IRSN
INSTITUT
DE RADIOPROTECTION
ET DE SÛRETÉ NUCLÉAIRE

UTILISATION DES METHODES D'ARBRES DE DISCRIMINATION EN RADIOECOLOGIE

04 septembre 2007 / Cadarache

Présenté par Bénédicte BRIAND, IRSN

Directeur de thèse : Gilles DUCHARME
Université de Montpellier II

Responsable de thèse IRSN : Catherine MERCAT-ROMMENS
DEI/SESURE/LERCM

PLAN DE L'EXPOSE

Contexte

Objectif

Méthodologie

Exemple illustratif

Axe de recherche

Conclusion

Les conséquences pour l'homme et l'environnement d'une pollution d'origine industrielle dépendent de l'importance et de la nature de celle-ci, mais aussi du territoire qui la reçoit

Projet SENSIB (Mercat-Rommens et al, 2005)

La sensibilité radioécologique peut se définir comme la composante environnementale qui détermine la réponse d'un territoire à une contamination radioactive

Identifier les spécificités des territoires français qui influent fortement sur le devenir d'un contaminant radioactif dans l'environnement

La connaissance sur les caractéristiques des territoires pourra alors être utilisée, de façon anticipée par rapport aux situations accidentelles :

- Emettre des recommandations en matière de gestion des territoires contaminés
- Hiérarchiser la prise de décision

Développer une méthodologie permettant d'identifier les facteurs qui vont influer sur les niveaux de contamination radioactive des végétaux

Comment ?

1 A partir de mesures effectuées dans l'environnement

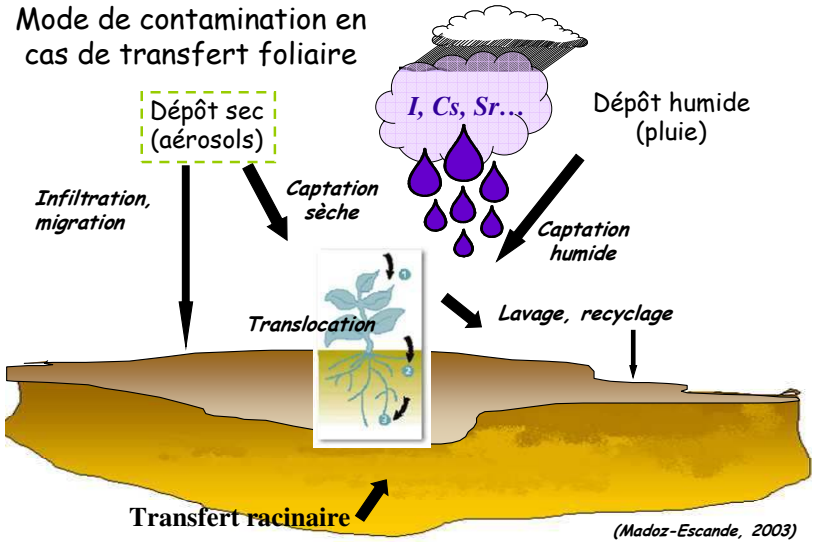
- Nombreuses mesures de radioactivité effectuées dans l'environnement Français (notamment milieux agricoles)
- Caractéristiques de prélèvement associées à ces mesures souvent peu nombreuses et imprécises

2 A partir d'un modèle radioécologique de transfert

$$Z=h(X_1, X_2, \ldots, X_p) \qquad Y=g(Z)=\begin{cases}1 & \text{if } Z \leq k_0\\ 2 & \text{if } Z > k_0\end{cases}$$

- Non-linéaire
- Dépendances entres certaines entrées du modèle
- Sortie de type qualitative (niveau de contamination)

Utilisation des techniques d'arbres de discrimination comme méthode d'analyse de sensibilité globale



Etape 1

Echantillonnage des valeurs d'entrées et codage de Z

Variables d'entrées du modèle radioécologique

$$\begin{cases} X_1 \approx F_1 \\ \vdots \\ X_p \approx F_p \end{cases}$$

Modèle radioécologique

$$z = h(X_1, \ldots, X_p)$$

Fonction qui permet le codage de la sortie du modèle en différentes classes

$$y = g(z) = \begin{cases} 1 & \text{si } z \le k_0 \quad P(y=1) \\ 2 & \text{si } z > k_0 \quad P(y=2) \end{cases}$$

Etape 2

Modélisation du phénomène par des arbres de discrimination

$$A(F_1, \ldots, F_p) = \begin{cases} 1 & P(A=1) \\ 2 & P(A=2) \end{cases}$$

- Simplicité de représentation et d'interprétation
- Pouvoir explicatif important
- Pouvoir prédictif

Mise en évidence des variables et/ou association de variables explicatives qui conduisent au différentes classes de la variable à expliquer (niveau de contamination)

Etat de l'art - *Les techniques d'arbres de discrimination et de régression en AS globale*

- Domaine du nucléaire (Mishra et al, 2003)
  - ➢ Etude d'un modèle d'évaluation des performance d'un site de stockage de déchet radioactif
    - → prédiction , à long terme, du débit de dose pour une personne vivant à 20 km à l'aval du site
  - ➢ Arbres de discrimination → identification des variables d'entrées clés responsables des valeurs de doses extrêmes pour différents scénarii d'applications (dose à 70000 et à 100000 ans)

- Domaine des risque alimentaires (Mokhtari et al, 2006)
  - ➢ Etude d'un modèle de prédiction des risque alimentaires, E. coli O157: H7
    - → estimer, de la production à la préparation, la présence de la bactérie E. coli
    - → Modèle assez complexe

      Non linéaire - Interaction entres les entrées - Entrées quantitatives et qualitatives - Points de saturations
  - ➢ Sorties quantitatives → Arbres de régression → identification des variables d'entrées conduisant à des valeurs particulières de la sortie de chaque module étudié (abattage, production)

Arbre de discrimination et de régression

Déterminer les combinaisons de plages de valeurs spécifiques des entrées du modèle qui engendrent des sorties particulières

Meilleure compréhension du phénomène modélisé en soulignant certaines relations entres les variables d'entrées et la variable de sortie du modèle

La méthode CART

Classification And Regression Trees (Breiman et al, 1984)

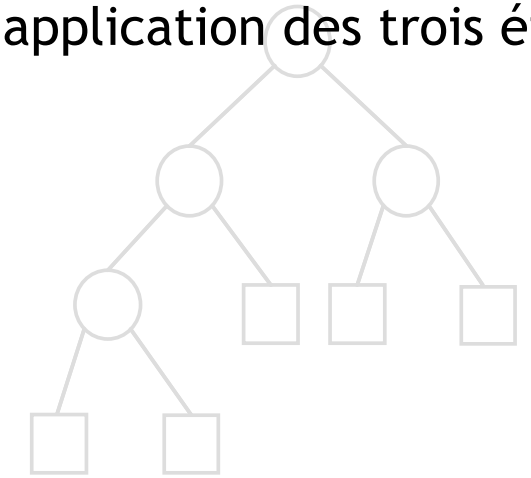
→ Construction d'arbre de régression ou de discrimination selon la nature de la variable à expliquer

NOTATIONS

→ X_k $(k=1,\ldots,p)$ variables explicatives

→ Y variable à expliquer qualitative

La construction d'un arbre de discrimination par la méthode CART repose sur l'application des trois étapes suivantes :



1. Construction de l'arbre maximal
2. Etape d'élagage
3. Sélection de l'arbre optimal

Exemple d'arbre de discrimination issu du domaine médical

Patient victime d'une crise cardiaque

San Diego Medical Center
University of California

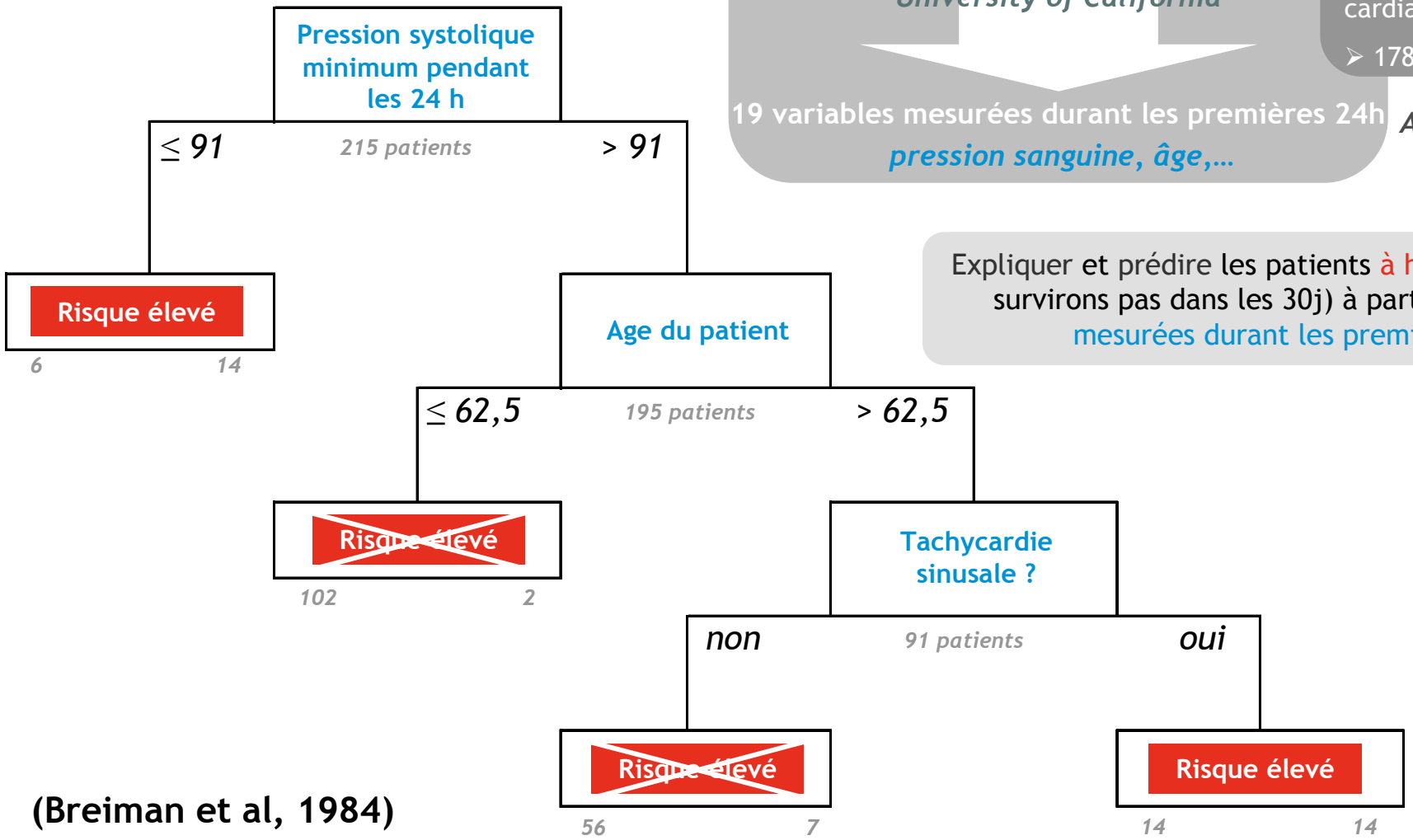
19 variables mesurées durant les premières 24h
pression sanguine, âge,…

