Apprentissage Machine / Statistique

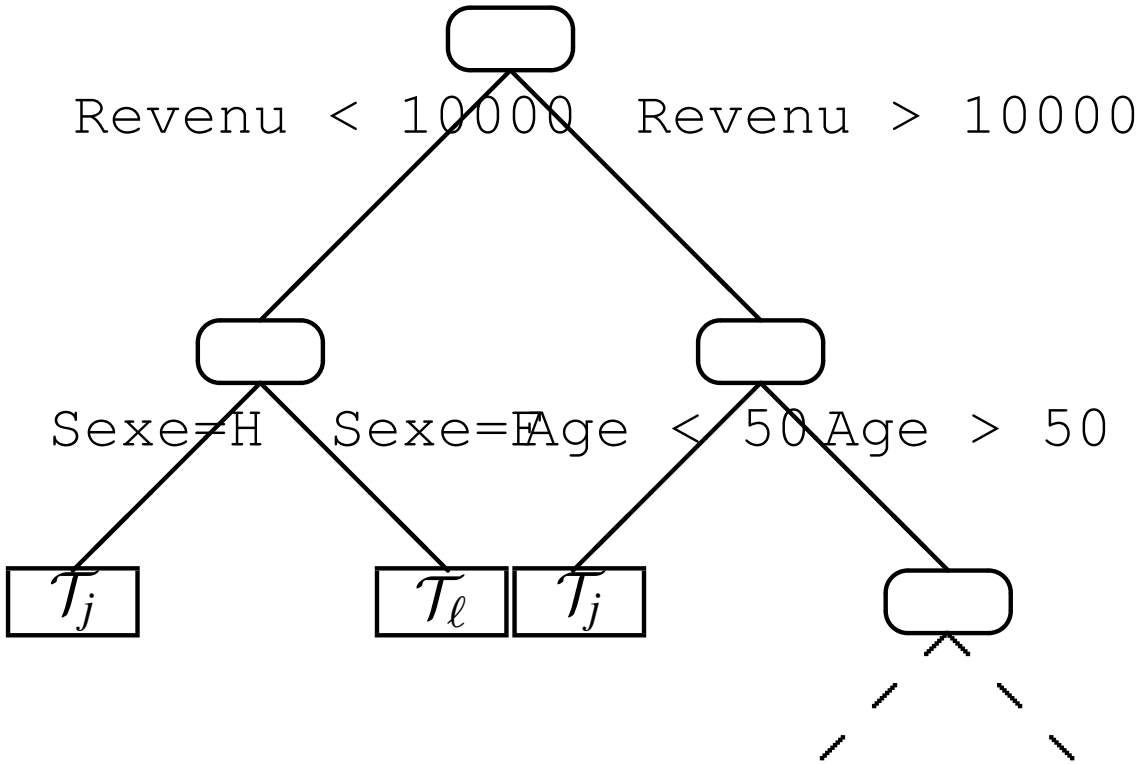
Arbres binaires de décision

PHILIPPE BESSE

INSA de Toulouse
Institut de Mathématiques

Introduction

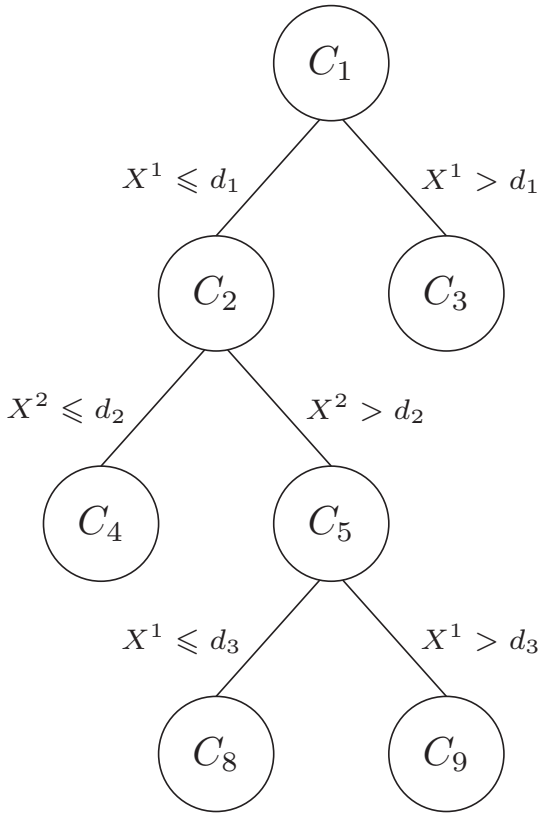
- Classification and regression trees (CART)
- Breiman et col. (1984)
- X^j explicatives quantitatives ou qualitatives
- Y quantitative : regression tree
- Y qualitative à m modalités $\{\mathcal{T}_\ell; \ell = 1 \ldots, m\}$: classification tree
