée dans la plupart des codes de calcul opérationnels de l'impact d'un rejet chronique, notamment FOCON (Rommens et al., 1999), mais également ABRICOT (Santucci, 1995). Elle correspond également à l'intégration de l'expression de la contamination des végétaux suite à un rejet aigu figurant dans ASTRAL (Mourlon et Calmon, 2002) ou ECOSYS (Müller et Pröhl, 1993). L'objectif de ce travail n'est pas de discuter la pertinence de cette modélisation mais de la confronter au mieux avec les résultats de mesures environnementales disponibles.

La première partie de l'équation calcule la contamination moyenne, due au transfert foliaire, d'un végétal durant son temps de croissance. Le dépôt est sommé sur les mois composant le temps de croissance du végétal ($\Delta_{t_d-\alpha+1},\ldots,\Delta_{t_d}$, t_i représente la date à la sortie de l'intervalle de temps). La deuxième partie de l'équation permet de calculer la contamination par voie racinaire. Pour cela, l'ensemble des dépôts antérieurs doit être pris en compte. En effet, il ne faut pas négliger le dépôt cumulé des années précédentes et remonter au dépôt le plus ancien (si cela est possible). En notant Δ_{t_1} l'intervalle de temps « origine », c'est-à-dire le pas de temps à partir duquel le sol a été contaminé pour la première fois, le calcul de l'activité du végétal (par transfert racinaire) se fait par la multiplication du facteur de transfert racinaire par la somme des dépôts antérieurs (multipliés par une exponentielle décroissante) et du dépôt durant l'intervalle de temps Δ_{t_d} . Dans les deux cas de contamination, le dépôt est supposé constant sur le mois.

De plus, les calculs sont effectués pour un végétal à maturité (prêt à être consommé). L'expression de cette équation est la suivante :

$$C_{r,v}^{chron} = \underbrace{\frac{R_c^{chron}}{Rdt}\left(\frac{D_{\Delta_{t_d}}^{tot}}{\Delta_{t_d}}\frac{1-e^{-(\lambda_b+\lambda_r)\Delta_{t_d}}}{\lambda_b+\lambda_r} + \sum_{i=t_{d-\alpha+1}}^{t_{d-1}} \frac{D_{\Delta_i}^{tot}}{\Delta_i}\frac{1-e^{-(\lambda_b+\lambda_r)\Delta_i}}{\lambda_b+\lambda_r} e^{-(\lambda_b+\lambda_r)\sum_{j=i+1}^{t_d}\Delta_j}\right)}_{transfert\ \ foliaire} \quad \text{(B.1)}$$

$$+ \underbrace{\frac{FT_r}{\mu h}\left(\sum_{i=t_1}^{t_{d-1}} D_{\Delta_i}^{tot} e^{-(\lambda_s+\lambda_r)\sum_{j=i+1}^{t_d}\Delta_j} + D_{\Delta_{t_d}}^{tot}\right)}_{transfert\ \ racinaire}$$

Avec :

R_c^{chron} (sans dimension) : rapport de captation,

Rdt (kg.m^{-2}) : rendement cultural,

λ_b (j^{-1}) : constante de décroissance biomécanique du radionucléide pour le végétal,

λ_r (j^{-1}) : constante de décroissance physique du radionucléide,

T_c (j) : temps de croissance du végétal ou durée de phase végétative,

$D_{\Delta_i}^{tot}$ (Bq par intervalle de temps) : activité déposée durant l'intervalle de temps Δ_i,

Δ_i : intervalle de temps de l'étude,

FT_r (kg de sol sec par kg de végétal frais) : facteur de transfert racinaire,

h (m) : hauteur de l'horizon racinaire,

μ (kg de sol sec par m^3) : masse volumique du sol (sec),

λ_s (j^{-1}) : constante de décroissance de la biodisponibilité du radionucléide dans le sol. Cette constante prend en compte l'ensemble des phénomènes de décroissance écologique (migration en profondeur, érosion, fixation irréversible aux particules minérales, …) mais ne prend pas en compte la décroissance radioactive.

1.1 Les données disponibles

1.1.1 Mesures des activités massiques

Les séries de mesures à disposition sont relatives à la période des essais atmosphériques d'armes nucléaires, elles proviennent des rapports trimestriels produits par l'IRSN entre 1961 et 1980 (Cf. paragraphe 2.2.1). Il s'agit de mesures de radioactivité mensuelles effectuées sur divers végétaux. Les différentes régions françaises dans lesquelles ces prélèvements ont été effectués se retrouvent sur la carte droite de la figure 2.a, présentée dans le chapitre 2. Cette figure montre que ces lieux de prélèvement que l'on

appelle « régions » dans la suite sont en fait des regroupements de départements qui ne correspondent pas exactement à la définition des régions administratives françaises. Comme il a été énoncé précédemment, ces mesures sont anciennes et certaines caractéristiques de prélèvement sont mal renseignées. Les lieux précis sont très rarement indiqués, seule la « région » est identifiée. De plus, les dates de prélèvement sont imprécises : uniquement le mois et l'année sont connus. Il a été choisi d'affecter l'ensemble des résultats de mesures au 28 de chaque mois (Vray, 2002).

Les radionucléides et les végétaux sélectionnés pour réaliser cette étude sont ceux présentant le plus grand nombre de mesures disponibles. Deux grands groupes de radionucléides sont considérés : les radionucléides à vie courte (^{95}Nb+^{95}Zr, ^{144}Ce+^{144}Pr, ^{141}Ce, ^{106}Ru+^{106}Rh et ^{103}Ru) mesurés en région Languedoc et les radionucléides à vie longue (ou moyennement longue) (^{90}Sr et ^{137}Cs) mesurés dans diverses régions françaises. Les végétaux étudiés sont les légumes-feuilles et les périodes d'études ont été choisies en fonction des séries de données disponibles sur les dépôts afin que les périodes d'études correspondent à celles où les dépôts des divers radionucléides ont pu être reconstruits (Renaud et al., 2003). Le nombre de données à disposition pour ces deux grands groupes est présenté dans les tableaux 1 et 2.

LANGUEDOC	^{95}Nb+^{95}Zr	^{144}Ce+^{144}Pr	^{141}Ce	^{106}Ru+^{106}Rh	^{103}Ru
Bette	35	33	25	15	16
Chou	27	19	-	-	-
Chou-fleur	20	12	15	-	-
Poireau	35	22	12	13	-
Salade	100	87	50	50	38
TOTAL	***217***	***173***	***77***	***63***	***54***

Tableau 1 : Données disponibles relatives aux radionucléides à vie courte

	Régions	*Chou*		*Chou-fleur*		*Poireau*		*Salade*	
		^{90}Sr	^{137}Cs	^{90}Sr	^{137}Cs	^{90}Sr	^{137}Cs	^{90}Sr	^{137}Cs
1	***BOURGOGNE LYONNAIS***	47	-	-	-	91	-	73	-
2	***BRETAGNE***	-	-	-	-	153	-	153	26
3	***CÔTE D'AZUR***	-	-	-	-	104	-	108	15
4	***GARONNE***	-	-	-	-	151	29	148	43
5	***LANGUEDOC***	-	36	-	31	116	44	116	193
6	***LORRAINE***	-	-	-	-	97	15	61	18
7	***NORD***	-	-	-	-	135	-	135	21
8	***PARIS***	-		-		150	14	152	20
9	***PROVENCE***	42	122	26	41	168	131	206	203
10	***ROUSSILLON***	-	-	-	-	15	-	102	16
11	***VAL DE LOIRE***	-	-	-	-	50	-	88	-
	TOTAL	***89***	***158***	***26***	***72***	***1230***	***233***	***1342***	***555***

Tableau 2 : Données disponibles relatives aux radionucléides à vie longue

Pour ces 7 radionucléides, environ 8000 résultats de mesures, effectuées dans différentes régions françaises et sur divers végétaux pour la période d'étude considérée, sont à disposition. Les données sélectionnées pour la réalisation de ces travaux correspondent à environ la moitié de l'ensemble des résultats disponibles (4289 prélèvements).
La figure 2.b du chapitre 2 présente une illustration de ces résultats de mesures de radioactivité en fonction des dates de prélèvement pour le légume-feuille poireau et le radionucléide ^{90}Sr. Les résultats de mesures semblent varier d'une région à l'autre, la figure fait apparaître de façon globale de plus fortes activités du ^{90}Sr dans les poireaux prélevés en régions Garonne, Bretagne et Bourgogne Lyonnais, par rapport à ceux prélevés en région Languedoc.

1.1.2. Activité déposée

Il s'agit de l'activité déposée (Bq.m^{-2}) durant un intervalle de temps donné. Les données présentées dans ce paragraphe reposent exclusivement sur les travaux réalisés par Vray (2002) et Vray et Renaud (2004). Pour l'ensemble des radionucléides étudiés, des séries de données d'activité déposée sont disponibles au pas de temps mensuel car les données utilisées pour effectuer les reconstructions ont été traitées à cette échelle de temps : sur la période novembre 1961-février 1980 pour le ^{90}Sr et sur la période juin 1961-juillet 1978 pour les autres radionucléides.
Grâce aux mesures de l'activité dans l'air et l'eau de pluie effectuées au pas de temps mensuel par le Service Central de Protection Contre les Rayonnements Ionisants (SCPRI) sur la période 1960-1980 (à Fontenay-aux-Roses puis au Vésinet), le dépôt mensuel en région parisienne peut être estimé par la formule suivante, utilisée dans la plupart des modèles radioécologiques :

$$D_{\Delta_{t_i}} = D^{\sec}_{\Delta_{t_i}} + D^{hum}_{\Delta_{t_i}} = A^{air}_{\Delta_{t_i}} V_d + A^{eau}_{\Delta_{t_i}} H_{eau}$$

Avec :

$D_{\Delta_{t_i}}$ (Bq.m^{-2}.mois^{-1}) : dépôt mensuel total,

$D^{\sec}_{\Delta_{t_i}}$ (Bq.m^{-2}.mois^{-1}) : dépôt mensuel par temps sec,

$D^{hum}_{\Delta_{t_i}}$ (Bq.m^{-2}.mois^{-1}) : dépôt mensuel par temps de pluie,

$A^{air}_{\Delta_{t_i}}$ (Bq.m^{-3}) : activité moyenne dans l'air durant le mois,

V_d (m.mois^{-1}) : vitesse de dépôt des aérosols radioactifs,

$A^{eau}_{\Delta_{t_i}}$ (Bq.l^{-1}) : activité moyenne dans l'eau de pluie durant le mois,

H_{eau} (l.m^{-2}.mois^{-1}) : hauteur mensuelle des précipitations.

Deux séries complètes de données sur l'activité dans l'air (en région parisienne) ont été constituées pour le $^{95}Zr+^{95}Nb$ et le ^{90}Sr (Vray et Renaud, 2004). Ces séries sont utilisées pour compléter les jeux de données relatifs aux radionucléides à vie courte (en utilisant les rapports d'activité [Radionucléide]/[$^{95}Zr+^{95}Nb$]) et la série de données relative au ^{137}Cs (en utilisant un rapport d'activité constant entre le ^{90}Sr et le ^{137}Cs, [^{137}Cs]/[^{90}Sr]=1,5 (Renaud et al., 2003)). En ce qui concerne le dépôt humide, les séries de données ont été constituées à partir des données de l'activité dans l'eau de pluie et la hauteur de pluie. Dès que celles-ci étaient inconnues, le dépôt humide a été considéré égal à deux fois le dépôt sec (Renaud et al., 2003).

