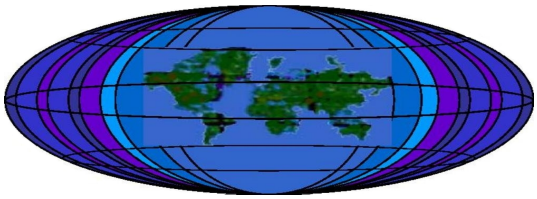
# Contributions de l'apprentissage statistique aux méthodes GLMM et LASSO: Application à la modélisation statistique de la morbidité liée au paludisme à Tori-Bossito (Bénin)

Bienvenue Kouwaye

# Contributions de l'apprentissage statistique aux méthodes GLMM et LASSO : Application à la modélisation statistique de la morbidité liée au paludisme à Tori-Bossito (Bénin)



Chaire Internationale en Physique Mathématiques et Applications
(ICMPA-UNESCO Chair)

**Thèse de Doctorat de l'Université d'Abomey-Calavi**

**Option :** Statistique Mathématique

**Spécialité :** Statistique appliquée au vivant

Présentée par


**Bienvenue Tanankpon Kouwayè**

Faculté des Sciences et Techniques (FAST)
Université d'Abomey-Calavi (UAC)
Cotonou, République du Bénin


Co-Directeurs de Thèse :
**Dr Gilles Cottrell (IRD-France)**
**Professeur Noël FONTON, Université d'Abomey-Calavi (Bénin)**
**Professeur Fabrice ROSSI, Université Paris 1 (France)**


**Université d'Abomey-Calavi, BENIN**
**Chaire Internationale en Physique Mathématique et Applications**
**(CIPMA - CHAIRE UNESCO )**

**PhD** n °006296 −2017/CIPMA/FAST/UAC

# Contributions de l'apprentissage statistique aux méthodes GLMM et LASSO : Application à la modélisation statistique de la morbidité liée au paludisme à Tori-Bossito (Bénin)

**Thèse de Doctorat de l'Université d'Abomey-Calavi**
**Option :** Statistique Mathématique
**Spécialité :** Statistique appliquée au vivant

Présentée par
**Bienvenue Tanankpon Kouwayè**

**Jury**

| | | |
|---|---|---|
| **Président** | Prof Mahouton Norbert Hounkonnou | (UAC-Bénin) |
| **Rapporteur** | Prof Judicael Déguénon | (UAC-Bénin) |
| **Rapporteur** | Dr André Garcia | (IRD-France) |
| **Examinateur** | Prof adanhounme Villevo | (UAC-Bénin) |
| **Co-Directeur** | Dr Gilles Cottrell | (IRD-France) |
| **Co-Directeur** | Prof Noël Fonton | (UAC-Bénin) |
| **Co-Directeur** | Prof Fabrice Rossi | (Paris 1-France) |

Cotonou, République du Bénin, 16 mars 2018

# Remerciements

Mes remerciements vont tout d'abord à l'endroit de mes directeurs de thèse, le Docteur Gilles COTTRELL, le Professeur Noël FONTON et le Professeur Fabrice ROSSI. Je leur exprime ma gratitude.

Je voudrais remercier particulèrement le Professeur Mahouton Norbert HOUNKONNOU qui a accepté de m'ouvrir les portes de la CIPMA pour l'achèvement du Master de Statistique appliqueé au vivant et de continuer en thèse à école Doctorale de la CIPMA. Sa rigueur scientifique, son soutien moral, matériel et psychologique ont été déterminants.

Je voudrais remercier spécialement le Docteur Gilles Cottrell pour tout ce qu'il a fait pour moi depuis le début du Master jusqu'à la fin de cette thèse.

Je remercie le Professeur Ezinvi BALOÏTCHA, Secrétaire scientifique de la CIPMA, pour tout ce qu'il a fait pour moi.

Je remercie le Professeur Simplice DOSSOU-GBETE, pour son soutien, ses conseils et pour sa contribution dans la thèse.

Je remercie les rapporteurs de ma thèse.

Je remercie le Professeur Jean Marc BARDET.

Je remercie le Professeur Elisabeth GASSIAT.

Je remercier le Professeur Marie COTTRELL, ancienne directrice du laboratoire SAMM.

Je remercie le Professeur Léonard TODJIHOUNDE pour son soutien.

Je remercie le Professeur Hippolyte HOUNNON pour son soutien.

Je remercie les membres du laboratoire SAMM pour leur accueil et leur soutien.

Je remercie les membres du loboratoire IRD/UMR216/CERPAGE de Cotonou.

Je remercie le service de coopération et d'action culturelle (SCAC) de l'Ambassade de France, l'Agence Universitaire de la Francophonie (AUF), le réseau STAFAV à travers le projet EDULINK, pour leur contribution dans le financement de cette thèse.

Je tiens à exprimer toute ma gratitude à Madame Marie Josephe Granier (MamiJO) pour toute son assistance et son soutien. Tu m'as accordé ta confiance en acceptant de m'héberger chez toi.

Je remercie William Kengne pour son aide et son soutien tout au long de mon séjour chez MamiJO.

Je voudrais remercier ma mère Mariama pour tous les efforts qu'elle a consenti pour nous ses enfants. Merci maman.

Je voudrais remercier mon épouse Lisette pour tous ses sacrifices. Merci Chérie.

Je voudrais remercier ma fille Ismay pour tout ce qu'elle endure, Dieu veillera sur toi ma petite chérie.

Je voudrais remercier mes frères et sœurs Yvette, Denise, Béatrice, Albertine, Adèle, Rosmonde, Ahissi, Elie et Dona.

Je voudrais remercier mes cousins Laurent Houngnibo et René Gbèwézoun

Je tiens à remercier tous les étudiants, les enseignants et les membres de l'administration de CIPMA.

Je tiens à remercier tous les amis et toutes les amies, tous les frères et sœurs ainsi que toutes les personnes (sans en oublier aucune) qui m'ont encouragé et soutenu.

# Table des figures

# Liste des tableaux

# Liste des algorithmes

# Table des matières

# Abstract

**GLMM, Lasso, variables selection, and prediction : Applications to malaria data of Tori-Bossito (Benin)**

The subject of this Thesis is the identification of environmental factors that may explain the variability of anopheline density at village and home scale and the determination malaria risk exposure in the study area. We consider these problems as variables selection and prediction problems in epidemiology context. Then, the main objective is the selection of an optimal subset of variables for the prediction of malaria risk exposure in the study area and also in an other area where the entomological data are not available. In the first part of the Thesis, we propose one method based on GLMM algorithm combined with a backward process for variables selection. Random effects are used at each hierarchy level of data for taking account the possible correlation because of the hierarchical structure of the data. This method provides an optimal subset of variables for prediction of malaria risk. But algorithm do not converge when some explanatory variables are too correlated or if data have a particular structure. For overcoming this, we propose in the second part an automatic machine learning method. We have generated automatically interactions between variables. The variables selection is performed by this automatic machine learning method based on Lasso and stratified two levels cross validation. Selected variables are debiased while the prediction is generated by simple GLM (Generalized linear model). The results of this method reveal to be qualitatively better, at selection, the prediction, and the CPU time point of view than those obtained in the first part. Finally, the best subset of prediction contains : Season ; interaction between Mean rainfall and openings ; interaction between Rainy days before mission and Number of inhabitants ; interaction between Rainy days during the mission and Vegetation.

**Keywords :** Malaria, variables selection, prediction, cross validation.

# Résumé

L'objectif principal de cette thèse est la détermination des facteurs environnementaux pouvant expliquer la variabilité de la densité anophélienne et la prédiction du risque d'exposition au vecteur palustre au niveau village et maison de la zone de Tori-Bossito. Dans ce travail, nous avons considéré ces deux problèmes comme des problèmes de sélection de variables et de prédiction dans le contexte épidémiologique. L'objectif principal est alors de sélectionner un sous ensemble optimal de variables pertinentes pour la prédiction du risque d'exposition au vecteur palustre dans le milieu d'étude ainsi que dans un autre milieu où les données entomologiques ne sont pas disponibles. Dans la première partie de cette Thèse, nous avons proposé une méthode basée sur un algorithme de type GLMM combiné avec une sélection de variables de type *backward*. Des effets aléatoires ont été mis au niveau de chaque hiérarchie des données pour prendre en compte les possibles corrélations à cause de la structure hiérarchique des données. Les résultats ont permis de déterminer un sous ensemble optimal pour la prédiction du risque palustre. Ces algorithmes deviennent non convergents lorsque les données possèdent une structure particulière ou sont très correlées. Dans la seconde partie de cette Thèse, nous avons donc proposé une méthode d'apprentissage machine automatique. Cette méthode combine le GLM, le Lasso et une validation croisée stratifiée à deux niveaux. Nous avons généré automatiquement les interactions entre les variables. La sélection de variables a été faite par la combinaison GLM, Lasso et validation croisée. Les variables sélectionnées sont débiaisées par le GLM pour faire de la prédiction. Les résultats obtenus montrent que les pré-traitements effectués par les experts sur les données peuvent être surmontés. Aussi, ces résultats montrent une amélioration au niveau de la sélection, de la sparsité du sous ensemble optimal pour la prédiction, la qualité des prédictions et le temps CPU d'exécution des calculs.

Finalement, le meilleur sous ensemble de prédiction comporte Saison, interaction entre Quantité moyenne de pluie et Ouvertures, interaction entre Jours de pluie avant la mission et Nombre d'habitants, interaction entre Jours de pluie pendant la mission et Végétation.
**Mots-clefs :** Paludisme, sélection de variables, prédiction, validation croisée.

# Introduction

## Contexte général

Le paludisme est une maladie d'origine parasitaire causée par la piqûre de la femelle d'un moustique appelé anophèle. Ce moustique infecté par le parasite, devient un potentiel vecteur de transmission. La forme la plus répandue en Afrique subsaharienne de ce parasite est le *Plasmodium falciparum*. Cette espèce de parasite est la cause des formes les plus graves du paludisme. Cette maladie constitue un important problème de santé publique dans le monde. Selon le rapport 2017 de l'OMS sur le paludisme dans le monde, en 2016, le paludisme est considéré endémique dans 91 pays et territoires ; le nombre de cas de paludisme et de décès associés a été respectivement estimé à 212 millions et à 429 000 au niveau mondial en 2015. Plus de 40% des habitants de la planète vivant majoritairement dans les zones les plus pauvres sont exposés au paludisme et plus de 3 milliards de personnes sont à risque [6, 7]. La région Afrique de l'OMS reste la plus touchée avec, à elle seule, quelque 90 % des cas de paludisme et 92 % des décès associés en 2015. Plus de 2% des décès infantiles survenant en Afrique sont dûs au paludisme [6, 7]. En Afrique la population n'ayant pas accès aux outils permettant de prévenir et de traiter la maladie se compte encore par million. Au Bénin, le paludisme représente en 2014, 39.6% des affections rencontrées en consultation, 29.2% des causes d'hospitalisation, 26.0% des causes de décès, ce qui fait que cette maladie est la première cause de décès au Bénin [8, 9, 10].

Les conséquences liées au paludisme sont plus lourdes au niveau des femmes enceintes et les enfants de moins de cinq ans [11]. Chez la femme enceinte, les globules rouges infectés vont adhérer au placenta ce qui entraîne entre autre des fausses couches, favorise l'anémie [12], et chez le nouveau-né un risque accru de petit poids à la naissance (poids de naissance inférieur à 2.5 kg), lui-même associé à un risque accru de morbidité et de mortalité précoce [13].
Les facteurs conditionnant la survenue de la première infection palustre chez le nouveau-né ne sont pas encore bien connus [12]. En particulier, trois études, au Cameroun entre 1993 et 1995 [14], au Gabon entre 2002 et 2004 [15], et en

Tanzanie entre 2002 et 2004 [16], ont montré que les enfants, nés d'un placenta infecté avaient un risque plus important de développer une première infection palustre plus tôt que les enfants nés d'un placenta non infecté [14, 16, 15]. Ces observations suggèrent deux hypothèses non exclusives.
La première suppose une tolérance immunitaire de la femme enceinte induisant le passage trans-placentaire d'antigènes du Plasmodium falciparum [14, 16, 15]. Cette tolérance immunitaire est un phénomène complexe mal connu qui est dû à l'enfant. On suppose que l'enfant né d'une mère dont les globules rouges infectés ont adhéré au placenta déclenche ce processus biologique mal compris et cela aboutirait à sa plus grande sensibilité au paludisme.

La seconde hypothèse qui est de type environnemental suppose que les femmes qui ont un placenta infecté subissent une exposition plus élevée au parasite que les autres du fait de leur environnement qui est plus propice à la transmission vectorielle. Les jeunes enfants, vivant dans le même environnement, auraient également un risque plus élevé d'infection et par conséquent un délai d'apparition des premières infections plus précoce. L'exposition au parasite dépend de plusieurs facteurs dont la présence du vecteur, le type d'habitat, l'environnement proche des populations exposées, et le comportement de ces dernières par rapport aux méthodes de protection contre les piqûres de l'anophèle [17]. La prise en compte de la transmission entomologique du paludisme dans les lieux de vie des jeunes enfants est donc indispensable pour mieux comprendre la part respective des hypothèses environnementale et immunologique qui pourraientt intervenir sur le délai d'apparition d'une première parasitémie chez le nouveau-né. Ceci constitue l'un des buts principaux de l' étude dans laquelle s'inscrit cette Thèse. Il s'agit d'un large programme de recherche, financé par l'ANR (Agence Nationale de la Recherche (France)) portant sur les déterminants des premières infections palustres chez le nouveau-né et englobant des volets épidémiologique, biologique et environnemental. Dans ce cadre, un suivi de cohorte a été réalisé du 04 juin 2007 au 1er février 2010 dans 9 villages de la commune de Tori-Bossito (Bénin) où 600 nouveau-nés ont été suivis de la naissance jusqu'à l'âge de 18 mois. L'objectif général de ce projet est donc d'étudier dans la zone des 9 villages de l'étude, les déterminants qui peuvent intervenir dans le délai de survenue de la première infection palustre du nouveau-né.
Dans les trois études qui se sont intéressées à la survenue des premières infections palustres chez le nouveau-né, l'exposition aux parasites dans les lieux de vie des jeunes enfants est généralement prise en compte de manière approximative [15]. Dans chacun des cas, l'évaluation de l'exposition au vecteur s'était basée sur des enquêtes antérieures [14, 15] ou sur une observation sommaire de l'environnement [16]. Or la variabilité de l'intensité de la transmission vectorielle à laquelle sont soumis les enfants est très probablement occasionnée

par l'hétérogénéité des caractéristiques de leur espace de vie. Il a été montré que certains facteurs environnementaux tels que la saison, la pluviométrie et l'environnement de la maison d'habitation ont une influence sur la reproduction et la survie des anophèles [18, 19]. De plus, il est connu que des facteurs météorologiques tels que la température ou la pluviométrie influent sur le risque spatial et temporel d'infection par le Plasmodium falciparum [18, 20, 21]. Il a été également mis en évidence le rôle joué par certaines caractéristiques du milieu de vie des individus comme la proximité d'un cours d'eau pour expliquer le niveau d'exposition aux vecteurs [22, 23]. Il est donc très probable que les caractéristiques de l'environnement de vie à la fois à l' échelle du village et de la maison d'habitation puissent expliquer la variabilité du risque d'exposition aux vecteurs. A notre connaissance, peu d' études se sont intéressées à analyser la variabilité de la densité vectorielle et à la caractérisation du risque d'exposition au vecteur en relation avec des facteurs environnementaux à l' échelle de l'habitation, et ce, dans une zone peu étendue telle que notre zone d' étude (moins de 20km x 20km). Les difficultés de mesure, d'obtention d'information et de modélisation de ces données de type environnemental pourraient justifier le manque d'analyse à ce niveau.

