IRSN
INSTITUT
DE RADIOPROTECTION
ET DE SÛRETÉ NUCLÉAIRE

USING CLASSIFICATION TREES TECHNIQUES LIKE SENSITIVITY ANALYSIS IN THE FIELD OF RADIOECOLOGY

June 2007 / Budapest

Presented by Bénédicte Briand, IRSN

Catherine Mercat-Rommens, IRSN

Gilles Ducharme, Université de Montpellier II

The radiological consequences of radioactive releases highlight the fact that the consequences of industrial pollution on man and the environment depend not only on the extent and nature of this pollution but also on the territory that is polluted.

SENSIB project (Mercat –Rommens et al, 2005)

Radioecological sensitivity represents the intensity of a region's global reaction to an accidental or chronic radioactive pollution

Identify the characteristics of French territories which strongly influence the fate of a radioactive contaminant released into the environment

This knowledge of the characteristics of French territories can then be used, in anticipation of accidental situations:

- to issue recommendations on the management of contaminated territories
- to structure decision-making

Develop a standardised tool with a single scale of indexes in order to decribe and compare the radioecological sensitivity of various territories to a pollution

Identify which factors or combinations of factors are on a prime influence on the radioactive contamination of agricultural productions

Which factors are most important in determining the radioactive contamination levels of plant

How ?

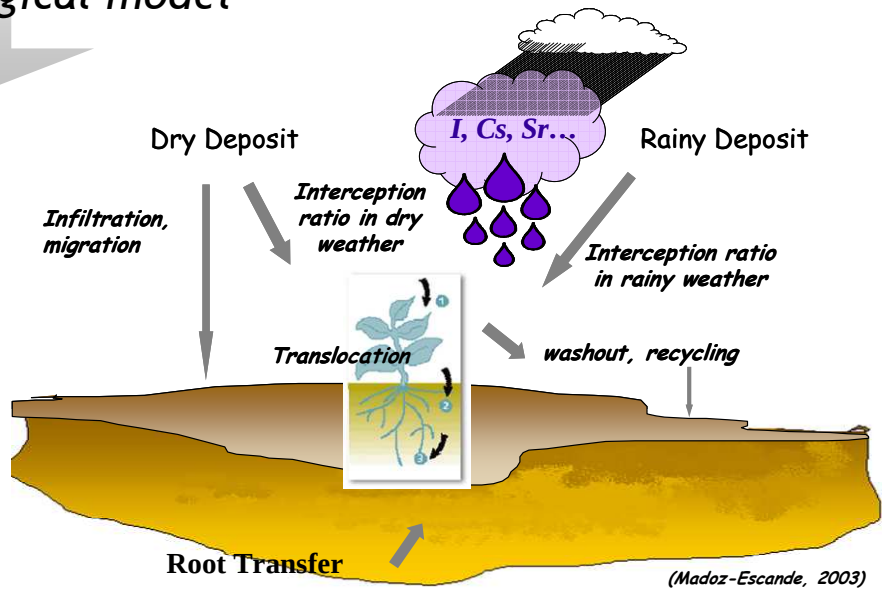
With a radioecological model

$$Z=f(X_1, X_2, \ldots, X_p)$$

Input → Agronomic and radioecological variables

Output → Radioactive contamination of the plant studied
Coded into classes according to applicable radioecological protection limits

$$Y=g(Z)=\begin{cases}1 & \text{if } Z \leq k_0\\ 2 & \text{if } Z > k_0\end{cases}$$



- Non-linear model
- Some radioecological and agronomic input are correlated

An original global sensitivity analysis is performed by using classification tree technique (Mishra et al, 2003)

Why do we use classification tree technique (CART) ?

- The model output is discrete
  - → The CART method enables the construction of classification or regression tree depending on whether the output is a continuous or a discrete variable
- Some input variables are correlated
  - → CART is a nonparametric method : it does not require any assumption regarding the structure of the input data
- In order to obtain more knowledge and precision of how the model works
  - → Contrary to the other method of global sensitivity analysis (Saltelli et al, 2000), the classification tree techniques allow to determine what input or interactions between the inputs drive the model output into particular classes

The pathways linking the input variables and the output of the model can be more precisely described and used to propose recommendations to mitigate the consequences of environmental radioactive contamination

Contamination of an agricultural production (lettuce) by the release of strontium-90 into the atmosphere

In order to estimate the radioactive contamination of the plant studied, we have chosen to use the following equation adapted from the ASTRAL radioecological model (Mourlon et al, 2002):

$$C_{veg} = \frac{DR_c e^{-(\lambda_b + 6{,}8E-05)\Delta}}{Yld}$$

C_{veg} (Bq.kg^{-1}) is the specific activity (at harvest time) resulting from foliar transfer

