Arbres binaires de décision

Résumé

*Méthodes de construction d'*arbres binaires *de décision, modélisant une discrimination (*classification tree*) ou une régression (*regression tree*). Principes et algorithmes de construction des arbres, critères d'homogénéité et construction des nœuds, élagage pour l'obtention d'un modèle parcimonieux.*

Retour à l'introduction.

Tous les tutoriels sont disponibles sur le dépôt :
`github.com/wikistat`

1 Introduction

Les méthodes dites de partitionnement récursif ou de segmentation datent des années 60. Elles ont été formalisées dans un cadre générique de sélection de modèle par Breiman et col. (1984)[1] sous l'acronyme de CART : *Classification and Regression Tree*. Parallèlement Quinlan (1993)[2] a proposé l'algorithme C4.5 dans la communauté informatique. L'acronyme CART correspond à deux situations bien distinctes selon que la variable à expliquer, modéliser ou prévoir est qualitative (discrimination ou en anglais *classification*) ou quantitative (régression).

Très complémentaires des méthodes statistiques plus classiques : analyse discriminante, régression linéaire, les solutions obtenues sont présentées sous une forme graphique simple à interpréter, même pour des néophytes, et constituent une aide efficace pour l'aide à la décision. Elles sont basées sur une séquence récursive de règles de division, coupes ou *splits*.

La figure 1 présente un exemple illustratif d'arbre de classification. Les variables `Age`, `Revenu` et `Sexe` sont utilisées pour discriminer les observations sous la forme d'une structure arborescente. L'ensemble des observations est regroupé à la racine de l'arbre puis chaque division ou coupe sépare chaque nœud en deux nœuds fils plus *homogènes* que le nœud père au sens d'un *critère* à préciser et dépendant du type de la variable Y, quantitative ou qualitative.

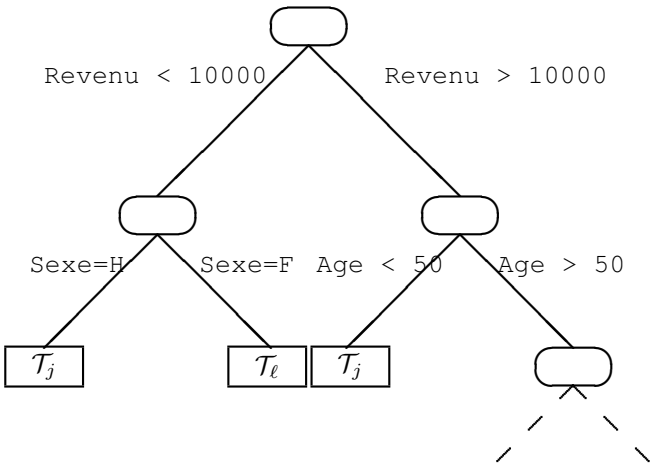

FIGURE 1 – Exemple élémentaire d'arbre de classification.

Lorsque le partitionnement est interrompu, chaque nœud terminal de l'arbre complet ainsi construit devient une feuille à laquelle est attribuée une valeur si Y est quantitative, une classe de Y si elle est qualitative.

La dernière étape consiste en un élagage correspondant à une sélection de modèle afin d'en réduire la complexité dans l'objectif, toujours répété, d'éviter le sur-ajustement. Depuis que Breiman et al. (1984)[1] ont formaliser cette étape, CART connaît un succès important avec un l'atout majeur de la facilité de l'interprétation même pour un néophyte en Statistique. La contrepartie est que ces modèles sont particulièrement instables, très sensibles à des fluctuations de l'échantillon.

Par ailleurs, pour des variables explicatives quantitatives, la construction d'un arbre constitue un *partitionnement dyadique* de l'espace (cf. figure 2). Le modèle ainsi défini manque par construction de régularité même, et surtout, si le phénomène à modéliser est lui-même régulier.