Au niveau régional, l'activité dans l'air et celle dans l'eau de pluie sont considérées partout identiques à celles mesurées en région parisienne. Ainsi, à l'échelle de la France le dépôt sec est le même, et le dépôt humide varie uniquement en fonction des hauteurs de pluies régionales (Vray, 2002). Il est calculé en multipliant l'activité dans l'eau de pluie en région parisienne (quotient du dépôt humide en région parisienne par la hauteur de pluie dans cette même région) par la hauteur de pluie régionale. Les données pluviométriques utilisées pour réaliser ces calculs sont issues du CD-Rom de météo France (1996), « Le climat de la France ». Il s'agit de moyennes mensuelles établies sur 30 ans (1961-1990) pour 113 villes françaises. Ainsi, des moyennes mensuelles régionales des hauteurs de pluies ont pu être calculées à partir des données issues de différentes stations (Tableau 3).

	Régions	*Janv*	*Fév*	*Mars*	*Avril*	*Mai*	*Juin*	*Juil*	*Août*	*Sept*	*Oct*	*Nov*	*Déc*
1	***BOURGOGNE LYONNAIS***	56,7	53,5	57,9	60	86,2	69,5	55,8	71,1	70,9	68,6	67,8	60,6
2	***BRETAGNE***	93,7	77,4	72	53,3	65,7	46,8	41,7	44,8	61,7	81,3	91,4	93,6
3	***CÔTE D'AZUR***	81,1	80,5	65,4	59,8	48,7	31,4	14,6	33,1	56,1	104,3	93,4	76
4	***GARONNE***	83,1	75,5	66,5	66	78,5	58,2	48,1	55,3	61,2	72,4	74,7	80,3
5	***LANGUEDOC***	69,3	68,7	55,6	53,5	53,2	34	19,5	41,1	59,3	115,7	65,2	60,7
6	***LORRAINE***	49,9	47,1	48,2	52,7	74	74,4	59,6	71,5	57,6	50,5	56,5	55,1
7	***NORD***	57,9	45,1	53,6	47,6	53,2	57,5	53,2	50,7	60,7	68,5	75,1	63,8
8	***PARIS***	54,9	47	54	48,8	60,6	51,8	55,4	46,7	53,9	55,2	59	54,8
9	***PROVENCE***	48,5	55,6	51,6	54,5	54,2	39,7	23,5	42,7	56	86,6	59,4	54,5
10	***ROUSSILLON***	49,4	45,2	43,4	51,1	51,5	28,3	17,1	34,4	48,8	92,6	59,9	53
11	***VAL DE LOIRE***	70,5	61,9	59,4	48,2	62,4	44,3	46,8	42,6	53,6	66,2	71,5	69,4

Tableau 3 : Hauteur de pluie moyenne mensuelle régionale (mm)

Pour le ^{90}Sr et le ^{137}Cs, les dépôts qui précèdent la période d'étude ont été pris en compte. Les dépôts de ^{90}Sr consécutifs des retombées atmosphériques des années antérieures ont été pris en compte à un pas de temps annuel en utilisant les données de l'UNSCEAR (1982) (Vray 2002). La valeur affectée à l'année 1957 représente l'ensemble des dépôts antérieurs à cette date. Le tableau 4 regroupe ces différentes données. En utilisant le rapport [^{137}Cs]/[^{90}Sr]=1,5 (UNSCEAR, 2002), les dépôts antérieurs de ^{137}Cs ont été reconstruits, à partir des valeurs issues de ce tableau.

Les séries de données temporelles disponibles pour réaliser les estimations sont les mesures de radioactivité effectuées dans les végétaux ainsi que les dépôts mensuels. Les autres composantes de l'équation à savoir les caractéristiques agronomiques et radioécologiques ne sont pas connues pour chaque observation de la variable réponse (mesure de radioactivité dans un végétal) et doivent être renseignées par des recherches bibliographiques. C'est le cas des entrées de l'équation présentées dans le paragraphe suivant (T_c, Rdt, h, μ et λ_b). Un travail de recherche a donc été effectué afin de trouver les valeurs les plus adaptées pour réaliser les estimations.

Année	Dépôt sur la période ($Bq.m^{-2}$)
1957	470,48
1958	164,11
1959	273,98
1960	68,32
1961 (janv-oct)	66,8

Tableau 4 : Dépôt de ^{90}Sr antérieur à la période d'étude

1.2 Les autres entrées de l'équation

Des valeurs génériques ont été utilisées pour renseigner certaines variables radioécologiques, considérées comme a priori les moins sensibles à un effet de la région (décroissance biomécanique), ainsi que pour l'ensemble des variables agronomiques dont la variabilité régionale n'est pas l'objet de la présente étude.

1.2.1 Les entrées de l'équation de nature radioécologique et radiophysique

1.2.1.1 Les entrées de l'équation qui sont fixées

- La constante de décroissance biomécanique du radionucléide pour le végétal (λ_b)

Cette décroissance rend compte principalement de la dilution de l'activité lors de la croissance du végétal. Pour la plupart des radionucléides, la valeur générique affectée à cette constante est de 0,05 j^{-1} (correspondant à une période d'environ 14 jours) (GRNC, 2001). Cette valeur, correspondant aux valeurs proposées dans les codes de calcul FOCON (Rommens et al., 1999) et FARMLAND (Brown et Simmonds, 1995), a été choisie pour le présent travail.

- La constante de décroissance physique du radionucléide (λ_r)

Elle est fonction du radionucléide. Les valeurs associées à ce paramètre sont regroupées dans le tableau 5.

	^{95}Nb+^{95}Zr	^{144}Ce+^{144}Pr	^{141}Ce	^{106}Ru+^{106}Rh	^{103}Ru	^{90}Sr	^{137}Cs
λ_r (j^{-1})	1,08E-02	2,44E-03	2,13E-02	1,90E-03	1,80E-02	6,80E-05	6,30E-05
T_r (an)	0,18	0,78	0,09	1,00	0,11	27,93	30,14

Tableau 5 : Constante de décroissance radioactive et période associée

1.2.1.2 Les entrées de l'équation à estimer

- Le rapport de captation (R_c^{chron})

Il désigne la fraction du dépôt exprimé en Bq.m^{-2}, qui est interceptée par la masse foliaire des végétaux se trouvant à la surface du sol. Il s'exprime donc par un rapport d'activité sans dimension. Dans le cas d'un dépôt continu durant toute la phase de croissance du végétal, le rapport de captation peut prendre différentes valeurs possibles car il évolue dans le temps, depuis la sortie hors du sol jusqu'à la maturité du végétal. Partant de l'hypothèse que les observations (mesures de radioactivité sur un végétal) ont été effectuées sur le végétal à maturité, c'est une valeur moyenne du rapport de captation qui sera estimée.

- Le facteur de transfert racinaire (FT_r)

Il permet de quantifier la fraction d'activité présente dans un sol qu'un végétal prélève par ses racines. Il s'exprime par le rapport entre l'activité d'un radionucléide dans les parties consommables d'un végétal et l'activité dans le sol.

- La constante de décroissance de la biodisponibilité du radionucléide dans le sol (λ_s)

Cette quantité intègre les phénomènes de migrations horizontale et verticale ainsi que le vieillissement du radionucléide (fixation irréversible sur les particules minérales) qui diminue sa biodisponibilité pour les racines des plantes.

1.2.2 Les entrées de l'équation de nature agronomique

- Le temps de croissance du végétal ou durée de la phase végétative (T_c)

Des recherches bibliographiques ont été effectuées afin de renseigner les temps de croissance de l'ensemble des légumes-feuilles étudiés. Pour chaque légume, une valeur moyenne a été calculée à partir des différentes données trouvées dans la littérature (Cf. Tableau 6).

Pour le poireau, des données régionales sont à disposition, issues de Le Bohec et al. (1993). Ainsi, pour certaines régions de l'étude des valeurs moyennes ont pu être calculées (à partir des différentes données proposées pour les périodes de production). Seules les régions Garonne, Languedoc, Lorraine et Roussillon n'ont pas pu être renseignées, c'est

alors la valeur moyenne, calculée à partir de l'ensemble des données proposées par Le Bohec et al. (1993), qui a été utilisée (Cf. Tableau 7).

Légume-feuille	T_c *(mois)*
Bette	6
Chou	5
Chou-fleur	5
Salade	3

Tableau 6 : Valeurs sélectionnées pour les temps de croissance des différents légumes-feuilles étudiés

Régions	T_c *(mois)*
1 Bourgogne Lyonnais	4
2 Bretagne	8
3 Côte d'Azur	6
4 Garonne	6
5 Languedoc	6
6 Lorraine	6
7 Nord	5
8 Paris	6
9 Provence	6
10 Roussillon	6
11 Val de Loire	6

Tableau 7 : Valeurs sélectionnées pour le temps de croissance du poireau dans les différentes régions étudiées

- Le rendement cultural noté (*Rdt*)

Comme pour le temps de croissance, le rendement cultural est une donnée agronomique très variable suivant l'espèce et pour laquelle le regroupement en catégories (légumes-feuilles, légumes-racines,...) ne permet pas de réduire la variabilité. De la même manière que pour le temps de croissance, des recherches bibliographiques ont été effectuées et les valeurs moyennes retenues sont présentées dans le tableau 8. D'autres données sont également disponibles dans les cahiers de la statistique agricole d'Agreste[28]. Ces rapports regroupent les données de rendements de différents légumes par département et à partir de l'année 1960. N'étant pas saisies dans une base de données, ces données sont assez longues à acquérir : il faut prendre en compte les différentes variétés et modes de culture pour chaque département de la zone d'étude. Ainsi, pour l'instant, des chroniques complètes ont été informatisées uniquement pour le poireau et la bette. Dans la suite,

[28] Statistiques et études sur l'agriculture, la forêt, les industries agroalimentaires, l'occupation du territoire, les équipements et l'environnement en zone rurale.

plusieurs estimations pourront donc être effectuées en fonction des deux séries de données disponibles sur les rendements du poireau et de la bette.

Légume-feuille	*Rendement ($kg.m^{-2}$)*
Bette	3,0
Chou	2,0
Chou-fleur	1,5
Poireau	3,0
Salade	2,5

Tableau 8 : Valeurs sélectionnées pour les rendements des différents légumes-feuilles étudiés

- La profondeur de labour (h)

En comparant les différentes valeurs proposées pour les légumes-feuilles dans les modèles radioécologiques ECOSYS (Müller et Pröhl, 1993), FARMLAND (Brown et Simmonds, 1995) et AIEA (IAEA, 1994), la valeur retenue est une valeur moyenne égale à 25 cm.

- La masse volumique du sol sec (μ)

Cette quantité dépend du type de sol et de son degré de tassement. Ne connaissant pas ces caractéristiques pour chaque mesure, une valeur moyenne a été utilisée égale à 1410 kg sec.m^{-3} (GRNC, 2001).

2 La méthode d'estimation

A partir de l'équation de la contamination des végétaux et des données disponibles d'activité massique et déposée, les trois paramètres radioécologiques (rapport de captation, facteur de transfert racinaire et constante de décroissance du radionucléide biodisponible dans le sol) vont être estimés.

Considérons le modèle suivant :

$$Y = f(X_1, X_2, \theta) + \varepsilon$$

où

Y est la variable à expliquer (activité (Bq.kg^{-1} frais) du radionucléide r dans le végétal v),

X_1, X_2 sont les variables explicatives,

θ est le vecteur des paramètres inconnus à estimer,

ε l'erreur ou résidu (on le suppose additif).

La nature de la fonction f est fixée par la structure de l'équation de transfert radioécologique étudiée et l'absence d'information sur l'erreur nous conduit à la supposer

additive. Nous présentons, dans le tableau 9, les méthodes d'estimation utilisées, selon le type de radionucléide étudié.

