**Journal of Mathematics et Statistics Studies (JMSS)**
**www. al-kindipublisher.com/index.php/jmss**
**ISSN: 2709-4200**

# Apprentissage statistique du modèle Cox-logistique : application à la survie des enfants de moins de 5 ans au Bénin

Mohamadou Salifou [1]* & N.S.A.F. Madjid A. Houessou [2]
[1] *Chaire Internationale en Physique Mathématique et Applications (CIPMA-Chaire UNESCO) de la Faculté des Sciences et Techniques (FAST) de l'Université d'Abomey-Calavi, Bénin.*
[2] *Département de Mathématiques et Applications, Université de Pau et des Pays de l'Andour, France.*
**Auteur correspondant :** Mohamadou Salifou, E-mail : salifoumouhamed.ms.ms@gmail.com

**ARTICLE INFO**

**Received**: September 02, 2020
**Accepted**: October 25, 2020
**Volume**: 1
**Issue**: 2

**MOTS CLES**

Apprentissage statistique, mortalité des enfants, analyse de survie, modèle Cox-Logistique, prédiction, Elastic-Net, arbre de survie, forêt aléatoire (de survie)

**RESUME**

La mortalité des enfants de moins de cinq ans est un phénomène connu et très rependu au Bénin. Rechercher les facteurs qui l'engendrent et caractériser les couches d'enfants les plus vulnérables serait un pas dans la réduction du taux de mortalité des enfants. Après avoir souligné les limites des modèles utilisés jusque-là pour appréhender les déterminants du phénomène, la présente étude s'est fixée comme principal objectif de construire un modèle de prédiction qui explique au mieux la mortalité des enfants de moins de cinq ans dans le contexte béninois. Ainsi, nous avons mis en évidence la pertinence du modèle à hasards proportionnels avec un surplus de zéro (Cox-logistique) pour analyser les déterminants de la mortalité des enfants de moins de cinq ans à travers l'âge au décès. Nous comparons les performances prédictives du modèle Cox-logistique pénalisé par la procédure Elastic-Net avec celles obtenues à partir de la méthodologie des forêts aléatoires (respectivement de survie) connue pour donner la plus petite erreur de prédiction, mais difficile à interpréter par les non-statisticiens. Nous comparons également leur facilité d'interprétation. Enfin, les covariables sélectionnées par chaque procédure sont comparées et discutées. Les résultats suggèrent que la méthode de sélection basée sur la pénalisation Elastic-Net appliquée à la régression logistique donne une bonne alternative aux forêts aléatoires de classification, en association avec un modèle final facile à interpréter pour les démographes, afin de distinguer le statut des enfants dès leur naissance face au décès. À l'opposé, la procédure d'arbre de survie offre un cadre d'interprétation visuelle à travers l'arbre optimal fourni, dont les variables constitutives sont toutes identifiées comme importantes pour la prédiction de l'âge au décès des enfants au Bénin.

## 1. Introduction

La mortalité en général, et celle des enfants de moins de cinq ans en particulier est l'un des problèmes majeurs de développement que connaissent les pays en voie de développement et surtout ceux d'Afrique subsaharienne, en dépit des stratégies de développement sanitaires mises en œuvre (OMS, 2005). C'est pourquoi l'importance accordée à la mortalité des enfants a connu plus d'intensité au cours des dernières décennies. Le Bénin, à l'instar de la quasi-totalité des autres pays d'Afrique, s'est ainsi engagé à mettre à exécution des politiques nationales pour réduire la mortalité des enfants de moins de cinq ans. En 2017, le taux de mortalité des enfants de moins de cinq ans est estimé à 96 pour 1 000 naissances vivantes (INSAE et ICF, 2019). Ce taux était de 160 % en 2001, de 125 % en 2006 et de 115% en 2014 (INSAE et ORC-Macro, 2002; INSAE et Macro-International-Inc, 2007; INSAE, 2016). Ces chiffres démontrent qu'effectivement, des stratégies d'action ont été mises en place en vue d'assurer une réduction de la mortalité infanto-juvénile au Bénin.

Toutefois, malgré les progrès en matière de santé au plan global, d'importants défis restent à relever pour une universalisation effective des soins de santé, susceptible d'accroitre la survie des enfants en bas-âge. Agir dans le sens de réduction du taux de mortalité des enfants de moins de cinq ans, c'est d'abord comprendre les déterminants du phénomène et définir les particularités rattachées aux individus qui en sont touchés. D'ores et déjà, en plus des recensements, et, des Enquêtes Démographique et de Santé (EDS), plusieurs études (Barbieri, 1991; Boco et Bignami, 2008; Dansou, 2016) ont été menées au

Bénin pour appréhender les facteurs déterminants de la mortalité des enfants. Ces études, comme la plupart réalisée en Afrique ont utilisé les méthodes statistiques de régression logistique de type logit (Aly et Grabowsky, 1990; Barbieri, 1991; Boco et Bignami, 2008 ; Mboko Ibara, 2009; Dansou, 2016) ou probit (Melligton et Cameron, 1999), qui, estime les risques ou la probabilité de survenance d'un décès d'enfant en fonction des covariables sur les 0-59 mois. Dans ce cas, la variable expliquée est binaire (prenant la valeur 1 lorsque l'enfant est décédé, 0 sinon). On peut cependant faire remarquer, qu'en donnant la même valeur à la variable expliquée sur un vaste intervalle, on perd une information précieuse (Ray, 1988). Par ailleurs, avec le développement des méthodes statistiques des dernières décennies, il pourrait s'avérer plus avantageux de répondre à la question *combien de temps s'est écoulé avant la survenue d'un décès d'enfant qu'à la question y a-t-il survenue ou non de décès d'enfant dans une période de temps.* Autrement, nous allons nous intéresser non pas au fait que l'enfant soit en vie ou décédé à un moment donné, mais plutôt à la transition d'un état à l'autre (de la vie à la mort). Plus précisément, on se donne comme objectif de modéliser l'âge auquel l'enfant décède. Il se pose alors un problème : qu'en est-il des enfants qui ne sont pas décédés durant la période de l'étude ? En effet, quel que soit le mode de collecte, prospectif ou rétrospectif, on est en présence de données incomplètes ; l'information sur les trajectoires individuelles des enfants s'interrompt à la date de l'enquête et on ne connaît pas leur avenir. De telles durées sont dites censurées à droite (Lelièvre, 2010). Il est évident compte tenu de ce phénomène que l'application en toute rigueur des modèles de régression multiple n'est plus admissible, car une des conditions de base n'est pas remplie (Ray, 1988) : l'absence d'erreurs de mesure sur la variable dépendante ; d'où des estimateurs non efficaces, bien que non biaisés si le terme d'erreur est réellement aléatoire. Cependant, cette caractéristique particulière des données censurées, source de difficultés, nécessite le développement des techniques alternatives à l'inférence usuelle : il s'agit des modèles de durée (hazard models) ou encore l'analyse des données de survie (survival data analysis) qui ont été pendant longtemps le propre des seuls démographes et actuaires et dont récemment les applications se sont étendues à d'autres domaines tels que la fiabilité, l'économie, la bio statistique, la psychologie, et les sciences sociales en général. C'est une méthode d'analyse plus informative car non seulement elle répond à la question *si oui ou non un événement s'est produit mais également combien de temps s'est écoulé avant sa réalisation*.

Nonobstant, les modèles classiques en analyse des données de survie ne sont définis que pour des durées strictement positives. Malheureusement, dans le cas des données de la cinquième édition de l'Enquête Démographique et de Santé (EDSB-V) réalisée au Bénin, il y a des enfants qui sont décédés au même moment précis où ils naissent, ce qui fait que la durée jusqu'à la réalisation de l'événement (décès) est nulle. Partant, l'objectif de ce papier est de construire un modèle de prédiction qui doit être en mesure de modéliser à la fois les durées strictement positives et les durées nulles. À cet effet, à travers une approche analytique basée sur les modèles à risques proportionnels avec un surplus de zéro, le présent travail essaiera de comparer différents modèles de prédiction pour quantifier les relations entre les différentes covariables et le temps de survie ou décès (ou probabilité de survie) des enfants au Bénin.

En nous inspirant des travaux de Grouwels et Braekers (2011), nous proposons un modèle de mélange où la masse à zéro est modélisée avec la régression logistique et les durées positives par le modèle à hasards proportionnels de Cox. Dans un premier temps, cette modélisation paramétrique est appliquée ici pour prédire l'âge au décès des enfants au Bénin. Dans un second temps, nous allons utiliser une modélisation non paramétrique, en l'occurrence les forêts aléatoires (respectivement les forêts aléatoires de survie) s'inspirant des forêts d'arbres décisionnels pour construire des ensembles de risque permettant de calculer les probabilités de survie des enfants. Bien entendu, on ne saura parler de forêts sans parler des arbres (respectivement des arbres de survie).
Pour analyser l'âge auquel l'enfant décède, on utilise, les données l'EDSB-V. Notre population d'étude est constituée de l'ensemble des naissances vivantes des femmes de 15-49 ans au cours des 5 années ayant précédé l'enquête. Au total, notre échantillon est composé des 13 589 naissantes vivantes survenues au plus tard en 2012, dont 938 cas de décès enregistrés.