- Objectif : construction d'un arbre de décision binaire simple à interpréter
- Méthodes calculatoires : pas d'hypothèses mais des données

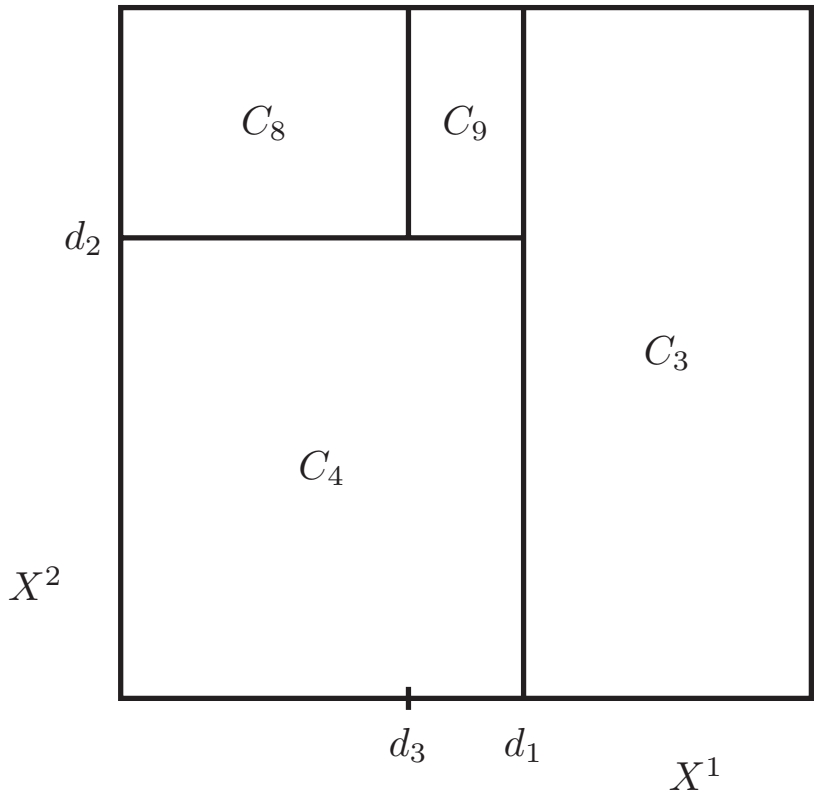


Exemple fictif : arbre binaire de classification

Définitions

- Déterminer une séquence itérative de *nœuds*
- *Racine* : nœud initial ou ensemble de l'échantillon
- *Feuille* : nœud terminal
- *Nœud* : choix d'une variable et d'une *division* sous-ensemble auquel est appliquée une dichotomie
- *Division* : valeur seuil ou groupes des modalités