215 patients

- 37 ne survivent pas dans les 30j qui suivent la crise cardiaque
- 178 « survivants »

Arbre de discrimination

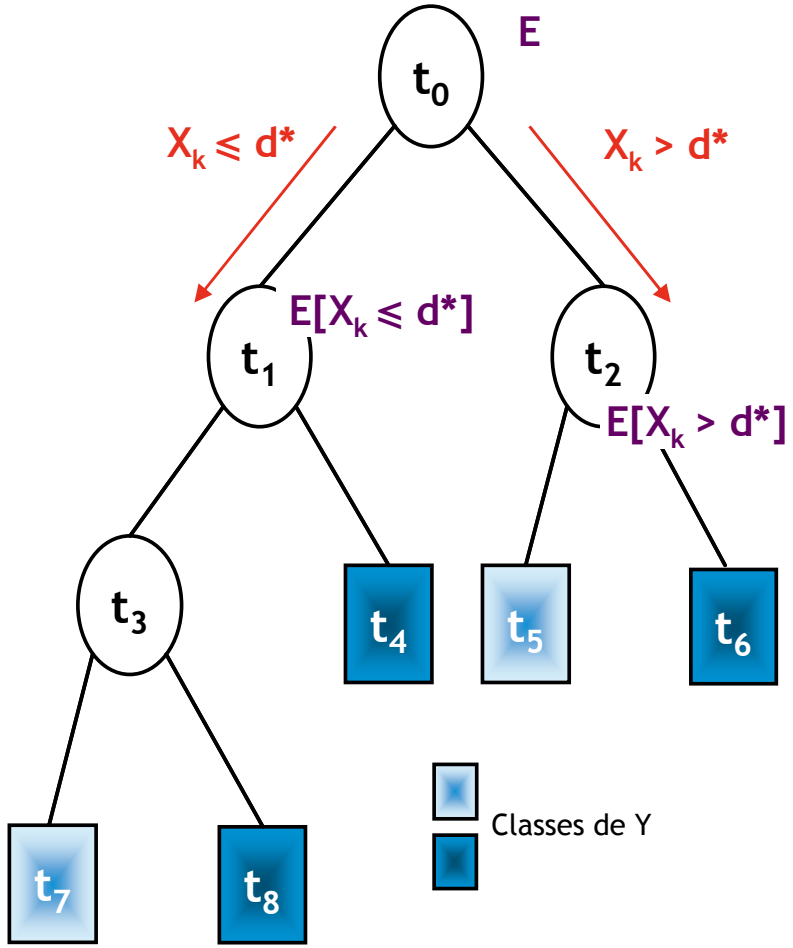
Expliquer et prédire les patients à haut risque (qui ne survirons pas dans les 30j) à partir des données mesurées durant les premières 24h



(Breiman et al, 1984)

La méthode CART

1 Construction de l'arbre maximal



Critère de division d'un nœud

→ basé sur une fonction d'hétérogénéité *i(t)* qui va mesurer le degré de mélange des classes *m* de Y dans un nœud.

$$i(t) = -\sum_{k=1}^{m} P(k/t)\log(P(k/t))$$

où P(k/t) est la proportion d'observations dans la classe k de Y au nœud t

Chaque division *d* au nœud t entraîne une réduction de l'hétérogénéité exprimée par :

$$\Delta i(d,t) = i(t) - p_g i(t_g) - p_d i(t_d)$$

où p_g et p_d sont les proportions d'observations du nœud t respectivement dans les nœuds descendants t_g et t_d

Meilleure division d* (d'un nœud t) :

$$\Delta i(d^*,t) = Max\{\Delta i(d,t); d \in D\}$$

(Breiman et al, 1984)

La méthode CART

2 Etape d'élagage

Construction d'une séquence d'arbre S entre A_{max} et sa racine

→ Suppression des branches de grande taille les moins informatives

Paramètre de complexité, critère d'élagage (Guéguen et Nakache, 1988)

$$k(t) = \frac{R(t) - R(A_i^t)}{\left|\tilde{A}_i^t\right| - 1}$$

R(t) est le taux de mauvais classement (échantillon d'apprentissage) au nœud t

$R(A_i^t)$ est le taux de mauvais classement (échantillon d'apprentissage) de l'arbre A_i^t

$\left|\tilde{A}_i^t\right|$ est le nombre de nœuds terminaux de l'arbre A_i^t

3 Sélection de l'arbre optimal

- Validation croisée (lorsque échantillon d'apprentissage petit)
- Échantillon de validation (lorsque échantillon d'apprentissage suffisamment grand)

→ Pour chacun des arbres de la séquence S, calcul du taux de mauvais classement à partir d'un nouvel échantillon (échantillon de validation) qui n'a pas participé à la construction de arbre

Application de la méthodologie : le cas (laitue / ^{90}Sr)

Scénario : contamination accidentelle de ^{90}Sr (voie foliaire) / laitue

Utilisation de l'équation dérivée du modèle radioécologique Astral (Mourlon et Calmon, 2002) pour modéliser la contamination d'un légume feuille à la suite d'un dépôt accidentel de ^{90}Sr :

$$C_{veg} = \frac{D\,R_c(t)\,e^{-(\lambda_b + 6.8E-05)(Tc-t)}}{R_{dt}}$$

D : dépôt (Bq/m^2)

R_{dt} : rendement cultural à la récolte (kg/m^2)

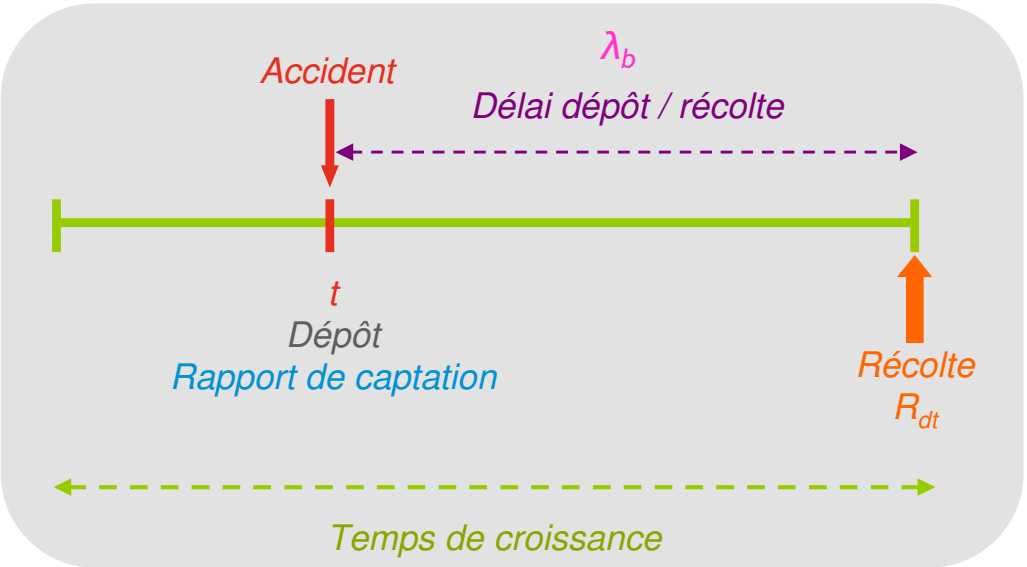
λ_b : constante de décroissance biomécanique (j^{-1})

T_c : temps de croissance du végétal (j)

t : date de l'accident (j)

$R_c(t)$: rapport de captation à la date t (sd)

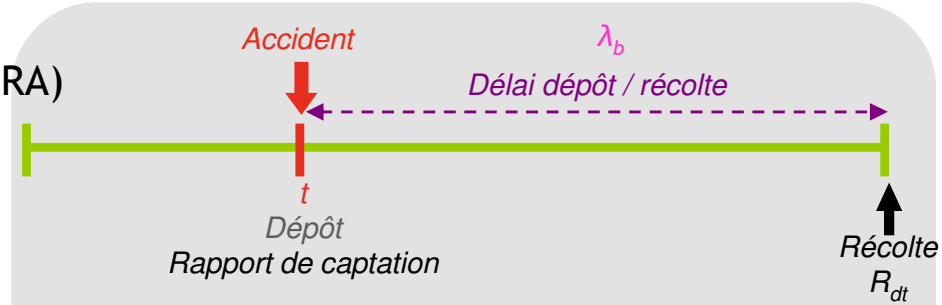
C_{veg} (Bq/kg) : concentration dans le végétal (à la récolte) résultant du transfert foliaire



Etape 1 | Echantillonnage des valeurs d'entrées et codage de Z

Le temps de croissance

➔ Loi triangulaire (min=30, mode=60, max=90) (AGRIAL, INRA)



$$C_{veg} = \frac{D\,R_c(t)\,e^{-(\lambda_b + 6.8E-05)(Tc - t)}}{R_{dt}}$$

La date de l'accident

➔ Loi uniforme [0, T_c]

Le dépôt

Borne max: valeurs proposées par les critères de zonages définies par la loi de 1991 : 111 kBq.m^{-2} valeur charnière entre zone de relogement « consécutif » et zone de relogement obligatoire et immédiat.