**2. Modèle de mélange de regressions logistique et Cox**

Nous présentons dans cette section, le modèle Cox-logistique issu des modèles de régression logistique et de régression de Cox. Pour une étude plus approfondie sur les contextes de la régression logistique, et des données censurées à droite ainsi que quelques fondements théoriques du modèle de régression logistique et des techniques d'analyse statistique de base des données de survie, le lecteur pourra se référer respectivement aux travaux de Czepiel (2018), Chesneau (2015), Hill et al. (1999) et Klein et Moeschberger (1997).

### *2.1 Présentation du modèle*

Désignons par *T*, une variable aléatoire mesurant l'âge au décès de l'enfant dans cette étude. Nous rappelons que dans notre échantillon, nous avons des enfants qui sont décédés à la naissance et de ce fait pour ces derniers *T* = 0. Ainsi, *T* prend une valeur nulle avec une loi de probabilité discrète et une valeur positive avec une loi de probabilité continue.

Soit $Y_i$ la variable aléatoire telle que $Y_i = 1$ si l'enfant *i* décède à la naissance, c'est-à-dire *T* = 0 et $Y_i$ = 0 sinon, c'est-à-dire *T* > 0. La distribution conditionnelle de *T* sachant les caractéristiques *X* (sexe, poids, taille, rang, gémellité, …) des enfants est donnée par :

$$\boldsymbol{F\,(t|x) = P(T \leq t|X = x) = \pi(x) + [1 - \pi(x)]\, F_{T>0}\,(t|x)} \qquad \textbf{(1)}$$

Où $\boldsymbol{\pi(x) = P(Y = 1|X = x)}$ et $\boldsymbol{F_{T>0}\,(t|x) = P(T \leq t|\, T > 0, X = x)}$ est la distribution conditionnelle de *T* sachant *T* > 0. Ainsi, nous allons modéliser l'un après l'autre $\boldsymbol{\pi(x)}$ et $\boldsymbol{F_{T>0}\,(t|x)}$.

Pour déterminer les facteurs qui influencent le temps de survie des enfants, nous exprimons $\boldsymbol{\pi}$ en fonction de $x_i$ (*i* = 1 *: n*) grâce à la fonction sigmoïde :

$$\boldsymbol{\pi(x_i) = \frac{\exp\,(\theta_0 + \sum_{j=1}^{p} x_{ij}\,\theta_j)}{1 + \exp(\theta_0 + \sum_{j=1}^{p} x_{ij}\,\theta_j)}} \qquad \textbf{(2)}$$

où $\boldsymbol{\theta_j}$ représente le logarithme du rapport de la cote qu'un enfant décède à la naissance par rapport à la même cote chez les enfants qui en ont survécus.

Étant donné que certains enfants sont encore en vie à la date de l'enquête, la durée *T* n'est pas observée pour tous les enfants. Elle est donc susceptible d'être censurée à droite. Partant, on suppose qu'il existe une variable aléatoire *C* telle qu'on observe seulement *Z* = *min* (*T*, *C*) et $\delta = \mathbf{1}_{T \leq C}$. A cet effet, nous estimons que $F_{T>0}\,(z|X)$, peut être ajusté par un modèle de Cox de risques proportionnels. Ainsi, nous supposons que la fonction de hasard conditionnelle pour tout *z* > 0, est de la forme suivante :

$$\boldsymbol{\lambda_{T>0}(z|x) := \frac{f_{T>0}(z|x)}{S_{T>0}(z|x)} := \lambda_0(z)\exp(\beta' x)} \qquad (\mathbf{3})$$

Où,

- $\boldsymbol{f_{T>0}(z|x)}$ est la densité de probabilité et correspond à la probabilité dite *instantanée* que le décès d'enfant survienne dans l'intervalle de temps infinitésimal $\boldsymbol{[z, z + dz[}$ ;
- $\boldsymbol{S_{T>0}(z|x)}$ est appelé fonction de survie, représentant la probabilité d'avoir survécu au-delà d'un instant *z*. Son expression est : $\boldsymbol{S_{T>0}(z|x) = P(Z > z|\, x) = 1 - F_{T>0}\,(t|x)}$ **;**
- $\boldsymbol{\lambda_0(z)}$ est appelée fonction de risque de base inconnue, il ne dépend que du temps, mais indépendant des covariables ;
- $\boldsymbol{x_i = (x_i^{(1)}, \ldots, x_i^{(p)})'}$ est un vecteur de valeurs particulières prises par le vecteur de *p* covariables associé à l'individu *i*, $\boldsymbol{X_i = (X_i^{(1)}, \ldots, X_i^{(p)})'}$, et indépendant du temps ;
- $\boldsymbol{\beta = (\beta_1, \ldots, \beta_p)'}$ est le paramètre de régression inconnu, indépendant du temps ;
- $\boldsymbol{\exp(\beta' x)}$ est le risque relatif, indépendant du temps, mais fonction des covariables.

Pour estimer les paramètres $\beta$ et $\vartheta$ du modèle ainsi défini, nous construisons la fonction de vraisemblance à partir des observations $\boldsymbol{(Z_1, \delta_1, X_1, Y_1), \ldots, (Z_n, \delta_n, X_n, Y_n)}$ *indépendantes et identiquement distribuées* issues de l'échantillon de variables $\boldsymbol{(Z, \delta, X, Y)}$**.** Par ailleurs, on peut analyser comment les ***n*** enfants de notre échantillon contribuent à la vraisemblance, comme suit :

- ✓l'enfant *i* qui est décédé à la naissance, c'est-à-dire celui pour qui $Y_i = 1$, contribue par $\boldsymbol{\pi(x_i)}$ à la vraisemblance ;
- ✓l'enfant *i* pour qui le décès est observé au temps $\boldsymbol{z_i}$, c'est-à-dire $Y_i = 0$ et $\delta_1 = 1$, contribue par [1- $\pi_i(x_i)$] $f_i\,(z_i|x_i)$ à la vraisemblance ;
- ✓l'enfant *i* qui est censuré au temps $\boldsymbol{z_i}$, c'est-à-dire celui pour qui $Y_i = 0$ et $\delta_1 = 0$, contribue par [1-$\pi_i(x_i)$] $S_i\,(z_i|x_i)$, car tout ce que nous savons est qu'il a survécu jusqu'à la date de l'enquête.

Ainsi, la vraisemblance prend la forme :

$$L(\beta, \theta) = \prod_{i=1}^{n} [\pi_i(x_i)]^{y_i} \{[1 - \pi_i(x_i)]^{1-y_i} f_i\,(z_i|x_i)\}^{\delta_i} \{[1 - \pi_i(x_i)]^{1-y_i} S_i\,(z_i|x_i)\}^{1-\delta_i}$$

$$\boldsymbol{L(\beta, \theta) = \underbrace{\prod_{i=1}^{n} [\pi_i(x_i)]^{y_i}\,[1 - \pi_i(x_i)]^{1-y_i}}_{(1)} \underbrace{\prod_{i=1}^{n} \{f_i\,(z_i|x_i)\}^{\delta_i} \{S_i\,(z_i|x_i)\}^{1-\delta_i}}_{(2)}} \quad (\mathbf{4})$$

On constate que (1) est la vraisemblance d'une régression logistique ordinaire pour l'ensemble de l'échantillon (Czepiel, 2018) et que (2) est équivalent à $\boldsymbol{\prod_{i=1}^{n} \{f\,(z_i|X_i)\}^{\delta_i} \{S\,(z_i|X_i)\}^{1-\delta_i}}$ qui est la vraisemblance pour un modèle de survie pour la variable

*Z* avec une censure à droite (Klein et Moeschberger, 1997). On peut donc remplacer ce terme (2) par la vraisemblance partielle du modèle de Cox pour inférer sur le paramètre $\beta$. Ainsi, ***la vraisemblance du modèle Cox-logistique est le produit de la vraisemblance d'une régression logistique et de la vraisemblance partielle d'une régression de Cox***. En conséquence, les statistiques de Wald ou du rapport de vraisemblance pour tester la significativité de $\theta$ et $\boldsymbol{\beta}$ dans leur modèle respectif peuvent être additionnées. La somme de ces statistiques permet de tester si $\theta$ et $\boldsymbol{\beta}$ sont simultanément significatifs dans le modèle. Autrement, c'est évaluer les facteurs déterminants de la durée de survie des enfants de moins de cinq ans au Bénin.

### *2.2 Sélection de variables pour éviter le sur-apprentissage*

Le sur-apprentissage se définit comme le fait d'apprendre parfaitement les observations présentes. Le modèle connaît donc de mémoire la variable réponse à prédire pour toutes les observations. Ainsi lorsqu'on présente une nouvelle observation au modèle, celui-ci n'est plus en mesure de prédire correctement la valeur. Afin d'éviter le sur-apprentissage, on cherche alors à sélectionner les variables les plus pertinentes afin d'obtenir les meilleurs modèles de prédiction possibles, en visant l'interprétabilité et la stabilité de ces modèles. Les techniques classiques de sélection de sous-ensembles telles que la sélection progressive (*stepwise*) ou la sélection pas-à-pas descendante/ascendante se sont révélées insatisfaisantes au cours du temps dans le cadre du modèle de Cox. Le point de vue du praticien est qu'aucune stratégie de sélection ne s'est montrée meilleure que la stratégie consistant à inclure toutes les covariables dans le modèle (après un filtrage préliminaire), car ces méthodes élimineraient ou ignoreraient facilement des facteurs importants (Greenland, 2008). Une approche alternative est donnée par les méthodes de pénalisation.