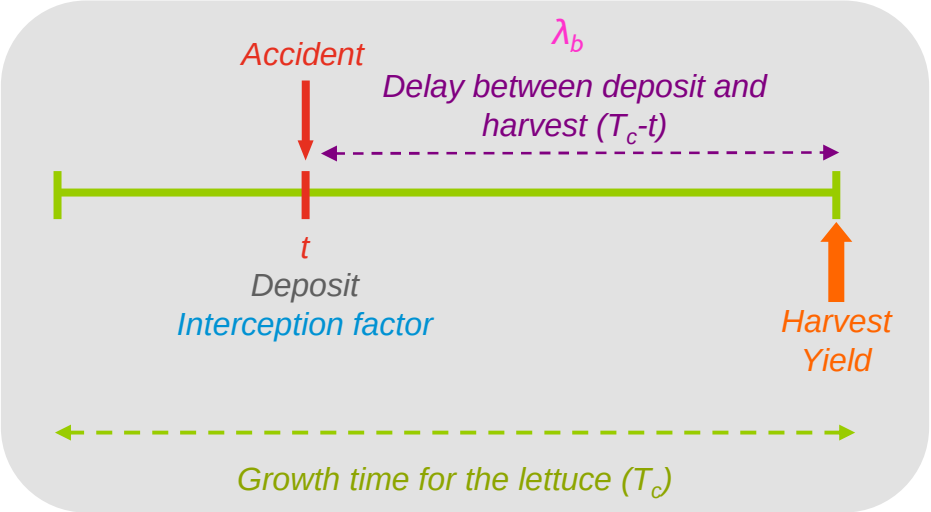
D (Bq.m^{-2}) is the deposit of radioactivity

R_c is the interception factor on the day of the accident

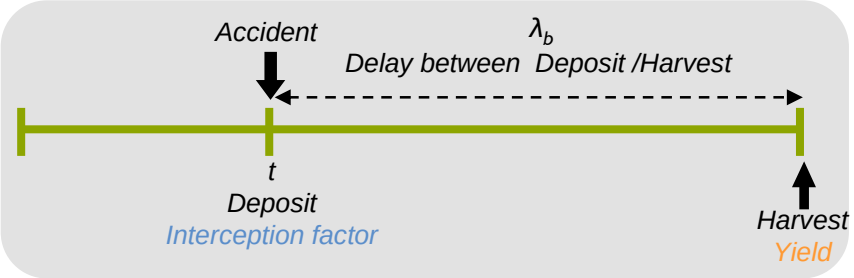
λ_b (d^{-1}) is the biomechanical decay constant of the radionuclide for the plant

Δ(d) is the delay between deposit and harvesting

Yld (kg.m^{-2}) is the crop yield at harvest time

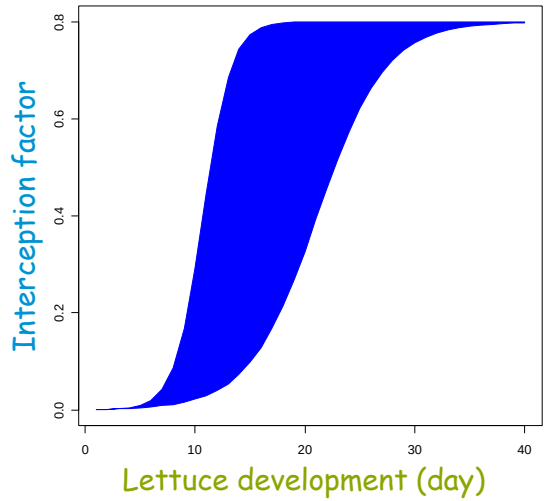


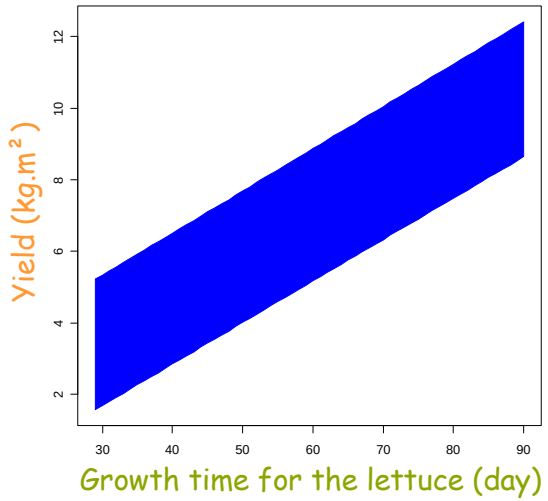
Sampling method for the sensitivity analysis