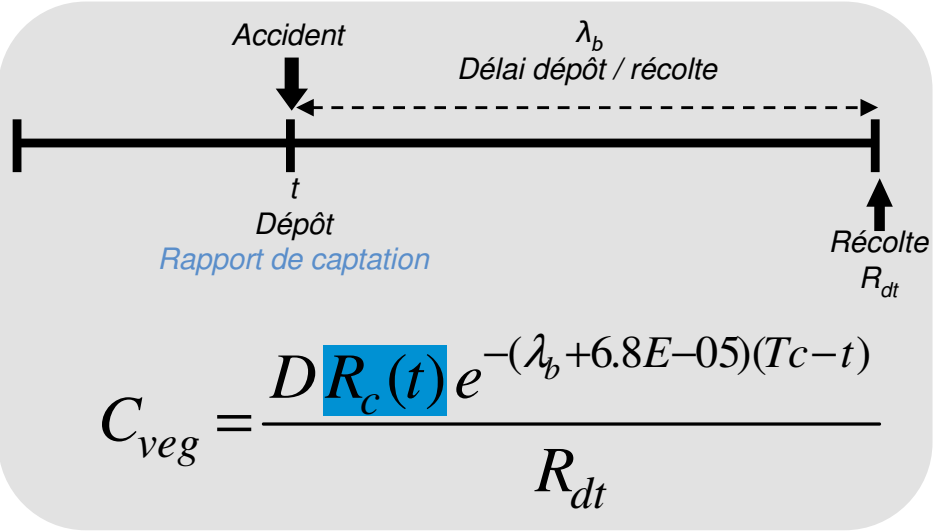
Borne min : en dessous de cette valeur, la concentration dans le végétal C_{veg} n'atteint jamais un seuil fixé à 10 Bq.kg^{-1} jugé négligeable.

➔ Loi uniforme [20, 100000]

La constante de décroissance biomécanique

➔ Loi triangulaire (min=0,03, mode=0,046, max=0,14) (GRNC, 2001) et (Renaud et al. 1999)

Le rapport de captation



Fraction du dépôt interceptée par les parties aériennes des végétaux

→ *Utilisation d'un modèle agronomique de culture (STICS) pour le modéliser*

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

ECOSYS-87 (Müller et Pröhl, 1993)

Vg_i : Vitesse de dépôt sur la plante i ($m.s^{-1}$),

Vg_{imax} : Vitesse de dépôt maximale sur la plante i ($m.s^{-1}$),

LAIi : 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.

De Tourdonnet (1998)

Le rapport de captation

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

Avec f=1,196 (valeur ajustée) et R=0,08 correspond à la réflectance du couvert (De Tourdonnet, 1998)

εamax : efficience d'absorption maximale du rayonnement

k : coefficient d'extinction du rayonnement dans le couvert végétal

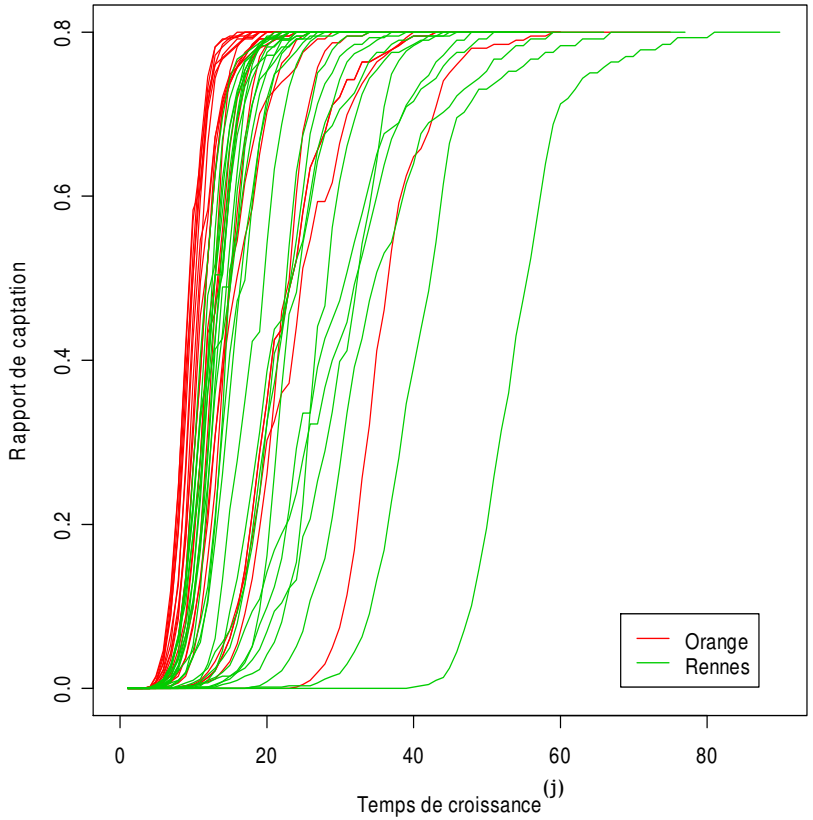
Le rapport de captation

Simulations réalisées avec STICS et données utilisées

(Simulateur mulTIdisciplinaire pour les Cultures Standard)

Sélection de deux stations :

- Orange *climat méditerranéen*
- Rennes *climat océanique*



Variables à renseigner :

✓ Climat

✓ Le sol

✓ L'itinéraire technique

Les données observées peuvent être approchées par une fonction sigmoïde d'équation :