Le principe de la pénalisation d'une régression est de contraindre les coefficients à ne pas être « trop grands », qui se traduisent par l'introduction d'une contrainte de la forme :

$$\sum_{j=1}^{p} |\beta_j|^{\xi} \leq \lambda \qquad (5)$$

Lorsque le modèle intègre une constante, elle est exclue de la contrainte. Les choix classiques pour $\xi$ sont 0 (pénalisation en fonction du nombre de coefficients), 1 (LASSO) et 2 (Ridge).

#### *2.2.1. Modèle LASSO*

Le LASSO (Least Absolute Shrinkage and Selection Operator) est une méthode de pénalisation $\boldsymbol{\ell_1}$ et de sélection de variables initialement proposée pour la régression linéaire. Adaptée par Tibshirani (1997) au modèle de Cox, cette méthode permet d'estimer les paramètres $\boldsymbol{\beta}$ via la maximisation de la log-vraisemblance partielle (définie en (2) dans l'équation (4)) sous la contrainte $\sum_{j=1}^{p}|\beta_j| \leq \lambda$ , où λ est un paramètre de régularisation. La contrainte LASSO sélectionne les variables en enlevant celles qui ont les plus petits coefficients estimés. Cela conduit à des coefficients exactement égaux à zéro et permet d'obtenir un modèle parcimonieux et interprétable. Le LASSO appliqué au modèle Cox-logistique consiste à maximiser la fonction de vraisemblance (4) pénalisée par la norme $\boldsymbol{\ell_1}$ du vecteur des coefficients inconnus (d'une part de la régression logistique et d'autre, du modèle de Cox).

#### *2.2.2. Modèle Ridge*

La régression Ridge nous permet d'éviter le sur-apprentissage en restreignant l'amplitude des coefficients de régression. Elle sert donc à l'estimation des paramètres $\boldsymbol{\beta}$ (du modèle de Cox) par la maximisation de la log-vraisemblance partielle sous la contrainte $\sum_{j=1}^{p}|\beta_j|^2 \leq \lambda$ , où λ est un paramètre de régularisation (la même procédure est appliquée dans le cas de régression logistique). Il apparait que la pénalisation utilisée ici est la norme $\boldsymbol{\ell_2}$. Contrairement au LASSO, la régression Ridge a un effet de *sélection groupée* : les variables corrélées ont le même coefficient.

Par ailleurs, si plusieurs variables corrélées contribuent à la prédiction de la variable réponse, le LASSO va avoir tendance à choisir une seule entre elles (affectant un poids de 0 aux autres), plutôt que de répartir les poids équitablement comme la régression Ridge. C'est ainsi qu'on arrive à avoir des modèles très parcimonieux. Cependant, *laquelle* de ces variables est choisie aléatoirement, et peut changer si l'on répète la procédure d'optimisation. Le LASSO a donc tendance à être instable. Partant, un compromis entre la parcimonie et la qualité du modèle est trouvé en la méthode *Elastic-Net*.

#### *2.2.3. Modèle Elastic-Net*

La méthode Elastic-Net combine les atouts des méthodes Ridge et LASSO. En particulier, elle pallie le défaut de l'estimation LASSO lorsque les covariables sont fortement corrélées. Pour des raisons de commodité, nous multiplions chacune des log-vraisemblances des modèles obtenues à travers (1) et (2) dans l'équation (4) par$2/n$. Par conséquent, si nous considérons la

formulation Lagrangienne, ajouter une pénalité de type *Elastic Net* au modèle Cox-logistique consiste à résoudre les deux problèmes suivants :

$$\hat{\theta}_{EN} = argmax_{\theta} \left\{ \frac{2}{n} \left[ \sum_{i=1}^{n} y_i \left( \boldsymbol{\theta_0} + \sum_{j=1}^{p} \boldsymbol{x_{ij}}\, \boldsymbol{\theta_j} \right) - \log \left[ 1 + exp \left( \boldsymbol{\theta_0} + \sum_{j=1}^{p} \boldsymbol{x_{ij}}\, \boldsymbol{\theta_j} \right) \right] \right] - \lambda\, P_{\alpha}(\theta) \right\} \quad (6)$$

où

$$P_{\alpha}(\theta) = \lambda \left[ \alpha \sum_{j=1}^{p} |\theta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} \theta_j \right] \quad (7)$$

$$\hat{\beta}_{EN} = argmax_{\beta} = \left\{ \frac{2}{n} \sum_{i=1}^{D} \left[ \beta' x_{(i)} - \log\left( \sum_{j \in R(t^*_i)} \exp(\beta' x_j) \right) \right] - \lambda \left[ \alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} \beta^2{}_j \right] \right\} \quad (8)$$

Cette pénalité dévient une pénalité *LASSO* avec $\alpha = 1$ et devient une pénalité *Ridge* pour $\alpha = 0$. L'*Elastic net* combine donc les normes $\boldsymbol{\ell_1}$ et $\boldsymbol{\ell_2}$ pour obtenir une solution *moins parcimonieuse* que le LASSO, mais plus stable et dans laquelle toutes les covariables corrélées pertinentes pour la prédiction de la variable réponse sont sélectionnées et reçoivent un poids identique. Cette pénalité est celle utilisée dans ce papier. Il se pose ensuite le problème de choix des paramètres de régularisation $\lambda$ et $\alpha$. Une alternative est de sélectionner $\lambda$ par une validation croisée, fixant $\alpha$ sur une grille de valeurs ($\alpha \in [0,1]$), qui tient compte de la nature dépendante des données.

### *2.2.4. Validation croisée*

Le paramètre $\lambda \geq 0$ contrôle la complexité du modèle, de sorte que si $\lambda \to \infty$ aucune variable n'est retenue dans le modèle, alors que si $\lambda = 0$, la solution est celle obtenue par la vraisemblance définie en (4). Le paramètre de régularisation $\lambda$ peut être estimé par une méthode de ré-échantillonnage telle que la validation croisée.

Nous estimons le paramètre $\lambda$ par la valeur qui minimise le critère de validation croisée à $K$ ensembles appliqué à la vraisemblance du modèle Cox-logistique, respectant l'appariement des données. En effet, on constitue des échantillons $E_k$ à peu près de même taille, c'est-à-dire $E = \bigcup_{k=1}^{K} E_k$ tel que les $E_k$ soient disjoints deux à deux. Tour à tour, chacun des k échantillons est utilisé comme l'*ensemble test* et $E \setminus E_k$ comme l'ensemble d'apprentissage. L'estimateur de validation croisée sur la vraisemblance du modèle Cox-logistique est :

$$\boldsymbol{CVL}\,(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{L}\left( \widehat{\boldsymbol{\beta}}^{\boldsymbol{K}}{}_{\boldsymbol{E \setminus E_k}}, \widehat{\boldsymbol{\theta}}^{\boldsymbol{K}}{}_{\boldsymbol{E \setminus E_k}}, \boldsymbol{E_k} \right) - \boldsymbol{L}\left( \widehat{\boldsymbol{\beta}}^{\boldsymbol{K}}{}_{\boldsymbol{E \setminus E_k}}, \widehat{\boldsymbol{\theta}}^{\boldsymbol{K}}{}_{\boldsymbol{E \setminus E_k}}, \qquad \boldsymbol{L}\left( \widehat{\boldsymbol{\beta}}^{\boldsymbol{K}}{}_{\boldsymbol{E \setminus E_k}}, \widehat{\boldsymbol{\theta}}^{\boldsymbol{K}}{}_{\boldsymbol{E \setminus E_k}}, \boldsymbol{E_k} \right) \right) \quad (\mathbf{9})$$

L'objectif principal est ici la sélection de variables, avant la prédiction ou l'estimation. Ce critère fait intervenir le terme de pénalité, ainsi les grandes dimensions sont pénalisées, favorisant l'élimination des variables considérées moins pertinentes.

### *2.3 Des arbres aux forêts aléatoires*

Les arbres de décision peuvent être autant de l'ordre de la régression, donc la variable réponse doit être continue ou encore de l'ordre de la classification, ainsi la variable réponse est catégorielle. Dans les deux cas, les modèles ont pour but d'estimer la valeur de la variable réponse. Dans cette section, nous présentons des méthodes à base d'arbres de régression adaptées aux données de survie (le principe est le même pour ce qui est de la classification), qui, ont connu un essor important durant ces dernières décennies. Il faut cependant noter que les résultats présentés ici sont grandement inspirés de l'article de Walschaerts et al. (2011) portant sur la sélection de modèles dans le cas où la variable réponse est censurée à droite et celui de Genuer and Poggi (2017) pour la sélection des variables avec les forêts aléatoires.

