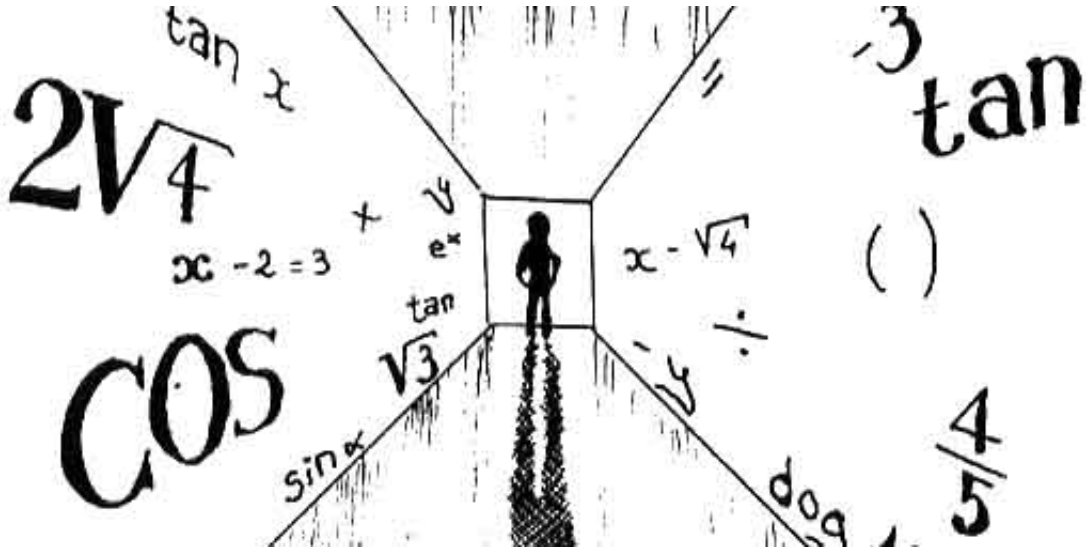
UNIVERSITÉ PARIS SUD

Département de mathématiques d'Orsay

Mémoire
Magistère $3^{\text{ème}}$ année



Résumé. Ce mémoire présentera dans une première partie, un descriptif de mon cursus au sein du magistère et ce qu'il m'a apporté. Ensuite, une seconde partie sera consacrée à une présentation généraliste (historique, motivations, grandes questions ouvertes) du domaine de recherche dans lequel j'effectuerai ma thèse. Enfin, on trouvera en annexe les divers mémoires que j'ai réalisé lors de ma scolarité au sein de cette formation.

Valérie Robert

17 octobre 2013

Table des matières

1 Introduction

Durant mes deux années en classe préparatoire au lycée Leconte de Lisle à la Réunion, les mathématiques se sont révélées être une discipline réellement passionnante avec l'introduction de nouveaux concepts et théories jamais abordées lors de la Terminale.

Grâce à ces deux années, j'ai redécouvert les mathématiques sous un nouvel aspect, elles me sont apparues comme indispensables et pleines d'avenir et j'avoue que certains théorèmes ont été pour moi source d'émerveillement.

Par ailleurs, cela m'a donné envie d'enseigner ces mathématiques et de susciter ainsi chez les élèves, le même engouement que j'ai ressenti pour cette matière. C'est pour cela que j'ai décidé d'intégrer en 2006 la Licence troisième année intitulée *Mathématiques Fondamentales et Appliquées* (L3 MFA) de l'université Paris-SUD ainsi que la première année du Magistère de mathématiques. Je voulais par ce biais, développer au mieux ma culture mathématique, et je savais par ailleurs, que ces deux formations me fourniraient un environnement mathématique fécond, et me permettraient ensuite, de concrétiser mon projet professionnel .

2 Parcours au sein du magistère

2.1 Première année de magistère

2.1.1 Cours suivi

La première année de magistère a été l'occasion d'assister à un cours très intéressant intitulé *Espaces de Banach et mesures complexes* dispensé par M. Patrick Gérard. Ce cours s'est révélé être un des cours les plus utiles lors de la poursuite de mes études. En effet, les théorèmes de Banach, de compacité et de convergence faible abordés dans ce cours magistral, ont été au centre d'un cours d'analyse de M1 et m'a permis ainsi, d'avoir de solides bases pour appréhender la première année de master. J'ai également retrouvé ces notions lors de ma préparation à l'agrégation, et j'ai pu alors retravailler ces concepts en toute sérénité, et me concentrer sur d'autres points du programme d'agrégation moins familiers.

2.1.2 Projet TER

Comme je le disais précédemment, j'ai toujours été émerveillée et intriguée par les mathématiques et lorsque ce sujet de TER *Transcendance de e et de* π a été proposé par M. Stéphane Fischler, je n'ai pas hésité une seconde. En effet, j'avais déjà entendu parler de la transcendance de π qui a notamment de jolies implications en géométrie. Mais pouvoir le démontrer rigoureusement était une aubaine afin de satisfaire ma curiosité mathématique. Ainsi, ce premier projet de recherche n'a fait que confirmer mon attrait pour les mathématiques et j'ai entrevu les mécanismes et les qualités requises pour effectuer un travail de recherche : une bibliographie préliminaire sur le sujet, une motivation sans faille et de l'opiniâtreté pour pouvoir mener à terme une démonstration récalcitrante par exemple.

2.1.3 Stage

Lors de la première année de magistère, j'ai également effectué un stage obligatoire dans un laboratoire de recherche. Etant originaire de la Réunion, j'ai donc contacté le laboratoire d'Informatique et de Mathématiques (LIM) de là-bas afin d'y faire mon stage.