- Define distribution for the input $X_i \sim F_i \ (i=1,\ldots,p)$

  → Constraints on some input variables





(Briand et al., 2007)

- Generation of a sample of size N by random sampling
- Evaluation of the model → N output values $[C_{veg,1},\ldots, C_{veg,N}]$

These realizations are categorized according to indicative concentration limits for marketed foodstuffs (according to the Codex Alimentarius Commission (1989))

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

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

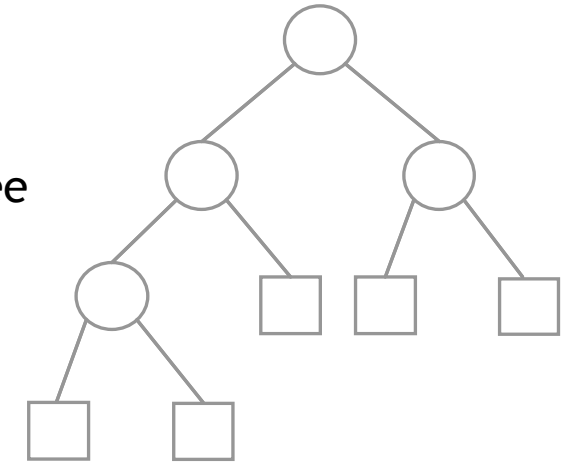
(Breiman et al, 1984)

*C*lassification *A*nd *R*egression *T*rees

- Nonparametric statistical methodology

- Alternative to linear and additive models / additive logistic models for regression problems / for classification problems

- The results are presented in the form of a binary tree



Notations

Y response variable / model output

$X_{i,i=1,\ldots,p}$ *explanatory variables / model inputs*

Example of classification tree in the field of medical decision-making

Heart attack patient

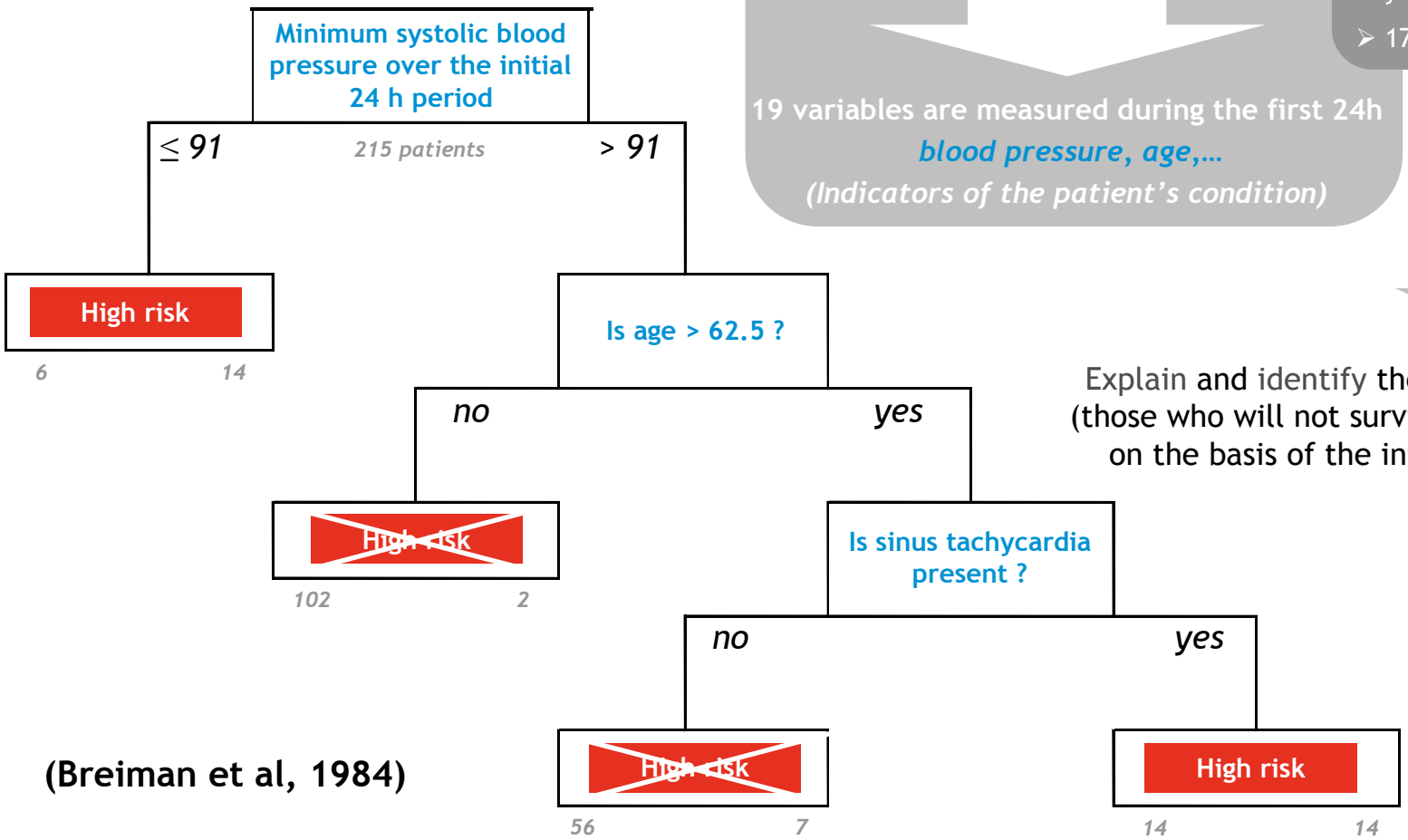
San Diego Medical Center
University of California