Exemple fictif : pavage dyadique de l'espace

Règles

- Choix nécessaires :
  1. Critère de la "meilleure" division parmi celles *admissibles*
  2. Règle de nœud terminal : *feuille*
  3. Règle d'affectation à une classe $\mathcal{T}_\ell$ ou une valeur de Y
- Division *admissible* : descendants $\neq \emptyset$
- X^j réelle ou ordinale : $(c_j - 1)$ divisions possibles
  Attention effets incontrôlables si c grand
- X^j nominale : $2^{(c_j-1)} - 1$ divisions
- Fonction d'hétérogénéité D_κd'un nœud
  1. Nulle : une seule modalité de Y ou Y constante
  2. Maximale : modalités de Y équiréparties ou grande variance

Division optimale

- Notation
  - κ : numéro d'un nœud
  - κ_G et κ_D les nœuds fils
- L'algorithme retient la division rendant minimales

$$D_{\kappa_G} + D_{\kappa_D}$$

- Chaque étape κ de construction de l'arbre :

$$\max_{\{\textit{Divisions de } X^j; j=1,p\}} D_\kappa - (D_{(\kappa_G} + D_{\kappa_D})$$

Feuille et affectation

- Un Nœud est terminal ou feuille, si :
  - Homogène
  - Plus de partition admissible
  - Nombre d'observations inférieur à un seuil
- Affectation
  - Y quantitative, la valeur est la moyenne des observations
  - Y qualitative, chaque feuille est affectée à une classe $\mathcal{T}_\ell$ de Y en considérant le mode conditionnel :
    - la classe la mieux représentée dans le nœud
    - la classe a posteriori la plus probable si des a priori sont connus
    - la classe la moins coûteuse si des coûts de mauvais classement sont donnés

Y quantitative : hétérogénéité en régression

Hétérogénéité du nœud κ :

$$D_\kappa = \frac{1}{|\kappa|} \sum_{i \in \kappa} (y_i - \bar{y}_\kappa)^2$$

où $|\kappa|$ est l'effectif du nœud κ

Minimiser la variance intra-classe

Les nœud fils κ_G et κ_D minimisent :

$$\frac{|\kappa_G|}{n} \sum_{i \in \kappa_G} (y_i - \bar{y}_{\kappa_G})^2 + \frac{|\kappa_D|}{n} \sum_{i \in \kappa_D} (y_i - \bar{y}_{\kappa_D})^2.$$

Hétérogénéité et *déviance* dans le cas gaussien (Breiman et al. 1984)

Y qualitative : hétérogénéité en discrimination

Hétérogénéité du nœud κ :

- Entropie avec la notation $0\log(0) = 0$

$$D_\kappa = -2\sum_{\ell=1}^{m} |\kappa| p_\kappa^\ell \log(p_\kappa^\ell)$$

  p_κ^ℓ : proportion de la classe $\mathcal{T}_\ell$ de Y dans κ.
- Concentration de Gini : $D_\kappa = \sum_{\ell=1}^{m} p_\kappa^\ell (1 - p_\kappa^\ell)$
- Statistique du test du χ^2 (CHAID)

Entropie et déviance d'un modèle multinomial (Breiman et al. 1984)

Discrimination : extensions

- Les probabilités conditionnelles sont définies par la règle de Bayes lorsque les probabilités *a priori* π_ℓ sont connues
- Sinon, les probabilités de chaque classe sont estimées sur l'échantillon et donc les probabilités conditionnelles s'estiment par des rapports d'effectifs :

$$p_{\ell k} \quad \text{est estimée par} \quad n_{\ell k}/n_{+k}$$

- Des coûts de mauvais classement connus conduisent à la minimisation d'un risque bayésien