J'ai été alors encadré par M. Khalid Addi, professeur au sein de ce laboratoire sur le sujet suivant : *Influence des paramètres environnementaux sur le comportement des prédateurs supérieurs dans le canal du Mozambique.*

Ainsi, ce sujet mêlant mathématiques et environnement m'a beaucoup plu et pendant un mois, j'ai fait partie d'une équipe et j'ai touché du bout des doigts ce microcosme qu'est le monde de la recherche.

Une des expériences pendant ce stage que j'ai trouvé très formatrice, est cette participation au sixième colloque de WIOMSA (Western Indian Ocean Marine Science Association) qui s'est tenu à la Réunion du 24/08/09 au 29/08/09.

En effet, j'ai exposé en anglais mon travail effectué lors du stage, et cela m'a donné l'opportunité d'avoir une première expérience d'un passage en public et ce dans une langue étrangère, ce qui était nouveau pour moi à cette époque. Je garde un très bon souvenir de ce stage car les expériences y ont été variées et très enrichissantes.

2.2 Deuxième année de magistère

2.2.1 Cours Semestre 1

La deuxième année de magistère en parallèle avec le master 1 de *Mathématiques Fondamentales et Appliquées* a consisté au premier semestre à suivre un cours intitulé *Théorie spectrale et analyse harmonique* enseigné par M. Frédéric Paulin.

Là aussi, ce fut un cours très instructif et surtout très utile, étant donné que les opérateurs compacts, les fonctions harmoniques et les espaces de Hilbert sont un pan central du programme d'agrégation.

Ainsi, ce cours complet sur cette thématique, a permis de se familiariser avec des notions incontournables par la suite, mais surtout d'assimiler quelques techniques de démonstration réapplicables dans d'autres situations.

2.2.2 Cours Semestre 2

Le cours du semestre 2 dispensé par M. Emmanuel Breuillard quant à lui, consistait en une *Introduction à la théorie analytique des nombres.*

Ce cours s'est révélé vraiment intéressant et a réveillé ma curiosité mathématique, même si parfois il était d'un niveau assez relevé.

En effet, les nombres premiers m'ont toujours fasciné, et de pouvoir comprendre le lien avec la fonction ζ de Riemann grâce à l'hypothèse de Riemann, a été source d'une grande satisfaction.

Mieux encore, on a pu lors d'un devoir à la maison, aboutir à une formulation équivalente de l'hypothèse de Riemann.

2.2.3 Projet TER

Le projet TER que j'ai réalisé en seconde année de magistère, se rapproche du thème de celui effectué en première année puisqu'il s'intitule *Constructibilité à la règle et au compas.* Il a été encadré par M. Daniel Perrin qui m'a beaucoup conseillé et aidé dans la suite de ma scolarité, notamment pour les concours d'enseignement.

Ainsi, ce sujet autour de la constructibilité à la règle et au compas est historiquement très ancien. Mais c'est avec l'arrivée de la théorie des groupes, des corps et celle de Galois que l'on a réussi à résoudre ces problèmes anciens que sont la quadrature du cercle, la duplication du cube ou encore la constructibilité de polygones réguliers à la règle et au compas.

Ainsi, ce projet TER a été riche d'enseignements, m'a fait découvrir la théorie de Galois et aborder cette fois-ci un pan fondamental de l'algèbre étudié pour l'agrégation. Je reste de plus, très satisfaite d'avoir réussi à réaliser moi-même la construction d'un polygone régulier à 17 côtés à la règle et au compas (voir annexes).

Après l'obtention de mon master 1, j'ai décidé naturellement de passer le concours de l'agrégation et donc d'interrompre pendant un an le magistère, afin d'intégrer le M2 *Formation de Professeurs Agrégés en mathématiques* (FPA).

2.3 Troisième année de magistère

2.3.1 Cours supplémentaire suivi

L'agrégation en poche, j'ai alors effectué un report de stage pour suivre le M2 recherche *Probabilités et statistiques* d'Orsay.

J'ai repris ainsi le magistère en cours, et la troisième année de magistère consistait alors à suivre et à valider un cours supplémentaire du M2. Etant vraiment passionnée par ce M2, j'ai donc suivi et validé 5 cours supplémentaires : *Grandes déviations*, *Transitions de phase*, *Concentration et sélection de modèles*, *Statistiques et théorie de l'information*, *Modèles pour la classification non supervisée*.

Même si j'avais choisi le M2 orienté plus vers les statistiques car c'est une discipline enclin à de nombreuses applications variées, les cours de probabilités m'ont permis d'élargir ma culture scientifique et de garder une certaine ouverture d'esprit. Par exemple, grâce au deux cours *Grandes déviations* et *transitions de phase* de M. Raphaël Cerf, j'ai pu approfondir certains notions survolées dans le programme de statistiques de l'agrégation et avoir les réponses aux quelques questions que je me posais.

De plus, j'y ai rencontré celui qui allait devenir mon futur maître de stage.

2.3.2 Stage

Au second semestre de ce M2, j'ai alors effectué un stage obligatoire de 4 mois en partenariat avec le CEA sous la direction de M. Gilles Celeux, M. Patrick Pamphile et Mme Krystyna Biletska. Celui-ci portait sur l'*application des méthodes CART et forêts aléatoires au problème d'estimation de l'état de charge d'une batterie électrique*.

Ce stage a fait le lien entre la théorie apprise lors des cours de ce M2 et la mise en pratique de celle-ci qui n'est pas toujours évidente. En effet, lors de ce stage, j'ai eu affaire à une très grande base de données et il a fallu d'une part, savoir les gérer et d'autre part, appliquer les méthodes d'apprentissages sur celle-ci. Un travail bibliographique s'est avéré alors nécessaire.