### *2.3.1 Arbre de survie*

Les arbres de survie sont la généralisation aux données censurées des arbres de régression et de classification, popularisés par l'algorithme CART (*Classification and Regression Tree*) de Breiman et *al*. (1984) qui est basé sur un partitionnement récursif binaire de l'espace engendré par les covariables de façon dyadique. C'est un processus itératif qui divise les données en deux sous-groupes (les nœuds fils) selon la valeur des prédicteurs. La règle de division maximise la différence entre les deux nœuds fils. Le processus continu jusqu'à ce que chaque nœud atteigne une taille minimale précisée par l'utilisateur ou soit homogène. On parle alors de nœud terminal ou feuille lorsque le nœud est homogène, c'est-à-dire 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 la valeur seuil définie à l'avance par l'utilisateur. Pour contrôler la taille de l'arbre et éviter le sur-ajustement, une règle d'arrêt est utilisée pour élaguer les grands arbres contenant des nœuds terminaux « purs ». La méthode employée par CART est basée sur une mesure de coût-complexité. La complexité d'un arbre est définie comme suit :

$$R_{cp}(T) = R(T) + cp\, |\tilde{T}|, \quad (10)$$

où $R(T)$ est l'erreur totale de mesure (somme des erreurs de mesure de tous les nœuds terminaux), $|\tilde{T}|$ est le nombre de nœuds terminaux et $cp$ est un paramètre de pénalité positif appelé paramètre de complexité à choisir entre *0* et *+∞*. Quand la valeur du paramètre de complexité $cp$ augmente, la division pour laquelle l'amélioration de $R(T)$ est plus faible ($R(T) \leq cp$), est supprimée et deux nœuds sont regroupés. En outre, une séquence d'arbres emboîtés est construite. L'arbre final est le plus petit arbre pour lequel $R_{cp}(T)$ est minimal. Des techniques de la validation croisée peuvent être utilisées afin de déterminer la valeur du paramètre $cp$ optimal qui conduit à plus petite erreur de prédiction.

Cependant, l'instabilité des arbres est bien connue. Elle peut provenir d'un sur-ajustement de l'arbre aux données, mais aussi du choix arbitraire du seuil pour dichotomiser les prédicteurs continus à chaque nœud (Breiman, 1996; Dannegger, 2000). Une méthode de stabilisation bien connue pour améliorer la performance de prédiction d'un arbre unique consiste à agréger une famille d'arbres construits sur des échantillons *bootstrap* avec une sélection aléatoire des covariables à chaque nœud. Cette procédure, proposée par Breiman (2001) et appelée « *forêts aléatoires* », a été adaptée au domaine de la survie par Ishwaran et *al*. (2008). C'est cette dernière que nous présentons dans la suite.

### *2.3.1 Forêts aléatoires de survie (RSF)*

Pour stabiliser les arbres obtenus par l'algorithme CART, Breiman (1996) a proposé la méthode du *bagging* (pour *bootstrap aggregating*), qui consiste à agréger des arbres construits sur des échantillons bootstrap pour obtenir un estimateur robuste. La technique du bagging a été adaptée au contexte de la survie par Efron (1981) et Akritas (1986). Un autre moyen d'améliorer la stabilité des arbres est le *boosting*, développé par Freund and Schapire (1999). Comme le *bagging*, le *boosting* consiste à agréger une famille d'arbres. Chaque arbre est construit de façon itérative à partir d'un échantillon pondéré (un individu mal classé gagne en poids tandis qu'un individu bien classé en perd) et évalué en fonction de sa capacité à classer les individus. Considéré comme plus efficace que le *bagging*, le *boosting* est cependant limité quand les données sont trop bruitées.

Breiman (2001) a proposé une méthode de sélection de variables qui combine la méthode du *bagging* avec une sélection aléatoire d'un ensemble de covariables à chaque nœud de l'arbre. Cette méthodologie, appelée « forêts aléatoires », s'est révélée plus stable que les deux méthodes précédentes (Breiman, 1996). Cette méthodologie a été adaptée au domaine de la survie par Ishwaran et *al*. (2008) et est appelée « forêts aléatoires de survie » (*Random Survival Forests* - RSF). Des échantillons bootstrap sont tirés de l'échantillon original. On peut noter que chaque échantillon bootstrap exclut environ 37 % des individus, un ensemble appelé données *out-of-bag* (*OOB*) (nous y revenons sur la notion à la fin de la section). Pour chaque échantillon bootstrap *b* (*b* variant de 1 à *ntree*), un arbre de survie est construit : à chaque nœud de l'arbre, un sous-ensemble de covariables est sélectionné aléatoirement parmi l'ensemble des covariables. Le nœud est divisé en deux nœuds fils à partir des covariables sélectionnées en maximisant le critère de division. Le processus de division continue jusqu'à ce que chaque nœud terminal contienne un nombre minimum fixé d'événements non censurés avec des temps distincts. L'algorithme RSF calcule alors un estimateur global qui est la moyenne des fonctions de risque cumulées estimées pour chaque arbre par l'estimateur de Nelson-Aalen.

### *2.3.2 Erreur OOB*

Fixons une observation $(X_i,\ Y_i)$ de l'échantillon d'apprentissage $\mathcal{L}_n$ et considérons l'ensemble des arbres construits sur les échantillons bootstrap ne contenant pas cette observation, c'est-à-dire pour lesquels cette observation est "*Out-Of-Bag*". Nous agrégeons alors uniquement les prédictions de ces arbres pour fabriquer notre prédiction $\hat{Y}_i$ de $Y_i$. Après avoir fait cette opération pour toutes les données de $\mathcal{L}_n$, nous calculons alors l'erreur commise : l'*erreur quadratique moyenne* en régression $\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2\right)$, et la *proportion d'observations mal classées* en classification $\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\hat{Y}_i \neq Y_i}\right)$. Cette quantité est appelée erreur OOB du prédicteur forêt aléatoire (Genuer et Poggi, 2017).

### *2.4 Critères de comparaison*

Dans cette section, nous allons définir quelques critères nous permettant de juger de la performance des différentes méthodes (classification ou régression) utilisées pour prédire l'âge au décès des enfants au Bénin. Partant, nous allons utiliser la technique de l'échantillon test. Elle consiste à diviser le jeu de données en deux sous-échantillons, l'échantillon d'apprentissage (2/3 de l'échantillon total) et l'échantillon test (1/3 de l'échantillon total).
Ainsi, l'échantillon d'apprentissage servira à construire le modèle et celui test à le valider.

### *2.4.1 Critères d'évaluation pour les modèles de classification*

En apprentissage automatique supervisé, on se sert de la matrice de confusion pour évaluer les performances d'un modèle de classification. En effet, chaque ligne de la matrice correspond à une classe réelle alors que chaque colonne correspond à une classe prédite. La cellule ligne ***L***, colonne ***C*** contient le nombre d'éléments de la classe réelle ***L*** qui ont été prédites comme appartenant à la classe ***C***. Un des intérêts de cette matrice est qu'elle montre rapidement si un système de classification parvient à classifier correctement. Dans le cadre de ce papier, nous avons une matrice binaire ***2×2***. Autrement, l'implémentation des différentes méthodes de classification nous permettra de distinguer les enfants décédés à la naissance et ceux qui en ont survécu (y compris les enfants décédés par la suite). À cet effet, la matrice de confusion sera obtenue grâce à l'ensemble de données de l'échantillon test. Les valeurs qu'on y retrouve nous aideront par la suite à définir les critères d'évaluation suivants : la ***sensibilité***, la ***spécificité***, et le ***taux d'erreur de prédiction***. Le Tableau 1 présente l'architecture de la matrice de confusion utilisée dans cette recherche.

**Tableau 1 : Matrice de confusion**

| **Classe prédite** / **Classe réelle** | *Décédé à la naissance* | *Survécu à la naissance* |
|---|---|---|
| *Décédé à la naissance* | *Vrai positif (VP)* | *Faux négatif (FN)* |
| *Survécu à la naissance* | *Faux positif (FP)* | *Vrai négatif (VN)* |

La sensibilité représente la probabilité que le modèle prédise qu'un enfant est décédé à la naissance sachant qu'il l'est réellement [$\boldsymbol{Sensibilit}$é $:= \boldsymbol{VP} / (\boldsymbol{VP} + \boldsymbol{FN})$].
La spécificité, quant à elle, est la proportion des enfants qui ont survécu à la naissance, qui sont correctement classés par le modèle [$\boldsymbol{Sp}$é$\boldsymbol{cificit}$é $:= \boldsymbol{VN} / (\boldsymbol{VN} + \boldsymbol{FP})$].
Le taux d'erreur de prédiction est la proportion des modalités prédites qui diffèrent des modalités observées (c'est la somme des 2 valeurs non-diagonales de la matrice de confusion divisée par *n*) [$\boldsymbol{Taux\ d'erreur de pr}$é$\boldsymbol{diction} := (\boldsymbol{FP} + \boldsymbol{FN}) / (\boldsymbol{VP} + \boldsymbol{VN} + \boldsymbol{FP} + \boldsymbol{FN})$].