Ces deux aspects ou faiblesses de CART : instabilité, irrégularités, sont à l'origine du succès des méthodes d'ensemble ou d'agrégation de modèles.

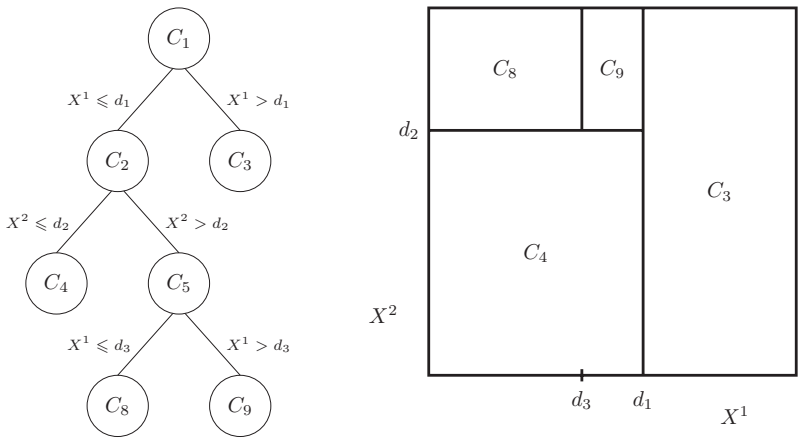

FIGURE 2 – *Construction d'un arbre avec variables explicatives quantitatives et pavage dyadique de l'espace. Chaque nœud père engendre deux fils conséquence d'une division ou coupe définie par le choix d'une variable : X^1 ou X^2 et d'une valeur seuil, successivement d_1, d_2, d_3. À chaque pavé de l'espace ainsi découpé, est finalement associée une valeur ou une classe de Y.*

2 Construction d'un arbre binaire maximal

2.1 Principe

Les données sont constituées de l'observation de p variables quantitatives ou qualitatives explicatives X^j et d'une variable à expliquer Y qualitative à m modalités $\{\mathcal{T}_\ell; \ell = 1 \ldots, m\}$ ou quantitative réelle, observées sur un échantillon de n individus.

La construction d'un arbre de discrimination binaire (cf. figure 1) consiste à déterminer une séquence de *nœuds*.

- Un nœud est défini par le choix conjoint d'une variable parmi les explicatives et d'une *division* qui induit une partition en deux classes. Implicitement, à chaque nœud correspond donc un sous-ensemble de l'échantillon auquel est appliquée une dichotomie.
- Une division est elle-même définie par une *valeur seuil* de la variable quantitative sélectionnée ou un partage en deux *groupes des modalités* si la variable est qualitative.
- À la racine ou nœud initial correspond l'ensemble de l'échantillon ; la procédure est ensuite itérée sur chacun des sous-ensembles.

L'algorithme considéré nécessite :

1. la définition d'un *critère* permettant de sélectionner la meilleure division parmi toutes celles *admissibles* pour les différentes variables ;
2. une règle permettant de décider qu'un nœud est terminal : il devient ainsi une *feuille* ;
3. l'affectation de chaque feuille à l'une des classes ou à une valeur de la variable à expliquer.

2.2 Critère de division

Une division est dite *admissible* si aucun des deux nœuds descendants qui en découlent n'est vide. Si la variable explicative est qualitative ordinale avec m modalités, elle fournit $(m-1)$ divisions binaires admissibles. Si elle est seulement nominale le nombre de divisions passe à $2^{(m-1)} - 1$. Une variable quantitative se ramène au cas ordinal.

Attention, l'algorithme tend à favoriser la sélection de variables explicatives avec beaucoup de modalités car celles-ci offrent plus de souplesse dans

la construction de deux sous groupes. Ces variables sont à utiliser avec parcimonie (e.g. le code postal) car susceptibles de favoriser un sur-apprentissage ; il est souvent préférable de réduire drastiquement le nombre de modalités (e.g. région géographique ou zone urbaine *vs.* zone rurale) par fusion de modalités comme c'est classique en analyse des correspondances multiple.