215 patients

- 37 died not more than 30 days following heart attack
- 178 « survivors »

19 variables are measured during the first 24h
blood pressure, age,…
(Indicators of the patient's condition)

Construction of a classification tree



Explain and identify the high risk patients (those who will not survive at least 30 days) on the basis of the initial 24 hour data

(Breiman et al, 1984)

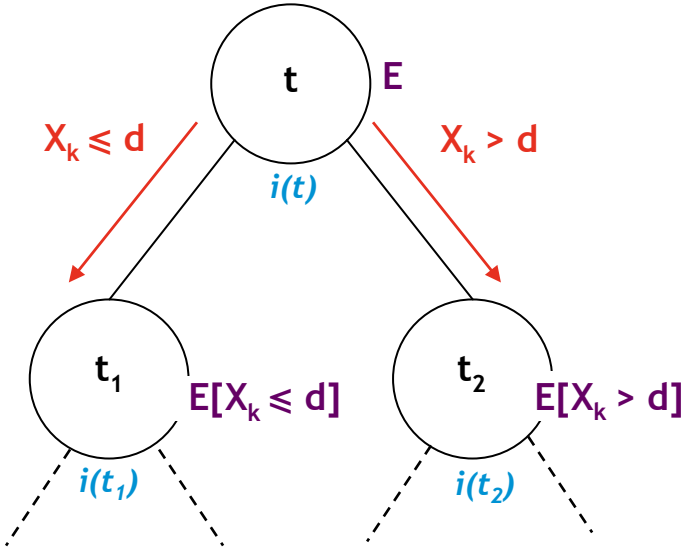
1 Building the maximal tree

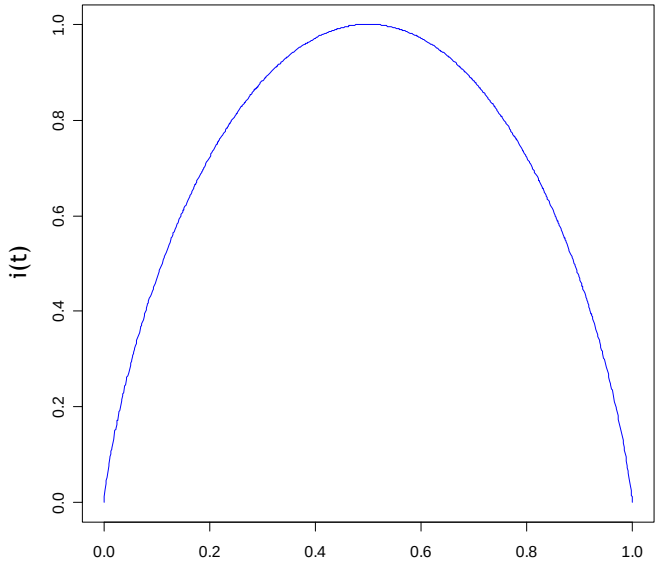
How to split a node ?

→ It is based on a impurity function $i(t)$ which measures the degree of mixture in a node

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

where $P(k|t)$ is the proportion of observations corresponding to class j of Y at node t





P(k/t) : proportion of observations corresponding to class j of Y

Each splitting d at node t leads to a decrease in impurity expressed as follows:

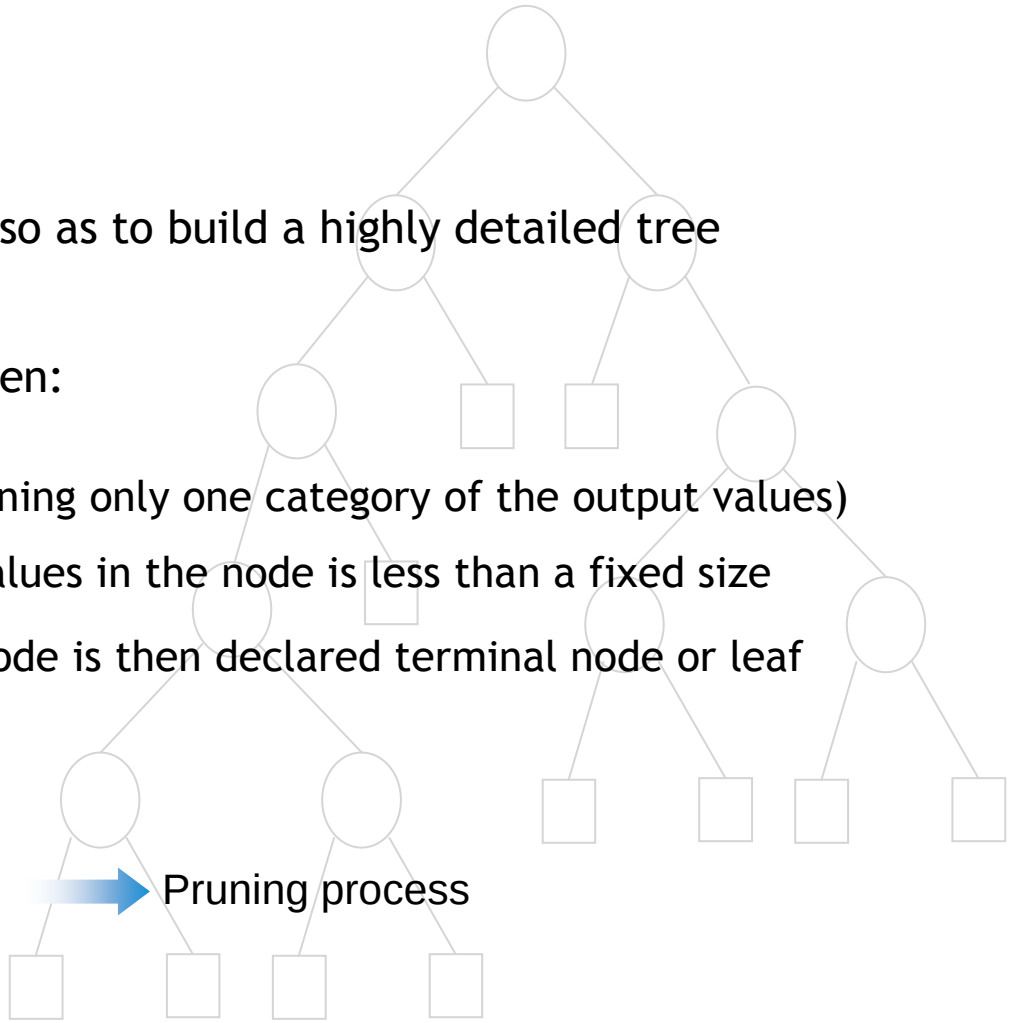
$$\Delta i(d,t) = i(t) - p_1 i(t_1) - p_2 i(t_2)$$

where p_1 and p_2 are the proportions of the node t population that go to child nodes t_1 and t_2,respectively

The selection of the best division d* (of a node t) corresponds to that maximizing the decrease in impurity:

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

1 Building the maximal tree



- The sample is successively split so as to build a highly detailed tree
- The splitting process is stop when:
  - The node is pure (containing only one category of the output values)
  - The number of output values in the node is less than a fixed size

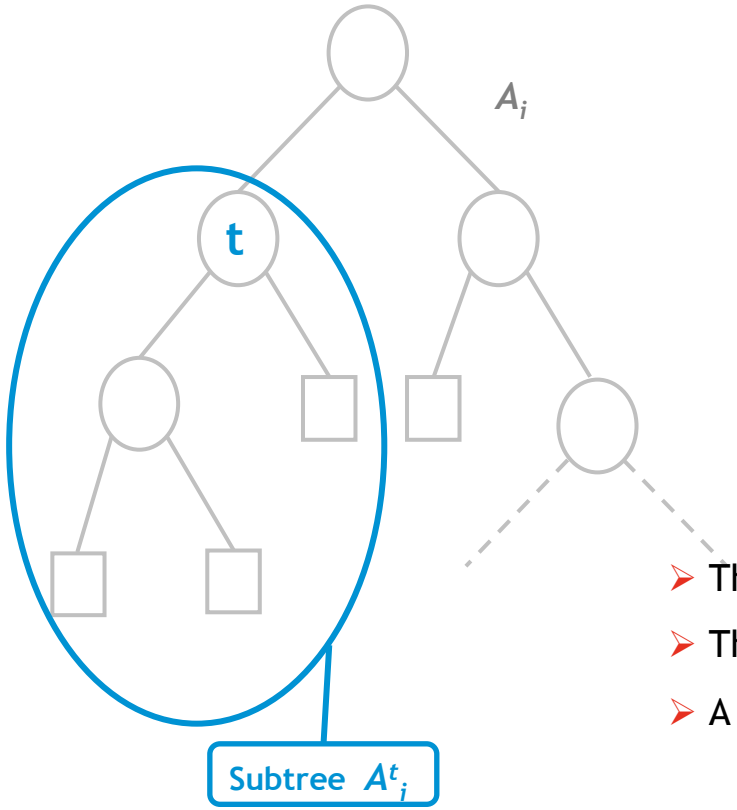
    → This node is then declared terminal node or leaf
- The maximal tree is very large