Élagage : notations

- Recherche d'un modèle parcimonieux
- *Complexité* d'un arbre : $K_A =$ nombre de feuilles de A
- Qualité d(ajustement de A :

$$D(A) = \sum_{\kappa=1}^{K_A} D_\kappa$$

D_κ : hétérogénéité feuille κ

Séquence d'arbres emboîtés

- Critère de qualité pénalisé par la complexité :

$$C(A) = D(A) + \gamma \times K_A$$

- Pour $\gamma = 0$: $A_{\max} = A_{K_A}$ minimise $C(A)$
- Lorsque γ croît, la division de A_H, dont l'amélioration de D est inférieure à γ, est annulée ; ainsi
  - deux feuilles sont regroupées (élaguées)
  - le nœud père devient terminal
  - A_{K_A} devient A_{K_A-1}
- Après itération du procédé :

$$A_{\max} = A_{K_A} \supset A_{K_A-1} \supset \cdots A_1$$

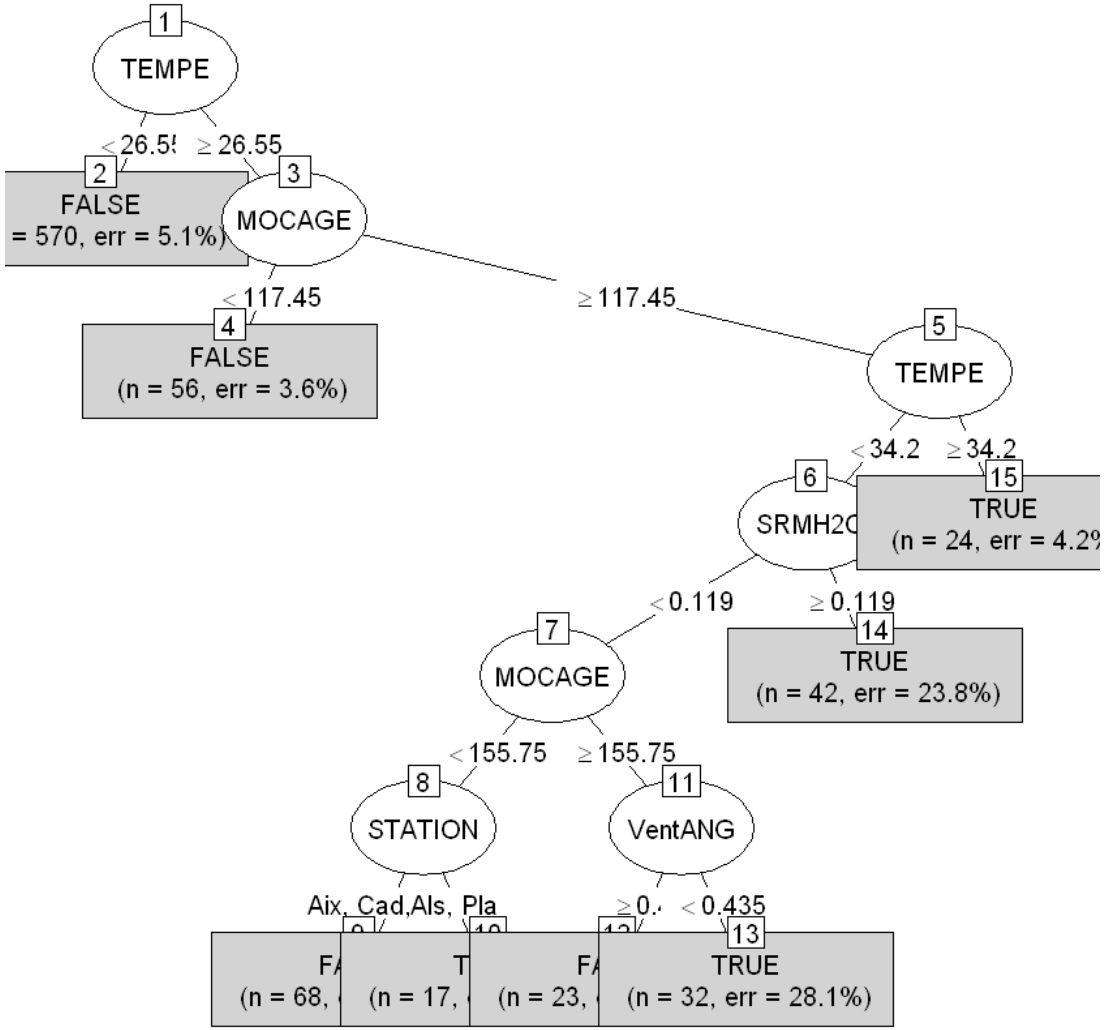
Algorithme de sélection de l'arbre optimal