Radionucléide	Radionucléides à vie courte ^{95}Nb+^{95}Zr, ^{144}Ce+^{144}Pr, ^{141}Ce, ^{106}Ru+^{106}Rh, ^{103}Ru	Radionucléides à vie longue ^{90}Sr, ^{137}Cs
Nature de f	f est linéaire $Y = aX_1 + \varepsilon \;\; où \;\; \theta = a$	f est non linéaire $Y = aX_1 + bX_2[c] + \varepsilon \;\; où \;\; \theta = {}^t(a,b,c)$
Méthode d'estimation	**Méthode des moindes carrés ordinaires** $\hat{\theta} = (X_1^t X_1)^{-1} X_1^t Y$ sous réserve que $(X_1^t X_1)$ soit inversible	**Méthode des moindes carrés séparables (Golub et Pereyra, 1973)** 1. a et b sont estimés par la méthode des moindres carrés ordinaires en fonction de c : $\begin{pmatrix} \hat{a}(c) \\ \hat{b}(c) \end{pmatrix} = (X^t X)^{-1} X^t Y$ où $X = (X_1 \; X_2[c])$ 2. La valeur du paramètre c minimise : $\sum_{i=1}^{n} (y_i - \hat{a}(c)x_{1,i} - \hat{b}(c)x_{2,i}(c))^2$

Tableau 9 : Méthodes d'estimation utilisées en fonction de la nature de la fonction f

Dans le cas d'une contamination par des radionucléides à vie courte, la fonction f est linéaire. En raison de leur courte période, ces radionucléides ne sont pas accumulés dans les sols et ne donnent pas lieu à un transfert racinaire. L'équation de la contamination des végétaux se restreint alors à la partie transfert foliaire. Les séries de mesures relatives aux radionucléides à vie courte sont donc particulièrement intéressantes car elles permettent de travailler sur un seul paramètre : le rapport de captation (paramètre a, dans le tableau 9). Ainsi, en première analyse, le rapport de captation sera étudié seul avec pour objectif de vérifier s'il peut être considéré comme indépendant du radionucléide et/ou du légume-feuille. Il ne s'agira pas d'étudier l'effet région (seules des mesures de la région Languedoc sont traitées) mais de tester l'hypothèse classiquement utilisée d'indépendance du rapport de captation vis-à-vis du radionucléide. L'idée est de simplifier la suite du travail pour la prochaine étape d'estimation à partir des séries de données relatives aux radionucléides à vie longue.

Dans le cas d'une contamination par des radionucléides à vie longue, les transferts par voie foliaire et racinaire sont pris en compte. Estimer le rapport de captation, le facteur de transfert racinaire et la constante de décroissance de la biodisponibilité du radionucléide dans le sol revient à estimer les paramètres a, b et c présentés dans le tableau 9. La fonction f est non linéaire et à une structure particulière : elle est linéaire sur les paramètres a et b et non linéaire sur c. La méthode d'estimation classiquement utilisée dans ce cas de figure est celle des moindres carrés séparables (Golub et Pereyra, 1973).

Cette méthode a été programmée sous le logiciel *Mathematica* (version 4). Pour certains échantillons, nous n'avons pas eu assez de mémoire disponible pour effectuer les calculs. Ce problème de mémoire vient principalement de la deuxième partie de l'équation qui calcule la contamination relative au transfert racinaire. En effet, le dépôt est sommé de la date initiale de contamination du sol jusqu'à la date de mesure et s'il y a 20 ans entre ces deux dates, 12mois*20=240 dépôts sont sommés pour une observation (mesure d'activité), etc. pour les autres observations. Le paramètre c qui figure parmi les paramètres à estimer apparaît dans chaque composante de cette somme. Lors de l'étape d'estimation, lorsque des opérations sont effectuées sur le vecteur $X_2(c)$, où $X_2(c) = \left(x_1(c) \quad \cdots \quad x_k(c) \quad \cdots \quad x_n(c)\right)$, le logiciel *Mathematica* manque d'espace disponible pour stocker l'information nécessaire à la réalisation du calcul. Pour pallier cet inconvénient, il faudrait réduire le pas de temps de l'étude et considérer le semestre ou l'année au lieu du mois. Mais cela induirait une perte de précision. En effet, l'hypothèse initiale faite sur le modèle serait encore moins vérifiée (constance du dépôt) et cela dégraderait les prédictions d'activités massiques. Nous avons donc utilisé une méthode de minimisation numérique. Les valeurs de a, b et c minimisant la quantité

$$E = \sum_{i=1}^{n} (y_i - ax_i + bz_i[c])^2$$

ont été recherchées grâce à un algorithme de minimisation issu du logiciel *Mathematica* (fonction *FindMinimum*). Un inconvénient de cette technique est l'exactitude de la solution. En effet, contrairement à la méthode précédente (moindres carrés séparables) qui fournit un minimum global, il est possible que le résultat obtenu ici soit un minimum local.

Néanmoins, la méthode des moindres carrés séparables a fonctionné pour quelques échantillons de faible taille : les triplets (chou, ^{90}Sr, Provence), (chou-fleur, ^{90}Sr, Provence), (chou-fleur, ^{137}Cs, Provence), (salade, ^{137}Cs, Roussillon), (salade, ^{137}Cs, Paris), (salade, ^{137}Cs, Nord), (salade, ^{137}Cs, Lorraine) et (salade, ^{137}Cs, Côte d'Azur) et les résultats obtenus étaient identiques à ceux de la méthode de minimisation numérique.

Pour les différents jeux de données étudiés, l'ajustement entre modèle et mesures est évalué lors d'une phase de confrontation en observant les différences entre les valeurs prédites par le modèle et la réalité observée. Selon Bernier et al. (2000), trois situations peuvent se présenter à l'environnementaliste selon que ces différences sont : insignifiantes, gênantes ou inacceptables.

3 Bilan du traitement des différentes séries de données

Les estimations obtenues permettent de calculer, grâce à l'équation (B.1), les valeurs d'activité dans les végétaux et de les confronter aux valeurs observées.

La phase de confrontation a mis en évidence des différences entre les prédictions du modèle et la réalité observée : les résultats des calculs d'ajustement se situent entre deux des situations énoncées précédemment ; pour certains échantillons, les prédictions ne sont ni clairement bonnes, ni clairement mauvaises (différence gênante) et pour d'autres, elles s'écartent fortement de la réalité observée (différence inacceptable).

Les échantillons étudiés étant assez nombreux, nous avons choisi de présenter uniquement les résultats graphiques (valeurs prédites en fonction des valeurs observées et analyse graphique des résidus) pour les cas (chou, $^{95}Zr+^{95}Nb$, Languedoc), (poireau, ^{90}Sr, Val de Loire) et (salade, ^{90}Sr, Côte d'Azur) car ils illustrent relativement bien les différents résultats obtenus (Cf. Figure 1).

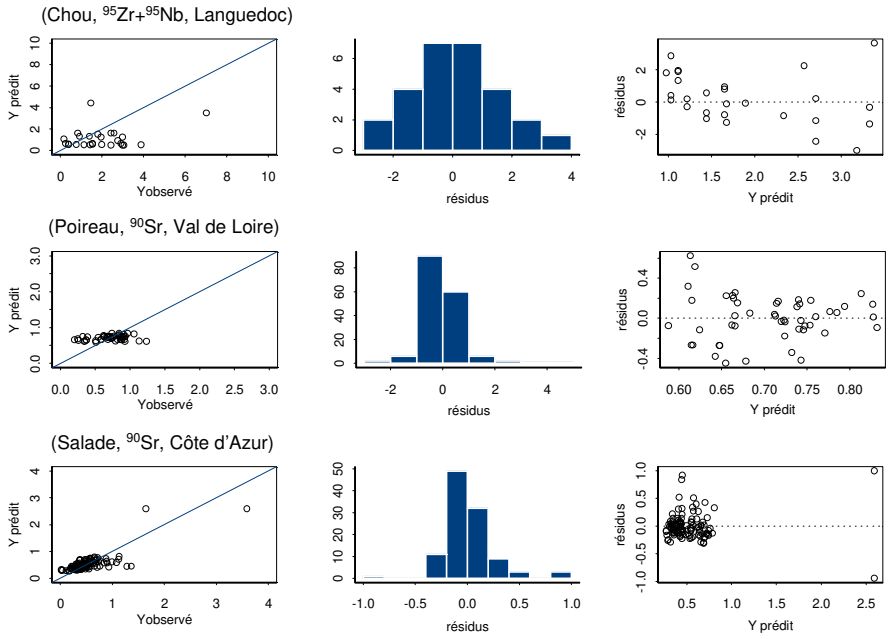


Figure 1 : Valeurs prédites en fonction des valeurs observées (en $Bq.kg^{-1}$ frais), histogramme des résidus et représentation des résidus en fonction des valeurs prédites pour les (chou, $^{95}Zr+^{95}Nb$, Languedoc), (poireau, ^{90}Sr, Val de Loire) et (salade, ^{90}Sr, Côte d'Azur)

D'une manière générale, la confrontation des valeurs prédites et observées met en évidence une mauvaise qualité de l'ajustement. Aucune tendance n'apparaît clairement dans les représentations des résidus. Le plus souvent, les faibles valeurs observées sont surestimées tandis que les plus fortes ont tendance à être sous-estimées. Pour certains échantillons étudiés, les prédictions s'écartent fortement de la réalité observée, c'est le cas du triplet (poireau, ^{90}Sr, Val de Loire) (deuxième série de graphes présentés en figure 1). Que ce soit pour les faibles ou les fortes valeurs d'activité, les prédictions sont

quasiment toujours identiques, le nuage de points a la forme d'une « bande » parallèle à l'axe des abscisses.

L'ensemble des estimations précédentes a été réalisé à partir des données sur les rendements présentées dans le tableau 8 (valeur fixe). Pour les bettes et les poireaux, des séries de données plus précises sont disponibles. De nouvelles estimations ont été effectuées à partir de ces jeux de données. Que ce soit pour le traitement des séries de données relatives aux radionucléides à vie courte ou celles relatives aux radionucléides à vie longue, il n'y a pas d'amélioration dans les prédictions. Vu le temps d'acquisition des séries de données plus précises, il a été décidé de se contenter des premières valeurs moyennes acquises sur le rendement.

Les estimations précédentes ont été réalisées en considérant un dépôt total (les dépôts sec et humide sont considérés comme sommés). Le dépôt pouvant s'exprimer comme la somme d'un dépôt sec et d'un dépôt humide, d'autres estimations ont été effectuées en distinguant le dépôt sec du dépôt humide. Ces estimations ont été réalisées pour les séries de données relatives aux radionucléides à vie longue.