Ainsi, ce stage m'a apporté une expérience des plus concrètes dans le monde de la recherche, de mettre en lumière l'importance des partenariats et dialogues entre les différents acteurs du développement de la recherche en mathématiques appliquées.

Enfin, et non des moindres, ce stage m'a ouvert les portes de la thèse que j'effectuerai pendant les trois ans à venir, sous la direction de M. Gilles Celeux et de M. Patrick Pamphile que je remercie chaleureusement de m'avoir fait confiance. Elle sera par conséquent, dans le prolongement de mon stage et s'intitulera *Modèles de classification pour l'analyse de données temporelles*. La prochaine partie de ce mémoire sera par conséquent consacrée à la présentation du contexte de ma thèse et je reste très motivée pour la suite de mes aventures dans le monde envoûtant que constituent les mathématiques.

3 Présentation du domaine de recherche : Application des méthodes CART et forêts aléatoires au problème d'estimation de l'état de charge d'une batterie électrique

3.1 Contexte appliqué

Une batterie élémentaire est un système électrochimique complexe.

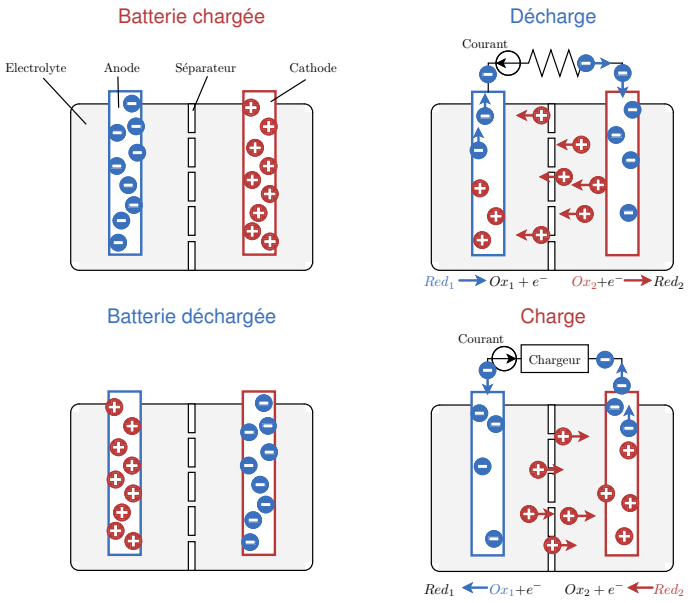


FIGURE 1 – Charge/Décharge d'une batterie élémentaire

En effet elle se compose de deux électrodes (anode et cathode) et d'un électrolyte qui permet de faire passer les ions d'une électrode à une autre (voir figure 1).

Lorsque la batterie est chargée, les ions se retrouvent à la cathode et inversement lorsque la batterie est déchargée. Lors du phénomène de décharge, un dispositif est branché et un courant circule. Les électrons se déplacent alors dans le sens contraire donné au courant. Pour rejoindre la cathode, les ions utilisent l'électrolyte et l'anode est dit "réducteur". De manière analogue, on peut décrire le phénomène de charge.

Ensuite, rappelons quelques définitions utiles pour la suite dans la compréhension de la problématique.

Définition 1. ***Tension***. *La tension U de la batterie est la différence des potentiels entre deux électrodes.*

Définition 2. ***Courant***. *Le courant est la charge électrique traversant un conducteur par une unité de temps et portée par les électrons dans une électrode (courant électronique) ou par les ions dans un électrolyte (courant ionique).*

Définition 3. ***Charge, décharge***. *La quantité de charge, fournie par la batterie durant sa décharge, est calculée comme une intégrale de courant de décharge sur une période de décharge $[t_1; t_2]$:*

$$Q = \int_{t_1}^{t_2} I_t \ dt.$$

Elle est mesurée en ampère-heure (Ah).

Définition 4. ***Capacité C***. *C'est la quantité maximale de charge, qui peut être obtenue en déchargeant complètement à courant constant, la batterie initialement complètement chargée.*

Définition 5. ***Régime de courant***. *Il est défini comme la vitesse à laquelle une quantité de charge équivalente à $\frac{C}{n}$ est extraite de la batterie (décharge) ou fournie à la batterie (charge) durant n heures.*

Définition 6. ***Capacité nominale :*** C_{nom}. *C'est une capacité obtenue sous le régime de décharge nominal i.e. sous certaines conditions de décharge préconisées par le constructeur.*

Capacité disponible : C_{disp}. *Elle est définie comme la capacité fournie par une batterie, qui n'est pas complètement chargée initialement, sous le régime de décharge non nominal.*

Remarque. La température, le régime du courant et l'historique de l'utilisation de la batterie influencent la capacité nominale.

Définition 7. *On définit l'état de charge de la batterie ([14]) notée SOC (State Of Charge) par :*

$$SOC = \frac{C_{disp}}{C_{nom}}$$

Remarque. Cette quantité est souvent difficile à estimer car il y a :
- une influence de la température interne : la capacité nominale augmente lorsque la température interne augmente,
- une influence de l'état de santé SOH (*State Of Health*) : la capacité nominale diminue au cours du vieillissement de la batterie,
- une influence du régime de courant.

Cependant, nous pouvons répertorier plusieurs méthodes existantes pour estimer l'état de charge d'une batterie ([15],[8]).

Nous citerons la plus usuelle qui est une méthode physique, la méthode coulométrique. En effet, elle consiste à estimer le SOC à l'instant t de la manière suivante :

$$SOC_t = SOC_0 + \frac{1}{C_{nom}} \int_{t_0}^{t} I(\tau) d\tau \tag{1}$$

Ainsi, cette méthode décrit le SOC à l'instant t comme le SOC à l'instant initial t_0 plus la somme des courants distribués entre t et t_0.