Arbre maximal $A_{\max}$
Séquence $A_K \dots A_1$ emboîtée associée à
Séquence des valeurs γ_κ
for do $v = 1, \dots, V$
    Estimation de la séquence d'arbres associée à γ_κ
    Estimation de l'erreur
end for
Séquence des moyennes de ces erreurs
γ_{Opt} optimal
Arbre associé à γ_{Opt} dans $A_K \dots A_1$

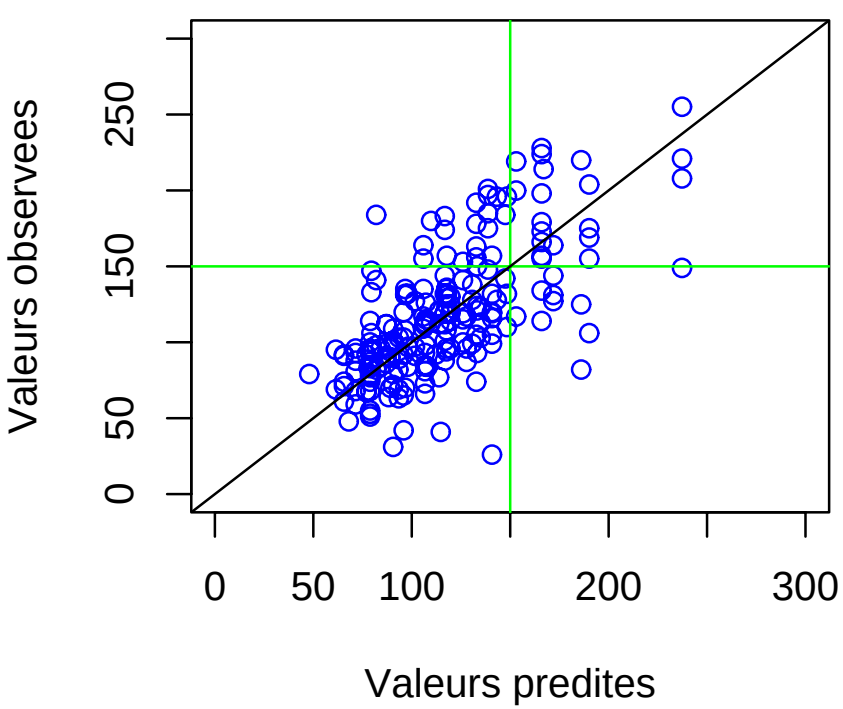
Attention : Séquences d'arbres différentes, même séquence γ_κ