# Contexte statistique

L'apprentissage statistique tout comme l'apprentissage machine sont des domaines de recherche situés à la frontière entre les mathématiques, en particulier la statistique et l'informatique. L'objectif principal est de rendre automatiques des procédures élaborées en vue de prendre des décisions complexes. Dans plusieurs domaines, des algorithmes sont mis en œuvre pour déterminer les facteurs de tout genre qui pourrait favoriser certains processus. Par exemple, les algorithmes développés pour les processus de séquençages génomiques sont en perpétuelle mutation. Ces processus permettent de déterminer les gènes ou allèles portant le code d'une maladie. Ce qui permettra de trouver des solutions pour agir directement sur la transcription ou le codage de ces gênes ou allèles. L'apprentissage statistique et l'apprentissage machine sont de nos jours très pratiques et couvrent de nombreux domaines d'études. Dans cette thèse, l'étude des facteurs environnementaux, climatiques et comportementaux qui expliquent le risque palustre sera considéré comme un problème de sélection de variables. Aussi, la caractérisation de ce risque dans le milieu d'étude et dans un autre milieu où les données entomologiques ne sont pas disponibles sera considérée comme un problème de prédiction.

# Objectifs

L'objectif général de cet travail est l'étude de l'évolution du risque de transmission du paludisme par le vecteur Anophèle aux populations humaines à partir des données entomologiques (capture de moustiques) et environnementales (saison, niveau de précipitation, comportements, habitats, etc.).
L'objectif spécifique de cette thèse est d'élaborer des méthodes statistiques basées sur l'apprentissage statistique et l'apprentissage machine pour dégager les facteurs environnementaux liés au risque d'exposition palustre et pour pouvoir répondre à la question biostatistique objet de cette thèse qui est de déterminer la meilleur méthode permettant de prédire au mieux de l'information entomologique pour d'autres études de même type pour lesquelles les données entomologiques ne sont pas disponibles.

# Contribution

Cette Thèse a réalisé les contributions suivantes : la justification de la variabilité spatio-temporelle à la fois au niveau village et au niveau maison du risque palustre [3], la proposition d'une nouvelle méthode de validation croisée stratifiée à deux niveaux [24, 25, 26, 27, 28], la construction de deux fonctions de prévision du risque palustre dans la commune de Tori-Bossito, la proposition d'une méthode pour surmontér des pré-traitements opérés par les experts avant l'analyse de notre jeu de données [28] et la mise à disposition des experts d'un nouvel ensemble de facteurs climatiques et environnementaux plus parcimonieux et plus performant pour la prévision du risque lié au paludisme dans cette commune [28]. Cet travail a ainsi permis de montrer que le risque d'exposition au vecteur palustre peut être mesurer de façon précise et efficace.

# Applications

A l'issue du suivi de cohorte comportant 600 enfants de Tori-Bossito, une importante base de données a été recueillie. Le jeu de données utilisé dans la thèse est extrait de cette base de données. Il comporte 612 observations et 21 variables originales. Les algorithmes développés dans cette thèse ont été appliqués à ces données. Ces algorithmes peuvent être appliqués à tous types de données possédant les mêmes structures.

# Organisation du rapport

Le manuscrit comporte une introduction, cinq chapitres et une conclusion.
Le chapitre 1 traite des éléments de l'apprentissage statistique dont nous aurons besoin au cours des travaux. Ce chapitre est un cadre général de la Thèse. Dans le chapitre 2, nous avons présente la méthodologie de construction d'une fonction de prédiction par une combinaison des modèles Modèle linéaire généralisé mixte (GLMM) et un processus de sélections de variables *Backward elimination*. Cette méthode permet de construire des modèles avec des effets aléatoires pour prendre en compte les possibles corrélations à cause de la structure hiérarchique des données. Le chapitre 3 qui est une application du chapitre 2, a permis de construire une fonction de prévision du risque palustre sur le jeu de données liées au paludisme de Tori-Bossito. Dans le chapitre 4, nous avons présenté la méthodologie de construction d'une fonction de prévision en combinant les modèles GLMM-Lasso et une validation croisée à deux niveaux. Dans le chapitre 5, nous avons fait une application de la méthode développée en 4 pour construire une autre fonction de prévision du risque palustre pour le même jeu de données que dans le chapitre 3.

Chapitre 1

# Eléments sur l'apprentissage statistique

## 1.1 Introduction

Le développement des moyens informatiques, la multiplication des appareils de mesure (capteurs atmosphériques, appareils de séquençage génomique, etc.), la collecte de données sur des patients, dans le cas des études cliniques, épidémiologiques, etc, permet le stockage, le traitement, l'analyse d'ensemble de grandes bases de données. Le développement et le perfectionnement des logiciels de calculs offrent aux utilisateurs des possibilités de mettre en œuvre des méthodes d'analyse et de traitement sur ces volumineuses bases de données. Ces données sont pour la plupart du temps entachées de bruit pouvant fausser leur analyse. Pour perfectionner l'analyse de ces types de données, on préfère connaître la loi qui les régit en extrayant cette loi des données elles-mêmes pour de nouvelles prévisions, c'est l'apprentissage statistique.

## 1.2 Objectifs de l'apprentissage statistique

L'apprentissage statistique est considéré comme un problème d'inférence basé sur un nombre limité d'observations [29]. Le principe d'induction automatique qui constitue le raisonnement fondamental de l'apprentissage statistique, a pour but de créer des systèmes automatiques pouvant passer d'observations particulières à des lois générales. Cette approche est différente des approches classiques parce qu' elle permet de fournir des outils de contrôle non asymptotiques tels que les bornes sur la confiance de l'estimation de l'erreur de généralisation du modèle par l'erreur empirique [30]. L'apprentissage statistique permet d'étudier un phénomène et de trouver le modèle mathématique sous jacent pour faire des prévisions. L'un des objectifs de la théorie de l'apprentissage statistique est l'étude d'un modèle conceptuel basé sur le problème d'optimisation que consiste la minimisation du risque empi-

rique. Les estimateurs des paramètres ou des fonctionnelles obtenus doivent posséder des propriétés asymptotiques. Un enjeu principal de l'apprentissage statistique consiste à concevoir des systèmes de classification performants à partir d'un ensemble d'exemples représentatifs d'une population de données. C'est l'un des buts de l'apprentissage statistique qui possède deux grands aspects : le classement et la prévision. Les choses les plus importantes dans l'étude du phénomène sont : le nombre $n$ d'observations ou la taille de l'échantillon disponible et le nombre $p$ de variables observées sur cet échantillon. Lorsque les méthodes statistiques traditionnelles se trouvent mises en défaut en grandes dimensions $p \gg n$, les méthodes récentes d'apprentissage sont des recours pertinents car efficaces. On distingue deux types d'apprentissage : l'apprentissage statistique supervisé (la classification supervisée, la régression, etc.), l'apprentissage statistique non supervisé (la classification non-supervisée, l'estimation non-paramétrique, etc.).

## 1.3 Apprentissage non supervisé

Dans l'apprentissage non-supervisé, il n'y a pas de variable à expliquer. Il faut inférer une fonction des données pour décrire la variabilité des descripteurs et mettre en évidence des structures dans l'espace des observations : segmentation ou clustering, estimation de densité, mélange de densité, détection d'anomalie, etc. Il s'agit de rechercher une typologie ou taxinomie des observations. Comment regrouper les observations en classes homogènes où les classes obtenues sont les plus dissemblables possible entre elles. L'apprentissage non-supervisé est un problème de classification ou *Clustering* et fait appel à des méthodes de classification ascendante hiérarchique, des algorithmes de réallocation dynamique (*kmeans*) ou encore des cartes auto-organisatrices (*Kohonen*) [31].

## 1.4 Apprentissage supervisé

### 1.4.1 Objectifs

Dans l'apprentissage supervisé, on s'intéresse à la distribution de $(Y|X = x)$, c'est un problème de modélisation. Il s'agit de trouver une fonction $f$ susceptible, au mieux selon un critère à définir, de reproduire $Y$ ayant observé $X$.

$$Y = f(X) + \varepsilon,$$

où $\varepsilon$ est le bruit ou erreur de mesure. On suppose que l'erreur est additive, dans le cas où elle est multiplicative on applique la fonction logarithme pour revenir

au cas additif. L'apprentissage supervisé est un problème de régression. Soit un ensemble d'apprentissage $d_1^n$ constitué de données d'observations $(x_i, y_i)$ tel que :

$$d^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

où $x_i \in \mathbf{R}^p$, $y_i \in \mathbf{R}$. L'objectif est de construire, à partir de cet échantillon d'apprentissage, un modèle qui permettra de prévoir la sortie $y_{n+1}$ associée à une nouvelle entrée $x_{n+1}$ qui n'était pas dans l'ensemble d'apprentissage. Si la variable de sortie $y$ est quantitative on parlera de régression et si $y$ est qualitative on parlera de $classification$ (discrimination ou classement), reconnaissance de forme.

### 1.4.2 Stratégies

Avant même l'apprentissage, il serait correct et raisonnable de mener sur le jeu de données une exploration.

#### 1.4.2.1 Exploration des données

Dans les études, les variables explicatives ou prédictives ont été observées sur un ensemble de $n$ individus ou unités statistiques. Le premier travail difficile mais incontournable, consiste à mener une exploration statistique de ces données : allures des distributions, présence de données atypiques, corrélation et cohérence, incohérence, transformations éventuelles, description multidimensionnelle, réduction de dimension, classification, etc.

#### 1.4.2.2 Apprentissage

L'objectif de l'apprentissage est d'apprendre pour prévoir par l'intermédiaire d'un bon modèle prédictif. Pour réussir, cet apprentissage se fait suivant des étapes qui normalement doivent systématiquement s'enchaîner de la manière suivante :

1. Extraction des données avec ou sans échantillon faisant référence à des techniques de sondage appliquées ou applicables à des bases de données.
2. Exploration des données pour la détection de valeurs aberrantes ou seulement atypiques, d'incohérences, pour des distributions des structures de corrélation, pour la recherche de typologie, pour des transformations des données, etc..
3. Dans le cas d'un échantillon de grande taille, on pourra utiliser la méthode (apprentissage, validation, test), dans le cas d'un échantillon de petite taille on pourra utiliser la méthode de validation croisée

4. Pour chacune des méthodes considérées : modèle linéaire général (gaussien, binomial, poissonnien, etc.), discrimination paramétrique (linéaire ou quadratique) ou non paramétrique, $k$ plus proches voisins, arbres, combinaison de modèles (bagging, boosting), etc.
   — estimer le modèle pour une valeur donnée d'un ou de plusieurs paramètres.
   — optimiser ces paramètres en fonction de la technique d'estimation de l'erreur retenue : validation croisée, critères $C_p$, AIC, BIC, etc.
5. Choix de la méthode retenue en fonction de ses capacités à prédire, de sa robustesse, de l' interprétabilité du modèle obtenu.
6. Ré-estimation du modèle avec la méthode, le modèle et sa complexité optimisée à l'étape précédente sur l'ensemble des données.
7. Exploitation du modèle sur le jeu de données complet et de nouvelles données.

## 1.5 Méthode de construction de fonction de prévision

### 1.5.1 Sélection de variables

Les jeux de données fournies de nos jours sont essentiellement de grande dimension (petit nombre d'observations et le nombre de variables explicatives est supérieur au nombres d'observations). Il serait intéressant de sélectionner parmi toutes les variables explicatives un ensemble le plus parcimonieux possible et qui permet de construire une fonction de prévision. On se donne un ensemble de variables explicatives au départ. On pourra remplacer certaines variables par de nouvelles obtenues après transformation des variables de départ (changement d'échelle, transformation de variables numériques en non numériques, etc.). On pourra aussi intégrer d'autres variables obtenues par interactions entre certaines ou toutes les variables explicatives.

Il existe des méthodes statistiques permettant de déterminer de façon efficace un sous-ensemble stable ou non de variables explicatives offrant la possibilité de reconstruire presque exactement la variable réponse $Y$. Dans la littérature, il a été proposé diverses méthodes.

— La *sélection de sous-ensemble* qui consiste à déterminer le meilleur sous-ensemble $\mathcal{H}_l$ de variables explicatives qui minimise la somme des carrés des résidus [32]
— La méthode de *seuillage* qui consiste à sélectionner une variable explicative lorsque son coefficient $\beta_j$ est supérieur à un seuil défini et fixé. Nous pouvons citer le seuillage dur et le seuillage doux [33, 34].

— La méthode de *sélection par les tests* pour l'identification des variables pertinentes a été étudiée dans la littérature [35, 36, 37, 38, 39].

— La pénalisation $l_0$ basée sur le sous-ensemble des variables à coefficients non nuls. Pour cela on définit l'ensemble des variables actives ou l'ensemble des indices actifs associé au vecteur $\beta$ par :

$$\mathcal{A} = \{j : \beta_j \neq 0\} \tag{1.1}$$

On définit l'indice de sparsité de $\beta$ par $|\mathcal{A}|$ qui est le cardinal de l'ensemble $\mathcal{A}$. L'hypothèse fondamentale est que l'ensemble des variables actives peut être contrôlé. Il existe un entier $r$ tel que : $|\mathcal{A}| \leqslant r$ avec $r = o(n)$. Ce qui indique que le nombre de variables explicatives ayant un poids pertinent sur la variable réponse $Y$ n'est pas grand.
Cette hypothèse est d'une grande utilité en grande dimension parce qu'elle indique que peu de composantes du vecteur des paramètres $\beta$ seront non nulles.
On définit la pénalité $l_0$ introduite en 1994 par Foster et George [40, 41] par :

$$pen(\beta, n)_{l_0} = \lambda_n ||\beta||_0 = \lambda_n \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) \tag{1.2}$$

où $\mathbb{I}(.)$ est la fonction indicatrice, $\lambda_n$ le paramètre de régularisation. On pourra noter $pen(\beta, n)_{l_0} = pen(\beta, n)$ s'il n'y pas d'ambiguïté. La procédure décrite pour la sélection de variables est la *Prodédure de sélection canonique* qui se base sur le compromis entre de bonnes prédictions et la complexité du modèle encore appelé compromis entre le biais et la variance. Des estimateurs interprétables et utilisables sont construits à partir de cette pénalité, on peut citer :

— le critère d'information $C_p$ de Mallows [42] et le critère AIC (Akaike Information Criterion)[43] définis par :

$$pen^{AIC}(\beta, \lambda_n) = pen^{C_p}(\beta, \lambda_n) := \frac{2\sigma^2 ||\beta||_0}{n} \tag{1.3}$$

— le critère BIC (Bayesian Information Criterion) de Schwartz [44] défini par :

$$pen^{BIC}(\beta, \lambda_n) := \sigma^2 log(p) \frac{||\beta||_0}{n} \tag{1.4}$$

— le critère $C_p$ pour le modèle gaussien introduit par Birgé et Massart [45]

Ces critères sélectionnent les estimateurs parmi une collection $\mathcal{G} = \{\beta_1, \ldots, \beta_{\mathbf{T}}\}$ de taille $\mathbf{T}$ d'estimateurs de $\beta$ [46].
Le critère $BIC$ impose beaucoup plus de contraintes aux estimateurs et donc

tend à sélectionner dans la famille $\mathcal{G}$ d'estimateurs ceux qui sont plus parcimonieux c'est-à-dire des $\beta$ avec beaucoup de composantes nulles par rapport aux critères $C_p$ et $AIC$.