  → Pruning process

2 Tree-pruning

→ A sequence S of subtree is constructed between A_{max} and its root, $S=\{A_0,\ldots,A_{max}\}$

→ Removing the large branches that provide the least information



Pruning criterion:

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

$R(t)$ misclassification rate at the node t

$R(A_i^t)$ misclassification rate at the tree A_i^t

$\left|A_i^t\right|$ the number of terminal nodes of A_i^t

- The value of this parameter is calculated for each node of the tree A_i
- The node t for which $k(t)$ is a minimum is selected
- A new tree is build by pruning away the branch A_i^t

3 Selection of the final tree

From the obtained sequence of subtrees, the optimal tree has to be selected

→ Evaluation on the predictive error

- Cross-validation

  The sample used to build the tree is also used to choose the optimal tree

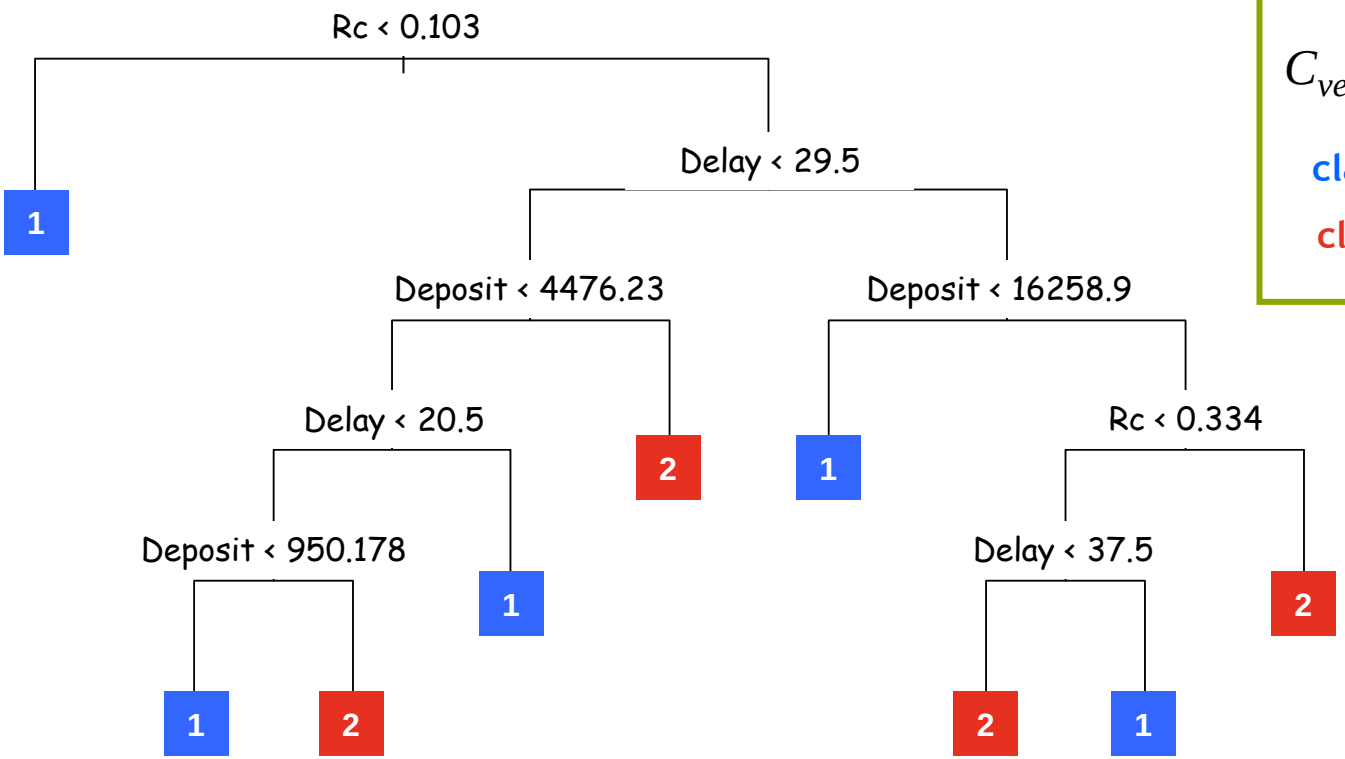
- Pruning-sample

  → A new sample (which has not been used to build the tree) is used

  → Evaluation on the misclassification rate of each subtree in the sequence

  → The tree with the lowest misclassification rate is selected

➢ Application: Radioactive contamination of ^{90}Sr to lettuce



$$C_{veg} = \frac{DR_c e^{-(\lambda_b + 6{,}8E-05)\Delta}}{Yld}$$

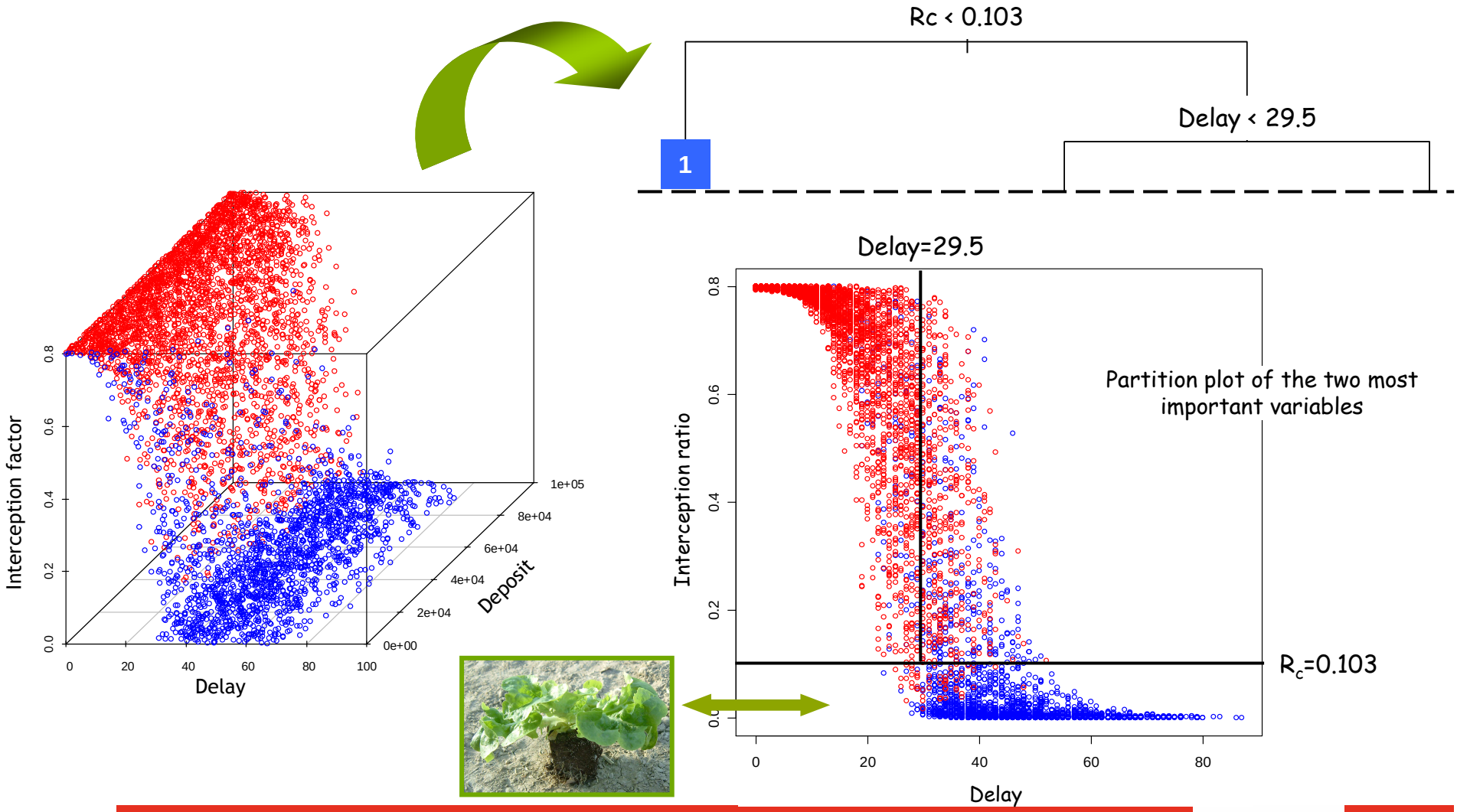
class 1: $C_{veg} \leq 100$ Bq.kg^{-1}

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

Misclassification percentage : 5,46%

→ Identify the input variables responsible to the different classes of the model output

- Application: Radioactive contamination of ^{90}Sr to lettuce



Stability of results ?