Le critère de division repose sur la définition d'une fonction d'*hétérogénéité* ou de désordre explicitée dans la section suivante. L'objectif étant de partager les individus en deux groupes les plus homogènes au sens de la variable à expliquer. L'hétérogénéité d'un nœud se mesure par une fonction non négative qui doit être

1. nulle si, et seulement si, le nœud est homogène : tous les individus appartiennent à la même modalité ou prennent la même valeur de Y.
2. Maximale lorsque les valeurs de Y sont équiprobables ou très dispersées.

La division du nœud κ crée deux fils, gauche et droit. Pour simplifier, ils sont notés κ_G et κ_D mais une re-numérotation est nécessaire pour respecter la séquence de sous-arbres qui sera décrite dans la section suivante.

Parmi toutes les divisions admissibles du nœud κ, l'algorithme retient celle qui rend la somme $D_{\kappa_G} + D_{\kappa_D}$ des hétérogénéités des nœuds fils minimales. Ceci revient encore à résoudre à chaque étape k de construction de l'arbre :

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

Graphiquement, la longueur de chaque branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division.

2.3 Règle d'arrêt

La croissance de l'arbre s'arrête à un nœud donné, qui devient donc terminal ou *feuille*, lorsqu'il est homogène ou lorsqu'il n'existe plus de partition admissible ou, pour éviter un découpage inutilement fin, si le nombre d'observations qu'il contient est inférieur à une valeur seuil à choisir en général entre 1 et 5.

2.4 Affectation

Dans le cas Y quantitative, à chaque feuille ou pavé de l'espace, est associée une valeur : la moyenne des observations associées à cette feuille. Dans le cas qualitatif, chaque feuille ou nœud terminal est affecté à une classe $\mathcal{T}_\ell$ de Y en considérant le mode conditionnel :

- celle la mieux représentée dans le nœud et il est ensuite facile de compter le nombre d'objets mal classés ;
- la classe *a posteriori* la plus probable au sens bayésien si des probabilités *a priori* sont connues ;
- la classe la moins coûteuse si des coûts de mauvais classement sont donnés.

3 Critères d'homogénéité

Deux cas sont à considérer, les arbres de régression ou de classification.

3.1 *Y* quantitative

Dans le cas de la régression, l'hétérogénéité du nœud κ est définie par la variance :

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

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

L'objectif est de chercher pour chaque nœud la division, ou plus précisément la variable et la règle de division, qui contribuera à la plus forte décroissance de l'hétérogénéité des nœuds fils à gauche κ_G et à droite κ_D. Ce qui revient à minimiser la variance intraclasse ou encore :

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

On peut encore dire que la division retenue est celle qui rend, le plus significatif possible, le test de Fisher (analyse de variance) comparant les moyennes entre les deux nœuds fils. Dans leur présentation originale, Breiman et al. (1984)[1] montrent que, dans le cas d'une distribution gaussienne, le raffinement de l'arbre est associé à une décroissance, la plus rapide possible, d'une

déviance, d'où la notation D_κ ou écart à la vraisemblance du modèle gaussien associé.

3.2 *Y* qualitative

Soit Y variable qualitative à m modalités ou catégories $\mathcal{T}$ numérotées $\ell = 1, \ldots, m$. Plusieurs fonctions d'hétérogénéité, ou de désordre peuvent être définies pour un nœud : un critère défini à partir de la notion d'*entropie* ou à partir de la *concentration de Gini*. Un autre critère (CHAID) est basé sur la statistique de test du χ^2. En pratique, il s'avère que le choix du critère importe moins que celui du niveau d'élagage, c'est souvent *Gini* qui est choisi par défaut mais le critère d'entropie s'interprète encore comme un terme de déviance par rapport à la vraisemblance mais d'un modèle multinomial saturé cette fois.