La première partie de l'équation (B.1) déterminant l'activité du végétal par transfert foliaire devient alors :

$$C_{r,v,fol}^{chron} = \frac{R_{c,\sec}^{chron}}{Rdt}\left(\frac{D_{\Delta_{t_d}}^{\sec}}{\Delta_{t_d}} \frac{1-e^{-(\lambda_b+\lambda_r)\Delta_{t_d}}}{\lambda_b+\lambda_r} + \sum_{i=t_{d-\alpha+1}}^{t_{d-1}} \frac{D_{\Delta_i}^{\sec}}{\Delta_i} \frac{1-e^{-(\lambda_b+\lambda_r)\Delta_i}}{\lambda_b+\lambda_r} e^{-(\lambda_b+\lambda_r)\sum_{j=i+1}^{t_d}\Delta_j} \right)$$

$$+ \frac{R_{c,hum}^{chron}}{Rdt}\left(\frac{D_{\Delta_{t_d}}^{hum}}{\Delta_{t_d}} \frac{1-e^{-(\lambda_b+\lambda_r)\Delta_{t_d}}}{\lambda_b+\lambda_r} + \sum_{i=t_{d-\alpha+1}}^{t_{d-1}} \frac{D_{\Delta_i}^{hum}}{\Delta_i} \frac{1-e^{-(\lambda_b+\lambda_r)\Delta_i}}{\lambda_b+\lambda_r} e^{-(\lambda_b+\lambda_r)\sum_{j=i+1}^{t_d}\Delta_j} \right)$$

Le rapport de captation se décompose alors en un rapport de captation par temps sec ($R_{c,\sec}^{chron}$) et un rapport de captation par temps de pluie ($R_{c,hum}^{chron}$), ces deux quantités étant sans dimension. De ce fait, quatre paramètres radioécologiques ($R_{c,\sec}^{chron}$, $R_{c,hum}^{chron}$, FT_r et λ_s) ont été estimés. Les résultats obtenus sont identiques aux précédents. Le fait de distinguer le dépôt sec du dépôt humide et d'estimer alors deux rapports de captation n'améliore pas la qualité de l'ajustement.

4 Discussion sur les estimations obtenues

Dans ce paragraphe, nous présentons les différentes valeurs estimées et nous les confrontons aux valeurs issues de la bibliographie. Les valeurs sur lesquelles nous discutons sont celles relatives à la dernière estimation effectuée (décomposition du dépôt total en dépôt sec et humide). Les résultats sont présentés dans le tableau 10. La valeur E,

correspond à la quantité qui a été minimisée et nous avons choisi de présenter les résultats relatifs à la décroissance totale de la biodisponibilité du radionucléide dans le sol sous la forme d'une période notée T_s (en année). Dans certain cas, des estimations négatives ont été obtenues, les paramètres ont alors été contraints de manière à obtenir des valeurs positives. Les estimations correspondantes sont présentées en gris dans le tableau 10.

	^{90}Sr					^{137}Cs				
	$R_{c,\text{sec}}^{chron}$	$R_{c,hum}^{chron}$	FT_r	T_s	E	$R_{c,\text{sec}}^{chron}$	$R_{c,hum}^{chron}$	FT_r	T_s	E
	estimations relatives au légume-feuille POIREAU									
Bourgogne L.	1,84E-08	0,35	0,18	28,58	30,76	-	-	-	-	-
Bretagne	0,40	0,05	0,25	12,20	22,36	-	-	-	-	-
Côte d'Azur	1,70E-07	0,04	0,13	10,64	2,72	-	-	-	-	-
Garonne	0,38	3,33E-10	0,29	7,72	33,62	7,96E-11	1,01E-10	0,04	15,02	1,25
Languedoc	5,44E-09	0,17	0,09	16,14	9,81	4,42E-09	2,97E-10	0,04	1,87E+14	6,83
Lorraine	0,33	0,20	0,17	16,13	7,90	2,50	4,47E-07	9,93E-02	9,46E-04	7,29
Nord	0,03	0,22	0,08	17,64	2,79	-	-	-	-	-
Paris	0,34	0,06	0,19	10,65	7,98	2,551E-16	4,48E-17	0,06	6,80	0,09
Provence	0,26	0,05	0,15	12,56	84,44	0,21	5,67E-10	0,80	0,68	243,03
Roussillon	0,41	1,25E-08	0,23	3,44	0,46	-	-	-	-	-
Val de Loire	0,20	0,09	0,11	45,87	2,51	-	-	-	-	-
	estimations relatives au légume-feuille SALADE									
Bourgogne L.	0,04	1,70E-09	0,12	21,39	7,74	-	-	-	-	-
Bretagne	1,25E-08	0,25	0,28	6,82	17,22	9,99E-11	0,23	0,02	2,18E+13	1,67
Côte d'Azur	0,03	0,15	0,12	11,70	6,47	0,24	0,10	0,02	10,48	0,17
Garonne	5,11E-10	0,10	0,26	6,98	34,31	1,176E-11	5,80E-11	0,05	35,66	7,60
Languedoc	0,03	0,07	0,10	18,97	7,54	0,44	1,53E-08	0,40	4,00	485,77
Lorraine	0,23	1,28E-06	0,18	11,08	2,42	0,07	0,09	0,14	3,14	0,27
Nord	5,90E-19	2,73E-18	0,35	4,62	12,46	1,004E-19	7,44E-18	0,20	5,28	14,04
Paris	0,10	0,133	0,19	7,49	6,53	0,02	1,36E-10	0,07	8,54	1,45
Provence	0,14	1,46E-08	2,28	0,72	439,22	0,12	0,08	3,20	0,45	1954,13
Roussillon	1,49E-06	0,52	0,10	15,70	7,66	0,05	0,23	0,01	52,17	0,14
Val de Loire	0,359	4,87E-12	0,13	14,32	2,10	-	-	-	-	-
	estimations relatives au légume-feuille CHOU									
Bourgogne L.	9,57E-21	1,19E-18	9,93E-02	2,47E+17	15,99	-	-	-	-	-
Languedoc	-	-	-	-	-	1,08E-06	0,08	0,12	5,196	4,99
Provence	0,06	0,06	2,32E-01	2,88E+16	95,51	0,36	0,08	0,03	5,99E+16	376,66
	estimations relatives au légume-feuille CHOU-FLEUR									
Languedoc	-	-	-	-	-	3,91E-09	1,53E-10	0,04	10,55	0,60
Provence	0,02	0,02	0,16	4,77E+15	22,02	0,11	0,01	0,06	5,57	29,80

Tableau 10 : Estimations régionales des quatre paramètres pour les différents cas de contamination considérés

Le rapport de captation par temps sec

Les valeurs estimées du rapport de captation par temps sec sont comprises dans l'intervalle [0,019 ; 0,412]. Les très faibles valeurs ont été exclues car obtenues lorsque les paramètres ont été contraints pour prendre des valeurs positives. Pour la captation par temps sec, les valeurs proposées pour les légumes-feuilles dans différents codes de calculs utilisés en radioécologie sont de :

- 0,3 dans FARMLAND (Brown et Simmonds, 1995),

- 0,33 dans ABRICOT (Santucci, 1995),
- 0,5 dans FOCON (Rommens et al., 1999).

Pour la captation par temps humide c'est la valeur de 0,1 qui est proposé dans FOCON (Rommens, 1999)). Ces valeurs génériques, qui sont utilisées dans les codes de calculs, sont généralement des valeurs plutôt majorantes. On peut donc considérer que les estimations obtenues sont globalement cohérentes avec les valeurs proposées par la littérature.

Le facteur de transfert racinaire

Les différentes estimations du facteur de transfert racinaire ont été comparées aux valeurs issues de Nisbet et al. (1999) qui fournit une estimation de la valeur moyenne du paramètre, exprimée en $Bq.kg^{-1}$ de végétal sec par $Bq.kg^{-1}$ de sol sec, pour les légumes-feuilles pour les radionucléides ^{90}Sr et ^{137}Cs et par type de sol, à partir de compilation de la base de données de l'UIR (International Union of Radioecologists). Des changements d'unité sont donc nécessaires pour ramener les valeurs du facteur de transfert racinaire en $Bq.kg^{-1}$ de végétal frais par $Bq.kg^{-1}$ de sol sec. C'est une valeur moyenne du rapport poids frais/poids sec pour les légumes-feuilles (18 kg $frais.kg^{-1}$ sec) qui a été utilisée pour effectuer ce changement d'unité (calculée à partir des données issues de la base de données SYLVESTRE[29] (Duffa et al., 2007)). Cette valeur correspond à une proportion d'eau dans le poids frais de végétal d'environ 94 %, cohérente avec la valeur de 92 % citée dans (Quinault et al., 1989). Les valeurs proposées par Nisbet et al. (1999) ont été converties et sont présentées dans le tableau 11.

De fortes valeurs du facteur de transfert racinaire ont été estimées pour la région Provence, respectivement 2,28 $Bq.kg^{-1}$ de végétal frais par $Bq.kg^{-1}$ de sol sec pour le couple (salade, ^{90}Sr), 3,2 $Bq.kg^{-1}$ de végétal frais par $Bq.kg^{-1}$ de sol sec pour le couple (salade, ^{137}Cs) et 0,8 de végétal frais par $Bq.kg^{-1}$ de sol sec pour le couple (poireau, ^{137}Cs). Excepté ces fortes valeurs, les estimations du facteur de transfert racinaire sont comprises dans les intervalles suivants :

- pour le ^{90}Sr : [0,080 ; 0,349] avec une valeur moyenne de 0,167 $Bq.kg^{-1}$ de végétal frais par $Bq.kg^{-1}$ de sol sec,
- pour le ^{137}Cs : [0,014 ; 0,397] avec une valeur moyenne de 0,081 $Bq.kg^{-1}$ de végétal frais par $Bq.kg^{-1}$ de sol sec.

Ces plages de valeurs ont été comparées à celles proposées dans le tableau 11, en considérant un sol limoneux car c'est le type de sol dominant en France. Pour le ^{90}Sr, il apparaît que les valeurs estimées sont entièrement comprises dans l'intervalle de valeurs

[29] Base de données de mesures de radioactivité effectuées dans l'environnement dans le cadre des études de terrains réalisées à l'IRSN.

proposé par Nisbet et al. (1999) : [0,036 ; 0,506]. Dans le cas du ^{137}Cs, la plupart des estimations sont assez proches de la valeur maximale (0,0667) issue de Nisbet et al. (1999). Dans certains cas, les valeurs estimées sont plus élevées et semblent donc plus proches d'un sol de type sableux ou organique. N'ayant pas d'information sur les sols associés à chaque prélèvement de végétal, la confrontation n'a pas pu aller plus loin.

Radionucléide	Type de sol	Facteur de transfert racinaire	Intervalle de confiance à 95%	
			Min	*Max*
^{90}Sr	Sables	0,178	0,025	1,222
	Limons	0,133	0,036	0,506
	Argiles	0,100	0,046	0,222
	Organiques	0,018	-	-
^{137}Cs	Sables	0,0117	0,0014	0,0944
	Limons	0,0067	0,0007	0,0667
	Argiles	0,0037	0,0004	0,0322
	Organiques	0,0161	0,0009	0,3056

Tableau 11 : Valeurs du facteur de transfert racinaire issues de Nisbet et al (1999) recalculées en Bq.kg^{-1} de végétal frais par Bq.kg^{-1} de sol sec pour les légumes-feuilles

La période de décroissance totale de la biodisponibilité du radionucléide dans le sol T_s

Excepté les très fortes valeurs estimées (obtenues lorsque les paramètres ont été contraints), les estimations de la période de décroissance de la biodisponibilité du radionucléide dans le sol sont comprises dans les intervalles suivants :

- pour le ^{90}Sr : [0,725 ; 46,870] avec une valeur moyenne de 12,443 ans,
- pour le ^{137}Cs : [0,452 ; 52,171] avec une valeur moyenne de 10,932 ans.

Pour le modèle FOCON, les valeurs préconisées sont d'environ 35 ans pour le ^{90}Sr et 139 ans pour le ^{137}Cs (Rommens, 1997). Les estimations obtenues avec les mesures traitées ici sont donc globalement plus faibles que les chiffres de la littérature, en particulier pour le ^{137}Cs. Ce type d'écart a déjà été observé avec les mesures *in situ* du réseau OPERA (Roussel-Debet et al., 2005) et celles du suivi radioécologique autour des centrales EDF (Roussel-Debet et Duffa, 2005).