### 1.5.2 La méthode *Backward elimination*

La méthode *Backward elimination* pour la sélection de variables consiste à procéder à une élimination progressive des variables explicatives. On fixe une *p-value* critique $\alpha_{crit}$, on introduit dans le modèle tous les prédicteurs. Par rapport à $\alpha_{crit}$, on élimine étape par étape les prédicteurs. Son algorithme se présente comme suit :

**Algorithme 1.1** *Backward elimination.*

1: Introduire toutes les variables explicatives dans le modèle.

2: Eliminer la variables ayant la plus grande *p-value* supérieure à $\alpha_{crit}$

3: Reprendre le modèle avec les variables restantes et retourner à 2.

4: Arrêter dès que toutes les *p-value* sont plus petites que $\alpha_{crit}$

### 1.5.3 Construction d'une fonction de Prévision

On considère le $n$-échantillon $D^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, et une observation
$d^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ de cet échantillon. On suppose que $D^n$ possède une loi conjointe inconnue $P$ sur $\mathcal{X} \times \mathcal{Y}$, $x$ une observation de $X$, $(X, Y)$ un couple aléatoire de loi conjointe $P$ indépendant de $D^n$. $(X, Y) \in (\mathcal{X} \times \mathcal{Y})$ où $\mathcal{X}$ et $\mathcal{Y}$ sont des espaces mesurables. $D^n$ est un échantillon d'apprentissage.

**Définition 1** *On appelle prédicteur toute fonction $\phi$ définie par :*

$$\phi : \mathcal{X} \longrightarrow \mathcal{Y}, \quad avec \ \ \phi(X) = \hat{Y} \tag{1.5}$$

**Définition 2** *On appelle algorithme d'apprentissage toute fonction $\Phi$ définie par :*

$$\Phi : \mathcal{X} \times \bigcup_n (\mathcal{X} \times \mathcal{Y})^n \longrightarrow \mathcal{Y}, \quad avec \ \ \Phi(X, D_n) = \hat{Y} \tag{1.6}$$

**Définition 3** *On appelle Erreur de généralisation la quantité définie par :*

$$\tilde{R}_n = \tilde{R}_n(\phi) \quad = \quad \mathbb{E}_P[l(Y, \phi(X, D_n))|D_n]$$

$$R_n(\phi) \quad = \quad \mathbb{E}_P[l(Y, \phi(X))] \tag{1.7}$$

**Définition 4** *On appelle règle de prévision une fonction mesurable $f$ définie par :*

$$\begin{array}{rcl} f & : & \mathcal{X} \longrightarrow \mathcal{Y} \\ & & x \longmapsto f(x) \end{array}$$

Pour mesurer la qualité de prédiction, on définit une fonction dite de perte

**Définition 5** *On appelle fonction de perte, toute fonction mesurable $l$ telle que :*

$$\begin{array}{rcl} l & : & \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbf{R}_+ \\ & & (y, y') \longmapsto l(y, y') \\ & avec & l(y,y) = 0 \;\; et \;\; l(y,y') > 0 \;\; si \;\; y \neq y' \end{array}$$

Dans ce cas, si $x$ est une entrée, $y$ une sortie réellement associée à $x$, $f$ une règle de prévision, alors $l(y, f(x))$ mesure une perte encourue lorsque l'on associe à $x$ la sortie $f(x)$.

En régression réelle, on définit les pertes $\mathbf{L}^p, (p \geq 1)$

$$l(y, y') = ||y - y'||^p$$

Si $p = 1$, on parle de perte absolue, si $p = 2$ on parle de perte quadratique, voir chapitre 4.
En discrimination binaire : $\mathcal{Y} = \{-1, 1\}$

$$l(y, y') = \frac{|y - y'|}{2}$$

Il est aussi intéressant d'étudier la moyenne de cette fonction de perte qui est le risque.

**Définition 6** *Étant donnée une fonction de perte $l$, le risque ou l'erreur de généralisation d'une règle de prévision $f$ est définie par*

$$R_P(f) = \mathbb{E}_{(X,Y)\sim P}\big[l(Y, f(X))\big] \tag{1.8}$$

Notons bien que $(X, Y)$ ne dépend pas de l'échantillon $D^n$ qui a permit de construire la règle de prévision $f$.

**Définition 7** *Soit $\mathcal{F}$ la famille de toutes les règles de prévision possibles. Une*

*règle $f^*$ de prévision est dite optimale si*

$$R_P(f^*) = \inf_{f \in \mathcal{F}} R_P(f)$$

#### 1.5.3.1 Quelques règles de prévision optimale

**Définition 8** *On appelle fonction de régression la fonction $\eta^*$ définie par :*

$$\begin{array}{rcl} \eta^* & : & \mathcal{X} \longrightarrow \mathcal{Y} \\ & & x \longmapsto \eta^*(x) \\ \textit{avec} \quad \eta^*(x) & = & \mathbb{E}[Y|X=x] \end{array}$$

Dans le cas d'une régression réelle :

1. $$\mathcal{Y} = \mathbb{R}, \quad l(y, y') = (y - y')^2$$

   La fonction de régression $\eta^*$ avec $\eta^*(x) : x \longrightarrow \mathbb{E}[Y|X=x]$, vérifie :

   $$R_P(\eta^*) = \inf_{f \in \mathcal{F}} R_P(f)$$

2. $$\mathcal{Y} = \mathbb{R}, \quad l(y, y') = |y - y'|$$

   La fonction de régression $\mu^*(x) : x \longrightarrow$ médiane$([Y|X=x])$, vérifie :

   $$R_P(\mu^*) = \inf_{f \in \mathcal{F}} R_P(f)$$

#### 1.5.3.2 Minimisation du risque empirique

Lorsqu'on se trouve dans le cadre non paramétrique, en l'abscence de toute informtion sur la loi $P$, il est naturel de remplacer la loi $P$ par la loi empirique $p_n$ de l'échantillon $D^n$ et de déterminer le risque empirique.

**Définition 9** *Soit $D^n = \{(X_i, Y_i), 1 \leq i \leq n\}$. Le risque empirique associé à $D^n$ d'une règle de prévision $f \in \mathcal{F}$ est donné par :*

$$\hat{R}_n(f, D^n) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i)) \tag{1.9}$$

La minimisation du risque empirique a commencé depuis les travaux de Vapnik [47].

**Définition 10** *Étant donné un sous-ensemble $F$ de $\mathcal{F}$ (un modèle), l'algo-*

*rithme de minimisation du risque empirique sur F est défini par :*

$$\hat{f}(D^n) = Arg \min_{f \in F} \hat{R}_n(f, D^n)$$

.

Dans le cas où $F = \mathcal{F}$ (ensemble de tous les prédicteurs possibles) : la minimisation se fera sur $\mathcal{F}$, le modèle obtenu conduira à un sur-apprentissage.
La règle de prédiction $f^*$ optimale introduite dans la définition (7) est une règle dite Oracle. L'objectif est de déterminer un modèle $F$ pour lequel le risque de l'estimateur $\hat{f}_F(D^n)$ est proche de celui de l'oracle.
Pour se faire on calcule :

$$\begin{aligned} R_P(\hat{f}_F(D^n)) - R_P(f^*) &= \underbrace{\left\{R_P(\hat{f}_F(D^n)) - \inf_{f \in \mathcal{F}} R_P(f)\right\}}_{\textbf{Variance}} \\ &\quad + \underbrace{\left\{ \inf_{f \in \mathcal{F}} R_P(f) - R_P(f^*\right\}}_{\textbf{Biais}} \end{aligned} \tag{1.10}$$

La quantité $\left\{R_P(\hat{f}_F(D^n)) - \inf_{f \in \mathcal{F}} R_P(f)\right\}$ est appelée **Erreur d'estimation** ou (Variance) et $\left\{ \inf_{f \in \mathcal{F}} R_P(f) - R_P(f^*\right\}$ est appelée **Erreur d'approximation** ou (Biais).
Les deux termes sont de natures différentes et il faut des considérations issues de la statistique pour évaluer le premier et des considérations issues de l'approximation pour évaluer le second. Pour obtenir un modèle $\hat{F}$ parmi une collection de modèle $\mathcal{C}$, pour lequel le risque de $\hat{f}_{\hat{F}}(D^n)$ est proche de celui de l'oracle, il faut minimiser un critère pénalisé de type :

$$\hat{F} = arg \min_{F \in \mathcal{C}} \{\hat{R}_n(\hat{f}_F(D^n), D^n) + pen(F)\} \tag{1.11}$$

La méthode d'estimation avec pénalisation a été développée dans le chapitre 4 et dans le document [48].
La façon la plus simple d'estimer l'erreur de prévision sans biais consiste à calculer le risque empirique sur un échantillon indépendant n'ayant pas participer à l'estimation du modèle. Pour cela il faut procéder à une correction pour l'estimation de l'erreur cherchée. La forme de cette correction est liée à la structure de la variance dans la décomposition en biais et en variance de l'erreur ou on associe une pénalité à l'erreur compte tenu de la complexité du modèle.
En considérant les hypothèses et les notations précédentes, celles de [49], on montre que :

$$C_p = \hat{R}_n(\hat{f}(d^n), d^n) + 2\frac{p}{n}\hat{\sigma}^2$$

$$\begin{aligned} AIC &= -2\mathcal{L} + 2\frac{p}{n} \\ BIC &= -2\mathcal{L} + log(n)\frac{p}{n} \end{aligned} \tag{1.12}$$

Dans le cas gaussien, $C_p = AIC$.

## 1.6 La méthode de validation croisée

### 1.6.1 Définition

C'est une méthode de re-échantillonnage utilisée lorsque le jeu de données disponible comporte un faible nombre d'observations. Comme quelques types de validation croisée, nous pouvons citer le k-folds ou k-blocs et le leave-one-out. Le k-folds consiste à diviser les observations en k sous-ensembles, d'apprendre sur (k-1) sous-ensembles et de tester sur le dernier sous-ensemble. Le leave-one-out consiste à apprendre sur un sous-ensemble constitué de (n-1) observations et de prédire la dernière observation. On peut aussi décider de faire une validation croisée sur l'ensemble d'apprentissage. Ceci donne une validation croisée à deux niveaux ou double validation croisée.

### 1.6.2 La validations croisée *K-folds*

Son algorithme se présente comme suit :

**Algorithme 1.2** Validation croisée k-folds.

1: Découper l'echantillon en $K$ blocs ($K$-fold) de taille approximativement égales selon une loi uniforme.
2: Pour $k$ allant de 1 à $K$
   a Mettre de côté l'un des blocs,
   b Estimer le modèle sur $K-1$ blocs restantes,
   c Calculer l'erreur d'estimation sur chacune des observations n'ayant pas participer à l'estimation,
3: Obtenir un vecteur d'erreur d'estimation sur chaque observation de l'échantillon.
4: Estimer l'erreur de généralisation de validation sur l'échantillon à partir du vecteur d'erreur d'estimation.

### 1.6.3 Validation croisée *Leave one out*

On suppose $n$ observations. Le *Leave one out* est le cas particulier de la validation croisée *K-folds* avec $K = n$

**Algorithme 1.3** Validation croisée *Leave one out.*

1: Découper l'echantillon en $n$ blocs (un bloc contient une observation) .
2: Pour $k$ allant de 1 à $n$

   a Mettre de côté une observation,

   b Estimer le modèle sur $n-1$ observations,

   c Calculer l'erreur d'estimation pour l'observation n'ayant pas participer à l'estimation,

3: Obtenir un vecteur d'erreur d'estimation sur chaque observation de l'échantillon.
4: Estimer l'erreur de généralisation de validation sur l'échantillon à partir du vecteur d'erreur d'estimation.

### 1.6.4 Validation Croisée à deux niveaux

La validation croisée à deux niveaux est une double validation croisée qui consiste à faire une seconde validation croisée à l'intérieur de la première. Elle est nommée LOLO-DCV (*Leave-One-Level-Out Double Cross-Validation*). En clair lorsque pour la première CV l'ensemble des données a été scindé en deux parties, ensemble d'apprentissage ($E_A$) ensemble de test ($E_T$), on fait une validation croisée complète sur $E_A$ afin de déterminer le modèle optimal de prédiction sur $E_T$. Les strates de la seconde CV sont les ensembles d'apprentissage à chaque étape de la première CV.

**Algorithme 1.4** Validation croisée à deux niveaux

1. A chaque étape du premier niveau de la validation-croisée
   (a) Les données sont divisées en $N$-blocs
   (b) Les blocs sont regroupés en deux parties : $E_A$ et $E_T$, $E_A$ : l'ensemble d'apprentissage qui contient les observations de $(N-1)$-blocs,
   $E_T$ : l'ensemble de test, contenant les observations du dernier bloc.
   (c) On met de côté $E_T$
   (d) Deuxième niveau de validation croisée.
      i. Les données de $E_A$ sont divisées en $(N-1)$-blocs
      ii. Les blocs sont regroupés en deux parties : $E_{A_2}$ et $E_{T_2}$, $E_{A_2}$ : l'ensemble d'apprentissage qui contient les observations de $(N-2)$-blocs de $E_A$,
      $E_{T_2}$ : l'ensemble de test, contenant les observations du dernier bloc de $E_A$.
      iii. On met de côté $E_{T_2}$
      iv. On reprend le processus 1(d)ii $(N-1)$ fois afin que chacun des $(N-1)$ blocs soit $E_{T_2}$
   (e) On reprend le processus 1b $N$ fois afin que chacun des $N$ blocs soit $E_T$

Dans le cadre de notre travail, cette validation croisée tient compte de la structure des données. Ainsi les blocs utilisés dans la CV ne sont pas obtenus de manière aléatoire, ils sont déterministes. Un bloc est l'ensemble de toutes les observations d'une maison de capture. Cette méthode de constitution des blocs nous permet de rester cohérent avec l'objectif final qui est de faire des prédictions optimales dans des zones dont aucune information n'a été utilisée dans l'apprentissage.