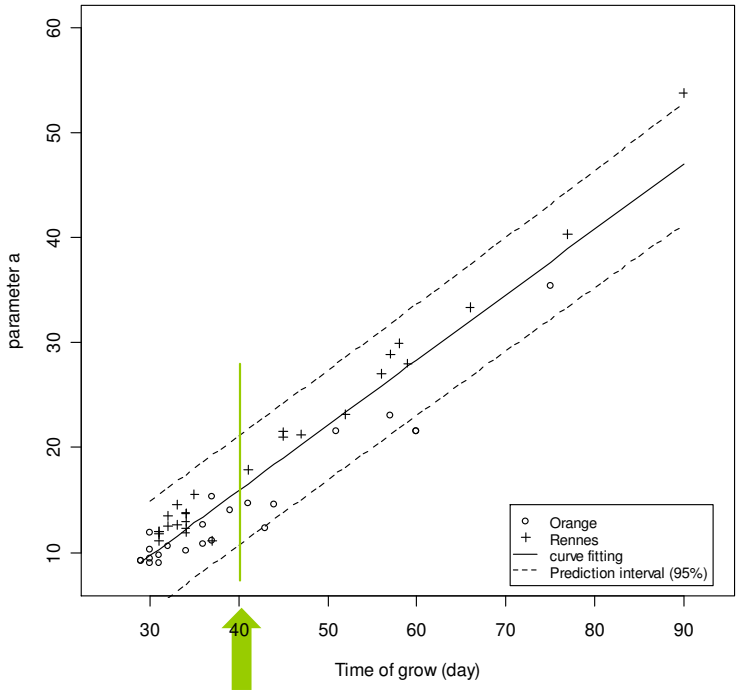
$$f(x) = \frac{0,8}{1 + e^{-b(x+a)}}$$

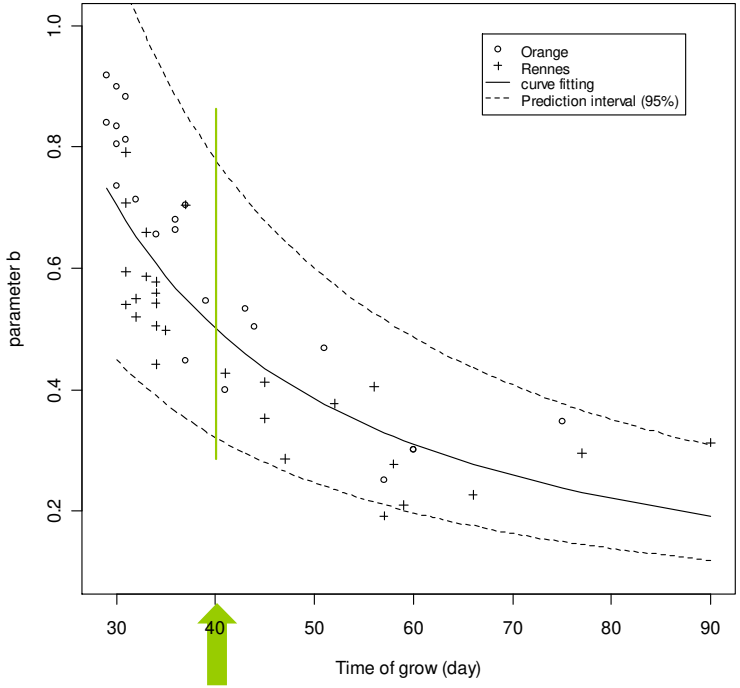
où

b contrôle la pente de la courbe

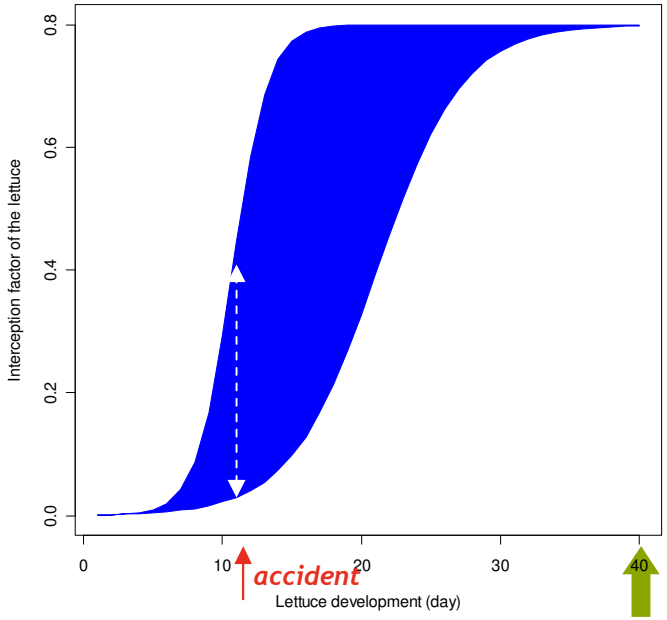
a est l'abscisse du point d'inflexion (centre de symétrie) de la courbe

Estimation de a et b : mise en évidence de relations avec le temps de croissance



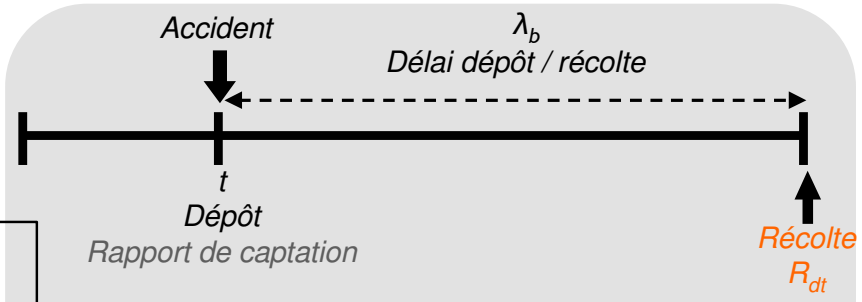


Proposition d'un « encadrement » à Tc fixé :

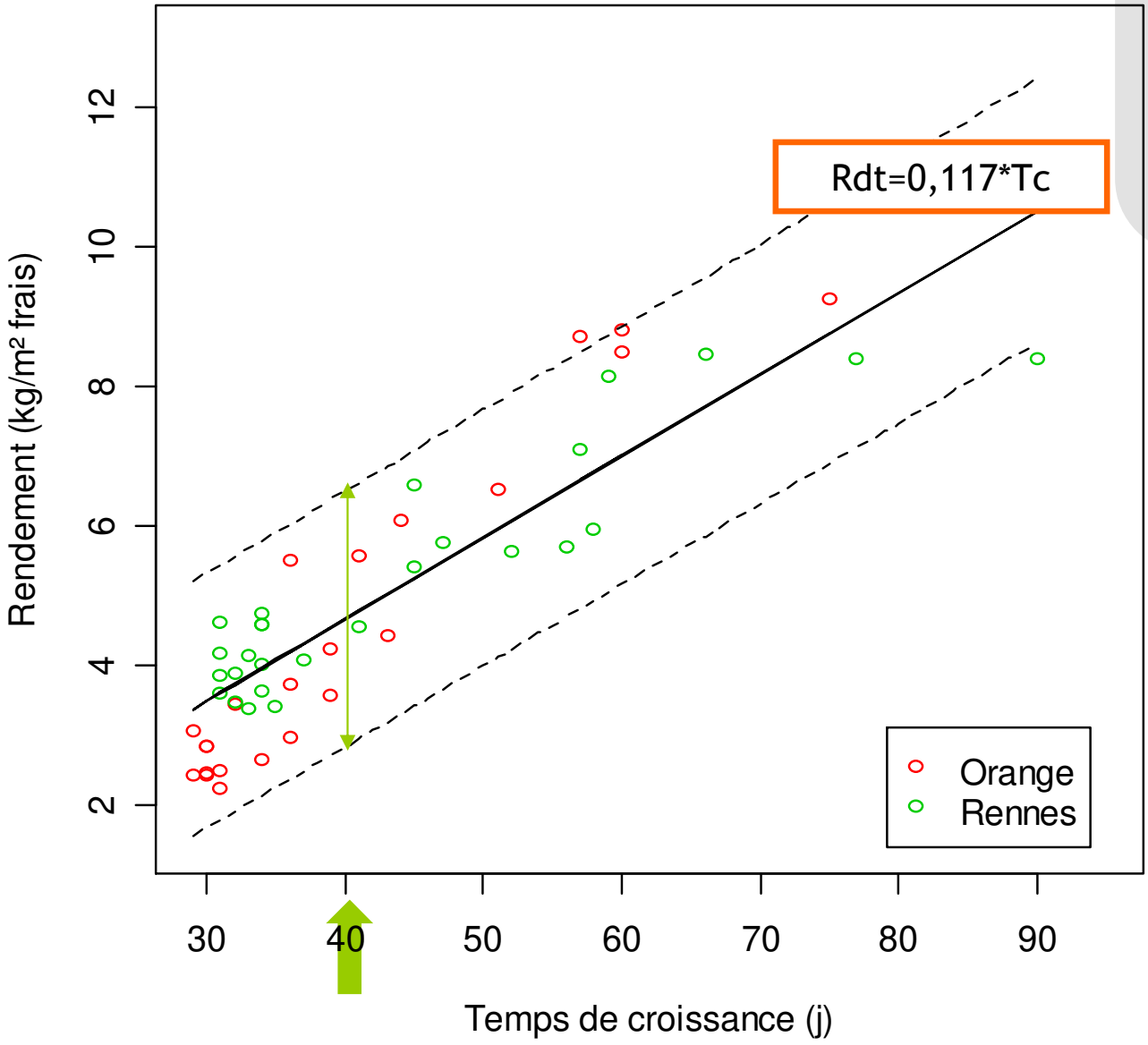


Le rapport de captation

Le rendement à la récolte



$$C_{veg} = \frac{D\,R_c(t)\,e^{-(\lambda_b + 6.8E-05)(Tc-t)}}{R_{dt}}$$



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

BILAN

$$C_{veg} = \frac{D\,R_c(t)\,e^{-(\lambda_b + 6.8E-05)(Tc-t)}}{R_{dt}}$$

Evaluation de C_{veg} et codage en variable qualitative

Connaissance des distributions et des relations entre les entrées

Génération de réalisation de ces variables

Codage selon les limites réglementaires sur les concentrations des denrées commercialisées :

Radionucléides dans les aliments	Limites en Bq/kg
^{238}Pu, ^{239}Pu, ^{240}Pu, ^{241}Am	10
^{90}Sr, ^{106}Ru, ^{129}I, ^{131}I, ^{235}U	100
^{35}S, ^{60}Co, ^{89}Sr, ^{99}Tc, ^{103}Ru, ^{134}Cs, ^{137}Cs, ^{144}Ce, ^{192}Ir	1000
^{3}H, ^{14}C	10 000

D'après *Codex Alimentarius Comission*

Création d'échantillons artificiels de données :

Variable de sortie → C_{veg} codée en 2 modalités,

Variables d'entrées → Dépôt (Dep), Rapport de captation (Rc), Décroissance biomécanique (λ_b), Délai dépôt/récolte (delai) et Rendement cultural (Rdt)