5 Synthèse

L'objectif initial de ce travail était d'obtenir des renseignements (intervalle de variation et distribution) sur trois paramètres radioécologiques de l'équation de la contamination des végétaux (rapport de captation, facteur de transfert racinaire et constante de décroissance de la biodisponibilité du radionucléide dans le sol) à partir de diverses séries de données disponibles pour différentes régions françaises.

L'ensemble des travaux effectués n'a pas permis d'obtenir les informations souhaitées. L'équation de transfert de la contamination radioactive vers les végétaux a été ajustée à

partir de divers jeux de données. Dans l'ensemble, les différents cas étudiés ont mis en évidence une mauvaise qualité de l'ajustement. Les nombreuses valeurs estimées ne peuvent donc pas être utilisées pour obtenir les renseignements désirés. En effet, les paramètres d'un modèle ont une signification lorsque celui-ci s'ajuste correctement aux données observées. Dans notre étude, les graphes des valeurs prédites en fonction des valeurs observées ont démontré que ce n'était pas le cas. Ainsi, même si la plupart des valeurs estimées sont cohérentes avec la bibliographie, elles ne peuvent pas être utilisées pour proposer des informations régionalisées.

ANNEXE C : Construction d'un arbre de régression par la méthode CART

Lorsque la variable à expliquer est quantitative, la méthode CART permet de construire des arbres de régression. La structure d'un arbre de régression est identique à celle d'un arbre de discrimination (Cf. 3.2) à l'exception des feuilles qui, dans ce cas, sont affectées à une valeur prédite de Y. Cette annexe présente de manière générale le principe de construction d'un arbre de régression par la méthode CART.

Avant de définir les prochaines notions, les notations et définitions suivantes sont nécessaires :

- $n(t)$ représente le nombre d'observations au nœud t,
- $p(t)$ représente la fréquence des observations au nœud t,
- La moyenne de Y au nœud t de l'arbre est définie par :

$$\bar{y}(t) = \frac{1}{n(t)} \sum_{i=1}^{n(t)} y_i$$

- La variance de Y au nœud t est définie par :

$$s(t)^2 = \frac{1}{n(t)} \sum_{i=1}^{n(t)} (y_i - \bar{y}(t))^2$$

- La mesure de la qualité de prédiction d'une feuille t est fournie par la moyenne des carrés des erreurs de prédiction :

$$R(t) = \sum_{t \in \tilde{T}} p(t) s^2(t)$$

- La mesure de la qualité de prédiction d'un arbre A est donnée par :

$$R(A) = \sum_{t \in \tilde{T}} R(t) \qquad \text{(C.1)}$$

- La variation de l'erreur quadratique due à la division d d'un nœud t est donnée par :

$$\Delta R(d,t) = R(t) - R(t_g) - R(t_d)$$

Soit d une division qui scinde le nœud t en deux nœuds enfants t_g et t_d. Comme dans le cas de la discrimination, la division recherchée est celle permettant d'obtenir t_g et t_d les plus homogènes relativement à la variable à expliquer. Cette condition se traduit, pour le cas

de la régression, par la notion de dispersion : la variance dans chacun des deux nœuds enfants doit être la plus petite possible (variance intra-nœud). Par conséquent, la meilleure division d^* est celle qui minimise la somme pondérée des variances :

$$V_I(d^*,t) = \arg\min_{d\in D}\{V_I(d,t)\} \qquad \text{(C.2)}$$

où $V_I(d,t) = p_g s^2(t_g) + p_d s^2(t_d)$ représente la somme pondérée des variances aux nœuds descendants.

Le principe de construction d'un arbre de régression repose sur les 3 étapes décrites au paragraphe 3.4.3 :

- Construction de l'arbre maximal

L'échantillon d'apprentissage est divisé successivement à l'aide du critère (C.2), de manière à construire l'arbre le plus grand possible. Comme pour le cas de la discrimination, il y deux règles d'arrêt dans la construction de l'arbre :

- lorsque la taille du nœud est inférieure à un effectif fixé : $N(t) \le N_{\min}$,
- lorsque toutes les valeurs de la variable à expliquer dans le nœud sont les mêmes (condition correspondant à la notion d'homogénéité ou de pureté d'un nœud en discrimination).

Lorsque l'un de ces deux cas est rencontré, le nœud est déclaré feuille et est affecté à la valeur moyenne de la variable à expliquer, calculée à partir de toutes les observations présentes dans le nœud.

- Elagage : Construction d'une série d'arbres

Il s'agit de construire une série d'arbres emboîtés entre l'arbre maximal et sa racine, à l'aide du critère d'élagage défini au paragraphe 3.4.3.2 :

$$\alpha = \frac{R(t) - R(A^t)}{\left|\tilde{A}^t\right| - 1}$$

- Sélection de l'arbre optimal

Comme pour le cas de la discrimination, il y a deux méthodes pour estimer l'erreur associée aux arbres : par validation croisée (Cf. Annexe D) ou à partir d'un nouvel échantillon qui n'a pas participé à la construction de l'arbre. Le principe de la sélection de l'arbre est le même que celui défini au paragraphe 3.4.3.3 en remplaçant le coût de mauvais classement par l'erreur quadratique définie en (C.1).

L'importance des variables

En reprenant les notations précédentes, l'importance de la variable X_k est définie par :

$$I(X_k) = \sum_{t \in A} \Delta R(\tilde{d}, t)$$

où $\tilde{d}$ représente la division de substitution sur la variable X_k (Cf. 3.4.4).

ANNEXE D : Méthode de validation croisée

Lorsque l'échantillon d'apprentissage est de petite taille, la méthode de validation croisée est utilisée pour estimer le coût de mauvais classement associé aux arbres obtenus lors de l'élagage défini par Breiman et al. (1984).

Nous proposons de décrire cette méthode au travers de différentes étapes :

Etape 1

A partir de l'échantillon d'apprentissage (E), les deux premières étapes de l'algorithme CART sont réalisées (Cf. 3.4.3) :

- construction de l'arbre maximal,
- construction de la série d'arbre de coût-complexité minimum :

$$S = \{T_0 = T_{\max}, T_1, ..., T_L\}$$

Etape 2

L'échantillon E est divisé aléatoirement en V sous-ensembles, E_v, $v = 1,...,V$. Le $v^{\text{ième}}$ échantillon, noté $E^{(v)}$, contient l'ensemble E privé du sous-ensemble E_v. Pour chaque échantillon d'apprentissage $E^{(v)}$ obtenu, l'étape 1 est réalisée : une série d'arbre $S_v = \{T_v^1, ..., T_v^H\}$ est construite. Chaque arbre de cette séquence est associé à une valeur du critère d'élagage. L'échantillon E_v qui n'a pas participé à la construction de l'arbre maximal, est utilisé pour estimer le coût de mauvais classement de chaque arbre de la séquence.

Etape 3

L'estimation du coût de mauvais classement associé à l'arbre T_i (arbre de la séquence S), est obtenue à partir des estimations de V autres arbres. Pour chaque série S_v, $v = 1,...,V$, l'arbre le plus proche de T_i, au sens du paramètre de complexité, est sélectionné. Breiman et al. (1984) associent à α_i , paramètre de complexité de l'arbre T_i, la moyenne géométrique $\alpha_i' = \sqrt{\alpha_i \alpha_{i+1}}$. Il s'agit alors de sélectionner, dans chaque séquence S_v, $v = 1,...,V$, l'arbre dont la valeur du paramètre de complexité est la plus proche de α_i' par valeur inférieure. L'estimation par validation croisée du coût associé à l'arbre T_i de la séquence S est calculée en moyennant les V estimations obtenues.

ANNEXE E : Présentation des différentes simulations réalisées sous STICS et résultats préliminaires

Dans cette annexe, nous présentons les différentes simulations réalisées à l'aide du modèle de culture STICS. En particulier, nous présentons les choix effectués pour renseigner les nombreuses variables d'entrées du modèle ainsi que les résultats préliminaires (sorties du modèle STICS). Ces références sont issues de Durand et al. (2006). Par la suite, ces résultats ont été analysés et ont permis de mettre en évidence deux relations, la première entre le rendement cultural et le temps de croissance de la laitue et la deuxième entre le rapport de captation et le développement de la laitue (Briand et al., 2008). Ces relations sont présentées, respectivement, dans les paragraphes 5.1.1.2.1 et 5.1.1.3.1 du chapitre 5.

Le logiciel STICS a été téléchargé à partir du site Internet de l'INRA d'Avignon, il donne accès à plusieurs fichiers d'entrée qui proposent une base de valeurs de paramètres indispensables pour faire fonctionner le modèle. Il importe de modifier certaines de ces valeurs par défaut afin de tester plusieurs scénarii et de prendre en compte différentes sources de variabilité (effet du climat, de la variété, des conditions de cultures, etc.). Pour les renseigner le plus précisément possible, la prise de contact avec des personnes spécialisées relevant du Centre Technique Interprofessionnel des Fruits et des Légumes (CTIFL) et d'autres organismes agricoles telles que l'Association Provençale de Recherche et d'Expérimentation Légumières (APREL), la Fédération Départementale des Groupes Etudes Techniques légumiers (FDGETAL) et le Groupe coopératif agricole et agroalimentaire (AGRIAL) a été nécessaire.

1 Présentation des jeux de simulations

1.2 Les données climatiques

Afin de mettre en évidence le maximum de variabilité, deux stations, opposées par leur climat, ont été sélectionnées : Orange pour le climat méditerranéen et Rennes pour le climat océanique (Cf. Figure 1). Le climat océanique se caractérise par des hivers doux (10°C en moyenne) et très humides marqués par des pluies intermittentes et surtout de la bruine ; l'été, le temps est beaucoup plus sec mais très frais (pas plus de 23°C en moyenne). Le climat méditerranéen est par contre un climat inégal sur le plan des précipitations ; en effet, elles sont très fortes au printemps et en automne et peuvent engendrer des inondations ; le reste de l'année, il n'y a quasiment pas de précipitations. Les étés sont chauds et secs (jusqu'à 40°C) et les hivers sont doux et humides. Les données

climatiques journalières portant sur une année calendaire (2005) et spécifiques aux villes concernées par la simulation ont été utilisées (Météo France, 2005). Il s'agit des températures minimales et maximales journalières, du rayonnement global journalier, de l'évapotranspiration et des précipitations journalières.