### *2.4.2 Critères d'évaluation pour les modèles de régression*

Les trois procédures suivantes ont été comparées : la procédure de Cox régularisé Elastic-net, la procédure de l'arbre de survie et la procédure de forêt aléatoire de survie. Nous avons regardé quelle était la meilleure méthode, c'est-à-dire celle associant un faible taux d'erreur de prédiction et permettant ainsi de quantifier la qualité prédictive des modèles finaux obtenus.
Pour les modèles de survie, le taux d'erreur de prédiction généralement utilisé est basé sur l'indice de concordance de Harrell, appelé l'indice *C* (Harrell et Davis, 1982). Afin de calculer cet indice, les temps de survie observés et les issues prédictives sont comparés selon l'algorithme suivant :

1. Former toutes les paires possibles entre les durées de survie des individus.
2. Éliminer les paires où la plus courte durée de survie est censurée, ainsi que celles où les deux durées de survie et les deux indicatrices de censure sont égales.
3. Pour chaque paire retenue, compter **1** si les prédictions sont concordantes avec les observations c'est-à-dire si la plus courte durée de survie correspond à la prédiction la plus pessimiste et compter **0.5** si les issues prédictives sont les mêmes. Sommer sur l'ensemble des paires retenues.
4. L'indice ***C*** est le rapport de cette somme sur les paires retenues.
5.

Le taux d'erreur de prédiction est défini comme **1 – C** et appartenant à **[0,1].** Il est à noter qu'une valeur de **0.5** du taux d'erreur de prédiction indique que le modèle ne fait pas mieux que le hasard.

### *2.4.3 Comparaison des covariables sélectionnées*

Nous allons comparer l'ensemble des covariables sélectionnées par les trois approches sur l'ensemble du jeu de données. Pour la procédure *Elastic-Net*, nous avons gardé les covariables dont les coefficients sont non nuls selon les résultats que fournit l'algorithme. Par contre, on peut classer les impacts des variables sélectionnées par les procédures des *arbres et forêts aléatoires* (de survie ou classification) sur l'événement d'intérêt grâce à la mesure de l'importance (notée VIM) de chaque covariable. Cependant, pour la procédure RSF, l'importance des variables est calculée par la différence entre l'erreur de prédiction obtenue

avec la forêt originale et l'erreur de prédiction obtenue en utilisant des affectations aléatoires à chaque nœud où la covariable considérée est rencontrée (Ishwaran et *al*., 2008).

### *2.4.4 Importance des variables pour une forêt aléatoire*

Fixons $j \in \{1, \dots, p\}$ et explicitons le calcul de l'indice d'importance de la variable $X_j$. Considérons un échantillon bootstrap ${\mathcal{L}^{\theta_l}}_n$ et l'échantillon $OOB_l$ associé, c'est-à-dire l'ensemble des observations qui n'apparaissent pas dans ${\mathcal{L}^{\theta_l}}_n$. Calculons $errOOB_l$, l'erreur commise sur $OOB_l$ par l'arbre construit sur ${\mathcal{L}^{\theta_l}}_n$ (erreur quadratique moyenne en régression, proportion de mal classés en classification). Permutons alors aléatoirement les valeurs de la $j^{ième}$variable dans l'échantillon $OOB_l$. Ceci donne un échantillon perturbé, noté ${\widetilde{OOB}^j}_l$. Calculons enfin ${\widetilde{errOOB}^j}_l$, l'erreur commise sur l'échantillon ${\widetilde{OOB}^j}_l$. Nous effectuons ces opérations pour tous les échantillons bootstrap. L'importance de la variable $X_j$, $VIM(X_j)$, est définie par la différence entre l'erreur moyenne d'un arbre sur l'échantillon OOB perturbé et celle sur l'échantillon *OOB* (Genuer et Poggi, 2017) :

$$\boldsymbol{VIM(X_j) = \frac{1}{q}\sum_{l=1}^{q}\left({\widetilde{\boldsymbol{errOOB}}^{\boldsymbol{j}}}_{\boldsymbol{l}} - \ \boldsymbol{errOOB_l}\right)} \qquad (\mathbf{11})$$

Ainsi, plus les permutations aléatoires de la $j^{ième}$ variable engendrent une forte augmentation de l'erreur, plus la variable est importante. À l'inverse, si les permutations n'ont quasiment aucun effet sur l'erreur (voire même la diminuent ce qui fait que ***VIM*** peut être légèrement négative), la variable est considérée comme très peu importante.

## 3. Résultats et discussion

### *3.1 Performance des différentes methods de classification*

Cette section a pour but, la comparaison de différentes méthodes d'apprentissage à classifier les enfants dès la naissance au Bénin, au regard du phénomène de décès. On dispose d'un échantillon de taille *13 589* comprenant *20* covariables et la variable réponse ***Y*** qui prend pour valeur *1* si l'enfant est décédé à la naissance, *0* sinon. Nous séparons l'échantillon en un échantillon d'apprentissage de taille *10 191* pour construire les modèles et un échantillon test de taille *3 398* pour comparer leurs performances.

#### *3.1.1 Comparaison des modèles de classification*

Le Tableau 2 donne les résultats des critères de comparaison pour les modèles de type classification. Nous comparons les résultats obtenus en fonction des critères définis dans la section *2.4.1*. Nous avons donc testé l'échantillon avec les modèles de régression logistique régularisée, d'arbre de décision et de forêts aléatoires. Il n'y a pas présence de sur-apprentissage, car on remarque que la différence entre les résultats du fichier d'apprentissage et celui du test n'est pas significative.

La procédure de forêt aléatoire appliquée à nos données donne le plus faible taux d'erreur de prédiction alors qu'un simple arbre de classification en fournit le plus mauvais. Il faut cependant remarquer que, quelle que soit la procédure considérée, nous avons obtenu de faibles taux d'erreur de prédiction. Par ailleurs, seule la procédure de forêt aléatoire fournit une sensibilité relativement élevée. Autrement, c'est le seul prédicteur qui discrimine le mieux les enfants à la naissance face au phénomène de décès : avec une probabilité de *63.63%,* il prédit les enfants décédés à la naissance alors qu'ils le sont réellement. En ce qui concerne la spécificité, tous les prédicteurs prédisent au moins avec *97%* de chances les enfants ayant survécu à la naissance.

**Tableau 2 : Récapitulatif des critères de comparaison des modèles de classification.**

| ***Prédicteurs*** | ***Sensibilité*** | ***Spécificité*** | ***Taux d'erreur de prédiction*** |
|---|---|---|---|
| ***Régression logistique régularisée Elastic-Net*** | 0.03914591 | 0.9997982 | 0.02669022 |
| ***Arbre décision classification*** | 0.07777778 | 0.9960701 | 0.02825191 |
| ***Forêt aléatoire classification*** | 0.6363636 | 0.9754945 | 0.0256033 |

#### *3.1.2 Importance et selection des variables*

Sur les données des enfants de moins de cinq ans au Bénin, le classement des variables par ordre décroissant d'importance pour les procédures d'arbre de décision et forêt aléatoire conduit à la Figure 1. L'arbre de classification identifie 12 covariables importantes (les 8 autres ayant une importance nulle) pour discerner les enfants décédés à la naissance de ceux qui ne le sont pas alors que la forêt aléatoire sélectionne 15 covariables. Pour ces deux procédures, la première covariable discriminante et de surcroit distinctive est la même : *enfant_moins_5ans*. En nous restreignant aux huit premières variables importantes dans les deux procédures qui semblent se dégager, on peut remarquer que la procédure CART sélectionne également six autres covariables parmi les variables les plus importantes dans la procédure de forêt aléatoire : *poids, ethnie, gemellite, religion,*

*taille_naissance*. Aussi, *gestation, source_eau_bue* et *inter_genesique, rang* se révèlent importantes pour CART et forêt aléatoire respectivement. Pour ce qui est de la régression logistique régularisée, le modèle à base des données de l'échantillon d'apprentissage retient douze (12) des vingt (20) covariables : *assist_accouc, duree_alaitement, enfant_moins_5ans, ethnie, gemellite, gestation, grossesse_souhait, inter_genesique, poids, religion, sexe_enfant, taille_naissance*.
Par ailleurs, les variables classées importantes dans les deux procédures sont sélectionnées par la régression logistique régularisée Elastic-Net, à l'exception de *sexe_enfant*, (c'est-à-dire ayant des coefficients non nuls) et donc sont discriminantes pour prédire le statut d'un enfant à la naissance.