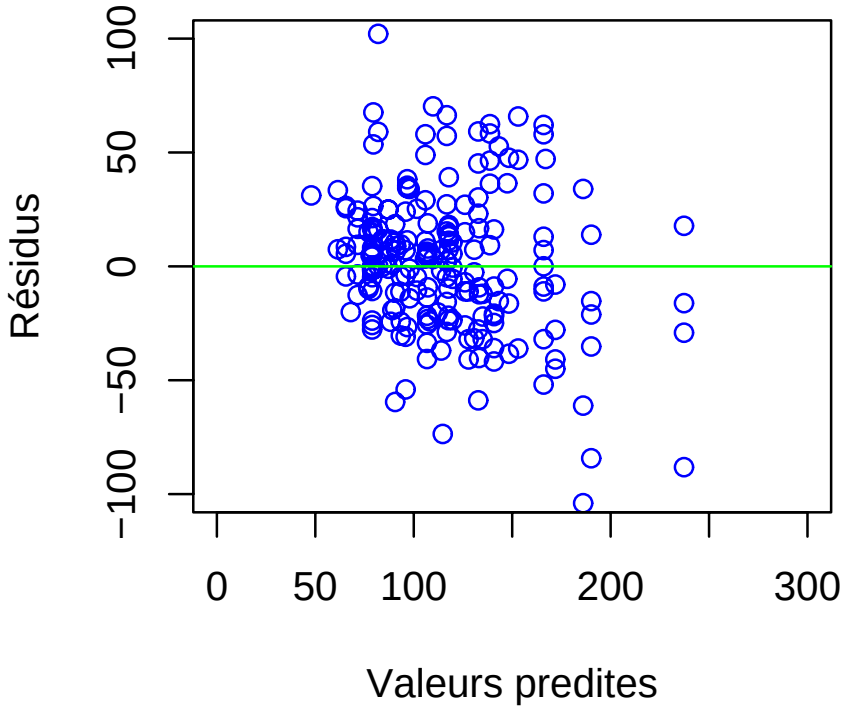
Remarques pratiques

- Sélection de variables et interactions
- Invariance par transformation monotone
- Hiérarchie et instabilité
- Découpes compétitives, *surrogate* et données manquantes
- Variantes : ternaire, linéaire...
- Approximation étagée de la régression : algorithme *MARS*