Chapitre 2

# Construction d'une fonction de prévision par les modèles linéaires généralisés mixtes (GLMM)

## 2.1 Introduction

En sciences de la vie, les données collectées dans les études possèdent souvant une structure hiérarchique. Les mesures peuvent être répétées sur le même sujet (on parle de données longitudinales). Dans le cas où les données appartiennent à un même niveau dans la hiérarchisation, elles ont tendance à se ressembler. De ce fait, il y a une non indépendance entre les obervations et il y a une possibilité de corrélation entre les données. Dans les analyses statistiques, pour éviter des incorrections en inférence statistique et obtenir de meilleures estimations et prédictions, il est nécessaire de prendre en compte cette corrélátion entre les données. La prise en compte de cette corrélation amène à introduire des effets aléatoires.

## 2.2 Modèles Linéaires Généralisés Mixtes

### 2.2.1 Définitions et notations préliminaires

Un modèle linéaire généralisé mixte (**GLMM**) se met sous la forme matricielle

$$g[\mathbb{E}[Y|(\mathcal{B} = b,\, \beta)]] = X\beta + Zb \tag{2.1}$$

Pour une capture $k$ d'une maison $j$ et d'un village $i$, on a :

$$g[\mathbb{E}[Y_{ijk}|(\, u_i, v_{ji}\, \beta)]] = \mathbf{X}_{ijk}\beta + u_i + u_{ij} \tag{2.2}$$

où i, $1 \leq i \leq 9$ est le numéro du village, j, $1 \leq j \leq 41$ est le numéro de la maison et k , $1 \leq k \leq 17$ est le numéro de la mission de capture avec

$$u_i \sim N(0, \sigma_u^2) \text{ et } v_{ji} \sim N(0, \sigma_v^2)$$

$Y$ est le vecteur des observations de dimension n $\times$ 1,
$n$ est le nombre d' observations,
$\mathcal{B}$ est le vecteur des coefficients des effets aléatoires de dimension q $\times$ 1,
b est une réalisation de $\mathcal{B}$,
$q$ est le nombre d'effets aléatoires,
$\beta$ est le vecteur des coefficients des effets fixes de p $\times$ 1,
$p$ est le nombre de covariables fixes dans le modèle,
$\mathbf{X}$ est la matrice des covariables de dimension n $\times$ p,
$Z$ est la matrice des covariables de dimensions n $\times$ q.
Nous allons nous placer dans le cas où $(Y_m|\mathcal{B} = b, X = x)$ suit une loi de poisson de paramètre $\mathbb{E}(Y_m|\mathcal{B} = b, X = x)$, $1 \leq m \leq n$.
Dans notre cas, la fonction de lien $g$ deux fois différentiable et monotone est le logarithme népérien et $g(x) = \ln(x)$, $x > 0$.
$\mathbb{E}(.)$ désigne l'espérance mathématique.
On pose :

$$\begin{aligned} \mu(x,u) &= \mathbb{E}(Y|\mathcal{B} = b, X = x)) \text{ ou } g(\mu) = x\beta + Zb && (2.3) \\ \eta(x,u) &= g(\mu(x,u)) && (2.4) \\ & && (2.5) \end{aligned}$$

Avec

$$\begin{aligned} (Y_m|\mathcal{B} = b, X = x) &\sim \mathcal{P}(\mu_m) && (2.6) \\ \mathcal{B} &\sim \mathcal{N}(0, \Gamma_\theta) && (2.7) \end{aligned}$$

$1 \leq m \leq n$,
$\theta$ est un vecteur appelé composante de la variance des effets aléatoires.
$\Gamma_\theta$ est la matrice de variance-covariance de $\mathcal{B}$, de dimension $q \times q$. Elle est définie semi-positive, c'est-à-dire :

$$\forall\, M \neq 0,\ M^T \Gamma_\theta M \geq 0$$

Par conséquent l'inversibilité de $\Gamma_\theta$ n'est pas assurée. Elle le serait si $\Gamma_\theta$ est définie positive. c'est-à-dire :

$$\forall\, M \neq 0,\ M^T \Gamma_\theta M > 0$$

Au cours de l'estimation des paramètres, une valeur non inversible de $\Gamma_\theta$ n'affecte pas la stabilité des méthodes de calcul et n'empêche pas la convergence des estimateurs.

### 2.2.2 Modèles GLMMs et notations principales

On considère le modèle GLMM défini plus haut.

**Définition 11** *On appelle facteur de covariance relative, toute matrice $\Lambda_\theta$ telle que :*

$$\Gamma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^T \tag{2.8}$$

où $\sigma$ est le paramètre d'échelle. Cette factorisation de $\Gamma_\theta$ n'est pas unique et il faut rechercher celle qui est plus adaptée à notre cas. En réalité, un estimateur $\hat{\sigma}^2$ de $\sigma^2$ est non nul parce qu'il existe toujours un écart entre les prédictions et les observations.

**Définition 12** *Une variable aléatoire spérique est toute variable aléatoire $\mathcal{U}$ telle que :*

$$\mathcal{B} = \Lambda_\theta \mathcal{U}, \text{ avec } \mathcal{U} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_q) \tag{2.9}$$

**Définition 13** *on appelle matrice de Cholesky associée au modèle GLMM, la matrice triangulaire inférieure notée $L_\theta$ de dimension $q \times q$ telle que :*

$$L_\theta L_\theta^T = \Lambda_\theta^T Z^T Z \Lambda_\theta + I_q \tag{2.10}$$

### 2.2.3 Estimation des quantités et des paramètres dans un GLMM

La méthode d'estimation des quantités $L_\theta$, $\mathcal{B}$, $Z$, $\Gamma_\theta$ et des paramètres $\theta$, $\beta$. Dans un modèle GLMM est un processus itératif qui combine à chaque itération l'algorithme Penalized Iterative Reweighted Least Squares (PIRLS) [1, 2, 50] et l'approximation de Laplace qui utilise l'intégration de Laplace [51]. Le PIRLS possède les caractéristiques de deux algorithmes : Iterative Reweighted Least Squares (IRLS) [1, 2, 50] et Penalized Least Squares (PLS) [50]. Dans le PLS, la pénalité utilisée est l'inverse de la matrice de variance-covariance des effets aléatoires. Le PIRLS prédit les Best linear unbiased predictors (BLUPs) par l'estimation du mode de la distribution conditionnelle des effets aléatoires sphériques conditionnnellement aux observations $(\mathcal{U}|Y = y_{obs})$ et les composantes de la variance. L'intégration de Laplace permet d'estimer

la vraisemblance du modèle et la maximisation de cette vraisemblance permet de déterminer les coefficients des effets fixes.
Si $L(\beta, \theta \mid y)$ est la vraisemblance du modèle fonction de $\beta$ et $\theta$, on a :

$$L(\beta, \theta|y) = f(Y|\beta, \theta, x) = \int_{\mathbf{R}^q} P(Y|\beta, b, x).h(b|\Gamma_\theta)\, db \tag{2.11}$$

Les éléments de cette équation sont :

Variables aléatoires

- $Y$ variable réponse
- $\mathcal{B}$ effets aléatoires
- $\mathcal{U}$ effets aléatoires sphériques tels que $\mathcal{B} = \Lambda_\theta \mathcal{U}$,

Paramètres

- $\beta$ coefficients des effets fixes
- $\sigma$ paramètre d'échelle
- $\theta$ composante de la variance des effets aléatoires telle que $\mathcal{B} = \Lambda_\theta \mathcal{U}$,

Matrices

$X$ de dimension $n \times p$ associée à $\beta$

$Z$ de dimension $n \times p$ associée à $b$

$\Lambda_\theta$ relative aux facteurs de covariances tels que $Var(\mathcal{B}) = \sigma^2 \Lambda_\theta \Lambda_\theta^T$

Quantités

$L(\beta,\theta|Y)$ est la vraisemblance fonction de ($\beta$,$\theta$), les observations $y_i$ étant données ;

$f(Y|\beta, \theta, x)$ est la densité marginale de $Y$ conditionnellementà $\beta$ et $\theta$ ;

$P(Y|\beta, b, x)$ est la fonction de probabilité de masse de $Y$ conditionnellement à $\beta$ et $b$.

$h(b|\Gamma_\theta)$ est la fonction densité gaussienne de $\mathcal{B}$ en $b$.

Les paramètres du modèle sont solutions de équation :

$$(\hat{\beta}; \hat{\theta}) = Arg \max_{(\beta,\theta)} L(\beta; \theta|Y)) \tag{2.12}$$

L'estimateur $\tilde{b}$ de $b$ est donné par :

$$\tilde{b}(\beta, \theta) = Arg \max_{b} [(P(Y|\beta, b).h(b|\Gamma_\theta)] \tag{2.13}$$

La fonction Logarithme népérien étant continue et strictement croissante, on obtient successivement :

$$\tilde{b}(\beta,\theta) = Arg\max_{b}\left[\ln P(Y|\,\beta,b) + \ln h(b|\,\Gamma_\theta\right] \tag{2.14}$$

$$\begin{aligned}
\tilde{b}(\beta,\theta) &= Arg\max_{b}\left[\ln P(Y|\,\beta,b) + \ln\left[\frac{1}{\sqrt{det(2\pi\Gamma_\theta)}}exp(-\frac{1}{2}b^T\Gamma_\theta^{-1}b)\right]\right]\\
\tilde{b}(\beta,\theta) &= Arg\max_{b}\left[\ln P(Y|\,\beta,b) - \frac{1}{2}\ln det(2\pi\Gamma_\theta) - \frac{1}{2}(b^T\Gamma_\theta^{-1}b)\right]\\
\tilde{b}(\beta,\theta) &= Arg\max_{b}\left[\ln P(Y|\,\beta,b) - \frac{1}{2}(b^T\Gamma_\theta^{-1}b)\right]
\end{aligned}$$

Ainsi,

$$\tilde{b}(\beta,\theta) = Arg\max_{b}\left[\ln P(Y|\,\beta,b) - \frac{1}{2}(b^T\Gamma_\theta^{-1}b)\right] \tag{2.15}$$

Mais l'estimation de $\tilde{b}$ par cette méthode reste non évidente car $\Gamma_\theta$ n'est pas connue. En réalité, l'estimation des paramètres comme nous l'avons précisé un peu plus haut se fait par un processus itératif utilisant l'algorithme Penalized Iterative Reweighted Least Squares (PIRLS) et l'approximation de Laplace [51, 52].

**Proposition 1** *L' estimateur $\hat{\beta}$ de $\beta$ est donné par la relation :*

$$\hat{\beta} = Arg\max_{\beta}\left[-d(Y|\,\beta,b) + (b^*)^T(b^*) + 2\ln(\det(D)\right] \tag{2.16}$$

**Preuve 1** *Utilisant la déviance du modèle définie par $d(Y,\beta,\theta) = -2l(\beta,\theta|\,Y)$ on a successivement :*

$$\begin{aligned}
-2l(\beta,\theta|\,Y) &= -2\ln\int_b P(Y|\,\beta,\,b).f(b|\,\Gamma_\theta)\,db\\
-2l(\beta,\theta|\,Y) &= -2\ln\int_b \exp\left[\ln P(Y|\,\beta,b).f(b|\,\Gamma_\theta)\right]\,db\\
-2l(\beta,\theta|\,Y) &= -2\ln\int_b \exp\left[\ln P(Y|\,\beta,b) + \ln f(b|\,\Gamma_\theta) - \frac{1}{2}b^TA^{-1}b\right]db
\end{aligned}$$

*ainsi*

$$-2l(\beta,\theta|\,Y) = -2\ln\int_b \exp\left[\ln P(Y|\,\beta,b) + \ln f(b|\,\Gamma_\theta) - \frac{1}{2}b^TA^{-1}b\right]db$$

*avec*

$$A = k\Gamma_\theta \; et \; k = \left[\frac{1}{2}\ln(det(\Gamma_\theta))\right]^{\frac{2}{q}}. \tag{2.17}$$

*En appliquant l'approximation de Laplace, on obtient successivement :*

$$\begin{aligned}
-2l(\beta,\theta|\,Y) &\approx -2\ln\int_b \exp\left[\ln P(Y|\,\beta,b) + \ln\ f(b|\,\hat{\Gamma}_\theta) - \frac{1}{2}b^T(k\hat{\Gamma}_\theta)^{-1}b\right] db \\
-2l(\beta,\theta|\,Y) &\approx -2\ln\int_b \exp\left[\ln P(Y|\,\beta,b)\right. \\
&+ \left.\ln\left[\frac{1}{(det2\pi\hat{\Gamma}_\theta)^{\frac{1}{2}}}e^{-\frac{1}{2}b^T\hat{\Gamma}_\theta^{-1}b}\right] - \frac{1}{2}b^T(k\hat{\Gamma}_\theta)^{-1}b\right] db \\
-2l(\beta,\theta|\,Y) &\approx -2\ln\int_b \exp\left[\ln P(Y|\,\beta,b)\ - \frac{1}{2}b^T\hat{\Gamma}_\theta^{-1}b\right. \\
&+ \left.\ln\left[\frac{1}{(det2\pi\hat{\Gamma}_\theta)^{\frac{1}{2}}}\right] - \frac{1}{2}b^T(k\hat{\Gamma}_\theta)^{-1}b\right]\ db \\
-2l(\beta,\theta|\,Y) &\approx -2\ln\int_b \exp\left[\ln P(Y|\,\beta,b)\ - \frac{1}{2}b^T\Delta^T\Delta b\right. \\
&- \left.\ln\left[\frac{1}{(det2\pi\hat{\Gamma}_\theta)^{\frac{1}{2}}}e^{-\frac{1}{2}b^T(k\hat{\Gamma}_\theta)^{-1}b}\right]\right]\ db \\
-2l(\beta,\theta|\,Y) &\approx -2\ln\int_b \exp\left[\ln P(Y|\,\beta,b)\ - \frac{1}{2}\ (\Delta b)^T(\Delta b)\right.
\end{aligned}$$

$$
\begin{aligned}
& - \ln\left[\frac{1}{(det 2\pi\hat{\Gamma}_\theta)^{\frac{1}{2}}} e^{-\frac{1}{2} b^T (k\hat{\Gamma}_\theta)^{-1} b}\right]\Bigg] db \\
-2l(\beta,\theta|\, Y) &\approx -2\ln \int_b \exp\left[\ln P(Y|\,\beta,b) - \frac{1}{2}(b^*)^T(b^*)\right. \\
& - \ln\left[\frac{1}{(det 2\pi\hat{\Gamma}_\theta)^{\frac{1}{2}}} e^{-\frac{1}{2} b^T (k\hat{\Gamma}_\theta)^{-1} b}\right]\Bigg] db \\
-2l(\beta,\theta|\, Y) &\approx -2\ln\left[\exp\left[\ln P(Y|\,\beta,b) - \frac{1}{2}(b^*)^T(b^*)\right]\right] \\
& - 2\ln\left[\exp \int_b \frac{k^{\frac{q}{2}}}{(det 2\pi\hat{\Gamma}_\theta)^{\frac{1}{2}}} e^{-\frac{1}{2} b^T (k\hat{\Gamma}_\theta)^{-1} b}\, db\right] \\
-2l(\beta,\theta|\, Y) &\approx -2\left[\ln P(Y|\,\beta,b) - \frac{1}{2}(b^*)^T(b^*)\right] - 2k^{\frac{q}{2}}(1) \\
-2l(\beta,\theta|\, Y) &\approx -2\left[\ln P(Y|\,\beta,b) - \frac{1}{2}(b^*)^T(b^*)\right] - 2k^{\frac{q}{2}} \\
-2l(\beta,\theta|\, Y) &\approx -2\left[\ln P(Y|\,\beta,b) - \frac{1}{2}(b^*)^T(b^*)\right] - 2\ln(det(\Gamma_\theta)) \\
2l(\beta,\theta|\, Y) &\approx -d(Y|\,\beta,b) + (b^*)^T(b^*) + 2\ln(det(\Gamma_\theta)) \\
2l(\beta,\theta|\, Y) &\approx -d(Y|\,\beta,b) + (b^*)^T(b^*) + 2\ln(det(D))
\end{aligned}
$$