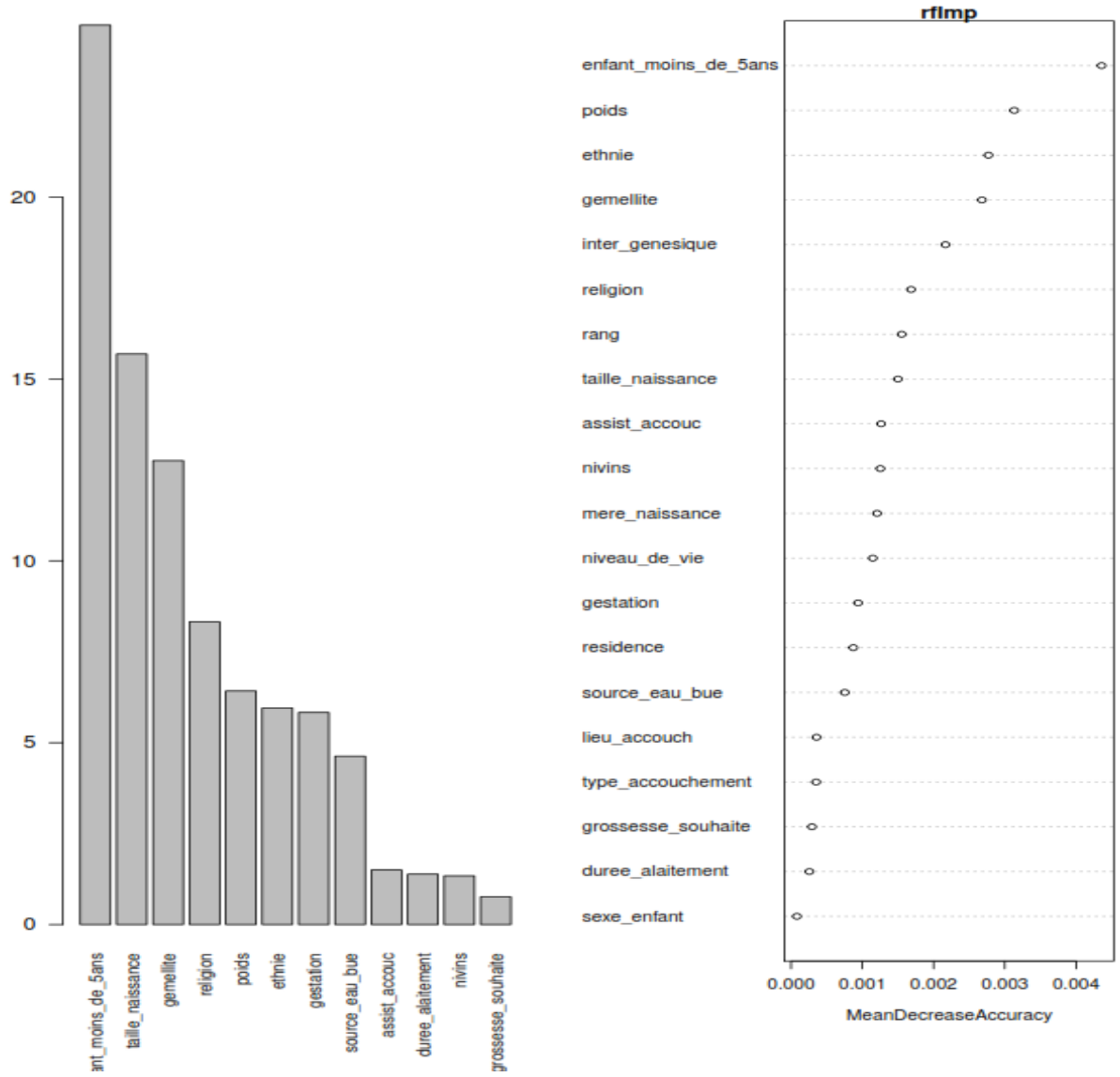



**Figure 1 : Importance des variables au sens de CART pour l'arbre optimal (à gauche) et Importance des variables selon la méthode de forêt aléatoire (à droite).**

### *3.2 Performance des différentes méthodes de régression*

L'objet de cette section est de comparer les différentes approches basées sur le modèle Cox permettant de prédire l'âge au décès des enfants au Bénin. En prélude, les données d'enquête indiquent que dans la population des enfants ayant un âge strictement positif, un total de *12384* enfants sur *12928* sont encore en vie à la date de l'enquête, ce qui conduit à un taux de censure de *95.79%.* Comme dans le cas des modèles de classifications, *20* covariables sont mis en jeu et la variable réponse est un objet de survie définit concomitamment avec la durée réelle de survie d'un enfant et une indicatrice qui vaut *0* ou *1* selon que l'observation correspond à une censure ou non. Un échantillon d'apprentissage de taille *9696* est utilisé pour entraîner les modèles alors que celui de taille *3232* a permis la validation et la comparaison des différents modèles entraînés.

### *3.2.1 Comparaison des modèles de régression*

Le Tableau 3 présentant les taux d'erreur de prédiction des trois procédures nous permet d'évaluer la performance prédictive de ces dernières. Nous pouvons observer tout d'abord que, les différents taux obtenus sont relativement élevés quel que soit la procédure considérée, contrairement au cas des modèles de classification. Toujours à l'opposé des modèles de classification, la procédure de l'arbre de survie se révèle meilleure que la forêt aléatoire de survie, ainsi que la régression régularisée Elastic-Net de Cox, avec nos données.

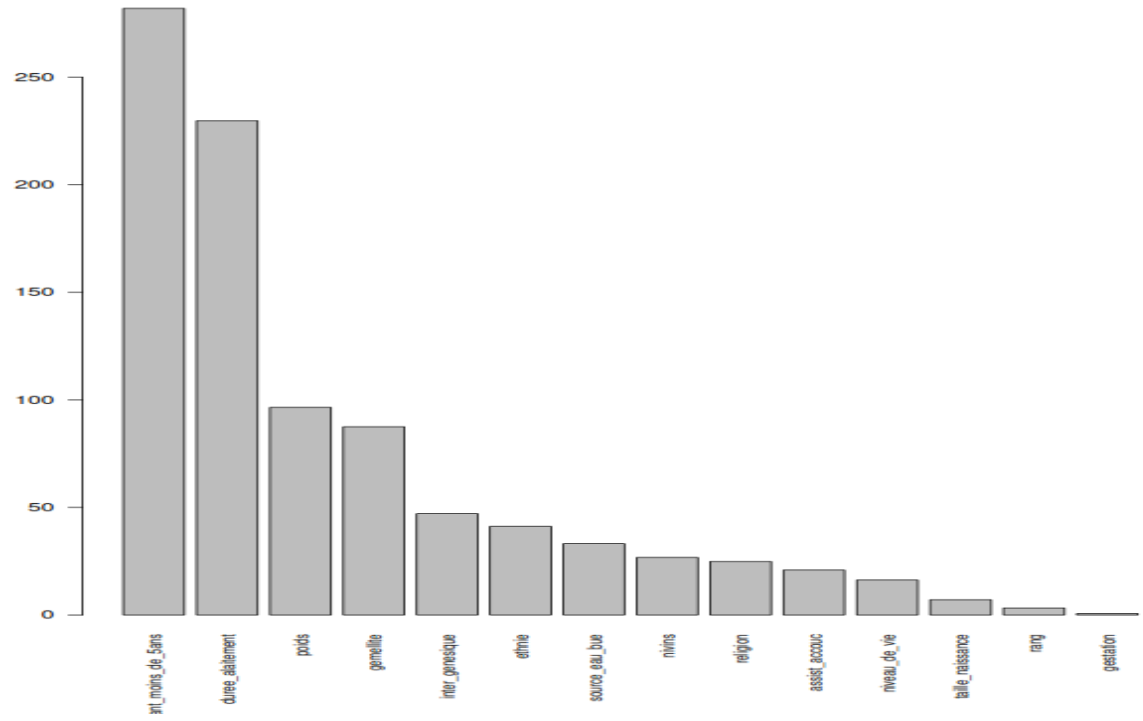
**Tableau 3 : Taux d'erreur de prédiction pour les procédures Cox-régularisée Elastic-Net, arbre de survie et RSF.**

| **Prédicteurs** | **Cox-régularisé** | **Arbre de survie** | **Forêt de survie** |
|---|---|---|---|
| **Taux d'erreur de prédiction** | 0.3379972 | 0.1937975 | 0.275743 |

### *3.2.2 Importance et selection des variables*

À l'instar de la section *3.1.2*, nous présentons les covariables classées par ordre décroissant d'importance pour les procédures d'arbre de survie et forêt aléatoire de survie, par la Figure 2. L'arbre de survie indique *14* variables ayant une importance non nulle et donc passibles de prédire l'âge au décès des enfants au Bénin. En outre, en comparant les deux procédures, on constate une similitude entre les covariables sélectionnées parmi les sept plus importantes. Notamment, en plus de la covariable la plus importante pour les deux procédures (*enfant_moins_5ans*), on retrouve dans tous les modèles finaux, les covariables suivantes : *duree_alaitement, poids, gemellite, inter_genesique, ethnie*.

Nonobstant, le modèle de Cox régularisé Elastic-Net conserve la quasi-totalité des covariables (à l'exception du *lieu d'accouchement*) pour la prédiction, et donc une similarité avec les autres procédures s'avère indiscutable.