Cette méthode a l'avantage d'être facile à implémenter et nécessite la mesure du courant uniquement.

Cependant, elle présente quelques inconvénients. En effet, si le capteur de courant est imprécis, on a une accumulation de l'erreur. De plus, comme nous l'avons dit précédemment, il y a un changement de la C_{nom}. Enfin le SOC_0 est difficile à estimer en temps réel.

On peut alors définir la problématique.

Une batterie d'un véhicule électrique est soumise à plusieurs charges/décharges partielles et complètes. Les mesures suivantes sont prélevées à chaque instant : température interne, température externe, courant et tension de la batterie.

La problématique est donc la suivante : estimer *l'état de charge* de la batterie d'un véhicule électrique en temps réel à partir d'une très grande base de données : mesures instantanées de la température interne, ambiante, du courant et de la tension. Il y a là une volonté de trouver une méthode qui puisse prendre en compte une très grande base de données. C'est pourquoi les enjeux soulevées par cette problématique vont être les suivants :

1. Appliquer les méthodes Classification And Regression Trees (CART) et forêts aléatoires aux problèmes d'estimation de SOC car d'une part ils permettent de prendre en compte une grande base de données issues de roulages réels et contenant des mesures instantanées de la température interne, de la température ambiante, du courant, et de la tension de la batterie.

   D'autre part, ce sont des méthodes applicables en temps réel.
2. Évaluer la robustesse des méthodes par rapport aux différentes utilisations de la batterie.
3. Réussir à adapter ces deux méthodes pour prendre en compte l'aspect temporel

3.2 Contexte théorique

3.2.1 Méthode d'apprentissage statistique

On considère un modèle de régression (dans notre cas non linéaire) : soit (X,Y) variable aléatoire dans $(\mathbb{R}^p, \mathbb{R})$ et satisfaisant :

$$Y = f(X) + \epsilon, \tag{2}$$

où ϵ bruit centré, de variance σ^2 et f fonction de régression à estimer.
Comme la loi P de (X,Y) est inconnue, on veut reconstruire f à partir d'un échantillon :

$$\mathcal{L} = \{(x_t, y_t)\}_{t=1,\cdots,n}$$

composé de n réalisations indépendantes de (X,Y) et où x_t sont les variables explicatives et y_t sont les variables à expliquer.

Le but de l'apprentissage supervisé ([9]) est de construire un modèle statistique en proposant un estimateur $\widehat{f}$ de la fonction f (2).

L'élaboration et la validation du modèle sont faites avec deux ensembles distincts de données :

$$\mathcal{L}_{app} = (x_t; y_t)_{t=1,\cdots,k}, \tag{3}$$

appelé ensemble d'apprentissage, qui est utilisé pour construire le modèle, l'ensemble, dit
de validation,

$$\mathcal{L}_{val} = (x_t; y_t)_{t=k+1,\cdots,j}$$

qui est utilisé pour calibrer les paramètres du modèle, et l'ensemble dit de test,

$$\mathcal{L}_{test} = (x_t; y_t)_{t=j+1,\cdots,n}$$

qui est utilisé évaluer la qualité du modèle choisi.

Ainsi, étant donné $\mathcal{L}_{app}$ et $\mathcal{L}_{test}$, le problème consiste à estimer la fonction f, en minimisant le risque de l'estimateur $\hat{f}$:

$$R(f, \hat{f}) = \mathbb{E}_P \left[\|f - \hat{f}\| \right]_\mu$$

où P est la loi marginale de X, $\|\|_\mu$ est choisie ici comme la norme 2 sur $\mathbb{L}^2(\mathbb{R}^p, \mu)$), ou parfois comme la norme 1.

Mais on ne connaît pas P, donc $\hat{f}$ est obtenu comme celui qui minimise la fonction contraste quadratique empirique ou erreur quadratique moyenne (MSE) :

$$R^*(\widehat{f}) = \frac{1}{k} \sum_{(x_t, y_t) \in \mathcal{L}_{app}} (y_t - f(x_t))^2 \tag{4}$$

Par ailleurs, si nos estimateurs $\widehat{f}$ dépendent d'un paramètre α, on utilise l'échantillon de validation pour calibrer au mieux ce paramètre en choisissant $\widehat{\alpha}$ qui minimise

$$\frac{1}{n-j+1} \sum_{(x_t, y_t) \in \mathcal{L}_{\text{val}}} (y_t - \widehat{f}(x_t, \alpha))^2.$$

Enfin, on valide notre modèle en souhaitant que la qualité de notre estimateur décrite par l'erreur quadratique moyenne des résidus ou erreur de généralisation (MSE residuals) (ou MAE residuals pour la norme 1) :

$$R_*(\widehat{f}) = \frac{1}{n-k} \sum_{t=k+1}^{n} (\widehat{f}(x_t) - y_t)^2 \tag{5}$$

soit la meilleure.

3.2.2 CART

La méthode CART est ainsi une méthode d'apprentissage statistique. Plus précisément, c'est une méthode de régression non paramétrique qui est utilisée pour construire des arbres de décision qui permettent de prendre un ensemble de décisions basé sur les données explicatives afin d'estimer la variable à expliquer. Ils servent ensuite, étant donné un nouvel échantillon de variables explicatives, à estimer la variable à expliquer relatif à celui-ci.