*Ainsi*

$$2l(\beta,\theta|\, Y) \approx -d(Y|\,\beta,b) + (b^*)^T(b^*) + 2\ln(det(D)). \tag{2.18}$$

*Dans ce cas $d(Y|\,\beta,b)$ est la fonction déviance obtenue à partir des prédicteurs linéaires uniquement, avec $d(Y|\,\beta,b) = -2\ln(p(Y|\beta,b))$. Il vient que $d(Y,\beta,\theta)$ est fonction uniquement de $\beta$. D'où*

$$\hat{\beta} = Arg\max_{\beta}\left[-d(Y|\,\beta,b) + (b^*)^T(b^*) + 2\ln(\det(D)\right]. \qquad \square$$

### 2.2.4 Algorithme PIRLS

Au début de l'algorithme, pour faciliter les calculs et assurer la convergence des estimateurs, on pose $\mu^{(0)} = Y$ et on a : $\eta^{(0)} = g(\mu^{(0)})$. Pour un $\beta$ fixé, et à l'itération d'ordre $r$ on a [50] :

$$\eta^{(r)} = X\beta + Zb^{(r)} \tag{2.19}$$

où on évalue aussi les grandeurs suivantes :

$\mu^{(r)} = g^{-1}(\eta^{(r)})$, $d\eta/dY = G^{(r)}$. On suppose $W^{(r)}$ une matrice diagonale des poids telle que : $(\mathbf{W}^{-1})^{(r)} = Var(Y|u) = d\mu/d\eta$

En posant $z = g\left(E(Y|\beta, \Gamma_\theta)\right)$, et en faissant un développement en série de Taylor à l'ordre 1 de $z$ en $\mu$ on a :

$$z^{(r)} \approx \eta^{(r)} + G^{(r)}(Y - \mu^{(r)}) \tag{2.20}$$

Alors le vecteur $b$ est solution de l'équation :

$$(Z^T W^{(r)} Z) b^{(r+1)} = Z^T W^{(r)} z^{(r)} \tag{2.21}$$

Bates et DebRoy ont démontré en 2004 [53, 54, 55] qu'on peut incorporer $\Gamma_\theta^{-1}$ à la distribution gaussienne de $\mathcal{B}$ en ajoutant q "pseudo-observations". Ainsi le mode conditionnel $\tilde{b}(\beta,\ \theta)$ est déterminé par le PIRLS

**Proposition 2** *On considère le modèle GLMM défini en (2.1). On suppose que l'équation (2.21) est pénalisée. La pénalité utilisée est $\Gamma_\theta^{-1}$ et elle est symétrique définie semi-positive. Alors il existe : $Z^*$, $b^*$ tels que :*

$$b^{(r+1)} = (Z^{*T} W^{(r)} Z^* + I_q)^{-1} (Z^{*T} W^{(r)} z^{(r)}). \tag{2.22}$$

**Preuve 2** *La matrice $\Gamma_\theta^{-1}$ étant symétrique et définie semi-positive, alors il existe une matrice $\Delta$ telle que : $\Gamma_\theta^{-1} = \Delta^T \Delta$ En pénalisant l'équation (2.21), on obtient l'équation suivante :*

$$(Z^{*T} W^{(r)} Z^* + \Gamma_\theta^{-1}) b^{(r+1)} = Z^{*T} W^{(r)} z^{(r)} \tag{2.23}$$

*On a successivement :*

$$\begin{aligned}
(Z^T W^{(r)} Z + \Gamma_\theta^{-1}) b^{(r+1)} &= Z^T W^{(r)} z^{(r)} && (2.24)\\
(Z^T W^{(r)} Z + \Delta^T \Delta) b^{(r+1)} &= Z^T W^{(r)} z^{(r)} && (2.25)\\
\left((\Delta^T \Delta)^{-1} Z^T W^{(r)} Z + I_q\right) b^{(r+1)} &= (\Delta^T \Delta)^{-1} Z^T W^{(r)} z^{(r)} && (2.26)\\
\left(\Delta^{-1} (Z\Delta^{-1})^T W^{(r)} Z + I_q\right) b^{(r+1)} &= \Delta^{-1} (Z\Delta^{-1})^T W^{(r)} z^{(r)} && (2.27)\\
\left((Z\Delta^{-1})^T W^{(r)} Z + \Delta\right) b^{(r+1)} &= (Z\Delta^{-1})^T W^{(r)} z^{(r)} && (2.28)\\
\left[(Z\Delta^{-1})^T W^{(r)} Z + \Delta\right] (\Delta^{-1}\Delta) b^{(r+1)} &= (Z\Delta^{-1})^T W^{(r)} z^{(r)} && (2.29)\\
\left[(Z\Delta^{-1})^T W^{(r)} Z\Delta^{-1} + I_q\right] (\Delta) b^{(r+1)} &= (Z\Delta^{-1})^T W^{(r)} z^{(r)} && (2.30)\\
& && (2.31)
\end{aligned}$$

*En posant : $Z^* = Z\Delta^{-1}$ et $b^* = \Delta b$, on obtient :*

$$\begin{aligned}
\left[(Z^*)^T W^{(r)} Z^* + I_q\right] (b^*)^{(r+1)} &= (Z^*)^T W^{(r)} z^{(r)} && (2.32)\\
(Z^{*T} W^{(r)} Z^* + I_q) b^{*(r+1)} &= Z^{*T} W^{(r)} z^{(r)} && (2.33)
\end{aligned}$$

*D'où :*

$$b^{*(r+1)} = (Z^{*T}W^{(r)}Z^* + I_q)^{-1}(Z^{*T}W^{(r)}z^{(r)}). \qquad \square$$

Les valeurs obtenues par itération, $b^{*(1)}$, $b^{*(2)}, \ldots, b^{*(n)}$, convergent vers $\tilde{b}^*(\beta, \theta)$ [1] , si la quantité

$$||\eta^{(r+1)} - \eta^{(r)}||/||\eta^{(r)}||$$

est inférieur à un seuil donné.

Au cours de chaque itération, la matrice de variance-covariance de $b^*$ conditionnnellement à $\beta$ et $\theta$ est approximée par :

$$Var(b^{(r)}|\beta, \theta, y) \approx D^{(r)} = \hat{\Gamma_\theta}^{(r)} = (Z^{*T}W^{(r)}Z^* + I_q)^{-1} \tag{2.34}$$

Cette méthode d'approximation de $Var(b^{(r)}|\beta, \theta, y)$ est similaire à l'utilisation de l'inverse de la matrice d'information de Fisher pour approcher la matrice de variance-covariance La vraisemblance de l'équation (2.11) ne peut être calculée parce que l'intégrale n'a pas une forme exacte. Une valeur approchée de cette intégrale peut être obtenue en considérant la méthode d'approximation de Laplace.

**Algorithme 2.1** PIRLS [1, 2] .

1. Initialisation

2. Calcul par la formule de $\Gamma^{(0)} = (Z^{*T}W^{(0)}Z^* + I)^{-1}(Z^{*T}W^{(0)}g(Y))$ et de $u^{(1)} = (Z^{*T}W^{(0)}Z^* + I)$.
   L'approximation de Laplace utilise $\Gamma^{(0)}$ et $u^{(1)}$ pour estimer $\beta^{(0)}$.
3. Itération d'ordre 1 :
   Le PIRLS utilise $\beta^{(0)}, u^{(1)}, \Gamma^{(0)}$ pour estimer $\Gamma^{(1)}$, $u^{(2)}$
   L'approximation de Laplace utilise $\Gamma^{(1)}$ et $u^{(2)}$ pour estimer $\beta^{(1)}$.
4. Itération d'ordre n :
   Le PIRLS utilise $\Gamma^{(n-1)}$, $u^{(n)}$ et $\beta^{(n-1)}$ pour estimer $\Gamma^{(n)}$, $u^{(n+1)}$
   L'approximation de Laplace utilise $\Gamma^{(n)}$ et $u^{(n+1)}$ pour estimer $\beta^{(n)}$.
5. Le processus s'arrête lorsque $\frac{\|\beta^{(n)} - \beta^{(n-1)}\|}{\|\beta^{(n-1)} - \beta^{(n-2)}\|} <$ c, où $c \approx 1$.

Au début de l'algorithme PIRLS, on sait que

$$\mu^{(0)} = y; \quad \text{et on a} : \eta^{(0)} = g(\mu^{(0)})$$

On évalue $\Gamma_\theta^{(0)} = (Z^{*T}W^{(0)}Z^* + I_q)^{-1}(Z^{*T}W^{(0)}g(y))$ et $b^{(1)} = (Z^{*T}W^{(0)}Z^* + I)$, l'approximation de Laplace permet d'estimer $\beta^{(0)}$.
A l'tération ordre $r$, l'algoritme PIRLS utilise $\Gamma_\theta^{(r-1)}$, $b^{(r)}$ et $\beta^{(r-1)}$ pour estimer $\Gamma_\theta^{(r)}$ et $b^{(r+1)}$, et l'approximation de Laplace permet d'estimer $\beta^{(r)}$.
A l'arrêt du processus, on a : $\hat{\beta} \approx \beta^{(n)}$, $\hat{u} \approx u^{(n+1)}$ et $\hat{\Gamma} \approx \Gamma^{(n)}$.

## 2.3 Méthodologie de contruction d'une fonction de prévision par combinaison du *Backward* et du GLMM

On suppose que le nombre de variables au départ est $n_v$. Le processus de construction de la focntion de prévision est le suivant :

1. On utilise le *Backward* pour construire un sous-ensemble de variables $\mathcal{X}_p$, $1 \leq p \leq n_v$.
2. On utilise la validation croisée *Leave one out* pour prédire toutes les observations. Dans ce cas, le modèle de régression utilisé est le GLMM, les effets aléatoires sont simulés dans la prévision.
3. On détermine l'erreur de prévision pour chaque $\mathcal{X}_p$.
4. On répète les étapes 1, 2 et 3 pour tous les sous ensembles de variables $\mathcal{X}_p$.
5. On retient le sous-ensemble $\mathcal{X}_p^{min}$ qui minimise la fonction de perte.
6. Aux variables de $\mathcal{X}_p^{min}$, on ajoute certaines interactions de $\mathcal{X}_p^{min}$ interprétables par les experts. Cet ensemble sera noté $\mathcal{X}_p^{pred}$
7. On reprend les étapes 1 à 5
8. On obtient ainsi le sous-ensemble optimal pour la prévision noté $\mathcal{X}_p^{opt}$
9. Par une validation croisée *Leave one out* et utilisant $\mathcal{X}_p^{opt}$, on réalise la prévision pour toutes les observations du jeu de données.

L'ensemble de variables optimal pour la prévision est constitué des variables de $\mathcal{X}_p^{opt}$. Le modèle de prévision est un modèle GLM construit avec les éléments de $\mathcal{X}_p^{opt}$ et les prévisions optimales sont obtenues par ce modèle de prédiction.

Chapitre 3

# Analyse de l'influence des facteurs environnementaux locaux sur la transmission du paludisme au Bénin à Tori-Bossito

Cette partie est basée sur l'article [3], publié dans le journal Plos One.

## 3.1 Présentation et analyse des données liées au paludisme de Tori-Bossito

### 3.1.1 Milieu d'étude

Les données dont nous disposons dans le cadre de ce travail, ont été recueillies dans la commune de Tori-Bossito, située à 40 km au Nord-Ouest de Cotonou (voir carte en annexe). Cette commune est caractérisée par un climat sub-équatorial avec deux saisons de pluies et deux saisons sèches. La transmission palustre est relativement identique chaque année, permanente et fortement influencée par les pluies. Les vecteurs du paludisme rencontrés le plus souvent sont les *Anopheles gambiae* et les *Anopheles funestus*. Neuf villages ont été retenus dans cette étude dans un ensemble de villages respectant certaines caractéristiques en particulier la proximité des trois centres de santé de la zone d'étude.

### 3.1.2 Protocol et objectifs du projet

Ce projet d'étude a été mis en place pour permettre aux responsables d'évaluer de manière précise l'exposition aux piqûres des vecteurs du paludisme pour 600 enfants suivis pendant le projet. Mesurer de l'information entomologique n'est pas envisageable au niveau individuel pour les 600 enfants suivis qui résident dans 600 maisons différentes. Pour ce faire, il fallait trouver une méthodologie statistique pour la prédire. Il a été mis en place un protocole de collecte d'informations entomologiques et environnementales sur 40 maisons n'appartenant pas aux 600 maisons des enfants. Ces 40 maisons permettront d'expliquer le lien entre la présence des vecteurs et les caractéristiques climatiques (saison, pluviosité, etc.), environnementales(type de sol, végétations, etc.), comportementales (utilisation de moustiquaire, de répulsif, nombre de personnes dormant dans la même pièce, etc.) et à partir de ce modèle explicatif de construire le modèle de prévision pour les 600 enfants du projet.

### 3.1.3 Dispositif expérimental

Les données entomologiques ont été recueillies du 08 juin 2007 au 17 juillet 2009. Toutes les six semaines, des missions de capture ont été réalisées selon le protocole OMS de capture sur sujets humains pendant trois nuits, et ont lieu à l'intérieur et à l'extérieur des maisons. La zone d'étude comprend neuf villages (Avamé centre, Gbédjougo, Houngo, Anavié, Dohinoko, Gbétaga, Tori Cada Centre, Zèbè et Zoungoudo). Ils ont été retenus de manière aléatoire parmi les villages possédant un centre de santé pour la prise en charge des cas de paludisme. Dans chaque village, quatre maisons ont été choisies de manière aléatoire parmi les maisons proches du centre de santé du village. Chaque maison comportant au moins deux pièces pour permettre au capteur, de travailler sans gêner la famille. Dans la période de recueil des données, certaines maisons ont été extraites du projet par leurs propriétaires pour raisons diverses. Ces maisons ont été remplacées par d'autres maisons sur base des mêmes critères de départ. Au total dix-neuf missions ont été faites sur l'ensemble de la période. Toutefois les données recueillies lors des deux premières missions n'ont pas été prises en compte car les pluviomètres n'avaient pas été installés en ce moment (les données pluviométriques sont importantes dans l'estimation du risque palustre). Il en a résulté des données sur 5 maisons (cas de Avamè, Cada-centre, Dohinonko) voire 6 maisons à Gbédjougo contre quatre maisons initialement prévues. La détermination des caractéristiques ( anophèle ou non) des moustiques capturés a été réalisée en laboratoire. Les données sont donc de type longitudinal et entomologique en parallèle du suivi de cohorte des enfants du projet.