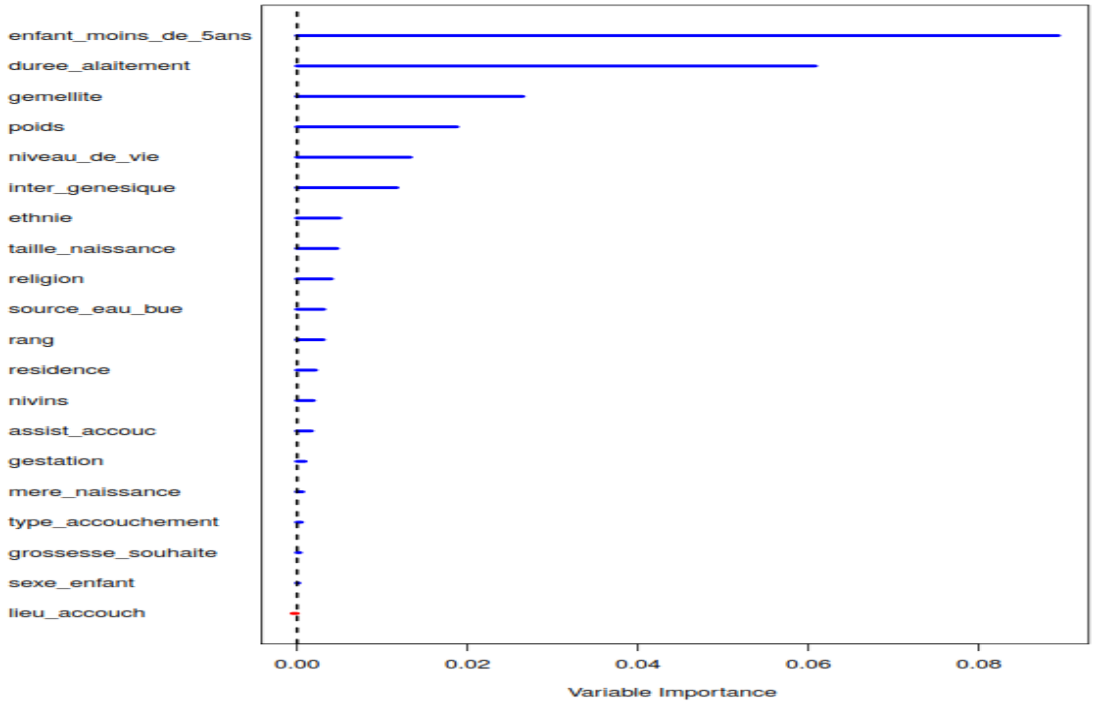



**Figure 2 : Importance des variables au sens de CART pour l'arbre optimal de survie (en haut) et selon la méthode de forêt aléatoire de survie (en bas).**

***3.3 Discussion***

Comparé aux modèles de classification, les taux d'erreur de prédiction obtenus à partir des modèles de régression de Cox sont largement plus élevés, suggérant la complexité et la difficulté d'étudier l'âge de décès des enfants au Bénin et de trouver un modèle de prédiction de très bonne qualité. De plus, la taille des deux jeux de données (c'est-à-dire pour les modèles de classification et de régression) est différente, ce qui peut entraîner les écarts importants dans les résultats. Cependant, nous trouvons que la procédure de forêt aléatoire donne le meilleur modèle de classification alors que l'arbre de survie est de loin le meilleur modèle prédictif de l'âge au décès des enfants au Bénin (pour les modèles de régression de Cox). Notons également que ces deux procédures permettent de présenter les variables les plus pertinentes pour expliquer les durées de survie des enfants ou les classer par groupe dès leur naissance.

Concernant les modèles de classification, bien que la procédure de forêt aléatoire ait donné le plus bas taux d'erreur de prédiction, les autres procédures (Elastic-Net et arbre de classification) ont autant de faible taux. De ce fait, nous pensons que nous avons bien fait de mettre à contribution la régression logistique pour répertorier les enfants qui décèdent à la naissance. Au demeurant, la pénalisation Elastic-Net nous a permis de garder un nombre plus restreint de variables (au nombre de 12, au lieu de 20 initialement) pour prédire les classes à affecter aux enfants dès leur naissance et surtout offre un cadre plus souple d'interprétation, comparativement à la forêt. Comparé aux études à caractère explicatif (Akoto and Hill, 1988; Ambapour and Moussana, 2008; Dansou, 2016) où certaines caractéristiques ont été habituellement identifiées comme offrant un climat favorable au risque de mortalité des enfants dans les contextes camerounais, sénégalais, tchadiens et béninois, notre étude suggère que l'âge de la mère à l'accouchement, le niveau d'instruction de la mère, le niveau de vie du ménage, le rang de naissance, ainsi que le milieu de résidence ne contribuent pas pour prédire efficacement la classe d'un nouveau-né face au décès au Bénin entre 2012 et 2018. Souvenons-nous que les données utilisées proviennent d'une enquête transversale rétrospective et que si l'enquête apporte une information rétrospective sur le statut de l'enfant, on ignore en revanche quelle était par exemple la situation socio-économique des parents, le niveau d'instruction de la mère, le milieu où résident les parents lors des différentes étapes de l'itinéraire de l'enfant.

Contrairement à ce que révèlent la littérature et certaines études comme l'excellent article de Walschaerts et *al*. (2011) sur la sélection stable des variables pour les données censurées à droite, dans le cadre de notre recherche, l'arbre de survie offre un

meilleur taux de prédiction que la forêt aléatoire de survie, qui, à son tour, est meilleur que la régression Cox régularisée, pour prédire l'âge au décès des enfants. En effet, le modèle de Cox et la méthode de RSF ont été déjà comparés par Omurlu et *al*. (2009) à l'aide de l'indice de Harrel. Sur la base de 1000 simulations de Monte Carlo, ils ont montré que le modèle de Cox a la meilleure performance prévisionnelle quelle que soit la taille de l'échantillon (pour n = 50, 100, 250, 500). Toutefois, sur un jeu de données réel sur le cancer du sein, ils ont constaté que la méthode RSF avait le plus faible taux d'erreur de prédiction. Ces résultats contradictoires ne sont pas surprenants si l'on considère le fait que les données simulées ont été générées à partir d'un modèle de Cox. En outre, la procédure de RSF est facile à utiliser et ne nécessite pas le choix de valeurs de réglage comme le font les procédures Elastic-Net et les arbres de survie. Cependant, même si les variables choisies sont identifiées et triées par leur importance, aucun arbre final n'est fourni et les résultats de RSF restent une "boîte noire", difficile à interpréter. Au contraire, la procédure d'arbre de survie offre un cadre d'interprétation visuelle à travers l'arbre optimal fourni, dont les variables constitutives sont toutes identifiées comme importantes pour la prédiction de l'âge au décès des enfants au Bénin. Du reste, les variables importantes sélectionnées par la procédure des arbres de survie, sont pour la plupart celles qui ont un effet significatif dans nombreuses études explicatives de la mortalité des enfants de moins de cinq ans (Rakotondrabe, 2004; Beninguisse, 1993 ; Mboko Ibara, 2009; Hamadou Daouda, 2012), respectivement au Madagascar, Cameroun, Congo et le Niger. La principale limite de notre étude réside dans les données que nous avons utilisées. Les limites de ce type de données sont de deux ordres. Premièrement, ce sont des données qui proviennent d'une enquête transversale rétrospective qui font appel à la mémoire. Cette information peut donc être considérée comme fragile. Deuxièmement, l'absence de biographie (sur le parcours et la garde des enfants d'une part et d'autre part sur le parcours de la situation des parents lors des différentes étapes de l'itinéraire de l'enfant) pour mieux cerner les variables prédictives de la mortalité des enfants, peut expliquer les mauvais taux de prédictions obtenus. En effet, pour modéliser l'âge au décès de l'enfant, nous avons adopté donc une hypothèse forte : celle qui consiste à étudier le lien entre cet âge et les principaux évènements de la vie familiale et professionnelle des parents en référence à la situation actuelle (ou date de l'enquête).

**4. Conclusion**

Dans l'étude de la mortalité des enfants de moins de cinq ans, les auteurs ont pour la plupart recours aux modèles de régression logistique qui génèrent des estimateurs non efficaces, quoique non biaisés. À cet effet, cette recherche s'est assignée l'objectif de construire un modèle de prédiction qui explique au mieux la mortalité des enfants de moins de cinq ans dans le contexte béninois. Pour aborder la problématique principale de ce mémoire, nous avons utilisé les données de la cinquième édition de l'Enquête Démographique et de Santé réalisée par l'INSAE en 2017-2018. Du fait de la configuration de ces données, où, l'on enregistre les enfants décédés à la naissance, nous nous sommes appuyés sur les méthodes statistiques classiques existantes pour proposer le modèle Cox-logistique. Ce modèle permet d'abord de spécifier les enfants décédés à la naissance, de ceux qui ont survécu à travers la régression logistique alors que pour ces derniers (enfants ayant survécu à la naissance), on modélise leur âge au décès par le modèle Cox (les enfants qui sont encore en vie à la date de l'enquête sont censurés). Nous comparons les performances prédictives du modèle Cox-logistique avec celles obtenues à partir de la méthodologie des forêts aléatoires (respectivement de survie) réputée pour donner la plus petite erreur de prédiction, mais difficile à interpréter pour les non-statisticiens.