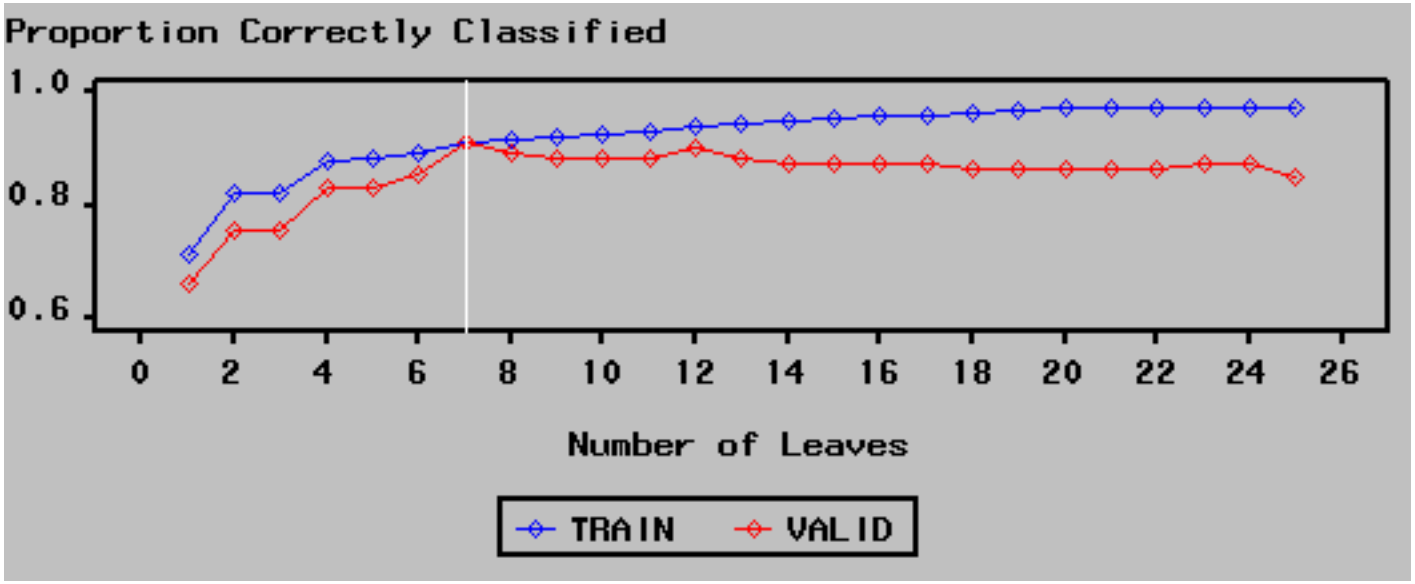


Ozone : arbre de discrimination élagué par validation croisée



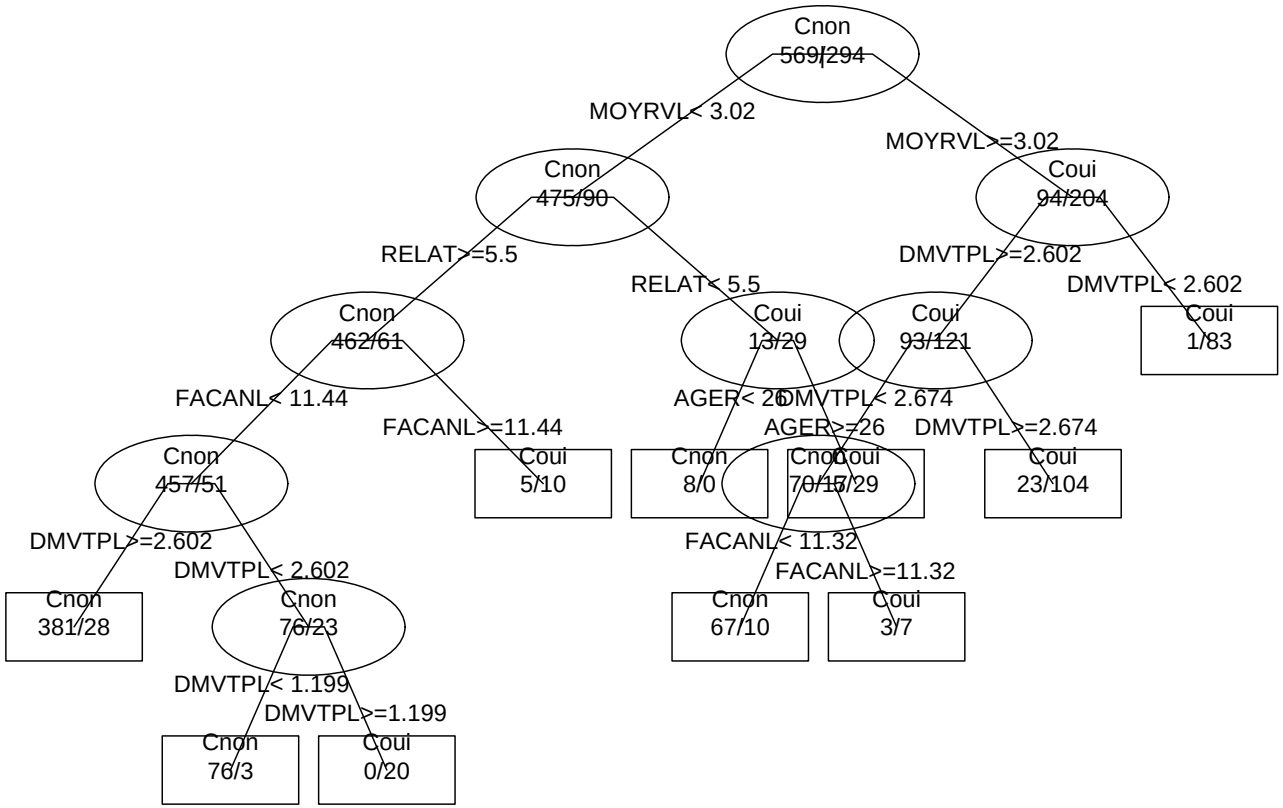


Ozone : observations et résidus en fonction des prévisions



Banque : élagage par échantillon de validation

Endpoint = CARVP



Banque : Elagage par validation croisée