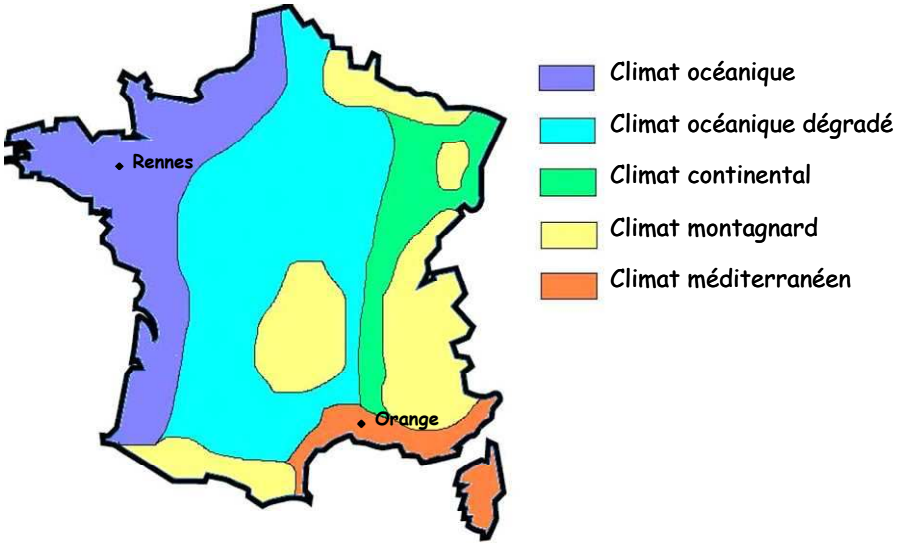


Figure 1 : Représentation schématique des différents climats de la France[30] et localisation des deux stations sélectionnées

1.2 Les données pédologiques

Les informations concernant les caractéristiques des sols utiles aux simulations (la teneur en argile de la couche de surface, le pH, l'albédo du sol nu à l'état sec, la limite d'évaporation de la phase potentielle d'évaporation du sol, etc.) sont fournies par défaut dans le modèle STICS. Ces valeurs sont rattachées à une culture de salade spécifique réalisée lors d'expérimentations de l'INRA. D'autres simulations ont cependant été effectuées pour validation à partir de valeurs régionales de la base de données de l'Association Française pour l'Etude des Sols (AFES). Les valeurs obtenues concernant le taux de recouvrement de la laitue sont très faibles et s'écartent assez des valeurs agronomiques (INRA, communication personnelle). Il semble que ces résultats soient liés au paramétrage du stress hydrique dans le modèle agronomique STICS. Il est donc préférable d'utiliser les caractéristiques pédologiques proposées par défaut par le modèle.

1.3 Les données culturales

Le paramétrage de l'itinéraire technique est très important car le développement de la culture dépend grandement des conduites techniques qui lui sont appliquées.

[30] D'après http://environnement.ecoles.free.fr/Vin/CARTE120.jpg.

Le travail du sol a lieu dans les deux régions, juste avant la plantation, généralement la veille (la réussite de la plantation est liée sensiblement à une plantation sur des sols fraîchement préparés). Pour la région d'Orange, un apport de matière organique est nécessaire (APREL, communication personnelle). Les historiques des parcelles étant variables et les données n'étant pas disponibles, l'apport de résidus pour la région d'Orange a donc été renseigné par les données par défauts issues du modèle agronomique STICS. L'apport de matière organique n'est pas nécessaire dans la région de Rennes (AGRIAL, communication personnelle).

Le semis de la laitue dans le modèle agronomique STICS correspond à la plantation d'un jeune plant de salade à deux feuilles. Le modèle agronomique STICS propose plusieurs techniques particulières dans l'itinéraire technique de la culture, notamment le paillage plastique (protection thermique). Cette option est donc utilisée dans les itinéraires des laitues plantées sous bâches plastiques entre février et mars, puis entre août et novembre. Les dates de plantation de notre étude ont été déterminées à partir des informations communiquées par les ingénieurs du CTIFL et à partir de Thicoïpé (1997). Au total 23 dates de plantations ont été retenues entre les semaines 9 et 36 pour la région d'Orange. Dans le cas de la région de Rennes, les plantations des laitues s'échelonnent entre la semaine 6 (début février) et la semaine 35 (fin août) (AGRIAL, communication personnelle). Il s'agit néanmoins de différents types de cycles : de la semaine 6 à la semaine 18, les laitues sont bâchées permettant une protection thermique ; des semaines 20 à 31, il s'agit des laitues de printemps et des semaines 32 à 35, il s'agit des laitues d'été. Ainsi, pour la région de Rennes, 27 dates de plantations ont été étudiées. Les simulations effectuées sous le modèle STICS permettent d'obtenir les dates des stades de développement de la laitue, notamment celles du stade où le taux de recouvrement est maximal. En lien avec la production commerciale, la récolte s'effectue au stade de la salade pommée. Ce stade n'est pas mentionné dans le modèle STICS mais il correspond au stade de recouvrement maximal. Ainsi, il suffit de spécifier que la récolte a lieu à maturité physiologique afin d'obtenir la date du stade où le taux de recouvrement est maximal et donc déduire le temps de croissance associé à la culture de laitue simulée.

Pour les deux régions, l'option de forçage de l'irrigation a été choisie. Pour Orange, les dates d'apport ont été calculées à partir des données de pluviométrie fournies par Météo France : l'irrigation (apport de 5 mm) aura lieu le jour de la plantation ainsi qu'aux jours où l'apport en eau deviendra nécessaire pour le bon développement de la plante. Pour la région de Rennes, les pratiques d'irrigations sont légèrement différentes. En effet, les maraîchers irriguent en fonction des évènements climatiques d'une part, mais également

tous les jours durant la semaine de plantation à raison de 10 mm/apports (AGRIAL, communication personnelle).

Pour la région d'Orange, les données concernant la fertilisation ont été communiquées par le CTIFL. Elles permettent de renseigner les jours Julien des apports et les quantités d'azote minéral fournies au végétal (en kg $N.ha^{-1}$), ce qui permet particulièrement d'ajuster ce paramètre technique à la conduite réelle en plein champ. En effet, dans le cas de la culture de laitue, les apports ont lieu presque exclusivement au début de la culture et notamment au moment du semis des jeunes plants de salade. Dans la région d'Orange, 30 unités d'azote sont apportées le jour de la plantation des laitues. D'après AGRIAL, les maraîchers de la région de Rennes fertilisent toujours au moment de la plantation des plants de laitue. Cependant, les quantités apportées diffèrent selon les cycles. Pour les laitues d'automne (bâchées), il s'agit généralement d'une fertilisation avec 650 kg/ha d'un engrais, correspondant à une solution azotée, composée à 15 % d'azote, soit un apport de 97 unités d'azote minéral. Pour les laitues d'été, un apport de 75 unités d'azote est réalisé (500 kg/ha d'engrais de 15/5/20[31]) et un apport de 60 unités d'azote est appliqué pour les laitues de printemps (400 Kg/ha d'engrais de 15/5/20). Ces quantités peuvent néanmoins varier selon les zones de production. Un exemple de valeurs, utilisées pour les simulations, est présenté dans le tableau 1.

2 Résultats des simulations

2.1 Les temps de croissance

Les simulations réalisées avec le modèle STICS permettent d'obtenir les dates de récolte associées à chaque plantation et donc de déduire les temps de croissance (Cf. Tableau 2). Les valeurs obtenues conviennent assez bien avec les données réelles puisque les temps de croissance de la laitue varient entre 30 et 90 jours (Thicoïpé, 1997). Néanmoins, les temps de croissance fournis par le logiciel STICS sont plus courts ou parfois plus longs que ceux relevés dans les pratiques réelles. En effet, pour la région de Rennes (AGRIAL, communication personnelle), les laitues plantées entre les semaines 20 à 31 sont récoltées durant la semaine 26 à 37, avec des cycles de 6 semaines en moyenne, soit une 40^{aine} de jours de croissance. STICS fournit des temps de croissance plus faibles, autour de 30 à 34 jours. Il en est de même pour les laitues plantées durant les semaines 32 à 35 et récoltées de la semaine 38 à 43, correspondant à des cycles culturaux de 42 à 56 jours, STICS donne des cycles plus courts, de l'ordre de 34 jours.

[31] Engrais N, P, K : engrais composés d'azote, de phosphore et de potassium.

Paramètres	Orange		Rennes	
Dates de plantation des laitues (date et jour Julien)	*25-févr-56*	25 juin - 176	*5 février – 36*	5 juin – 156
	15-mars-74	5 juillet – 186	*19 février – 50*	12 juin – 163
	21-mars-80	10 juillet - 191	*5 mars – 64*	19 juin – 170
	15 mars – 74	15 juillet – 196	*12 mars – 71*	26 juin – 177
	7 avril – 97	26 juillet – 207	*19 mars – 78*	3 juillet – 184
	20 avril – 110	5 août – 217	*26 mars – 85*	10 juillet – 191
	26 avril – 116	12 août – 224	*5 avril – 95*	17 juillet – 198
	5 mai – 125	20 août – 232	*12avril – 102*	24 juillet – 205
	15 mai – 135	25 août - 237	*19 avril – 109*	31 juillet - 212
	25 mai – 145	*25 août - 237*	*26 avril – 116*	7 août – 219
	5 juin – 156	*5 septembre - 248*	*5 mai – 125*	14 août – 226
	15 juin – 166		15 mai – 135	21 août – 233
			22 mai – 142	28 août - 240
			29 mai - 149	
Fertilisation	Dates de plantation / 30 unités d'azote		Dates de plantation / 97, 75, 60 unités d'azote	
Irrigation : dates de plantation (jour Julien) et jours selon pluies et ETP	25 février (56) / 1mm	18 juillet (199) /7mm	10 mars (69) / 2mm	7 juillet (188) / 5mm
	2 mars (61) / 5mm	23 juillet (204) / 7mm	15 mars (74) / 2mm	12 juillet (193) / 6mm
	8 mars (67) / 5mm	28 juillet (209) / 9mm	20 mars (79) / 3mm	17 juillet (198) / 6mm
	15 mars (74) / 5mm	2 août (214) / 5mm	30 mars (89) / 1mm	22 juillet (203) / 5mm
	19 mars (78) / 5mm	7 août (219) / 7mm	4 avril (94) / 1mm	1 août (213) / 4mm
	1 avril (91) / 5mm	12 août (224) / 7mm	9 avril (99) / 3mm	6 août (218) / 4mm
	29 avril (119) / 5mm	18 août (230) / 7mm	29 avril (119) / 4mm	11 août (223) / 4mm
	4 mai (124) / 5mm	24 août (236) / 7mm	4 mai (124) / 2mm	16 août (228) / 5mm
	10 mai (130) / 5mm	30 août (242) / 5mm	9 mai (129) / 4mm	21 août (233) / 4mm
	23 mai (143) / 5mm	4 sept. (247) / 3mm	18 mai (138) / 3mm	26 août (238) / 4mm
	29 mai (149) / 6mm	14 sept. (257) / 4mm	23 mai (143) / 3mm	31 août (243) / 5mm
	4 juin (155) / 7mm	24 sept. (267) / 3mm	28 mai (148) / 5mm	5 sept. (248) / 3mm
	10 juin (161) / 7mm	29 sept. (272) / 4mm	2 juin (153) / 5mm	15 sept. (258) / 3mm
	19 juin (170) / 7mm	4 octobre (277) / 3mm	7 juin (158) / 5mm	20 sept. (263) / 2mm
	25 juin (176) / 7mm	9 octobre (282) / 2mm	12 juin (163) / 5mm	25 sept. (268) / 2mm
	30 juin (181) / 7mm	14 oct. (287) / 2mm	17 juin (168) / 4mm	10 oct. (283) / 2mm
	6 juillet (187) / 7mm	29 oct. (302) / 1mm	22 juin (173) / 6mm	15 oct. (288) / 1mm
	12 juillet (193) / 9mm		27 juin (178) / 7mm	4 nov. (308) / 1mm
Récolte	Maturité physiologique (mais récolte réelle lorsque le taux de recouvrement est maximal)			

Tableau 1 : Description des paramètres techniques modifiés pour les simulations effectuées sous STICS, les dates de plantation en italique correspondent aux laitues plantées sous bâche plastique

Par ailleurs, les laitues plantées sous bâche plastique entre les semaines 6 à 19 sont récoltées de la semaine 17 à 25, conduisant à des cycles de 11 à 6 semaines soit à des temps de croissance de 42 à 77 jours. On retrouve ces temps de croissance avec le modèle STICS excepté pour la première plantation pour laquelle le cycle atteint 90 jours. Une différence d'une dizaine de jours s'établit donc entre les valeurs réelles des temps de croissance des laitues, de la plantation à la récolte, et les valeurs issues du logiciel STICS.