Pour conclure sur les résultats obtenus pour les modèles de classification, l'objectif est raisonnement bien atteint. La méthode de sélection basée sur la pénalisation Elastic-Net appliquée à la régression logistique fournit des qualités prédictives très comparables à celles obtenues par la méthode de forêt aléatoire de classification, en association avec un modèle final facile à interpréter pour les statisticiens et démographes, plus particulièrement dans l'étude de la mortalité des enfants. De même, la procédure des arbres de classification offre une alternative à la procédure de forêt aléatoire en termes de prédiction et permet également d'identifier les interactions les plus pertinentes entre les covariables. Par ailleurs, la comparaison entre ces approches sur le jeu de données des enfants au Bénin fournit des indications précieuses sur la stabilité de la sélection des variables.

Par contre, la prédiction s'avère moins bonne pour les modèles basés sur la régression de Cox. Néanmoins, vu la faible prédictivité des modèles envisagés, les variables collectées sont probablement trop faiblement prédictives pour être identifiées de manière stable par ces différentes approches, conduisant à de grandes variations dans les taux de prédiction observés. Ainsi, la mauvaise performance d'une approche n'est pas forcément liée à l'inadéquation du modèle choisi. Dans ce cas, les approches retenues incitent à la prudence quant à la portée des conclusions dans l'étude de la mortalité des enfants. On retient tout de même, les arbres de survie pour prédire l'âge au décès des enfants dans le contexte béninois entre 2012 et 2018.

En définitif, les résultats de cette étude préconisent que les procédures basées sur le modèle de Cox et les arbres de survie ont un rôle complémentaire à jouer afin d'identifier les covariables les plus pertinentes et fournir aux démographes un modèle stable et fiable pour étudier la mortalité des enfants. Chaque procédure étudiée montre un intérêt, soit dans ses résultats en

termes de prédiction, soit dans le choix des variables obtenues dans le modèle final par rapport à la procédure RSF qui est difficilement interprétable.

Nous terminons ce manuscrit en donnant quelques perspectives et prolongements de ce travail de recherche.

1. Nous comptons très prochainement comparer les performances des différentes méthodes utilisées et d'autres éventuellement telles que les réseaux de neurones sur des données simulées, en générant en particulier des covariables qui n'ont aucun effet sur la survie. Ceci permettra probablement de mettre en évidence la pertinence du modèle Cox-logistique dans l'étude de la mortalité des enfants.

2. Dans les techniques de régularisation et de sélection de variables via la vraisemblance pénalisée, la complexité du modèle et le taux de rétrécissement appliqué aux coefficients de régression sont fortement liés au choix des paramètres de régularisation. Il est remarqué que l'utilisation de la validation croisée pour sélectionner les paramètres de régularisation optimaux donne des modèles qui contiennent beaucoup de variables parasites. Nous allons nous intéresser à ce problème dans nos travaux futurs. Nous allons chercher sous quelles conditions les critères d'information AIC ou BIC peuvent être de bonnes alternatives pour choisir les paramètres de régularisation optimaux et par conséquent les variables et seulement les variables pertinentes.

**Réferences**


[1] Akoto, E. M. et Hill, A. G. (1988). Morbidité, malnutrition et mortalité des enfants. *Population et société au sud du Sahara*, pages 309–334. Sous la direction de Tabutin D., Paris, le Harmattan.

[2] Akritas, M. G. (1986). Bootstrapping the Kaplan-Meier estimator. *Journal of the American Statistical Association*, *81 :1032–1038*.

[3] Aly, Y. et Grabowsky, R. (1990). Education and child mortality in Egypt. *World Development*, *5 :733–742*.

[4] Ambapour, S. et Moussana, H. A. (2008). *Pauvreté et santé nutritionnelle de l'enfant au Congo*, *Document de Travail n° 15/2008*. Brazzaville. http : /www.cnsee.org.

[5] Barbieri, M. (1991). Les déterminants de la mortalité des enfants dans le tiers monde. *Les dossiers du CEPED*, *18*.

[6] Beninguisse, G. (1993). *Approvisionnement en eau et assainissement : effets sur la morbidité et la mortalité des enfants par maladies diarrhéiques, le cas du cameroun. Master's thesis, IFORD,* Yaoundé.

[7] Boco, A. G. and Bignami, S. (2008). Religions et survie des enfants de 0-5 ans en Afrique au sud du Sahara : l'exemple du Bénin. In *Démographie et cultures*, pages *1119–1137*. Actes des colloques de l'AIDELF. Université de Monréal, Département de Démographie C.P. 6128 (Québec).

[8] Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24 :123–140*.

[9] Breiman, L. (2001). Random Forests. *Machine Learning*, *45(1) :5–32*.

[10] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall, New York.

[11] Chesneau, C. (2015). *Modèles de régression*. Université de Caen Basse-Normandie, France. http : //www.math.unicaen.fr/ chesneau/.

[12] Czepiel, A.-S. (2018). *Maximum likelihood estimation of logistic regression models : Theory and implementation*. page 23. https ://czep.net/stat/mlelr.pdf.

[13] Dannegger, F. (2000). Tree stability diagnostics and some remedies for instability. *Statistics in Medicine, 19 :475–491*.

[14] Dansou, J. (2016). Influence des facteurs socioculturels sur la survie des enfants de moins de cinq ans au Bénin. In PENNEC Sophie, GIRARD Chantal, S. J.-P., editor, *Trajectoires et âges de la vie*. XVIIIe colloque international de l'AIDELF. Association internationale des démographes de langue française, ISBN : 978-2-9521220-5-4.

[15] Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, *76 :312–319*.

[16] Freund, Y. et Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, *14 :771–780*.

[17] Genuer, R. et Poggi, J.-M. (2017). *Arbres CART et forêts aléatoires, importance et sélection de variables*. https ://hal.archives-ouvertes.fr/hal-01387654v2.

[18] Greenland, S. (2008). Invited commentary : variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*, *167 :523–529*.

[19] Grouwels, Y. and Braekers, R. (2011). Zero-inflated semi-parametric Cox's regression model for left-censored survival data. *In 7th Conference on Statistical Computation and Complex System*, Padua, Italy.

[20] Hamadou Daouda, Y. (2012). Déterminants de la mortalité infantile et infanto-juvénile et la pauvreté au Niger. *Revue d'Economie Théorique et Appliquée*, 2(1) :23–47.

[21] Harrell, F. E. and Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, *69 :635–640*.

[22] Hill, C., Com-Nougué, C., Kramar, A., Moreau, T., O'Quigley, J., Senoussi, R., et Chastang, C. (1999). *Analyse statistique des données de survie*. Statistique en biologie et en médecine, médecine sciences publications edition. ISBN : 2257123107.

[23] INSAE (2016). *Enquête par grappes à indicateurs multiples 2014*, *Rapport final*. Technical report, Cotonou, Bénin : Institut National de la Statistique et de l'Analyse Économique. www.insae-bj.org.

[24] INSAE et ICF (2019). *Enquête Démographique et de Santé au Bénin, 2017-2018*. Technical report, Cotonou, Bénin et Rockville, Maryland, USA : Institut National de la Statistique et de l'Analyse Économique et ICF. www.insae-bj.org.

[25] INSAE et Macro-International-Inc (2007). *Enquête Démographique et de Santé (EDSBIII) – Bénin 2006*. Technical report, Calverton, Maryland, USA : Institut National de la Statistique et de l'Analyse Économique et Macro International Inc. www.insae-bj.org.

[26] INSAE et ORC-Macro (2002). *Enquête Démographique et de Santé au Bénin 2001*. Technical report, Calverton, Maryland, USA : Institut National de la Statistique et de l'Analyse Économique et ORC Macro. www.insae-bj.org.
[27] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random Survival Forests. *The annals of applied statistics*, pages *841–860*.
[28] Klein, J.-P. et Moeschberger, M.-L. (1997). *Surviral Analysis : Techniques for censored and truncated data*. Springer, New York.
[29] Lelièvre, E. (2010). *L'approche biographique des trajectoires individuelles. Analyse statistique des données longitudinales*.
[30] Mboko Ibara, S. (2009). *Influence des conditions d'existence sur la mortalité de moins de 5 ans en Afrique Centrale : Cas du Cameroun et du Congo*. Miméo. Brazzaville.
[31] Melligton, N. and Cameron, L. (1999). *Female education and child mortality in indonesia. Melbourne, Research Paper 693*. The University of Melbourne.
[32] OMS (2005). *Le rapport sur la santé dans le monde, 2005 - Donnons la chance à chaque mère et à chaque enfant.* Technical report, Organisation Mondiale de la Santé.
[33] Omurlu, I. K., Ture, M., et Tokatli, F. (2009). The comparisons of random survival forests and cox regression analysis with simulation and an application related to breast cancer. *Expert Systems with Applications*, *36 :8582–8588*.
[34] Rakotondrabe (2004). *Statut de la femme, prise de décision et santé des enfants à Madagascar*. PhD thesis, IFORD, Yaoundé.
[35] Ray, J. (1988). Données censurées et modèles de durée. *Recherche et Applications en Marketing*, 3(2) :*77–88*.
[36] Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Stat Med*, 16 : *385–395*.
[37] Walschaerts, M., Leconte, E., and Besse, P. (2011). *Stable variable selection for right censored data : comparison of methods*. https : //hal.archives-ouvertes.fr/hal-00681677.