Entropie

L'hétérogénéité du nœud κ est définie par l'entropie qui s'écrit avec la convention $0 \log(0) = 0$:

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

où p_κ^ℓ est la proportion de la classe $\mathcal{T}_\ell$ de Y dans le nœud κ.

Concentration de Gini

L'hétérogénéité du nœud est définie par :

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

Comme dans le cas quantitatif, il s'agit, pour chaque nœud de rechercher, parmi les divisions admissible, celle qui maximise la décroissance de l'hétérogénéité.

Comme pour l'analyse discriminante décisionnelle, plutôt que des proportions, des probabilités conditionnelles sont définies par la règle de Bayes lorsque les probabilités *a priori* π_ℓ d'appartenance à la ℓ-ième classe sont connues. Dans le cas contraire, les probabilités de chaque classe sont estimées sur l'échantillon et donc les probabilités conditionnelles s'estiment simplement par les proportions. Enfin, il est toujours possible d'introduire, lorsqu'ils sont connus, des coûts de mauvais classement et donc de se ramener à la minimisation d'un risque bayésien.

4 Élagage de l'arbre optimal

La démarche de construction précédente fournit un arbre $A_{\max}$ à K feuilles qui peut être excessivement raffiné et donc conduire à un modèle de prévision très instable car fortement dépendant des échantillons qui ont permis son estimation. C'est une situation de sur-ajustement à éviter au profit de modèles plus parcimonieux donc plus robuste au moment de la prévision. Cet objectif est obtenu par une procédure d'*élagage* (*pruning*) de l'arbre.

Il s'agit donc de trouver un arbre optimal entre celui trivial réduit à une seule feuille et celui maximal $A_{\max}$ en estimant leur performance par exemple sur un *échantillon de validation*. Tous les sous-arbres sont admissibles mais, comme leur nombre est de croissance exponentielle, il n'est pas envisageable de tous les considérer.

Pour contourner ce verrou, Breiman et col. (1984)[1] ont proposé une démarche consistant à construire une *suite emboîtée de sous-arbres* de l'arbre maximal puis à choisir, seulement *parmi cette suite*, l'arbre optimal qui minimise un risque ou erreur de généralisation. La solution ainsi obtenue est un optimum local mais l'efficacité et la fiabilité sont préférées à l'optimalité.

4.1 Construction de la séquence d'arbres

Pour un arbre A donné, on note K_A le nombre de feuilles ou nœuds terminaux κ, $\kappa = 1, \ldots, K_A$ de A ; la valeur de K_A exprime la complexité de A. La qualité d'ajustement d'un arbre A est mesurée par

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

où D_κ est l'hétérogénéité de la feuille κ de l'arbre A et donc, selon le cas : la variance interclasse, l'entropie, la concentration de Gini, le nombre de mal classés, la déviance ou le coût de mauvais classement.

La construction de la séquence d'arbres emboîtés repose sur une pénalisa-

tion de la complexité de l'arbre :

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

Pour $\gamma = 0$, $A_{\max} = A_{K_A}$ minimise $C(A)$. En faisant croître γ, l'une des divisions de A_{K_A}, celle pour laquelle l'amélioration de D est la plus faible (inférieure à γ), apparaît comme superflue et les deux feuilles obtenues sont regroupées (élaguées) dans le nœud père qui devient terminal ; A_{K_A} devient A_{K_A-1}.

Le procédé est itéré pour la construction de la séquence emboîtée :

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

où A_1, le nœud racine, regroupe l'ensemble de l'échantillon.

Il est alors facile de tracer le graphe représentant la décroissance ou éboulis des valeurs de D_κ en fonction du nombre croissant de feuilles dans l'arbre ou, c'est équivalent, en fonction de la séquence des valeurs décroissantes du coefficient de pénalisation γ.

4.2 Recherche de l'arbre optimal

Une fois la séquence d'arbres emboités construite, il s'agit d'en extraire celui optimal minimisant un risque ou erreur de généralisation. Si la taille de l'échantillon le permet, l'extraction préalable d'un échantillon de validation permet une estimation facile de ces risques.