Orange		Rennes	
Date de plantation	Temps de croissance	Date de plantation	Temps de croissance
25-févr	*75*	*05-févr*	*90*
15-mars	*60*	*19-févr*	*77*
21-mars	*57*	*05-mars*	*66*
15-mars	60	*12-mars*	*59*
07-avr	51	*19-mars*	*57*
20-avr	44	*26-mars*	*58*
26-avr	41	*05-avr*	*56*
05-mai	39	*12-avr*	*52*
15-mai	36	*19-avr*	*47*
25-mai	32	*26-avr*	*45*
05-juin	30	*05-mai*	*45*
15-juin	29	15-mai	41
25-juin	30	22-mai	37
05-juil	30	29-mai	35
10-juil	29	05-juin	33
15-juil	30	12-juin	31
26-juil	31	19-juin	31
05-août	31	26-juin	32
12-août	34	03-juil	31
20-août	36	10-juil	31
25-août	39	17-juil	33
25-août	*36*	24-juil	34
05-sept	*43*	31-juil	34
		07-août	34
		14-août	32
		21-août	34
		28-août	34

Tableau 2 : Temps de croissance des laitues pour les régions d'Orange et de Rennes, les valeurs en italique concernent les laitues sous bâche plastique

2.2 Le taux de recouvrement

Les valeurs de taux de recouvrement, fournies par le logiciel STICS, sont comprises entre 0 et 1 puisque cette variable s'exprime en pourcentage de recouvrement. La figure 2 présente l'évolution du taux de recouvrement en fonction du développement des laitues (en jour), respectivement pour les régions d'Orange et de Rennes. Pour la région d'Orange et pour un temps de croissance total de 75 jours, la courbe en pointillés correspondant à la plantation du 25 février présente un taux de recouvrement quasiment nul durant 21 jours car le plant de laitue sous bâche plastique à deux feuilles n'a pas les conditions climatiques, agronomiques et physiologiques idéales pour se développer. Une fois les conditions réunies, à partir du 21ème jour, il évolue pour approcher en une quarantaine de

jours son taux de recouvrement maximal de 81 %. Ce temps de « latence » est beaucoup moins marqué pour les laitues plantées durant le printemps et l'été : en effet, les conditions climatiques entraînent un développement très rapide des plants de laitue, dès le 5ème jour. Ceci se retrouve également pour la région de Rennes (Cf. Figure 2, graphe de droite).

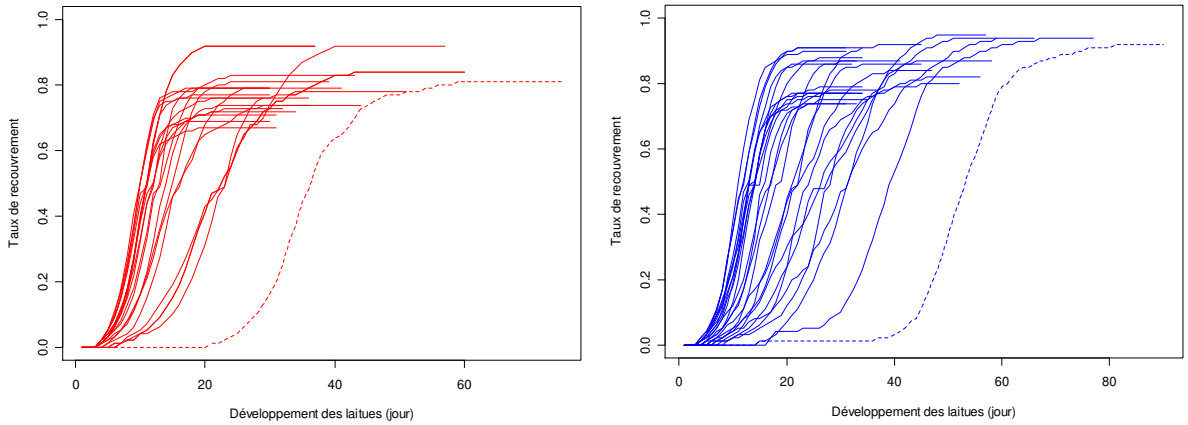


Figure 2 : Evolution du taux de recouvrement des laitues depuis le stade plantule jusqu'à la récolte pour les régions d'Orange et de Rennes

L'allure des courbes présentées sur la figure 2 est identique. Les valeurs du taux de recouvrement maximal varient entre 70 et 95 % et sont plus importantes dans la région de Rennes. Ceci est probablement lié aux apports de fertilisants, différents d'un facteur 2 à 3 entre les régions d'Orange et de Rennes. Notons qu'il faut près de 3 mois pour que la salade plantée le 5 février arrive à maturité sous le climat breton (courbe en pointillés).

2.3 Le rendement

Les résultats obtenus à la suite des simulations pour chaque date de plantation de la laitue plein champ, sont présentés sur la figure 3, respectivement pour la région d'Orange et celle de Rennes. Les valeurs des rendements obtenus à la récolte varient entre 1,25 et 5,13 t matière sèche.ha^{-1} pour la région d'Orange et elles s'échelonnent entre 1,87 et 4,66 t matière sèche.ha^{-1} pour la région de Rennes. La figure 3 montre que pour des temps de croissance courts, les rendements à la récolte sont globalement plus faibles que pour les temps de croissance plus longs. On retrouve en outre sur ces représentations graphiques les mêmes temps de « latence » évoqués précédemment pour le taux de recouvrement.

Ces valeurs sont globalement en accord avec les rendements réels. En effet, la statistique agricole AGRESTE donne pour l'année 2005, un rendement moyen pour la laitue de 281 q frais.ha^{-1}, soit 28,1 t.ha^{-1} de laitue fraîche. En prenant un rapport poids frais/poids sec pour la salade de 18 (Cf. Annexe B), cela nous ramène à un rendement annuel moyen de

1,56 t sec.ha^{-1}. On retrouve cette valeur et des valeurs l'approchant principalement dans la région d'Orange. Par ailleurs d'après Thicoïpé (1997), le rendement moyen d'une laitue avoisine les 42 t matière fraîche.ha^{-1}, soit 2,33 t matière sèche.ha^{-1}. La majorité de nos valeurs tournent autour de ces valeurs moyennes, excepté pour celles correspondant aux temps de croissance les plus longs. En outre, d'après AGRIAL, les rendements obtenus par les maraîchers de la région de Rennes, varient entre 23 t.ha^{-1} pour les premières laitues à 28 t.ha^{-1} de matière fraîche pour les dernières laitues. Ceci revient à des rendements en matière sèche variant de 0,8 à 1,5 t.ha^{-1}. Les valeurs données par le logiciel sont sensiblement plus élevées. Les valeurs les plus élevées concernent les premières plantations, en mars, qui ont un cycle du stade plantule au stade de recouvrement maximal beaucoup plus long (60 à 90 jours environ) contrairement à celles d'été pour lesquelles les cycles avoisinent 30 à 35 jours.

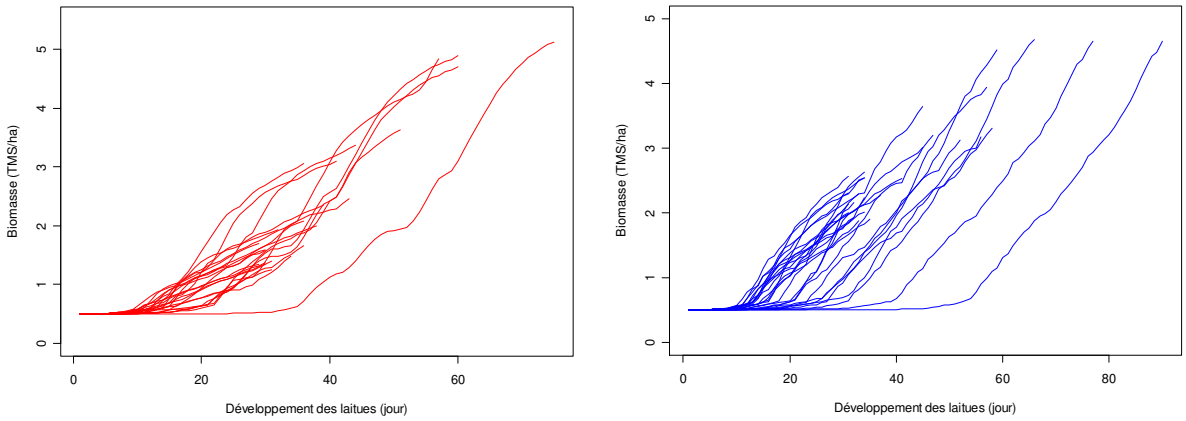


Figure 3 : Evolution de la biomasse des laitues depuis le stade plantule jusqu'à la récolte pour les régions d'Orange et de Rennes

ANNEXE F : Résultats des estimations des paramètres *a* et *b* pour chaque jeu de simulation étudié

Région d'Orange		
T_c	$\hat{a}$	$\hat{b}$
29	7,76	0,84
29	8,48	0,92
30	8,75	0,74
30	7,86	0,83
30	9,21	0,90
30	7,27	0,80
31	7,33	0,81
31	8,62	0,88
32	7,56	0,71
34	6,66	0,66
36	8,62	0,68
36	7,18	0,66
37	7,85	0,71
37	7,85	0,71
39	7,64	0,55
41	5,88	0,40
43	6,58	0,53
44	7,34	0,50
51	10,06	0,47
57	5,81	0,25
60	6,48	0,30
60	6,48	0,30
75	12,29	0,35

Région de Rennes		
T_c	$\hat{a}$	$\hat{b}$
31	6,99	0,60
31	7,87	0,71
31	9,51	0,79
31	6,48	0,54
32	6,51	0,52
32	7,42	0,55
33	9,60	0,66
33	7,41	0,59
34	6,25	0,51
34	7,51	0,58
34	6,09	0,44
34	7,44	0,54
34	6,69	0,56
35	7,74	0,50
37	6,88	0,45
41	7,65	0,43
45	7,61	0,35
45	8,67	0,41
47	6,10	0,29
52	8,74	0,38
56	10,97	0,41
57	5,56	0,19
58	8,31	0,28
59	5,91	0,21
66	7,56	0,23
77	11,89	0,30
90	16,82	0,31

ANNEXE G : Résultats relatifs aux estimations des coefficients des deux droites de régression