### 3.1.4 Structure des données

Les données présentent une structure hiérarchique à trois niveaux : niveau capture, niveau maison et niveau village comme le montre la Figure 3.1 où i, $1 \leq i \leq 9$, est le numéro du village, j, $1 \leq j \leq 4$, est le numéro de la maison de capture et k, $1 \leq k \leq 17$, est le numéro de la mission de capture figure 3.1

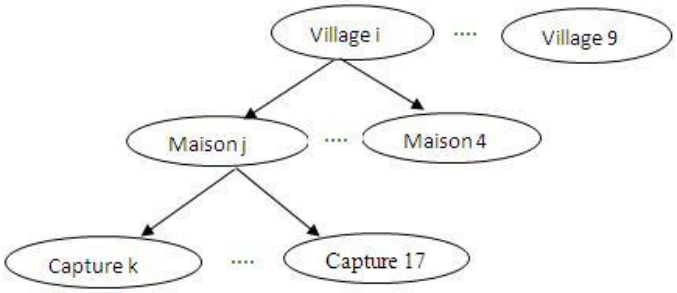



FIGURE 3.1 – **Hiérarchisation du dispositif expérimental**.

### 3.1.5 Variables mesurées

Les variables mesurées sont de plusieurs sortes : les variables de type entomologique (anophèle totaux capturés, anophèles infectés), les variables liées au comportement des habitants de la zone d'étude (possession de moustiquaire, utilisation de répulsif, nombre de personnes par chambre, présence de travaux, présence d'ustensiles), les variables liées aux caractéristiques de la maison de capture (nature du toit, nombre d'ouvertures), les variables de type environnemental (présence de cours d'eau, type de sol, indice de végétation) et les variables climatiques (pluie pendant la mission, nombre de jours de pluie, quantité de pluie,saison). Les détails sur ces variables sont donnés dans le tableau des variables mesurées en annexe.

#### 3.1.5.1 Variable expliquée

La variable expliquée et modélisée dans le cas de ce travail est la variable mesurant le nombre total d'anophèles capturés.

#### 3.1.5.2 Variables explicatives

Toutes les variables présentées dans le Tableau 5.7 en annexe (à l'exception des deux variables entomologiques relatives au nombre total d'anophèles

capturés et infectés), sont les variables explicatives qui nous permettront d'expliquer et de modéliser la variable expliquée. Parmi ces variables, il y en a de type binaire (nature du toit, type de sol, présence de cours d'eau, possession de moustiquaire, utilisation de répulsif, nombre de personnes par chambre, présence de travaux, présence d'ustensiles), à trois modalités (nombre de jours de pluie), à quatre modalités (nombre d'ouvertures, indice de végétation, quantité de pluie, saison). S'y ajoute la variable "nombre de jours de pluie" qui est une variable discrète.

## 3.2 Méthode statistique

### 3.2.1 Modèle explicatif

La variable dépendante $Y$ (le nombre d'Anophèles) est une variable de comptage. Après avoir réalisé un test d'adéquation de Chi-deux à la loi de Poisson pour $Y$, nous pouvons conclure que les $(Y_i)_{1\leq i\leq n}$ sont des réalisations d'une loi de Poisson de moyenne commune $\hat{Y}$. Pour prendre en compte la structure hiérarchique des données (captures répétées dans la même maison, quatre maisons par village) avec possible corrélation entre les mesures entomologiques, un modèle mixte de Poisson a été construit avec effets aléatoires au niveau village, au niveau maison et au niveau mission de capture. Il faut remarquer qu'il n'y a pas de répétition des mesures dans une maison au cours d'une mission de capture. Pour la mission de capture $k$ de la maison $j$ dans le village $i$ on a :

$$\ln[\mathbb{E}(Y_{ijk}|\, a_i, b_{ij}, c_{ijk}; \beta)] = \beta_0 + \sum_{l=1}^{p} \beta_l X_{ijkl} + a_i + b_{ij} + c_{ijk} \qquad (3.1)$$

où $Y$ est le nombre d'anophèles collectés, $X$ est un $p$-vecteur des variables environnementales, $\beta$ est le $(p+1)$-vecteur des paramètres du modèle y compris le coefficient fixe $\beta_0$. $a_i$ est l'effet aléatoire au niveau village, $b_{ij}$ est l'effet aléatoire au niveau maison et $c_{ijk}$ est l'effet aléatoire au niveau des missions de capture. On démontre que dans ce modèle on a [56, 57] :

$$Var(Y_{ijk}|\, a_i, b_{ij}) = \mathbb{E}(Y_{ijk}|\, a_i, b_{ij}) + (\mathbb{E}(Y_{ijk}|\, a_i, b_{ij}))^2[exp(\sigma_c^2) - 1]. \qquad (3.2)$$

Ainsi, $Var(Y_{ijk}|\, a_i, b_{ij}) \geq \mathbb{E}(Y_{ijk}|\, a_i, b_{ij})$ montre qu'en ajoutant un effet aléatoire au niveau des missions de capture réduirait significativement la sur-dispersion du modèle avec deux effets aléatoires au niveau village et maison. Toutes les variables environnementales ont été introduites dans le modèle et une procédure *Backward* a été appliquée pour sélectionner les variables significatives dans le modèle final.

#### 3.2.1.1 Modèle prédictif

Un modèle de régression a été construit pour prédire le nombre d'anophèles là où les données environnementales sont disponibles. Le type de validation croisé utilisé est le *Leave-one-out* [58] avec certaines particularités 3.1. Pour le vecteur des variables $X$ donné, les étapes suivantes sont répétées pour tous les sites (maisons) de 1 à 41. La fonction de perte utilisée est donnée par :

$$l(y, \phi(x)|\mathcal{D}_n) = \frac{|\hat{y} - y|}{|\hat{y} + 1|}.$$

où $\hat{y} = \phi(x)$. Si ce modèle est utilisé pour prédire les observations $y_{jk}$, $1 \leq k \leq 17$ de la $j^{eme}$ maison utilisant les données, environnementales de cette maison, l'erreur de prévision $E_{jk}$ sur chaque observation est donnée par :

$$E_{jk} = \frac{|\widehat{y}_{jk} - y_{jk}|}{|\widehat{y}_{jk} + 1|}$$

Cette erreur de prédiction a été retenue parmi celles classiques connues telles que : l'erreur absolue, erreur quadratique, etc. Aussi, le critère de positionnement utilisé est la médiane des erreurs de prévision et non la règle de prévision de l'équation 1.9.

**Algorithme 3.1** Leave-one-out niveau maison [3]

1: Les données sont séparées en au tant de blocs que de maisons.
2: On regroupe les blocs en deux sous-ensembles : l'ensemble d'apprentissage et l'ensemble de test.
3: Un modèle de régression du nombre d'anophèles collectés $y$ versus les variables environnementales $x$ est construit utilisant les observations de toutes les maisons sauf la maison numéro $i$ (c'est-à-dire qu'on exclut les observations de la $i^{eme}$ maison)
4: Ce modèle est utilisé pour prédire les observations $y_{jk}$, $1 \leq k \leq 17$ de la $j^{eme}$ maison utilisant les données environnementales de cette maison
5: L'erreur de prédiction est $E_{jk} = \frac{|\widehat{y}_{jk} - y_{jk}|}{|\widehat{y}_{jk}+1|}$
6: Les étapes (2, 3, 4 et 5) sont répétées jusqu'à prédiction et évaluation de l'erreur de prédiction pour toutes les 612 observations.
7: La moyenne des erreurs de prédiction est déterminée.

L'ensemble final de covariables retenu est celui qui minimise la médiane des erreurs de prédictions. Après cette étape de sélection, des termes d'interactions sont introduites et conservés dans le modèle si ils tendent à minimiser la médiane des erreurs de prédiction. La prédiction par ce modèle a été

améliorée en introduisant la variable "village" pour prendre en compte la possible corrélation dans les données à ce niveau. L'équation générale de ce modèle se met sous la forme

$$\ln[\mathbb{E}(Y_{ijk}|\beta)] = \beta_0 + \sum_{l=1}^{p} \beta_l X_{ijkl} \tag{3.3}$$

où $Y$ est le nombre d'anophèles collectés, $X$ est un $p$-vecteur des variables environnementales y compris la variable "village". Pour vérifier la capacité de ce modèle à faire de la prédiction, la comparaison entre les prédictions et les observations a été étudiée.

#### 3.2.1.2 Modèle pragmatique

Il a été aussi développé un modèle dit "pragmatique" qui permet d'estimer les paramètres ou fonctionnelles de la loi de probabilité de la réponse. Dans ce modèle, la prédiction de la mission de capture $k$, du site $j$ et du village $i$ est estimée par la moyenne du nombre d'anophèles collectés dans les trois autres maisons de ce même village durant la même mission de capture. Par exemple au cours de la troisième mission dans la ville Gbetaga, 4,7,7 et 26 anophèles ont été collectés sur les quatre sites respectivement. D'après ce modèle le nombre d'anophèles prédits pour les quatre sites est respectivement $(7 + 7 + 26)/3$, $(4 + 7 + 26)/3$, $(4 + 7 + 26)/3$ et $(4 + 7 + 7)/3$. Ainsi la comparaison de la distribution des erreurs de prédiction obtenues à partir de ces deux modèles (prédictif et pragmatique) a été faite selon la prédiction du nombre d'anophèles.

### 3.2.2 Résultats et discussion

#### 3.2.2.1 Vérification des hypothèses du modèle

##### 3.2.2.1.a La distribution de la variable d'intérêt conditionnellement aux variables explicatives et aux effets aléatoires

La figure 3.2 montre que tous les points sont proches de la première bissectrice mais au dessus. Nous pouvons déduire que les observations conditionnellement aux variables explicatives et aux effets aléatoires suivent une loi de Poisson avec une surdispersion. Ceci confirme l'information de l'équation (3.2) qui évoque une surdispersion dans les données.

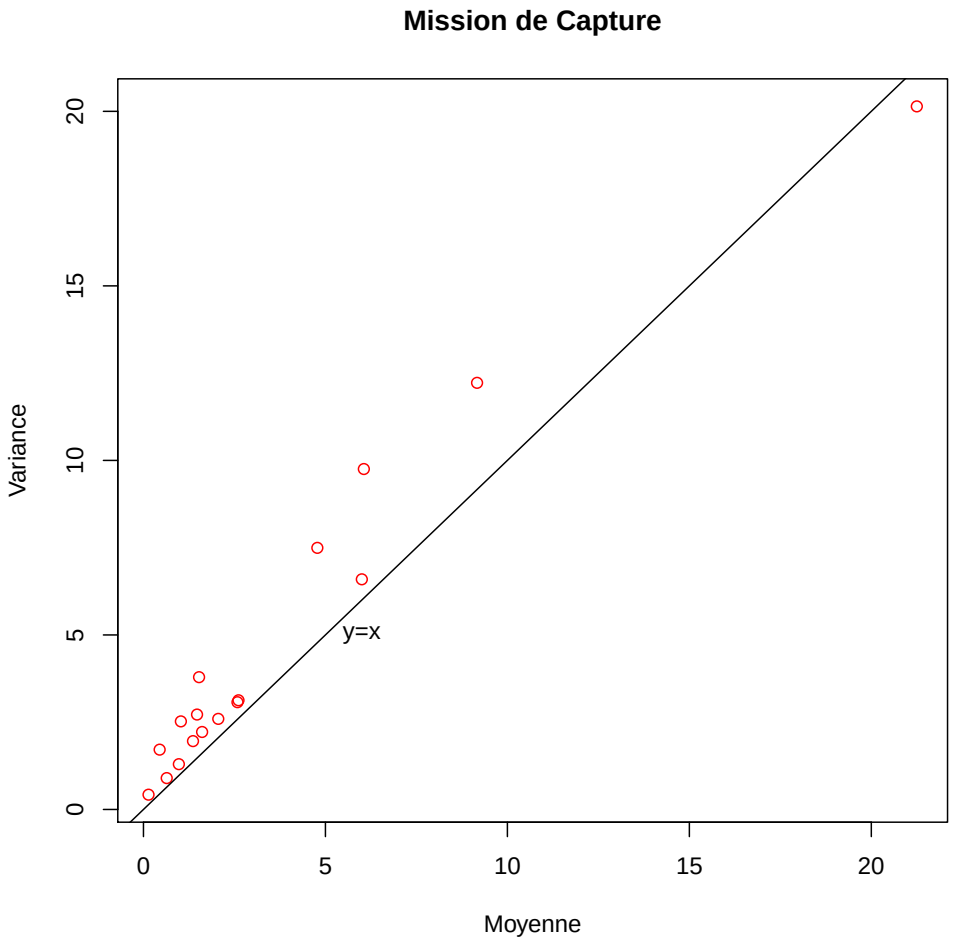



FIGURE 3.2 – Représentation de la variance des observations en fonction de la moyenne.

#### 3.2.2.1.b Normalité des aux effets aléatoires

L'une des hypothèses fondamentales du modèle est que les effets aléatoires suivent une loi normale conditionnellement aux variables explicatives. La figure 3.3 montre cette normalité.