Etape 2

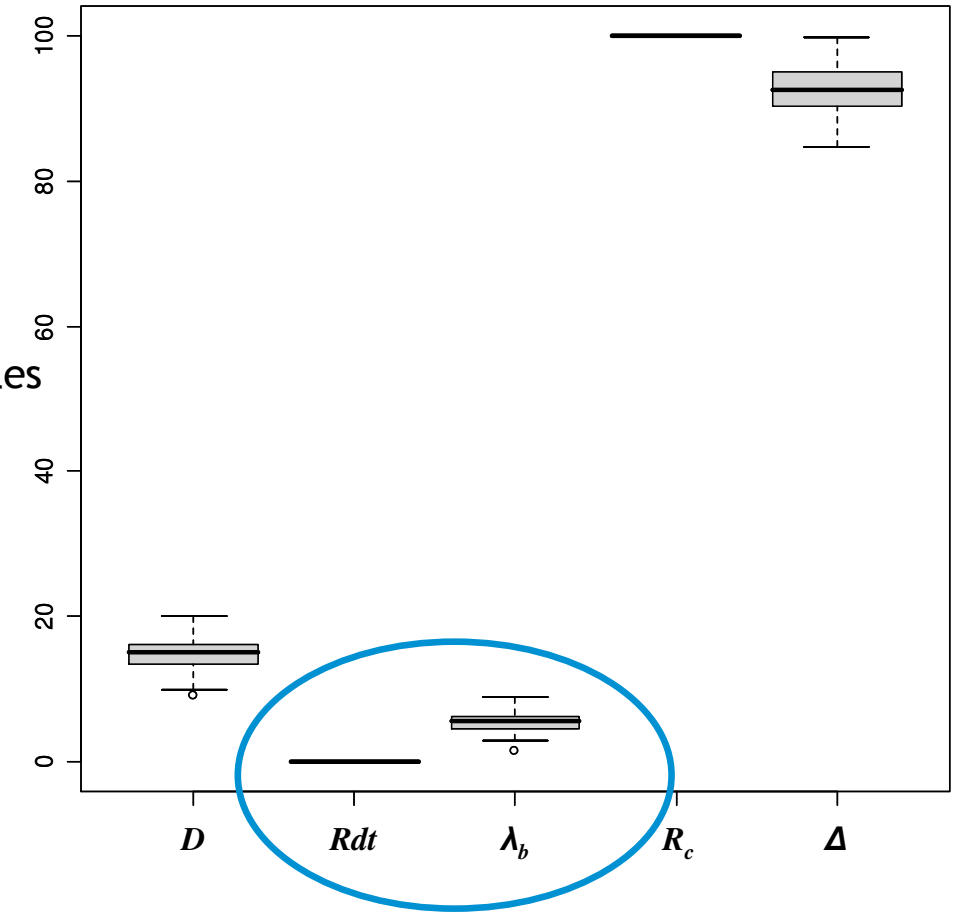
Modélisation du phénomène par des arbres de discrimination

Importance de la variable X_k : $I(X_k) = \sum_{t \in A} \Delta i(\hat{\delta}, t)$

- Application à notre cas de contamination

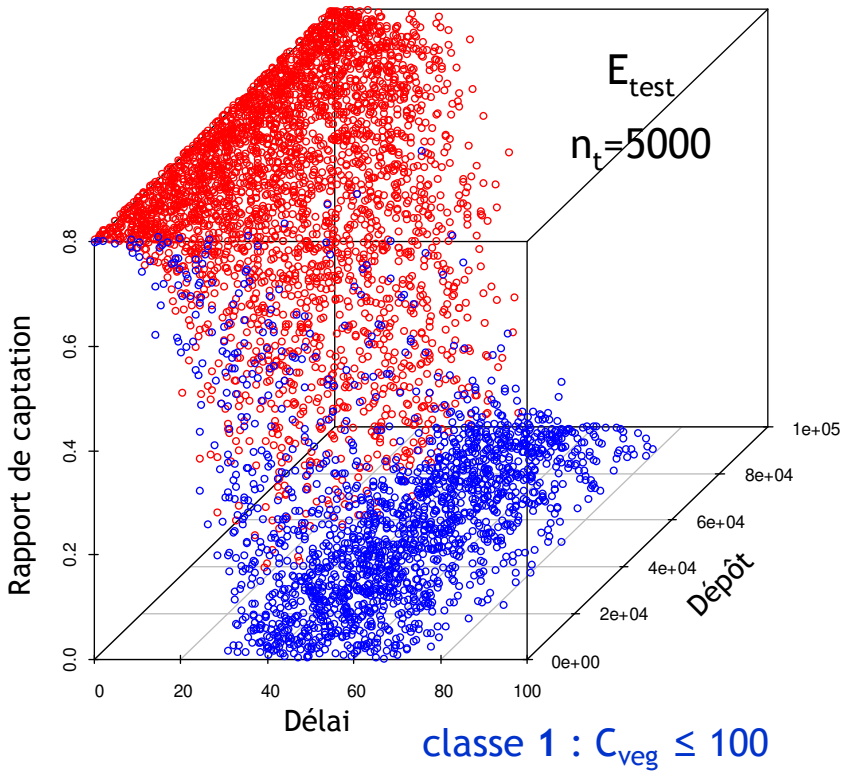
Ghattas (1999) → Instabilité dans l'importance des variables

- Génération de 100 couples d'échantillons (apprentissage, validation)
- Construction de 100 arbres de discrimination
- Détermination de l'importance des variables pour chaque arbre



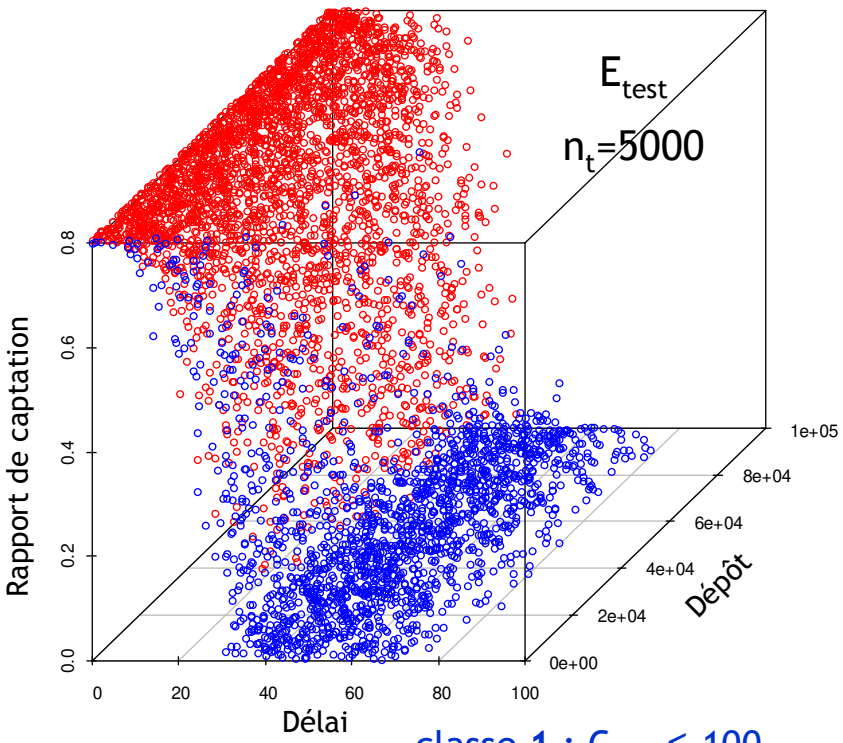
→ Sélection des variables Rc, Délai et Dépôt

Arbre de discrimination obtenu par la méthode CART

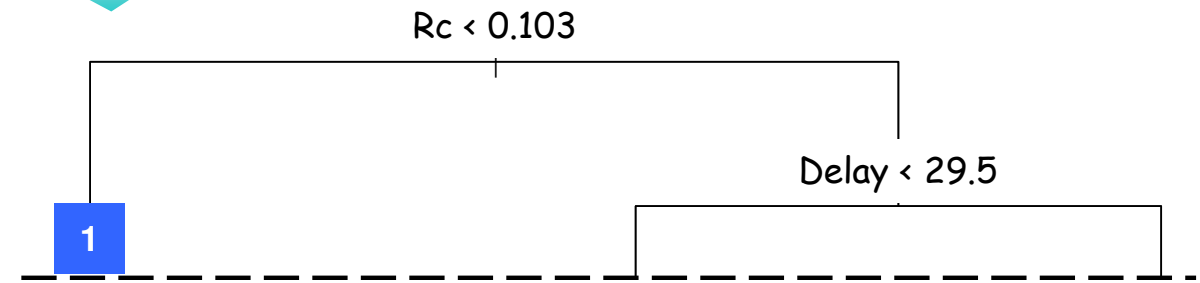


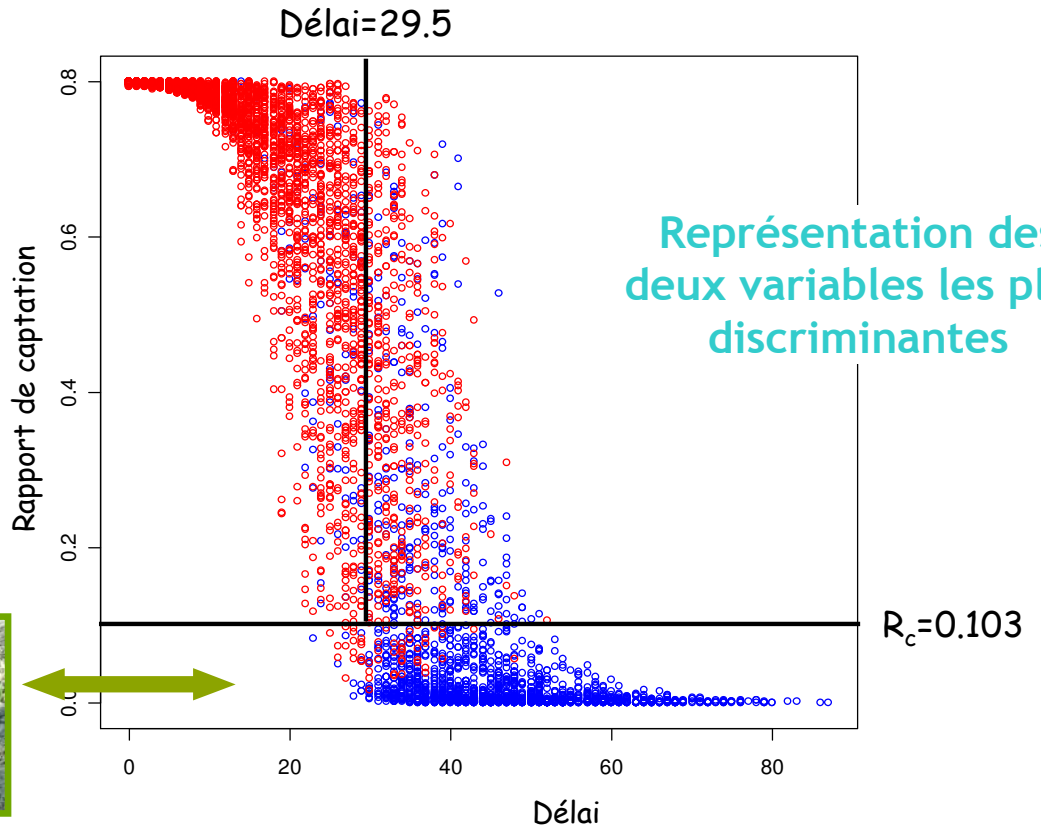
Pourcentage de mauvais classement : 5,46% (test)