La construction des arbres de décision à partir de données est une discipline déjà ancienne. Les statisticiens en attribuent la paternité à Morgan et Sonquist (1963) qui, les premiers, ont utilisé les arbres de régression dans un processus de prédiction et d'explication (AID Automatic Interaction Detection). Il s'en est suivi toute une famille de méthodes, étendues jusqu'aux problèmes de discrimination et classement, qui s'appuyaient sur le même principe de la représentation par arbres (THAID – Morgan et Messenger, 1973 ; CHAID – Kass, 1980). On considère généralement que cette approche a connu son apogée avec la méthode CART (Classification and Regression Tree) de Breiman et al. décrite en détail dans une monographie intitulée Classification And Regression Trees de 1984 ([4]) et qui fait encore référence aujourd'hui.

Elle a été ensuite implémentée pour le logiciel R en 1997 par T. Therneau, B. Atkinson et B. Ripley sous le paquet `rpart` ([12]) et contient des parties implémentées en C, ce qui optimise le code. Il existe par ailleurs d'autres paquets similaires sur R comme `tree` ou `party`, mais `rpart` reste le plus largement utilisé, notamment pour sa rapidité.

Par ailleurs, pour ce qui est de la représentation visuelle des arbres, un paquet `rpart.plot` a été implémenté pour le logiciel R en 2012 par Stephen Milleborrow ([13]) et permet d'améliorer la représentation déjà existante dans `rpart`, pour plus de lisibilité et d'esthétisme (voir figure 2).

Les arbres de régression sont des arbres de décision binaires. Rappelons la définition rigoureuse de ces derniers :

Définition 8. *Un arbre binaire est une structure de données qui peut se représenter sous la forme d'une hiérarchie dont chaque élément est appelé nœud, le nœud initial étant appelé racine.*

Dans un arbre binaire, chaque élément possède au plus deux éléments fils au niveau inférieur, habituellement appelés nœud fils gauche et nœud fils droit. Du point de vue de ces éléments fils, l'élément dont ils sont issus au niveau supérieur est appelé père.

Un nœud n'ayant aucun fils est appelé feuille. Le nombre de niveaux total, autrement dit la distance entre la feuille la plus éloignée et la racine, est appelé hauteur de l'arbre.

Le niveau d'un nœud est appelé profondeur.

Le principe des arbres de régression consiste à partitionner de manière récursive et dyadique l'espace d'entrée $\mathbb{R}^p$ dans lequel sont définies les variables explicatives .
Cette partition est réalisée à l'aide de conditions imbriquées sur les variables explicatives.

Définition 9. *Nous appelons coupure un élément de la forme*

$$S(j,d) \cup S(j,d)^c = \{x_t | x_t^j \leqslant d\} \cup \{x_t | x_t^j > d\}$$

où $j \in \{1, \ldots, p\}$ est le numéro d' une variable explicative et d un seuil (voir figure 3).

Nous noterons $S_1, \ldots, S_m, \ldots, S_M$ les régions formant cette partition. Sur la figure 3, nous avons schématisé un arbre où $S_1^1, \ldots, S_1^3$ représentent des coupures. Chaque région S_m est donc l'intersection

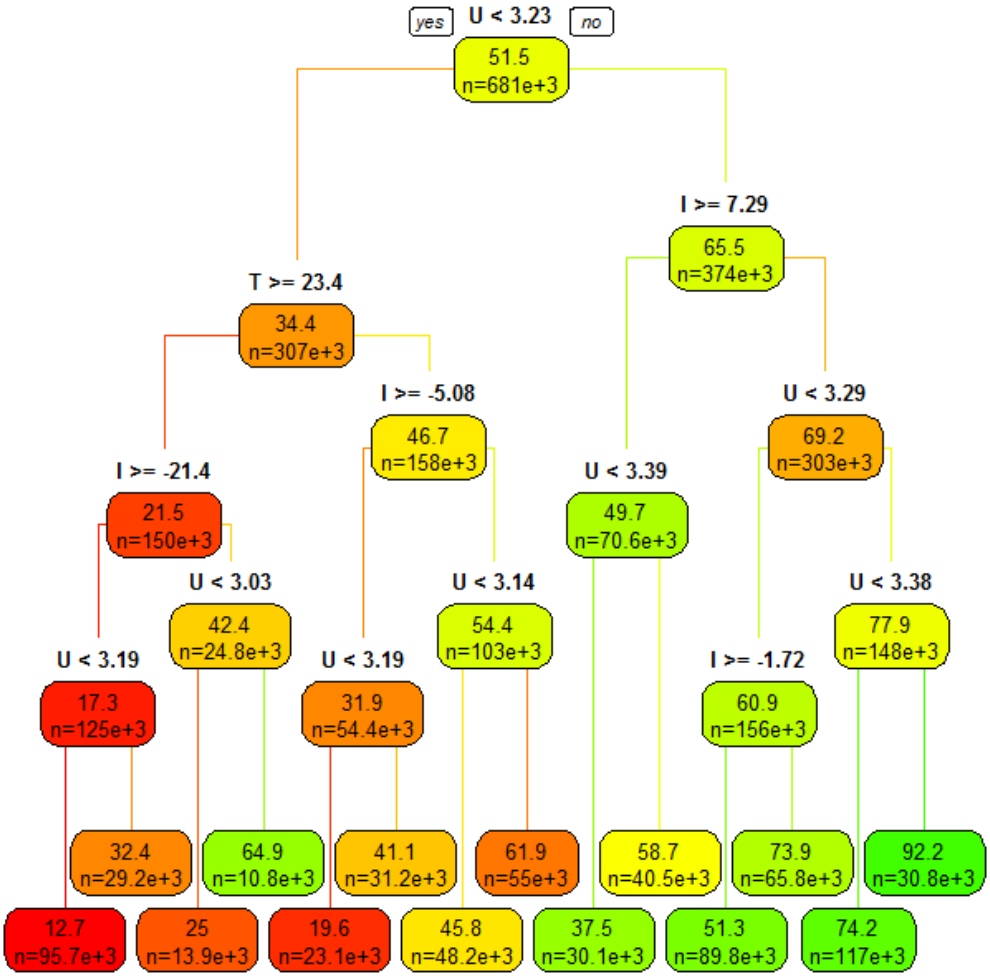


FIGURE 2 – Exemple d'arbre de régression avec le paquet rpart et rpart.plot de R (voir partie 3.3)

de coupures.