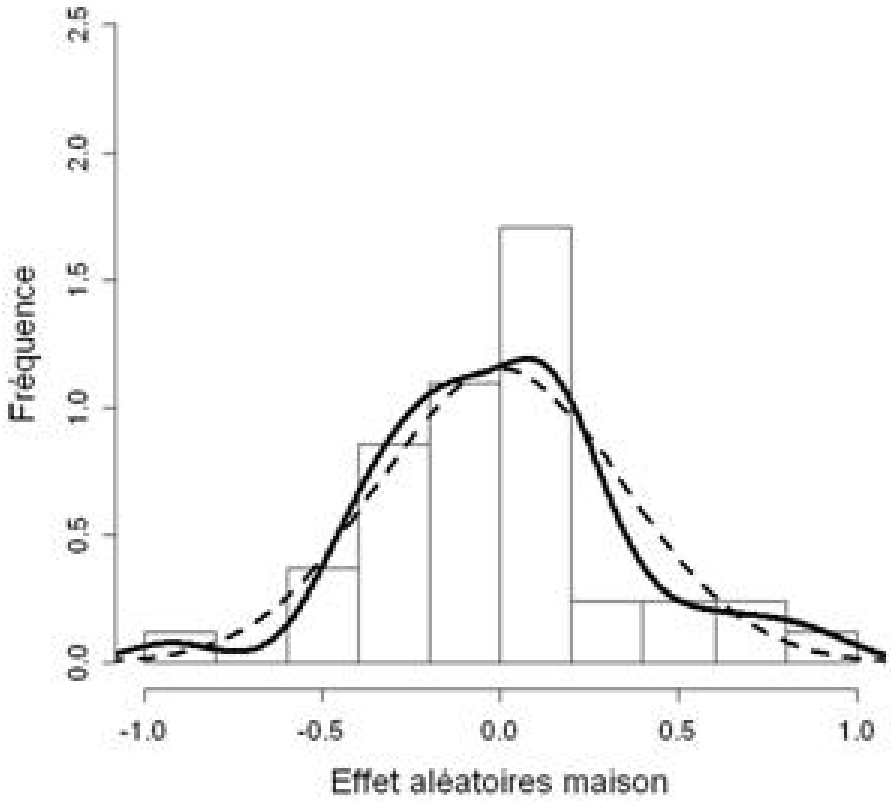



FIGURE 3.3 – **Distribution des BLUPs,**
*La ligne en pointillés représente la distribution théorique tandis que la ligne en trait plein représente la distribution empirique des BLUPs.*

#### 3.2.2.2 Estimation des paramètres de la distribution des effets aléatoires

##### 3.2.2.2.a Estimation de la matrice Z

Dans l'écriture matricielle du modèle tel que décrit dans l'equation (2.2), on a posé :

$$g[E(Y_{ijk}/a_i\,, b_j, \beta)] = X_{ijk}\beta + a_i + b_{ji} \tag{3.4}$$

On peut alors poser :

$$g[E(Y_{ijk}/a_i, b_{ji}, \beta)] = \eta_{ijk} = X_{ijk}\beta + a_i + b_{ji} \tag{3.5}$$

On a :

$$\eta_{ijk} = X_{ijk}\beta + Z_{ijk}\begin{pmatrix} a_i \\ b_{ji} \end{pmatrix} \text{ avec } Z_{ijk} = \begin{pmatrix} 1 & 1 \end{pmatrix} \tag{3.6}$$

où $Z_{ijk}$ est un vecteur $1 \times 2$.
Pour l'ensemble des 17 Captures d'une même maison i d'un même village j, on a :

$$\eta_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ij17} \end{pmatrix}\beta + Z_{ij}\begin{pmatrix} a_i \\ b_{ji} \end{pmatrix} \text{ avec } Z_{ij} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \tag{3.7}$$

où $Z_{ij}$ est une matrice $17 \times 2$.
Pour l'ensemble des captures de toutes les maisons d'un même village i, on a :

$$\eta_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ X_{i3} \\ X_{im_i} \end{pmatrix}\beta + Z_i\begin{pmatrix} a_i \\ b_{1i} \\ . \\ . \\ b_{m_i i} \end{pmatrix} \tag{3.8}$$

où $m_i$ est le nombre de maisons du village $i$.

$$\text{avec } X_{ij} = \begin{pmatrix} X_{i11} \\ X_{i12} \\ \vdots \\ X_{i17} \end{pmatrix} ; Z_i = \begin{pmatrix} A_{i1} \\ A_{i2} \\ A_{i3} \\ A_{im_i} \end{pmatrix} \text{ et } A_{ij} = \begin{pmatrix} & & & \textit{colonne } j & \\ 1 & 0 & . & 1 & 0 \\ 1 & 0 & . & 1 & 0 \\ . & . & . & . & . \\ 1 & 0 & . & 1 & 0 \end{pmatrix} \quad (3.9)$$

$A_{ij}$, $1 \leq j \leq m_i$ est une matrice composée de 1 sur la première et la j$^{ième}$ colonnes. Ainsi

$$Z_i = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ . & . & . & . & . \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ . & . & . & . & . \\ 1 & 0 & 1 & 0 & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ . & . & . & . & . \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (3.10)$$

Pour toutes les captures de toutes les maisons de tous les villages, la matrice Z est bloc diagonale et on a :

$$Z = \begin{pmatrix} Z_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Z_9 \end{pmatrix} \quad (3.11)$$

#### 3.2.2.2.b Estimation des paramètres de la distribution des effets aléatoires

Connaissant les paramètres des effets aléatoires, nous pouvons les simuler et les utiliser pour les prévisions.

### 3.2.2.3 Estimation de coefficients des effets fixes

Durant les 19 missions de capture entre juin 2007 et juillet 2009, au total 3074 vecteurs palustres ont été capturés (93,3% d'An. gambiae s.I et 6,7%

TABLE 3.1 – **Estimation des paramètres**

| **Effets aléatoires** | **Niveau** | **Variance** | **Erreur standard sur la variance** |
|---|---|---|---|
| Intercepts aléatoires | village | 0.71 | 0.19 |
| | maison | 0.21 | 0.11 |
| | capture | 1.04 | 0.06 |

TABLE 3.2 – **Estimation des effets fixes**

| **Effets fixes** | **Classes** | $\hat{\beta}$ | $\sigma_{\hat{\beta}}$ | p-**value** |
|---|---|---|---|---|
| Cours d'eau | Oui | - | - | |
| | Non | 1.869 | 0.63 | 0.003 |
| Sol | Humide | - | - | |
| | Sec | 2.27 | 0.72 | 0.002 |
| NDVI | Faible | - | - | |
| | Elevé | 0.46 | 0.23 | 0.05 |
| Saison | Fin saison sèche | - | - | $10^{-3}$ |
| | Début saison des pluies | 1.63 | 0.18 | |
| | Fin saison des pluies | 0.44 | 0.17 | |
| | Début saison sèche | 60.49 | 0.19 | |
| Quantité pluie | Faible | - | - | $10^{-3}$ |
| | Forte | 0.99 | 0.23 | |
| Nombre jours de pluie 10 jours avant la mission | [0,1] | - | - | $10^{-3}$ |
| | [2,4] | 0.34 | 0.17 | |
| | $> 4$ | 0.70 | 0.20 | |

d'An. funestus). le nombre médian de vecteurs capturés pour les 684 collectes (19 captures sur 4 sites de capture dans 9 villages), est dans l'intervalle [0–4], le nombre maximal est de 87. Deux quantités importantes seront étudiées : l'évolution dans la densité du vecteur définie comme le nombre de piqûres par personne et par nuit (m.a.) et la taux d'inoculation entomologique (EIR). La m.a donne le nombre moyen de piqûres d'anophèle par homme et par nuit tandis que l'EIR est le produit de la densité anophélienne (ma) par l'indice sporozoïtique (s). L'EIR donne la proportion de piqûre infectante (piqûres faites par des moustiques porteurs de sporozoïtes dans leur glande salivaire).

La donnée concernant la m.a est montrée par la figure 3.4. Cette figure met en exergue l'évolution spatiale et temporelle de la densité vectorielle. Les va-

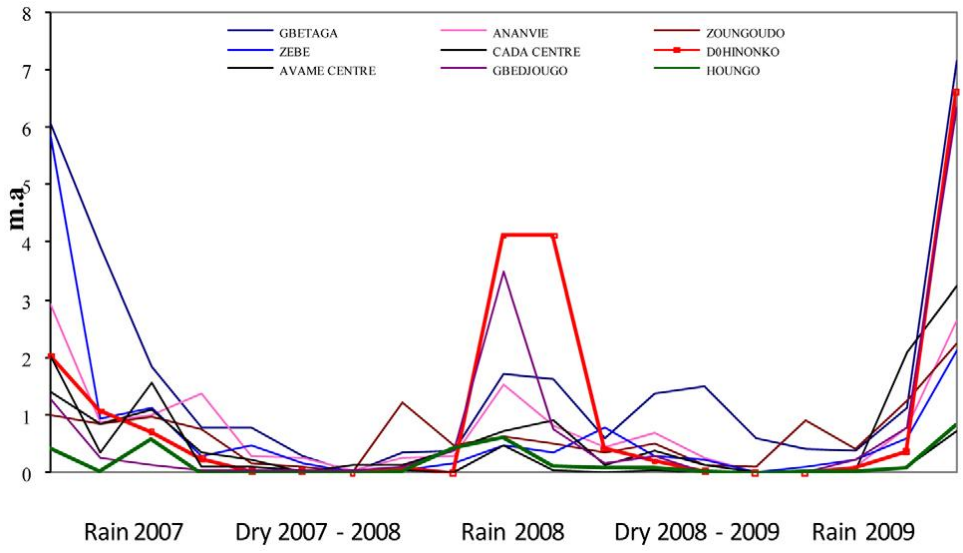



FIGURE 3.4 – **Nombre d'anophèles *gambiae* s.I. collectés par homme par jour dans les neuf villages pour chacune des 19 missions de capture.**

riations des m.a. dépendent des saisons et sont positivement associées avec la pluie. Les différences spatiales dans les m.a. sont bien observées entre les 9 villages particulièrement en saison des pluies (de juin à novembre), même à l'échelle du village. Il y a une grande différence dans les changements des m.a. entre les deux villages Houngo et Dohinonko qui sont à deux kilomètres l'un de l'autre. Le premier village montre une faible densité vectorielle durant l'étude et le second village une forte variation saisonnière avec une croissance substantielle en saison des pluies. Dans tous les villages, à l'exception de Houngo, on observe une différence marquée des m.a. entre les sites de capture, ce qui réflète une variation spatiale de la densité vectorielle au niveau des maisons figure 3.5,

Les analyses statistiques ont été conduites pour les 17 dernières missions de captures dans lesquelles au total 2292 vecteurs palustres ont été collectés. La table 3.2 montre le modèle explicatif multivarié final. Ce modèle comporte un effet aléatoire au niveau village, un effet aléatoire au niveau maison et un autre au niveau des missions de capture. Chaque effet aléatoire améliore la vraisemblance du modèle. La quantité moyenne de pluie missions de capture et le nombre de jours de pluie 10 jours avant la mission sont positivement correlés avec la densité vectorielle comme attendu. La saison est également correlée avec la densité vectorielle et ceci fortement en saison des pluies. Plusieurs caractéristiques des maisons sont aussi corrélées avec la densité vectorielle telle

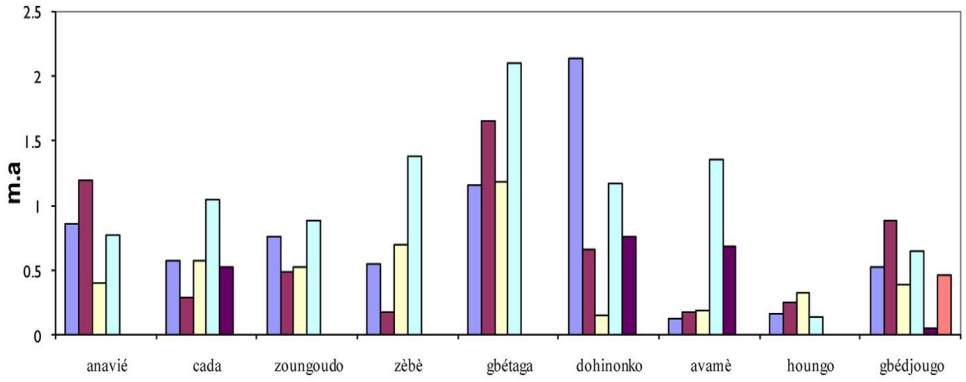



FIGURE 3.5 – **Moyenne des m.a dans les neuf villages**.
*Chaque barre représente la m.a. moyenne dans chaque maison dans le village correspondant.*

la proximité avec un cours d'eau, le sol sec, l'indice de végétation élevé. Tous ses résultats montrent la variabilité spatio-temporelle locale dans la transmission du paludisme. La figure 3.6 montre un bon ajustement entre les données collectées et les prédictions du modèle explicatif avec 3 effets aléatoires. Ainsi,

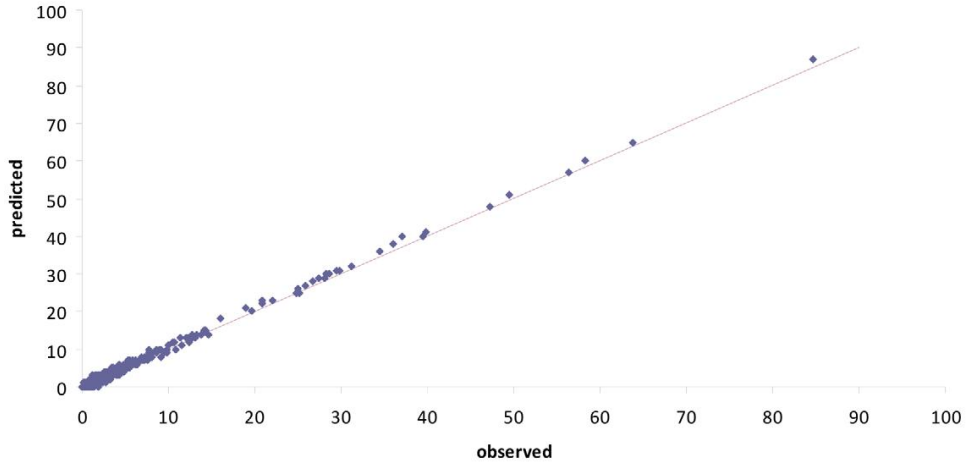



FIGURE 3.6 – **Relation entre le nombre d'anophèles collectés et prédits (modèle explicatif).**
*La ligne rouge représente la première bissectrice.*

là où les données entomologiques ne sont pas disponibles, le modèle prédictif peut utiliser les variables pour estimer le risque spatio-temporel entomologique dans une maison.
Le meilleur modèle prédictif contient les covariables suivantes : saison, quantité moyenne de pluie entre deux missions, nombre de jours de pluie 10 jours avant la mission. L'utilisation de répulsif, l'indice de végétation (NDVI) et l'interaction entre saison et le NDVI. La figure 3.7 montre la comparaison entre

les prédictions générées par le modèle de régression et les observations dans les 41 sites de capture. Le modèle s'ajuste bien aux données dans plusieurs

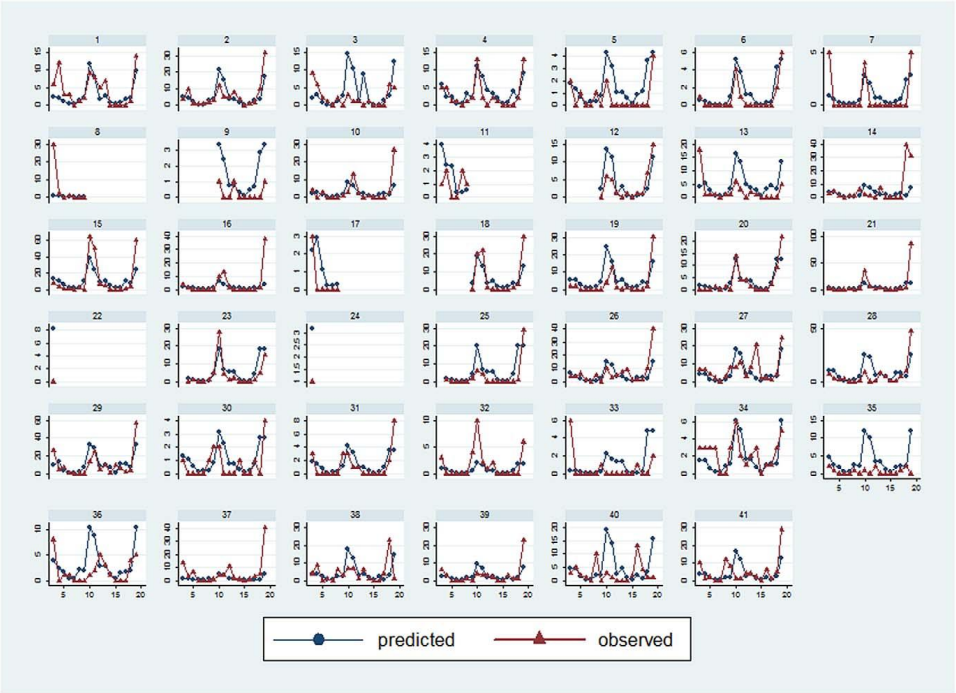



FIGURE 3.7 – **Comparaison entre les observations et les prédictions du modèle prédictif.**
*Sur chaque graphe, on observe en bleu les prédictions et en rouge les observations*

maisons mais pas toutes. La figure 3.8 montre la comparaison des erreurs de prédictions entre le modèle de régression et le modèle pragmatique suivant le nombre de vecteurs capturés. La distribution des erreurs de prédiction mais aussi le pouvoir prédictif des deux modèles sont sensiblement identiques.
Le nombre d'anophèles infectés est faible durant l'étude. Sagissant de l' (EIR), il est de 0.046 piqûre infectée par personne et par nuit. Lorsque l'EIR est utilisée comme variable dépendante au lieu de m.a., le modèle ne converge pas lorsque plusieurs variables sont introduites dans le modèle quoique l'EIR et la m.a. sont fortement correlés figure 3.9.