classe 2 : C_{veg} > 100



classe 1 : $C_{veg} \leq 100$





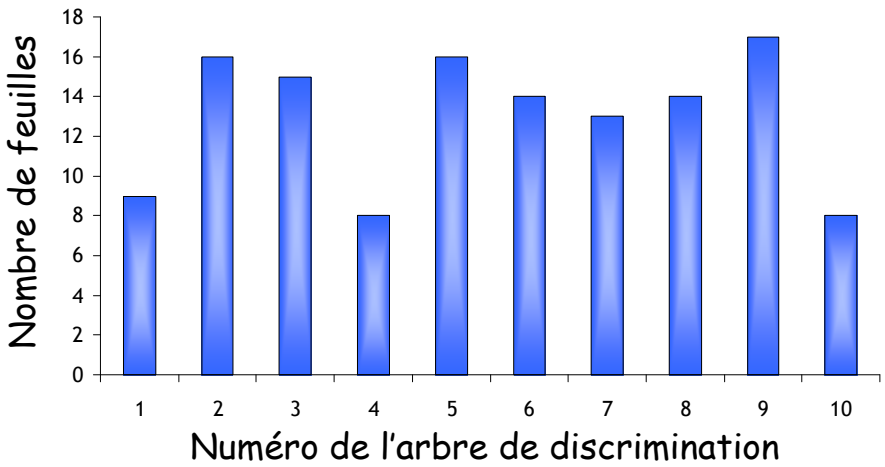
Représentation des deux variables les plus discriminantes

Axe de recherche : Réduire l'instabilité des arbres de discrimination

Illustration de l'instabilité

- Génération de 10 couples d'échantillons (apprentissage, validation) N=5000
- Construction de 10 arbres de discrimination par la méthode CART

➢ Taille des arbres



➢ Prédictions

Echantillon test n_{test}=1000

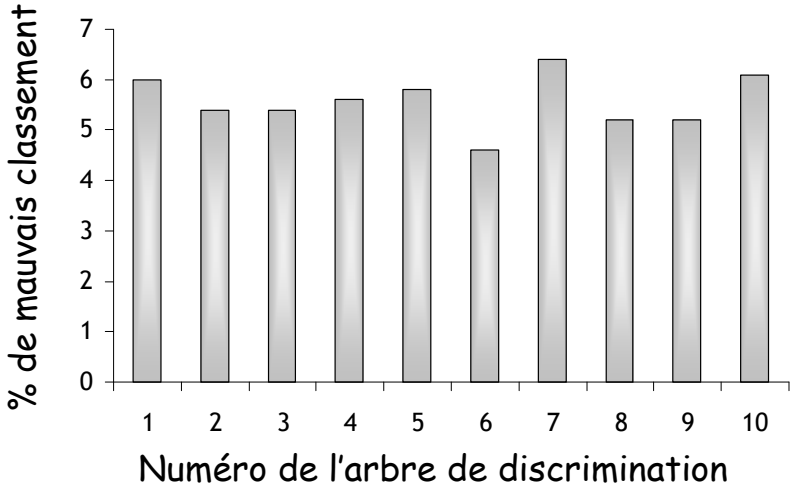
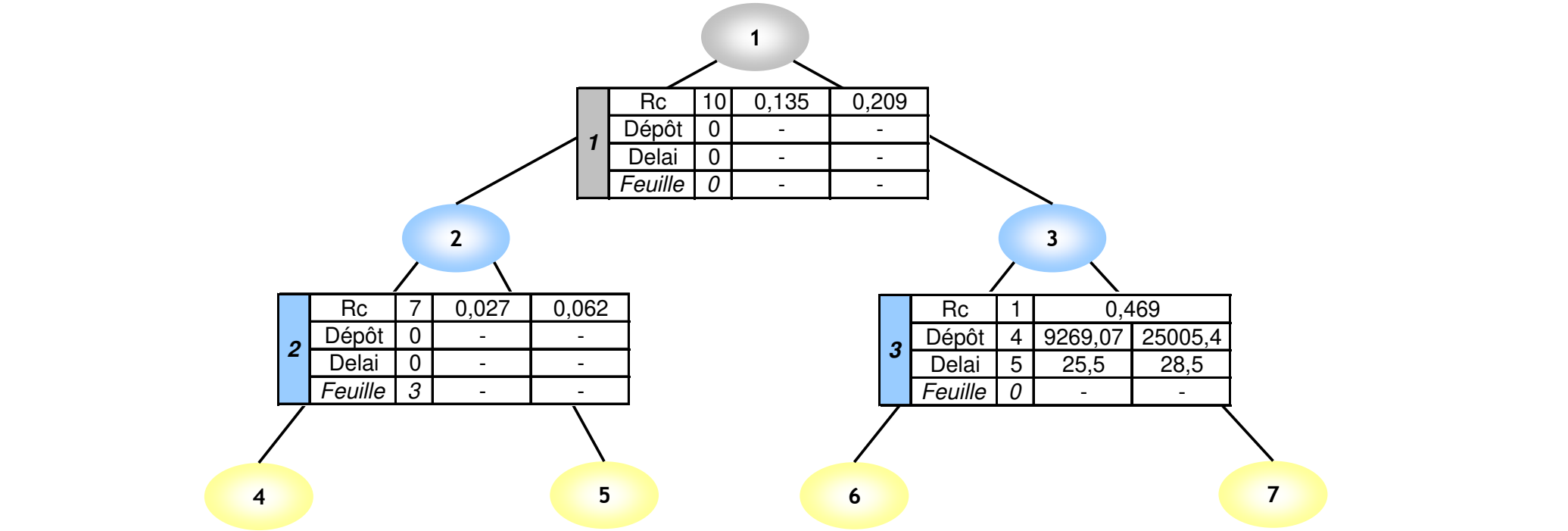


Illustration de l'instabilité



1				
	Rc	10	0,135	0,209
	Dépôt	0	-	-
	Delai	0	-	-
	Feuille	*0*	-	-

2				
	Rc	7	0,027	0,062
	Dépôt	0	-	-
	Delai	0	-	-
	Feuille	*3*	-	-

3				
	Rc	1	0,469	
	Dépôt	4	9269,07	25005,4
	Delai	5	25,5	28,5
	Feuille	*0*	-	-

4				
	Rc	0	-	-
	Dépôt	0	-	-
	Delai	0	-	-
	Feuille	*10*	-	-

5				
	Rc	0	-	-
	Dépôt	5	19438,3	50234
	Delai	2	34,5	36,5
	Feuille	*3*	-	-

6				
	Rc	0	-	-
	Dépôt	6	19,5	27,5
	Delai	4	4742,89	26528,6
	Feuille	*0*	-	-

7				
	Rc	0	-	-
	Dépôt	6	3847,85	30716,8
	Delai	3	26,5	35,5
	Feuille	*1*	-	-

Instabilité dans les règles de décision

Méthodes de stabilisation

Méthodes basées sur l'agrégation d'arbre de discrimination

(Bagging (Breiman, 1996), Random Forest (Breiman, 2001),…)

- 👍 Nette amélioration des prédictions
- 👎 Perte de la structure de l'arbre

Méthode de stabilisation par rééchantillonnage dans les nœuds (Dannegger, 2000)

ALGORITHME

Pour chaque nœud t de taille L_t *Faire* :

Pour b=1 à B *Faire* :

Générer 1 échantillon bootstrap $L_t^{(b)}$

Rechercher la division optimale sur chacune des variables explicatives

Fin Pour

Sélection (vote) de la variable qui sera utilisée pour effectuer la division

Détermination de la division pour la variable choisie:

division=médiane(réplications bootstrap)

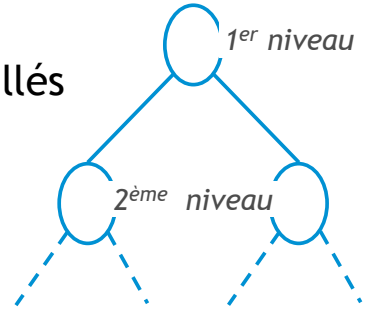
Fin Pour

Construction d'un arbre de discrimination par la méthode de stabilisation par rééchantillonnage dans les nœuds

1 Construction d'un arbre maximal

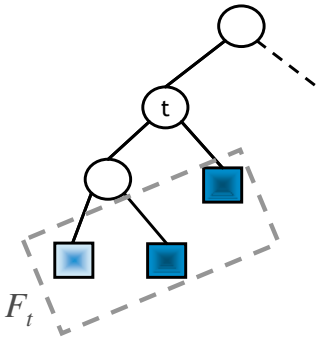
En lien avec le contexte de l'étude, nous ne voulons pas obtenir des arbres trop détaillés