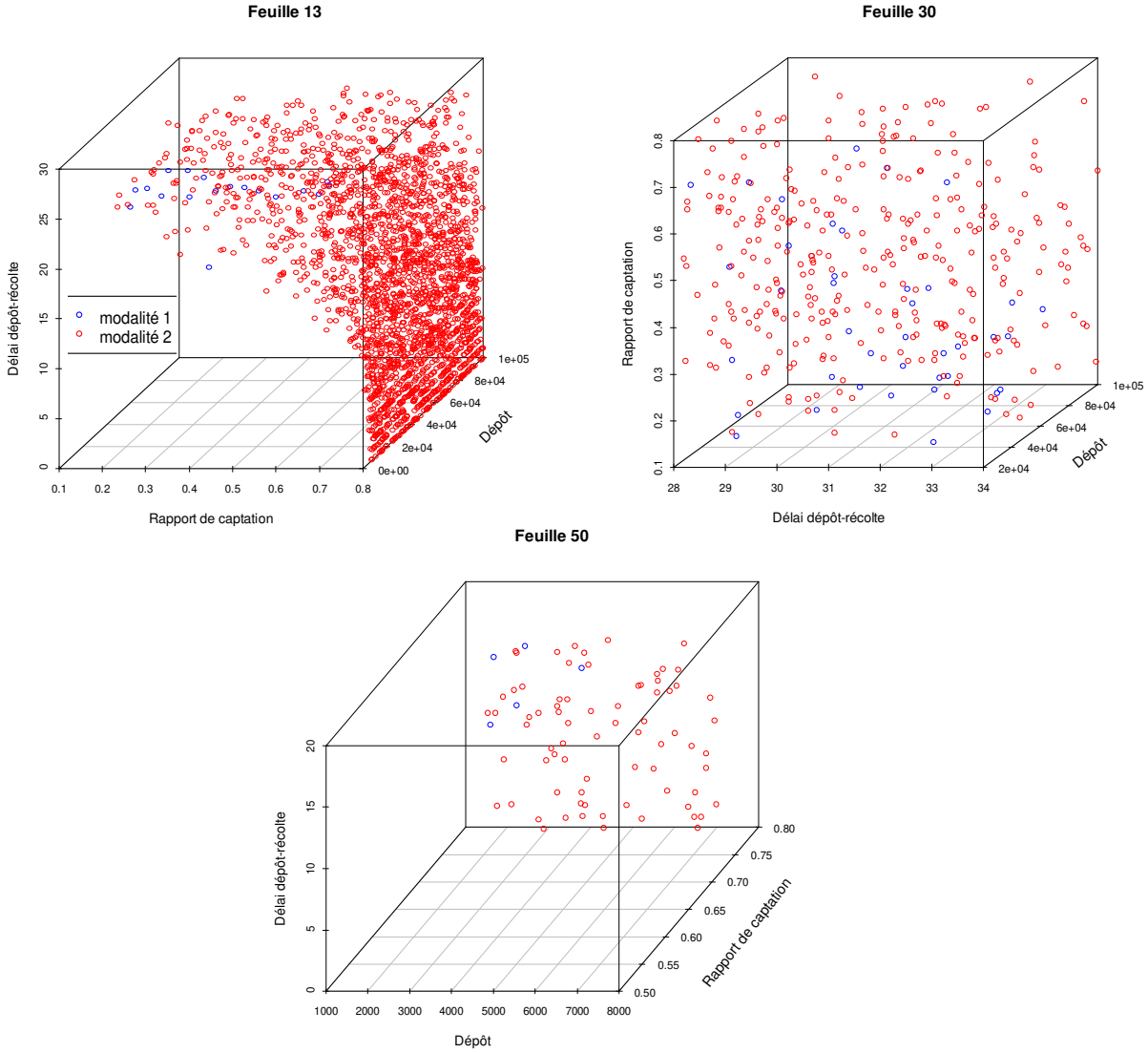
Résultats relatifs à la régression de a sur T_c :

	Estimate	Std. Error	t value	Pr(>|t|)
intercept	-8,94	1,14	-7,81	4,21E-10
slope	0,62	0,03	24,18	< 2e-16

Résultats relatifs à la régression de $ln(b)$ sur $ln(T_c)$:

	Estimate	Std. Error	t value	Pr(>|t|)
intercept	3,67	0,39	9,41	1,78E-12
slope	-1,18	0,11	-11,24	4,81E-15

ANNEXE H : Représentation des données relatives à l'échantillon test dans les feuilles n°13, 30 et 50 de l'arbre de discrimination relatif au légume-feuille laitue.



ANNEXE I : Arbres de discrimination construits par la méthode REN relatifs aux légumes-feuilles chou et épinard.

Cette annexe présente les arbres de discrimination, construits par la méthode REN, pour les deux légumes-feuilles chou et épinard (Cf. Figure 1 et Figure 2). La représentation graphique des arbres est la même que celle utilisée dans le paragraphe 8.1.1. Pour chaque légume-feuille, un échantillon test est généré. Ils sont utilisés pour estimer le taux de mauvais classement (%) des deux arbres : 6,7 % pour l'arbre de discrimination relatif au chou et 4,02 % pour l'arbre de discrimination relatif à l'épinard. Ils sont également utilisés pour estimer le taux de mauvais classement dans les feuilles des arbres.

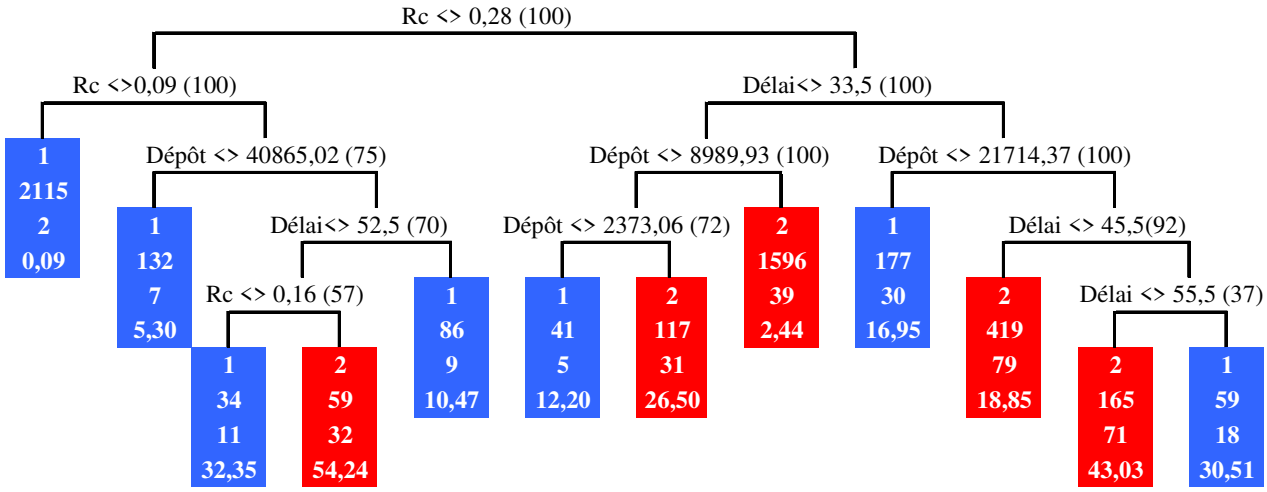


Figure 1 : Arbre de discrimination obtenu pour le scénario de contamination (chou, ^{90}Sr)

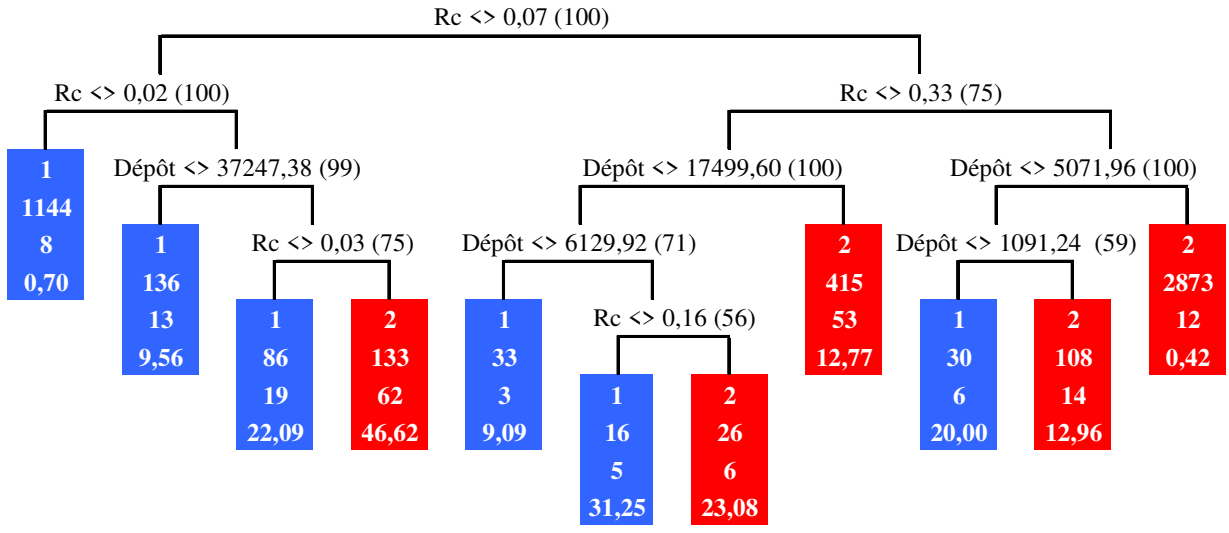


Figure 2 : Arbre de discrimination obtenu pour le scénario de contamination (épinard, ^{90}Sr)

ANNEXE J : Indicateurs statistiques des activités massiques du ^{90}Sr pour les feuilles n°4 et n°13 de l'arbre de discrimination construit pour les légumes-feuilles

Les valeurs (Bq.kg^{-1} frais), pour chaque légume-feuille étudié : laitue, chou, épinard et poireau, sont distinguées.

Feuille	Légume	Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
n° 4	laitue	8,32E-05	0,17	0,82	4,78	3,53	89,71
	chou	4,70E-08	7,44E-04	0,01	0,59	0,13	43,23
	épinard	1,57E-03	0,36	1,86	13,52	9,97	187,50
	poireau	1,30E-12	1,70E-04	0,01	0,68	0,13	84,25
n° 13	laitue	29,32	786,70	1868,00	3052,00	4574,00	17770,00
	chou	27,97	613,40	1594,00	2310,00	3085,00	12370,00
	épinard	61,08	1931,00	5105,00	7622,00	10730,00	36300,00
	poireau	54,43	2245,00	5113,00	7097,00	9282,00	45430,00

Résumé

L'objectif de cette thèse est le développement d'une méthode permettant l'identification de facteurs conduisant à différents niveaux de contamination radioactive des végétaux. La méthodologie proposée est basée sur l'utilisation d'un modèle radioécologique de transfert des radionucléides dans l'environnement (code de calcul ASTRAL) et une méthode de discrimination par arbre. En particulier, pour parer les problèmes d'instabilité des arbres de discrimination et conserver leur structure, une méthode de stabilisation par rééchantillonnage bootstrap dans les nœuds est utilisée. Des comparaisons empiriques sont effectuées entre les arbres de discrimination construits par cette méthode (appelée méthode REN) et ceux obtenus par la méthode CART. Une mesure de similarité, permettant la comparaison de la structure de deux arbres de discrimination, est définie. Cette mesure est utilisée pour étudier les performances de stabilisation de la méthode REN. La méthodologie proposée est appliquée à un scénario simplifié de contamination. Les résultats obtenus permettent d'identifier les principales variables responsables des différents niveaux de contamination radioactive de quatre légumes-feuilles (laitue, chou, épinard et poireau). Certaines règles extraites de ces arbres de discrimination peuvent être utilisables dans un contexte post-accidentel.

Mots clés

Arbre de discrimination, CART, instabilité, bootstrap, nœud, mesure de similarité, radioécologie.

Building classification trees to explain the radioactive contamination levels of the plants

Abstract

The objective of this thesis is the development of a method allowing the identification of factors leading to various radioactive contamination levels of the plants. The methodology suggested is based on the use of a radioecological transfer model of the radionuclides through the environment (ASTRAL computer code) and a classification-tree method. Particularly, to avoid the instability problems of classification trees and to preserve the tree structure, a node level stabilizing technique is used. Empirical comparisons are carried out between classification trees built by this method (called REN method) and those obtained by the CART method. A similarity measure is defined to compare the structure of two classification trees. This measure is used to study the stabilizing performance of the REN method. The methodology suggested is applied to a simplified contamination scenario. By the results obtained, we can identify the main variables responsible of the various radioactive contamination levels of four leafy-vegetables (lettuce, cabbage, spinach and leek). Some extracted rules from these classification trees can be usable in a post-accidental context.

Keywords

Classification tree, CART, instability, bootstrap, node, similarity measure, radioecology.