Dans notre cadre d'étude, nous associerons à chaque région S_m une valeur constante c_m. Nous cherchons donc un estimateur $\hat{f}$ de f (cf. (2)) sous la forme :

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \mathbb{1}_{x \in S_m}.$$

Par la suite, nous devrons déterminer c_m et S_m de telle sorte que $\widehat{f}$ minimise (4).

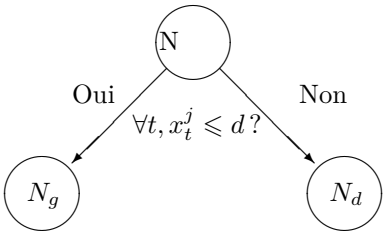


FIGURE 3 – Schématisation d'une coupure

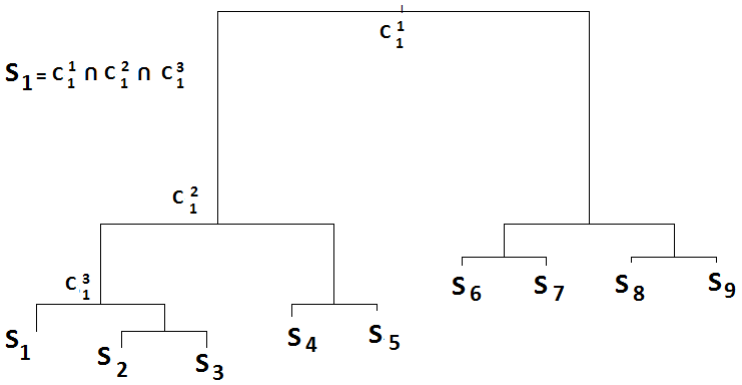


FIGURE 4 – Correspondance entre régions et feuilles de l'arbre

Coupure optimale.

On va préciser la règle de coupure qui permettra d'obtenir un tel estimateur :

Proposition 1. *Le découpage optimal $S(\widehat{j}, \widehat{d})$ est obtenu pour le couple $(\widehat{j}, \widehat{d})$ vérifiant :*

$$\begin{aligned}(\hat{j}, \hat{d}) \quad &\in \quad \underset{(j,d)}{\operatorname{argmin}} \left[\min_{c_1 \in \mathbb{R}} R^*(c_1 \mathbb{1}_{S(j,d)}) + \min_{c_2 \in \mathbb{R}} R^*(c_2 \mathbb{1}_{S(j,d)^c}) \right] \\ &\in \quad \underset{(j,d)}{\operatorname{argmin}} \left[\min_{c_1 \in \mathbb{R}} \sum_{x_t \in S(j,d)} (y_t - c_1)^2 + \min_{c_2 \in mathbbR} \sum_{x_t \in S(j,d)^c} (y_t - c_2)^2 \right]\end{aligned}$$

La méthode sélectionne alors la meilleure coupure, c'est-à-dire le couple (j, d) qui vise à diminuer à chaque étape l'erreur quadratique moyenne (MSE) (cf (4)) des nœuds fils.

Une fois la racine de l'arbre scindée en deux, on se restreint à chacun des nœuds fils et on recherche alors, suivant le même procédé, la meilleure façon de les découper en deux nouveaux nœuds, et ainsi de suite.

Remarques.

1. Une coupure est dite admissible si aucun des nœuds fils engendrés n'est vide.
2. Autrement dit, la meilleure segmentation $\widehat{S} \in \mathcal{S}$ d'un nœud N est celle qui fera le plus décroître

$R^*(\hat{f}_{|N})$ i.e. :

$$\widehat{S} = \underset{S \in \mathcal{S}}{\operatorname{argmax}}\ \Delta R(S, N), \tag{6}$$

où $\Delta R(S,N) = R^*(\widehat{f}_{|N}) - R^*(\widehat{f}_{|N_g}) - R^*(\widehat{f}_{|N_d})$.

3. Il existe d'autres versions de cet algorithme comme MARS, C4.5 ou CHAID qui utilisent des critères autres que l'erreur quadratique moyenne (MSE) pour choisir les meilleures coupures ([12]). Par exemple, CHAID (CHi-squared Automatic Interaction Detector) utilise un test du Chi-deux et ne prend en compte que des variables qualitatives comparé à CART qui permet d'utiliser indifféremment les deux sortes de variables : qualitatives ou quantitatives.

La proposition suivante permet de simplifier le double problème d'optimisation 6 et permet de déterminer enfin les c_m :

Proposition 2. *En assimilant un nœud N et la partie correspondante $S(j,d)$ (j et d fixés), la valeur associée à N sera :*

$$\underset{c \in \mathbb{R}}{\operatorname{argmin}}\ R^*(c\mathbb{1}_{S(j,d)}) = \frac{1}{|\{x_t \in S(j,d)\}|} \sum_{x_t \in S(j,d)} y_t = \overline{Y}_{S(j,d)}$$

où la somme est prise sur toutes les variables à expliquer y_t telles que $x_t \in S(j,d)$.