→ l'arbre maximal se limite à 5 niveaux



2 Elagage de l'arbre maximal

1) Les branches apportant aucune information sont supprimées



$$R_{mc}(t) = \frac{mc(t) - mc(F_t)}{N(t)}$$

$mc(t)$: nombre d'observations mal classées au nœud t

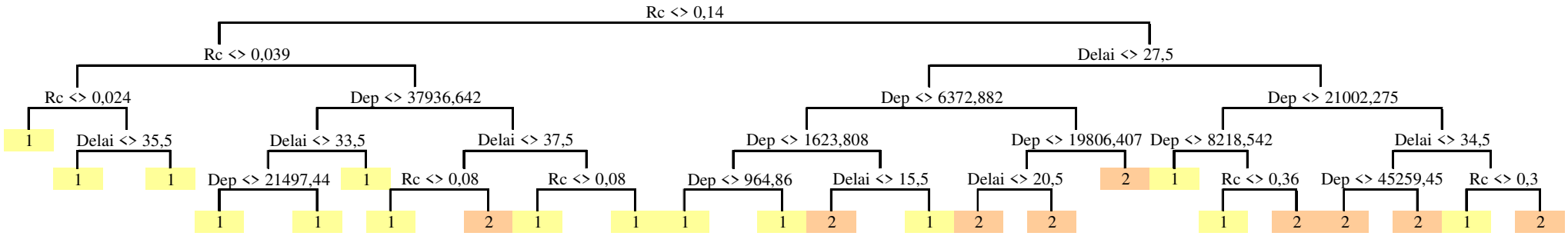
$mc(F_t)$: nombre d'observations mal classées dans les feuilles issues de t

$N(t)$: nombre d'observations au nœud t

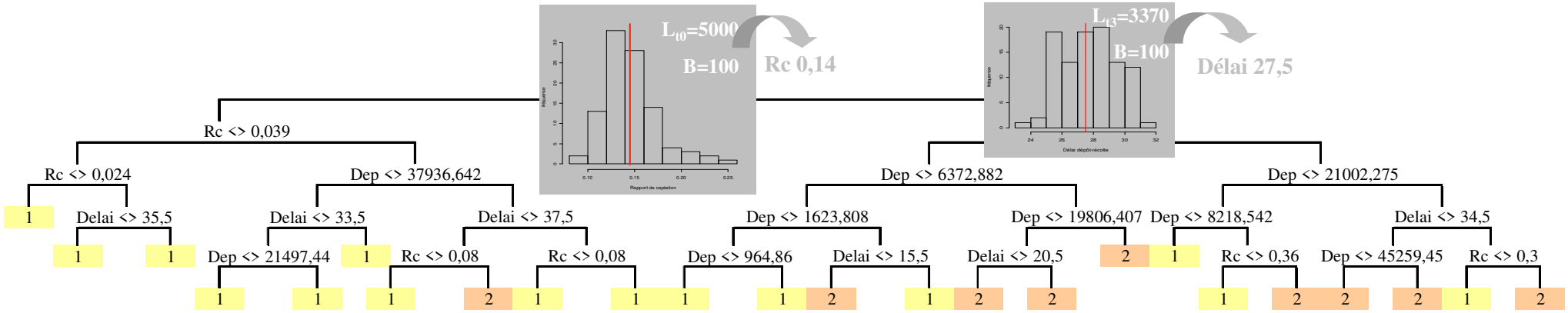
Choix des nœuds pour lesquels R_{mc} est | nul / négatif → l'arbre est alors élagué aux nœuds sélectionnés

2) L'expert peut intervenir dans la procédure d'élagage : suppression des branches qui lui paraissent peu pertinentes

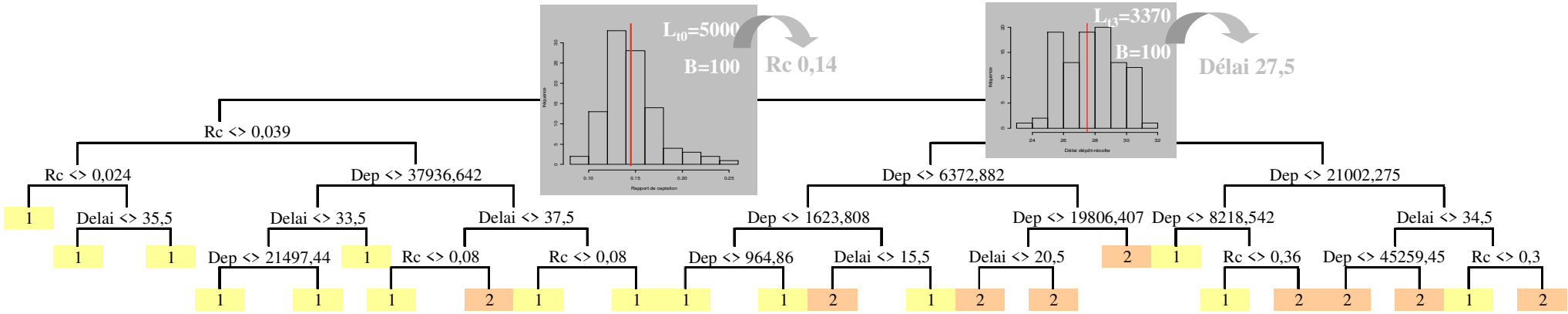
Application (laitue, ^{90}Sr)



Application (laitue, ^{90}Sr)



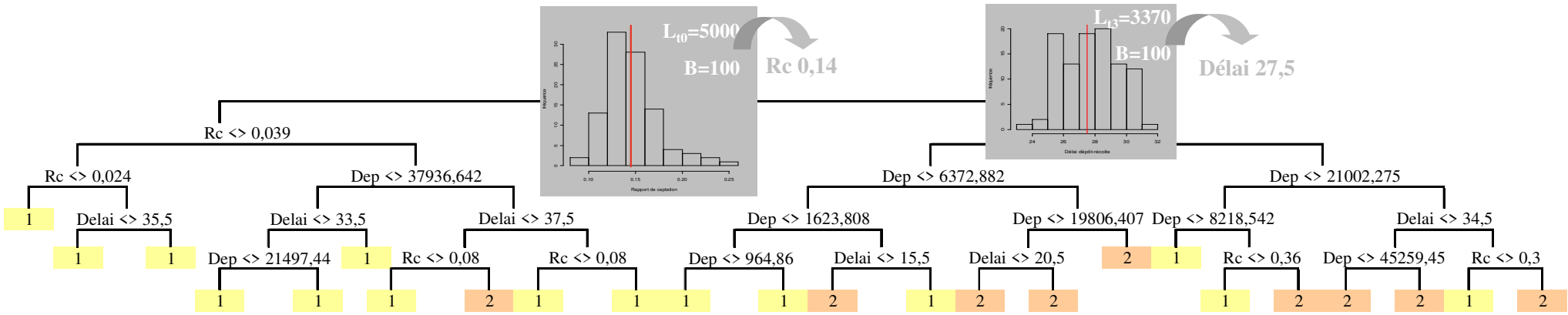
Application (laitue, ^{90}Sr)



nœud	*rac*	*g*	*d*	*gg*	*gd*	*dg*	*dd*	*ggd*	*gdg*	*gdd*	*dgg*
$R_{mc}(t)$	0,313	0,011	0,032	0	0,051	0,013	0,101	0	0	0,082	0,213

dgd	*ddg*	*ddd*	*gdgg*	*gddg*	*gddd*	*dggg*	*dggd*	*dgdg*	*ddgd*	*dddg*	*dddd*
0	0,013	0	0	0,106	0	0	0,063	0	0,020	0	0

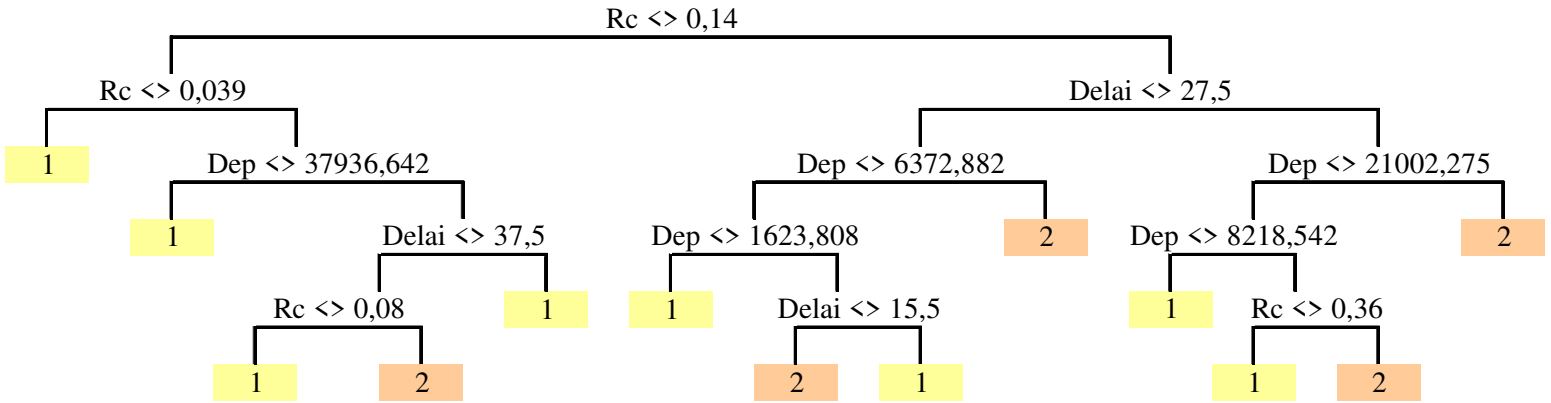
Application (laitue, ^{90}Sr)