Dans le cas contraire, c'est une stratégie de validation croisée en V segments qu'il faut mettre en place.

Celle-ci présente dans ce cas une particularité. En effet, à chacun des V échantillons constitués de $V-1$ segments, correspond une séquence d'arbres différente. L'erreur moyenne n'est pas, dans ce cas, calculée pour chaque sous-arbre avec un nombre de feuilles donné mais pour chaque sous-arbre correspondant à une valeur fixée du coefficient de pénalisation γ issue de la séquence produite initialement par tout l'échantillon. À chacun des V échantillons correspond un arbre différent pour chacune des valeurs de γ mais c'est cette valeur qui est optimisée.

À la valeur de γ minimisant l'estimation de l'erreur de prévision par validation croisée, correspond ensuite l'arbre jugé optimal dans la séquence estimée sur tout l'échantillon d'apprentissage.

Le principe de sélection d'un arbre optimal est donc décrit dans l'algorithme ci-dessous.

Algorithm 1 Sélection d'arbre ou élagage par validation croisée

Construction de l'arbre maximal $A_{\max}$
Construction de la séquence $A_K \ldots A_1$ d'arbres emboîtés associée à une
Séquence de valeurs de pénalisation γ_κ
for $v = 1, \ldots, V$ **do**
    Pour chaque échantillon, estimation de la séquence d'arbres associée la séquence des pénalisations γ_κ
    Estimation de l'erreur sur la partie restante de validation de l'échantillon
end for
Calcul de la séquence des moyennes de ces erreurs
L'erreur minimale désigne la pénalisation γ_{Opt} optimale
Retenir l'arbre associé à γ_{Opt} dans la séquence $A_K \ldots A_1$

4.3 Remarques pratiques

Sur l'algorithme :

- Les arbres ne requièrent pas d'hypothèses sur les distributions des variables et semblent particulièrement adaptés au cas où les variables explicatives sont nombreuses. En effet, la procédure de *sélection* des variables est *intégrée* à l'algorithme construisant l'arbre et les *interactions* sont implicitement prises en compte.
- Il peut être utile d'associer arbre et régression logistique. Les *premières division* d'un arbre sont utilisées pour construire une *variable synthétique* intégrée à une régression logistique afin de sélectionner les quelques interactions apparaissant comme les plus pertinentes.
- La recherche d'une division est *invariante* par transformation monotone des variables explicatives quantitatives. Cela confère une *robustesse* de l'algorithme vis-à-vis de possibles valeurs atypiques ou de distributions très asymétriques. Seuls les rangs des observations sont considérés par l'algorithme pour chaque variable quantitative.
- *Attention*, cet algorithme suit une stratégie pas à pas hiérarchisée. Il peut, comme dans le cas du choix de modèle pas à pas en régression, passer à coté d'un optimum global ; il se montre par ailleurs très *in-*

stable et donc sensible à des fluctuations d'échantillon. Cette instabilité ou variance de l'arbre est une conséquence de la structure hiérarchique : une erreur de division en début d'arbre est propagée tout au long de la construction.

- Plusieurs variantes on été proposées puis abandonnées : arbres ternaires plutôt que binaires, règle de décision linéaire plutôt que dichotomique. La première renforce inutilement l'instabilité alors que si une décision ternaire est indispensable, elle est la succession de deux divisions binaires. La deuxième rend l'interprétation trop complexe donc le modèle moins utile.
- *Attention* Dans le cas d'une régression, Y est approchée par une fonction étagée. Si Y est l'observation d'un phénomène présentant des propriétés de régularité, ce modèle peut ne pas être approprié ou moins approprié qu'une autre famille de méthodes. En revanche, si Y présente des effets de seuillage, des singularités, ou s'il s'agit de discriminer des classes non connexes, un arbre de décision peut s'avérer plus approprié. Cela renforce l'idée que, sans information précise sur la nature des donnée (variable Y) à modéliser, il n'y a pas d'autre stratégie qu'essayer plusieurs types de modèles en comparant leurs performances.