D'où le problème de coupure optimale devient :

$$\begin{aligned}(\hat{j}, \hat{d}) &\in \underset{(j,d)}{\operatorname{argmin}} \left[\min_{c_1 \in \mathbb{R}} \sum_{x_t \in S(j,d)} (y_t - c_1)^2 + \min_{c_2 \in \mathbb{R}} \sum_{x_t \in S(j,d)^c} (y_t - c_2)^2 \right] \\ &= \underset{(j,d)}{\operatorname{argmin}} \left[\sum_{x_t \in S(j,d)} (y_t - \overline{Y}_{S(j,d)})^2 + \sum_{x_t \in S(j,d)^c} (y_t - \overline{Y}_{S(j,d)^c})^2 \right].\end{aligned}$$

Conclusion. *Ainsi, l'arbre de régression T obtenu correspond à l'estimateur des moindres carrés $\widehat{f}_T$ de f sur l'espace des fonctions constantes par morceaux sur la partition formée par les feuilles définie par $(S_m)_{m=1,\dots,M}$. Ainsi, on a*

$$\widehat{f}_T(x) = \sum_{m=1}^{M} \widehat{c}_m \mathbb{1}_{x \in S_m} = \sum_{m=1}^{M} \overline{Y}_{S_m} \mathbb{1}_{x \in S_m}$$

La méthode CART peut être vue ainsi comme une sélection de modèle sur des sous-espaces de $\mathbb{L}^2(\mathbb{R}^p, \mu)$ de fonctions constantes par morceaux, définies sur des partitions finies de $\mathbb{R}^p$.

On peut maintenant exhiber un pseudo-code associé à l'algorithme CART :

1. *Commencer à la racine de l'arbre et associer à celle-ci toutes les observations $\{x_t = (x_t^1, \dots, x_t^p), y_t\}$ de l'échantillon d'apprentissage.*
2. *Choisir la coupure optimale $\widehat{S} = (x^{\widehat{j}} < \widehat{d})$ comme dans la proposition 1.*
3. *Une fois la coupure optimale choisie, associer à chaque nœud fils la valeur présentée dans la proposition 2*
4. *Si le critère d'arrêt choisi atteint un seuil fixé à l'avance, terminer l'algorithme et l'arbre construit sera dit maximal. Sinon, appliquer l'étape 2 à chaque nœud fils à son tour.*

Intéressons-nous maintenant aux critères d'arrêt évoqués dans la troisième étape du pseudo-code.

Critères d'arrêt.

Il existe plusieurs règles d'arrêt qui permettent de décider si un noeud est une feuille :

Noeud pur. *Un noeud devient une feuille lorsqu'il n'existe plus de coupure admissible. On dit alors que le noeud est pur.*

Profondeur maximale. *L'arbre atteint une profondeur maximale notée maxdepth fixée par l'utilisateur. La profondeur par défaut sur* `rpart` *est 30 et nous conserverons cette valeur pour ne pas tenir compte de ce critère.*

Taille du noeud. *Elle consiste à ne pas découper N si le nombre d'observations de $\mathcal{L}_n$ contenues dans N est inférieure à un certain nombre noté nodesize. La valeur par défaut est 25 sur* `rpart` *et nous la conserverons également.*

Paramètre de complexité c_p. *Elle consiste à ne pas découper le noeud N via la segmentation optimale $\widehat{S}$ si $\Delta R(S, N) \geqslant c_p$. La valeur par défaut dans* `rpart` *est 0.01. Par la suite, nous essaierons de calibrer au mieux ce paramètre.*

Les critères utilisés dans le package `rpart` peuvent être réglés par l'utilisateur.

3.2.3 Forêts aléatoires

La définition générale des forêts aléatoires est la suivante :

Définition 10. *Soit $(\hat{f}(.,\Theta_1), \ldots, \hat{f}(.,\Theta_L))$ une collection d'estimateurs par arbre, où $(\Theta_1, \ldots, \Theta_L)$ est une suite de variables aléatoires i.i.d., indépendante de l'échantillon d'apprentissage $\mathcal{L}_n$. L'estimateur par forêt aléatoire est obtenu par agrégation (c'est à dire en faisant la moyenne de cette collection d'estimateurs).*

Remarques. Le terme forêt aléatoire vient du fait que les estimateurs individuels sont, ici, explicitement des estimateurs par arbre.

Concernant le vocabulaire des forêts aléatoires, il existe une ambiguïté dans la littérature. En effet, Leo Breiman, dans son article de 2001 ([3]), définit les forêts aléatoires comme ci-dessus. Les forêts aléatoires sont donc pour lui une famille de méthodes.

Or, dans le même article, il présente un cas particulier de forêts aléatoires, appelées Random Forests-RI (Random Inputs), qu'il a ensuite implémentées et dont nous verrons juste après l'algorithme ([5],[11]).

Par suite, ce sont ces Random Forests-RI qui ont été utilisées dans de très nombreuses applications réelles. Et pour cause, le programme est accessible à tous, facile d'utilisation et la méthode atteint des performances exceptionnelles. Finalement, la dénomination "forêts aléatoires" désigne dans toute la suite les Random Forests-RI.

Comment construit-on de tels objets ? Pour ce faire, on a besoin de la définition d'un échantillon bootstrap :

Définition 11. *Un échantillon bootstrap $\mathcal{L}_n^{\Theta}$ est obtenu en tirant aléatoirement n observations avec remise dans l'échantillon d'apprentissage $\mathcal{L}_n$, chaque observation ayant une probabilité $\frac{1}{n}$ d'être tirée. La variable aléatoire Θ représente alors ce tirage aléatoire. L'algorithme définissant les Random Forest-RI est le suivant :*

1. Pour $\ell = 1, \ldots, L$ faire :
   (a) Engendrer un échantillon bootstrap $\mathcal{L}_n^{\Theta_\ell}$.

(b) Créer un arbre aléatoire T_ℓ associé à $\mathcal{L}_n^{\Theta_\ell}$ en répétant récursivement les trois étapes suivantes pour chaque nœud de l'arbre jusqu'à ce que *nodesize* soit atteint :

(i) Choisir aléatoirement et uniformément m variables explicatives parmi p.