- The random sampling method was replicated several times

  The results are unstable from sample to sample:

  - Complexity of the trees (number of terminal nodes)
  - Change in splitting variables and values

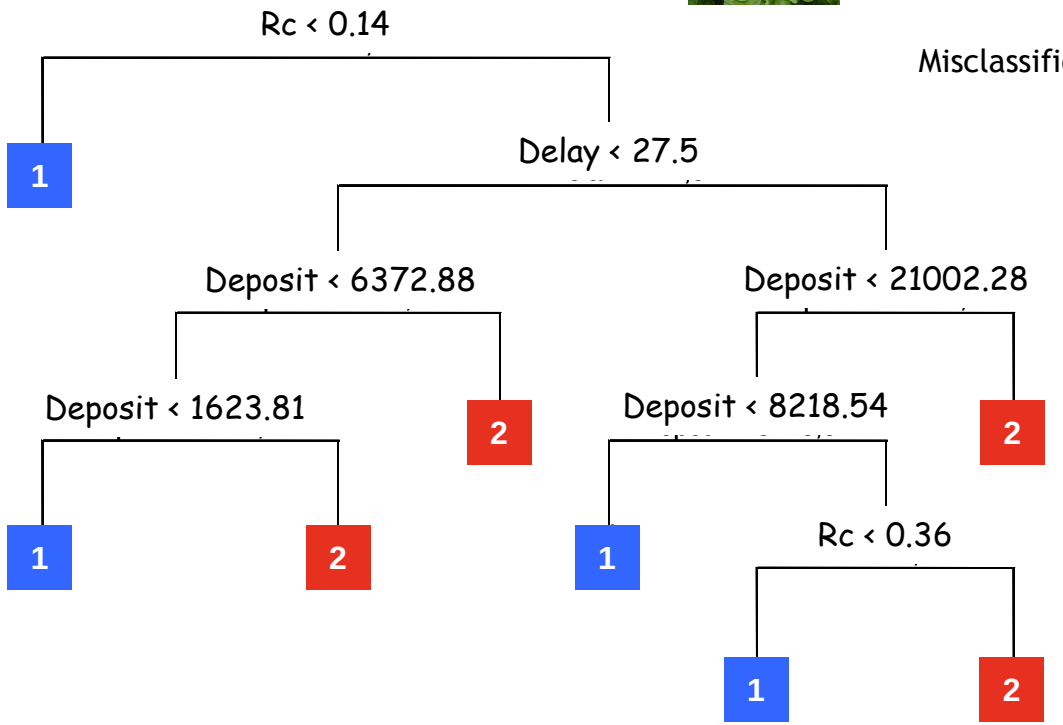
  Decision trees are known for they instability (Breiman, 1996), (Ghattas, 2000)

- In order to obtain more stable classification rules, we propose to use a node-level stabilizing procedure (Dannegger, 2000)

  - For every node t of size L_t, B samples $L_t(b)$ (b = 1, . . . ,B) are generated
  - For every sample, the best split is searched on every input variables
  - A voting procedure is performed to determine which variable is going to be used to split the node
  - For the chosen variable, the cutpoint is determined by taking the median of the replicate cutpoints

The node-level stabilization procedure

➢ Application: Radioactive contamination of ^{90}Sr to lettuce



Misclassification percentage : 5,44%

👍 This tree is more stable than the precedent and support a more robust identification of the most sensitive variables

👎 It is more expensive in computing times

Conclusion

- Application of CART
  - Determine main factors responsible for two radioactive contamination levels of the lettuce (indicative concentration limits for marketed foodstuffs)

Propose recommendation to mitigate the consequences of an environmental radioactive contamination

Classification tree techniques → *Focus on the combinations of model input values that drive the model output into particular categories (Eg.: extreme outcomes (Mishra et al., 2003),…)*

- CART has several advantages:
  - CART is nonparametric
  - CART results are invariant to monotone transformation of the input
  - CART allow more general interactions between the input variables
  - More adept at capturing nonadditive behavior
- Main disadvantage:
  - Unstable decision tree


Thank you for your attention !

References

Briand B., Durand V. et Mercat-Rommens C., (2007). Mise en évidence de relations entre variables agronomiques et radioécologiques par l'utilisation d'un modèle agronomique - Application au cas de la laitue. European Journal of Agronomy. Submit.

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

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

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., 2000. Agrégation d'arbre de classification. Revue de Statistique Appliquée, Vol.XLVIII (2), 85-98.

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*.

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

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