Suppression des branches non informatives

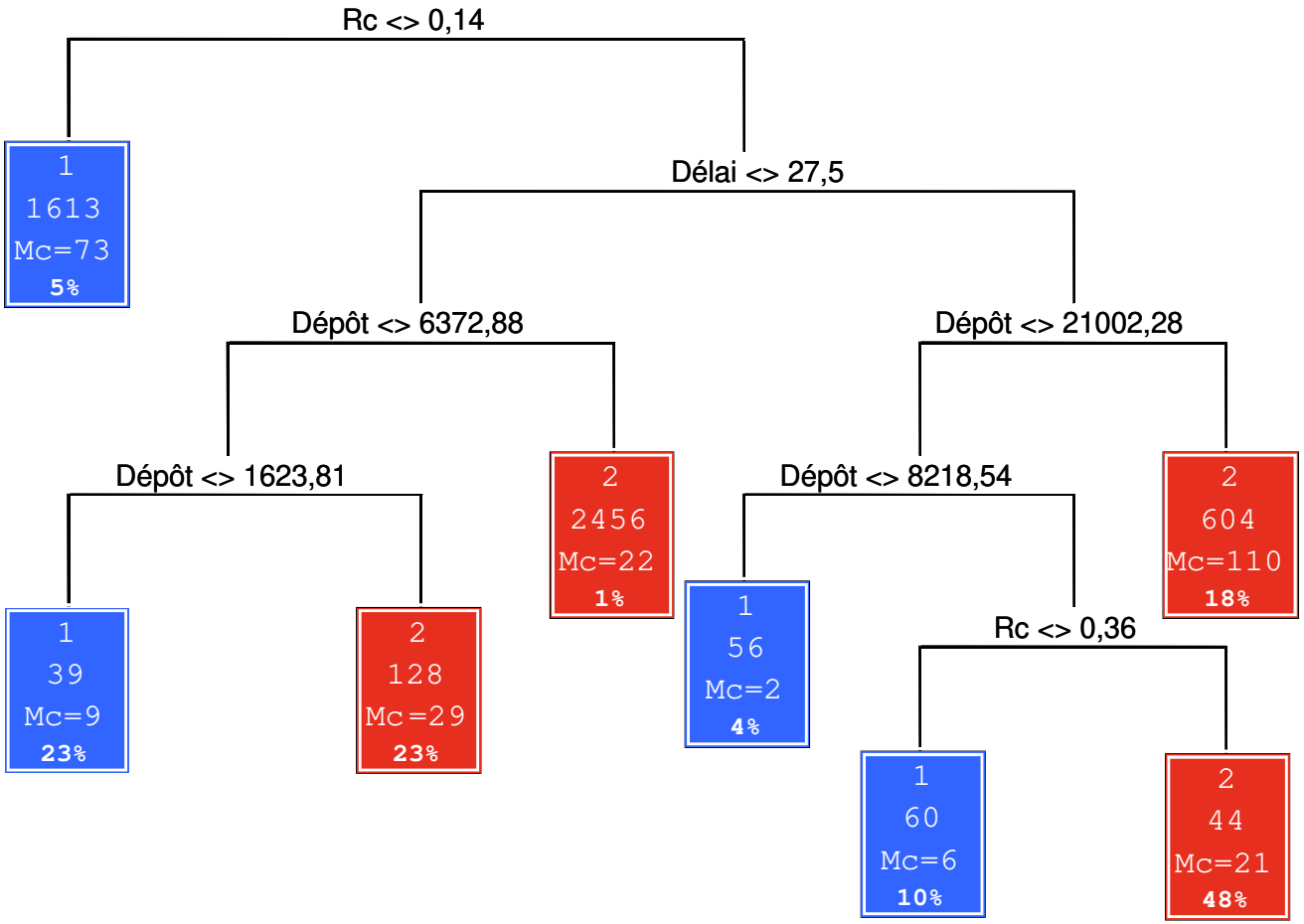
nœud	*rac*	*g*	*d*	*gg*	*gd*	*dg*	*dd*	*ggd*	*gdg*	*gdd*	*dgg*
$R_{mc}(t)$	0,313	0,011	0,032	0	0,051	0,013	0,101	0	0	0,082	0,213

dgd	*ddg*	*ddd*	*gdgg*	*gddg*	*gddd*	*dggg*	*dggd*	*dgdg*	*ddgd*	*dddg*	*dddd*
0	0,013	0	0	0,106	0	0	0,063	0	0,020	0	0



Application (laitue, ^{90}Sr)

Arbre de discrimination obtenu par la méthode de rééchantillonnage dans les nœuds



$$C_{veg} = \frac{D\,R_c(t)\,e^{-(\lambda_b + 6.8E-05)\Delta}}{R_{dt}}$$

class 1: $C_{veg} \leq 100$ Bq.kg^{-1}

class 2: $C_{veg} > 100$ Bq.kg^{-1}

CART	*Stabilisation par rééchantillonnage dans les nœuds*	*Bagging*	*Random Forest*
5,46%	5,44%	3,96%	3,94%

Conclusion

- Méthode CART (importance des variables)

  → Description des variables les plus discriminantes

  Rapport de captation, Délai, Dépôt

- Globalement, la méthode CART et la méthode de stabilisation dans les nœuds sont à performance équivalente

  ➲ L'arbre construit selon notre méthode propose des règles de décisions plus robustes

- La méthodologie développée nous a permis d'identifier les principaux facteurs responsables de deux niveaux de contamination radioactive de la laitue

Arbres de discrimination
Analyse de Sensibilité

Se concentrent sur les combinaisons des variables d'entrées du modèle qui conduisent à des catégories prédéterminées de la sortie
(Ex : valeurs extrêmes, (Mishra et al, 2003),…)

- Perspectives

Comparer, à l'aide d'une mesure, la stabilité de différent arbres de discrimination :

1/ construit par la méthode CART

2/ construit par la méthode de stabilisation dans les nœuds

Programmation réalisée sous R et S-PLUS

Construction des arbres (méthode CART)

rpart

```
Arbre<rpart(Y~X1+X2+…+Xn,data=YX,parms=list(split='information'),method="class")
```

→ Choix de l'arbre optimal uniquement par validation croisée

tree

```
Arbre<-tree(Y~X1+X2…+Xn,data=YX,split=c("deviance"),method="class")
Arbre<-prune.misclass(Arbre,newdata=test)
Afin<-prune.tree(Arbre,best=nombre_feuilles,method="misclass")
```

Représentation des arbres

```
plot(objet_arbre)
text(objet_arbre)
draw.tree(arbre,nodeinfo=T)
```

(package maptree)


Merci de votre attention

Références

Breiman L., Friedman J.H., Olshen R., and Stone C.J., (1984) Classification and Regression Trees, Wadsworth, Belmont CA.

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

Breiman L., (2001) Random Forest. *Machine Learning*, 45(1), 5-32.

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

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

Ghattas B., (1999) Importance des variables dans les méthodes CART, *Revue de Modulad*, 24, 29-39.

Gueguen A. et Nakache J.P. (1988). Méthode de discrimination basée sur la construction d'un arbre de décision binaire. *Rev. Stat. Appl.* XXXVI(1), 19-38.

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*.

La notion de division de substitution

• Probabilité qu'une observation appartienne aux nœuds t_g et t_g' :

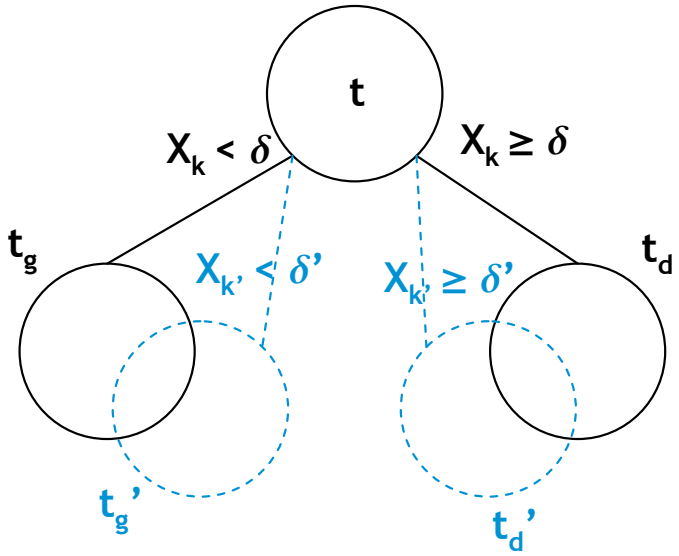
$$p(t_g \cap t_g') = \sum_{j=1}^{J} \pi_j \frac{n_j(t_g \cap t_g')}{n_j}$$

• Probabilité que δ et δ' conduisent une observation de t vers la gauche :

$$p_g(\delta, \delta') = \frac{p(t_g \cap t_g')}{p(t)}$$

• Estimation de la probabilité que la division δ' prédise correctement δ :

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



π_j : probabilité *a priori* d'appartenance à la classe *j*

$n_j(t_g \cap t_g')$: nombre d'observation appartenant à la classe j commun au nœud t_g et t_g'

$p(t)$: probabilité d'appartenir au nœud t

Une division $\hat{\delta}$ est une **division de substitution** de δ si $p(\delta, \hat{\delta}) = \max\{p(\delta, \delta'); \delta' \in D_{k'} \cup \overline{D}_{k'}\}$