(ii) Prendre le meilleur couple variable explicative/seuil pour ces m variables au sens de la proposition 2.

(iii) Découper le nœud en deux nœuds fils comme dans l'algorithme CART.

2. Renvoyer la collection d'arbres $(T_\ell)_{\ell=1,\dots,L}$ obtenue.

Conclusion. *L'estimateur par forêt aléatoire est obtenue de la façon suivante pour tout $x \in \mathbb{R}^p$:*

$$\widehat{f}_{RF}^{L}(x) = \frac{1}{L}\sum_{\ell=1}^{L} \widehat{f}_{T_\ell}(x).$$

Remarques.

1. Le tirage, à chaque noeud, des m variables se fait sans remise et uniformément parmi toutes les variables (chaque variable a une probabilité $\frac{1}{p}$ d'être choisie). Le nombre m ($m \leqslant p$) est fixé au début de la construction de la forêt et est donc identique pour tous les arbres. Généralement, pour des problèmes de régression on choisit $m = E(\frac{p}{3})$.
2. Pour les Random Forests-RI, il y a donc deux sources d'aléas pour engendrer la collection des estimateurs individuels : l'aléa dû au bootstrap et l'aléa dû au choix des variables pour découper chaque nœud d'un arbre. Ainsi, on perturbe à la fois l'échantillon sur lequel on lance la règle de base, et à la fois le cœur de la construction de la règle de base.

   Ainsi, l'explication des performances des forêts aléatoires réside dans le fait que rajouter un aléa supplémentaire pour construire les arbres, rend ces derniers encore plus différents les uns des autres, sans pour autant dégrader leurs performances individuelles. L'estimateur agrégé est alors meilleur. ([1],[4]).

3.3 Perspectives et questions ouvertes

En théorie, les méthodes proposées sont capables de prendre en compte une très grande base de données, c'est d'ailleurs pour cela que nous les avons choisies. Mais en pratique, nous nous sommes rendus compte que certes elles prenaient en compte plus de données que la plupart des autres méthodes, mais pas assez pour celles que nous avons à traiter. L'une des perspectives serait d'essayer d'améliorer cet aspect restrictif.

Par ailleurs, lors du descriptif de ces méthodes, nous avons pu voir qu'à aucun moment elles ne prennent en compte l'aspect temporel de nos données. Certes, lors de mon stage, nous avons appliqué telles quelles ces deux méthodes à notre très grande base donnée mais les résultats n'ont pas été entièrement satisfaisants. C'est pour cela qu'on pourrait se demander si l'on ne peut pas adapter ces méthodes de régression à l'analyse de données temporelles. Affaire à suivre...

Références

[1] Biau G., Devroye L., Lugosi G., *Consistency of random forests and other averaging classifiers.* Journal of Machine Learning Research 9 2015-2033, 2008.

[2] Breiman L., Friedman J.H., Olshen R.A., Stone J.H., *Classification and regression trees.* Chapman and Hall, 1984.

[3] Breiman L., *Random Forests.* Machine Learning, 45 (2001) 5–32. 2001.

[4] Breiman L., *Bagging predictors.* Machine Learning, 24(1996) 123—140.

[5] Breiman L., *RFtools-for predicting and understanding data.* Technical report Berkeley University USA, 2004.

[6] Genuer R., *Forêts aléatoires : aspects théoriques, sélection de variables et applications.* Thèse Université Paris-Sud, Orsay 2010.

[7] Gey S., *Bornes de risque, détection de ruptures, boosting : trois thèmes statistiques autour de CART en régression.* Thèse Université Paris-Sud, Orsay 2002.

[8] Hansen T., Wang C-H., *Support vector based battery state of charge estimator.* Journal of Power Sources 141, 351-358, 2005.

[9] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning.* Second edition, Springer 2009.

[10] Kim H., Loh WY, Shih YS, Chaudhuri P., *Visualizable and interpretable regression models with good prediction power.* IIE Transactions 39, 565-579, 2007.

[11] Liaw A., Wiener M., *Classification and regression by random forest.* R News, 2 (3) 18-22, 2002.

[12] Loh WY, *Classification and regression trees.* WIREs Data Mining Knowl Discov 1, 14-23, 2011.

[13] Milleborrow S., *Plotting rpart trees with prp.* http ://www.milbo.org/rpart-plot/prp.pdf. 2012.

[14] Pang S., Farrell J., Du J., Barth M., *Battery state-of-charge estimation.* Proceedings of the American Control Conference, June 2001, vol. 2, p. 1644-1649 2001.

[15] Piller S., Perrin M., *Methods for state-of-charge determination and their application.* Journal of Power Sources 96 113-120, 2001.

[16] Therneau TM., Atkinson B., *RPART : recursive partitioning.* R port by B. Ripley. R package version 3.1-41, 2008.

[17] Venables WN., Ripley BD., *Modern Applied Statistics with S.* Fourth Edition, p258, 2002.

[18] Verikas A., Gelzinis A., Bacauskiene M., *Mining data with random forests : a survey and results of new tests.* Pattern recognition, 2011

4 Conclusion

Les trois années de magistères ont été pour moi l'occasion de m'épanouir dans le monde des mathématiques, de confirmer mon engouement pour cette discipline et de mettre aussi toutes les chances de mon côté afin de concrétiser mon projet professionnel . Elles ont été également le théâtre de mes plus belles rencontres tant amicales que professionnelles et j'en garde un excellent souvenir rempli d'anecdotes, de travail acharné et d'amour des mathématiques.