Sur les implémentations :

- L'implémentation dans la librairie `rpart` de R mémorise la séquence des divisions ou *découpes compétitives*. Cela permet de préférer des arbres présentant des divisions certes moins optimales mais par exemple moins onéreuses à observer ou plus faciles à interpréter. Cela permet également de considérer des observations avec *données manquantes* pour certaines variables explicatives. Il suffit de déterminer pour chaque nœuds une séquence ordonnée de divisions possibles ou *surrogate splits* qui sont les division présentant le moins de *désaccords* avec celle meilleure. Au moment de calculer une prévision, si une donnée manque pour l'application d'une division ou règle de décision, la division suivante est prise en compte jusqu'à ce qu'une décision soit prise à chacun des nœuds rencontrés.
- *Attention* : dans la version actuelle (0.19) de `Scikit-learn`, seul le paramètre de profondeur maximale de l'arbre (`max_depth`) permet de contrôler le sur-apprentissage. Ce paramètre optimisé, encore par validation croisée, ne définit qu'une séquence d'arbres très grossière (croissance du nombre de feuilles par puissance de 2) par rapport à celle obtenue par pénalisation, feuille à feuille, dans `rpart` de R. Comme souvent dans les librairies de Python, l'efficacité algorithmique prévaut sur le sens "statistique".

5 Exemples

5.1 Concentration d'ozone

Arbre de régression

Un arbre de régression est estimé pour prévoir la concentration d'ozone. La librairie `rpart` du logiciel R prévoit une procédure d'élagage par validation croisée afin d'optimiser le coefficient de pénalisation. L'arbre (figure 3) montre bien quelles sont les variables importantes intervenant dans la prévision. Mais, compte tenu de la hiérarchisation de celles-ci, due à la structure arborescente du modèle, cette liste n'est pas similaire à celle mise en évidence dans le modèle gaussien. On voit plus précisément ici la complexité des interactions entre la prédiction par MOCAGE et l'effet important de la température dans différentes situations. Les résidus de l'échantillon test du modèle d'arbre de régression prennent une structure particulière (figure 5) car les observations communes à une feuille terminale sont affectées de la même valeur. Il y a donc une colonne par feuille. La précision de l'ajustement peut s'en trouver altérée ($R^2 = 0,68$) mais il apparaît que ce modèle est moins soumis au problème d'hétéroscédasticité très présent dans le modèle gaussien.

Arbre de discrimination

Un modèle est estimé (figure 4) afin de prévoir directement le dépassement d'un seuil. Il est de complexité similaire à celle de l'arbre de régression mais ne fait pas jouer le même rôle aux variables. La température remplace la prévision MOCAGE de l'ozone comme variable la plus importante. Les prévisions de dépassement de seuil sur l'échantillon test sont sensiblement moins bonnes que celle de la régression, les taux sont de 14,4% avec l'arbre de régression et de 14,5% directement avec l'arbre de discrimination. Les matrices de confusion présentent les mêmes biais que les modèles de régression en omettant un nombre important de dépassements.

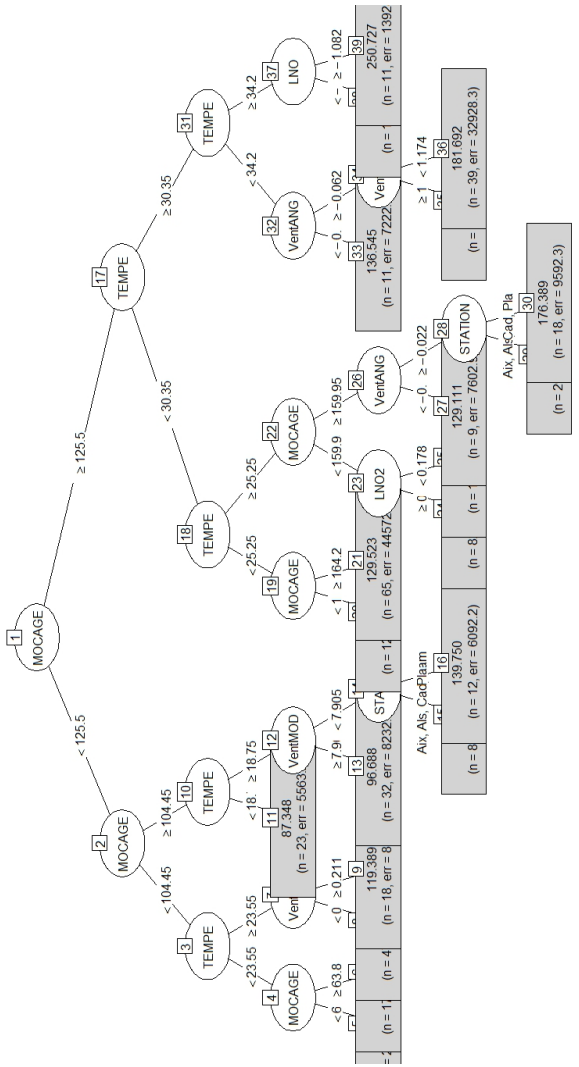

FIGURE 3 – *Ozone : arbre de régression élagué par validation croisée (R).*

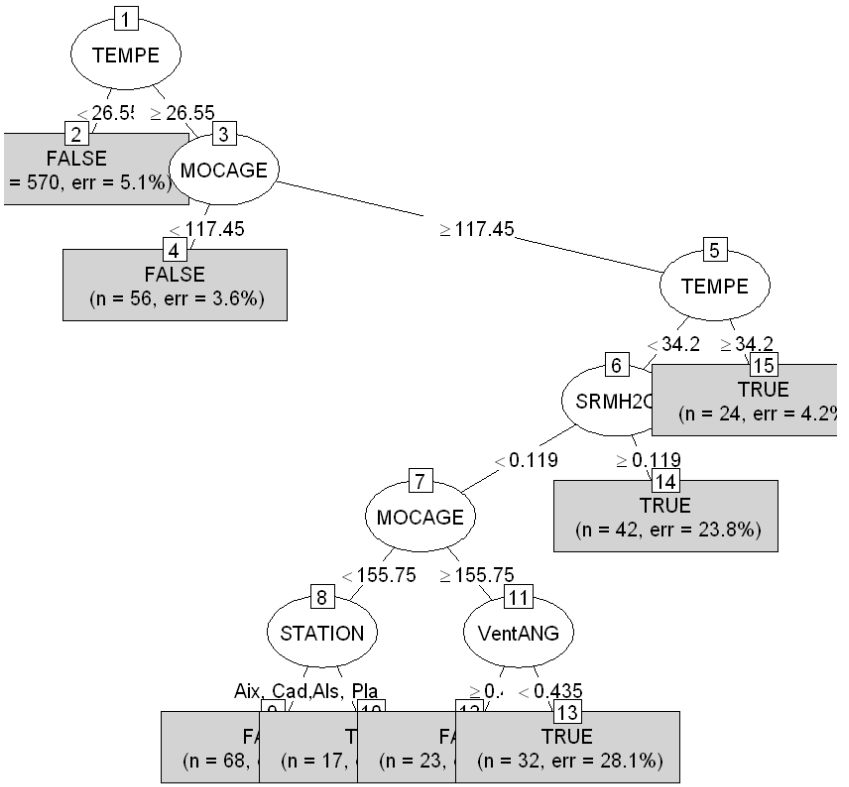

FIGURE 4 – *Ozone : arbre de discrimination élagué par validation croisée (R).*

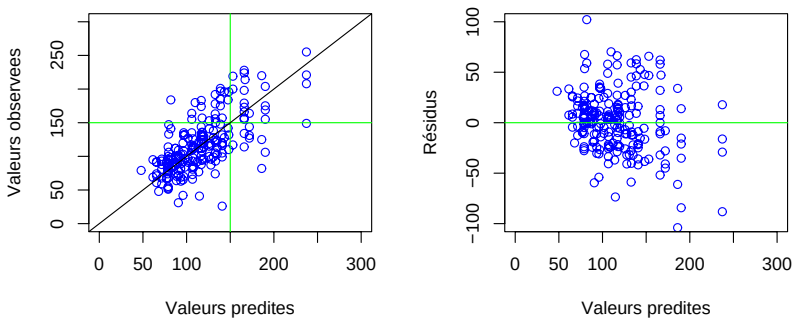

FIGURE 5 – *Ozone : Valeurs observées et résidus de l'échantillon test.*

5.2 Carte Visa Premier

L'étude des données bancaires s'intéresse soit aux données quantitatives brutes soient à celles-ci après découpage en classes des variables quantitatives. Ce découpage rend des services en régression logistique car le modèle construit s'en trouve plus flexible : plus de paramètres et moins de degrés de liberté, comme l'approximation par des indicatrices (des classes) de transformations non linéaires des variables. Il a été fait "à la main" en prenant les quantiles comme bornes de classe. C'est un usage courant pour obtenir des classes d'effectifs égaux et répartit ainsi au mieux la précision de l'estimation des paramètres mais ce choix n'est pas optimal au regard de l'objectif de prévision. Dans le cas d'un modèle construit à partir d'un arbre binaire, il est finalement préférable de laisser faire celui-ci le découpage en classe c'est-à-dire de trouver les valeurs seuils de décision. C'est la raison pour laquelle, l'arbre est préférablement estimé sur les variables quantitatives et qualitatives initiales.

La librairie `rpart` de R propose d'optimiser l'élagage par validation croisée pour obtenir la figure 6 Cet arbre conduit à un taux d'erreur estimé à 8% sur l'échantillon test, mieux que la régression logistique qui manque de flexibilité sur cet échantillon.

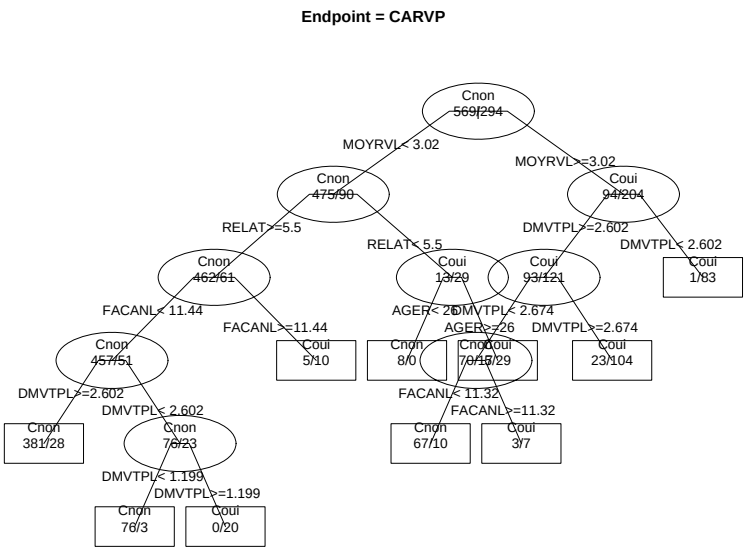

FIGURE 6 – *Banque : arbre de décision élagué par validation croisée dans R.*

Références

[1] L. Breiman, J. Friedman, R. Olshen et C. Stone, *Classification and regression trees*, Wadsworth & Brooks, 1984.

[2] J.R. Quinlan, *C4.5 – Programs for machine learning*, M. Kaufmann, 1993.