De plus, lorsque l'EIR est utilisée comme variable dépendante avec les variables climatiques, la quantité moyenne de pluie, le nombre de jours de pluie 10 jours avant la mission et la saison sont les seules variables indépendantes, la même structure est obtenue. Pour tout ceci seule la m.a. a été utilisée pour les analyses statistiques.

Pour la mise en œuvre de la méthode statistique, nous avons recodé certaines variables numériques en classe selon le Tableau 5.7. Ce recodage en un

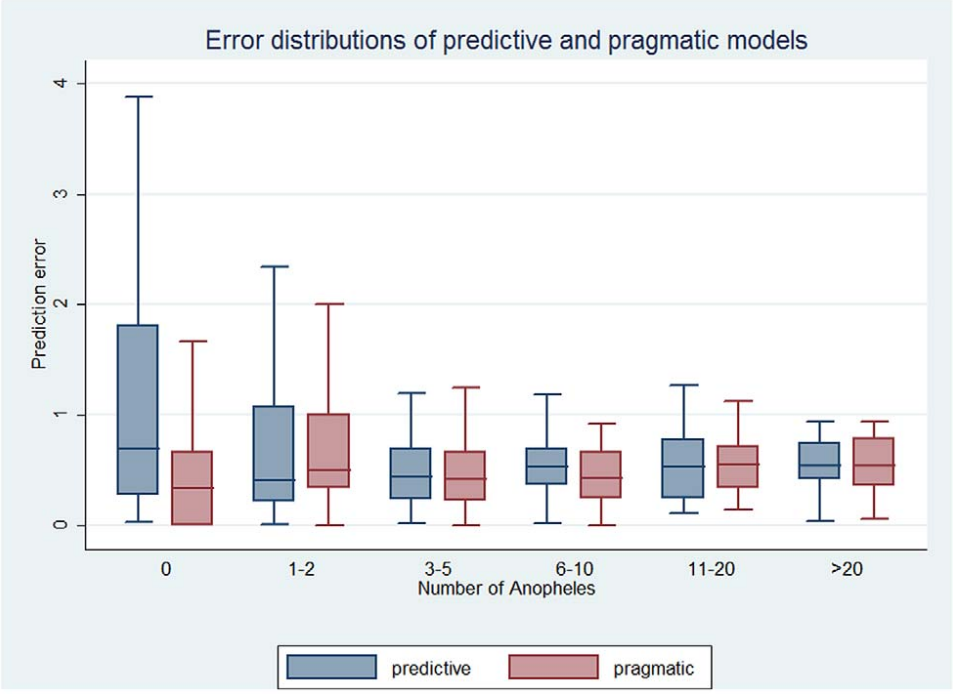



FIGURE 3.8 – **Distribution des erreurs du modèle pragmatique et du modèle prédictif selon le nombre d'anophèles observés.**,
*Dans chaque groupe (nombre d'anophèles), la boîte de gauche correspond au modèle de régression prédictif et celle de droite au modèle de régression pragmatique.*

nombre suffisant de classes présente le double avantage de s'affranchir de l'hypothèse de linéarité entre la variable à expliquer et la covariable, et de rendre les résultats facilement interprétables ; la contrepartie est la perte d'une partie de l'information contenue dans les variables, mais en catégorisant les variables sur au moins 3 ou 4 classes, nous pensons que le compromis est bon. Cette méthode est très souvent utilisée en épidémiologie.

D'une manière générale, les prédictions du modèle sont acceptables. En effet, globalement les valeurs observées sur le terrain et les prédictions du modèle ne présentent pas un grand écart entre elles. Néanmoins la qualité de l'ajustement du modèle aux valeurs observées est variable selon les villages, et à fortiori selon les maisons. Les écarts importants sont notés au niveau des petites valeurs et aussi des valeurs très grandes du nombre d'anophèles observés sur le terrain, figure 3.7. Il faut remarquer que le critère de convergence des estimateurs, le paramètre $c$ étant défini et fixé par le package, la convergence n'est pas assurée dans tous les cas. La convergence devient impossible lorsqu'on introduit certaines variables de type polynomiale ou certaines interactions entre les variables. Pour des variables explicatives continues variant

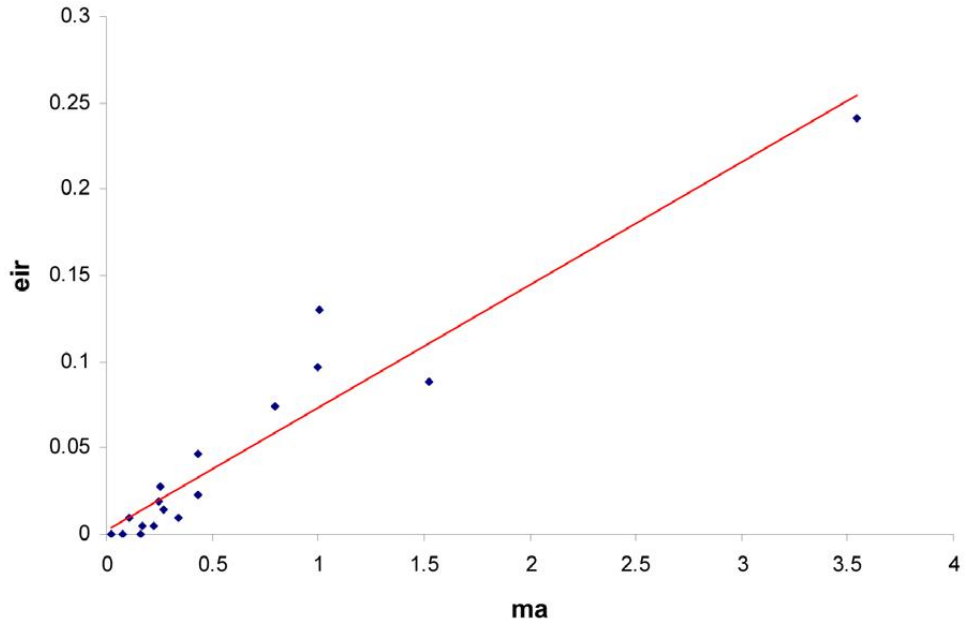



FIGURE 3.9 – **Relation entre la moyenne des m.a (le nombre de piqûres par personne et par nuit). et l'EIR**
*Sur l'axe des abscisses la moyenne des m.a. pour toutes les maisons durant une mission et en ordonnée la moyenne des EIR pour toutes les maisons durant la même mission.*

dans un petit intervalle, la convergence est presque impossible ceci à cause de la non inversibilité de la matrice $X'X$. L'algorithme utilisé combine plusieurs processus : l'algorithme IRLS, une pénalisation dont la matrice est évaluée suivant les données et l'approximation de Laplace pour le calcul des intégrales n'ayant pas une forme numérique définie. La méthode de Gauss-Hermite pourrait améliorer les calculs dans la précision mais elle est trop complexe et trop lente.

## 3.3 Conclusion

Ce chapitre nous a permis de parcourir les différents estimateurs les plus utilisés dans la sélection de variables, les différents types de pénalisation utilisées dans les modèles linéaires gaussiens, la méthode d'estimation des paramètres ainsi que les algorithmes utilisés dans un modèle GLMM. Les résultats obtenus sur les données montrent une variabilité spatio-temporelle dans les données entomologiques. La présence d'un effet aléatoire à chaque niveau d'hiérarchisation améliore la prédiction. Le meilleur ensemble de covariables pour la prédiction est composé de : saison, quantité moyenne de pluie entre deux missions, nombre de jours de pluie 10 jours avant la mission, l'utilisation de répulsif, l'indice de végétation (NDVI) et l'interaction entre saison et le NDVI.

Chapitre 4

# Construction d'une fonction de prévision par combinaison du GLM-Lasso et d'une double une validation croisée

## 4.1 Introduction

Les données de travail recueillies de nos jours répondent non seulement aux critères énumérés dans la section 2.1 mais sont aussi essentiellement en grande dimension. Il y a peu d'individus mais beaucoup d'informations sur chaque individu. Ces types de données ont succité la mise en place d'estimateurs à la fois stables, fiables dont l'interprétation est aisée. Un estimateur répondant à ces critères est l'estimateur LASSO (Least Absolute Shrinkage and Selection Operator). La nature de la variable d'intérêt et des variables explicatives permet de savoir quel type de Lasso ou quelle extension de cette méthode est adéquate pour ce qui nous intéresse. Cette méthode initiée par Tibshirani [59] a connu beaucoup d'extension de nos jours.

## 4.2 Méthode Lasso

### 4.2.1 Notions préliminaires

L'estimateur *LASSO* est introduit pour la première fois par Tibshirani [59]. Cet estimateur est défini comme un estimateur des moindres carrés sous contrainte de type $L_1$

$$\tilde{\beta}^L(t) = \begin{cases} arg\min_{\beta \in \mathbf{R}^p} ||Y - X\beta||_n^2 \\ \\ \text{s.c } \ ||\beta||_1 \leq t, \end{cases} \tag{4.1}$$

où $t$ est un paramètre réel positif. Cet estimateur a été déjà utilisé par Chen et Dohono dans le cas du traitement du signal [60] sous le nom de *Basis pursuit*. Sous cette forme, il est défini de la façon suivante :

$$\hat{\beta}^L(\lambda) = Arg \min_{\beta \in \mathbf{R}^p} ||Y - X\beta||_n^2 + \lambda ||\beta||_1 \tag{4.2}$$

Il faut noter que les deux estimateurs $\tilde{\beta}^L(t)$ et $\tilde{\beta}^L(\lambda)$ sont équivalents point par point. Pour un tout $t \in \mathbf{R}_+^*$ fixé, on peut trouver $\lambda$ qui va dépendre des données à traiter tel que $\tilde{\beta}^L = \hat{\beta}^L$. Egalement pour un $\lambda$ fixé, on peut trouver un $t$ tel que $\tilde{\beta}^L = \hat{\beta}^L$. Dans ce cas le réel positif $t$ est donné par $t = \sum_{k=1}^p |\hat{\beta}_k^L(\lambda)|$. Dans la suite nous supposerons que ces deux estimateurs sont équivalents, ils seront dénommés : estimateurs *Lasso* et notés $\hat{\beta}^L$.

### 4.2.2 Propriétes de l'estimateur Lasso

— Pour $\lambda = 0$, l'estimateur Lasso et celui des moindres carrés sont confondus. La méthode Lasso sélectionne toutes les variables explicatives sans exception.
Pour $\lambda = \infty$, le Lasso ne sélectionne aucune variable explicative, dans ce cas $\hat{\beta}^L = 0_p$.
Pour $\lambda \in ]0, \infty[$, le nombre de variables sélectionnées par le Lasso diminue lorsque $\lambda$ devient grand, c'est-à-dire, si $\lambda$ est grand, la contrainte exercée sur le vecteur $\beta$ l'est aussi.

— Dans le cas de l'inférence Bayésienne, l'estimateur Lasso est interprétable en terme d'estimation. Il peut être déterminé en considérant le modèle de régression linéaire avec un bruit gaussien et en supposant que le paramètre $\beta$ suit à priori la loi double exponentielle ou la loi de Laplace, c'est-à-dire $\beta_k$ admet pour densité à priori par rapport à la mesure de Lebesgue $\frac{\lambda}{2}\exp(-\lambda|\beta_k|)$.

— L'estimateur Lasso est linéaire par morceau si elle est une fonction de $\lambda$.

* Lorsque la matrice des variables explicatives $X$ est orthogonale (ou dans le cas trivial où p = 1), la résolution du problème posé par le Lasso revient à trouver la solution de $p$ problèmes de seuillage doux [61, 62]. Les composantes du vecteur $\hat{\beta}^L$ sont données par :

$$\hat{\beta}_k^L(\lambda) = sgn(\hat{\beta}_k^{OLS})(|\hat{\beta}_k^{OLS}| - \frac{\lambda}{2})^+, \quad \forall k \in \{1, \cdots, p\}, \tag{4.3}$$

où $(\gamma)^+ = (\gamma)_+ = max\{\gamma, 0\}$ pour tout réel $\gamma$ avec la fonction signe

définie par :

$$sgn(\gamma) = \begin{cases} 1 & \text{si } \gamma > 0 \\ 0 & \text{si } \gamma = 0 \\ -1 & \text{si } \gamma < 0 \end{cases} \tag{4.4}$$

Ainsi pour $\lambda^{(k)} = |\hat{\beta}_k^{OLS}|, k \in \{1, \cdots, p\}$, on voit aisément que la linéarité change.

* En considérant un peu plus généralement le cas où $X$ n'est pas orthogonale et que $p \neq 1$ la linéarité par morceau a été bien établie par Hastie, Johnstone et Tibshirani [49, 63]. Les valeurs du paramètre de régularisation qui changent la linéarité peuvent être définies en considérant les conditions d'optimalité de l'équation (4.42) comme des conditions de premier ordre ou condition de *Karush-Kuhn-Tucker (KKT)*.

Utilisant l'ensemble des variables actives $\mathcal{A}_L(\lambda)$ de l'estimateur $\hat{\beta}^L$ pour une valeur de $\lambda$, nous pouvons écrire les conditions $(KKT)$ comme suit :

$$\begin{cases} 2X_j^T(Y - X\hat{\beta}^L(\lambda)) = \lambda sgn(\hat{\beta}_j^L) & \forall j \in \mathcal{A}_L(\lambda) \\ \\ 2\left|X_j^T(Y - X\hat{\beta}^L(\lambda))\right| < \lambda & \forall j \notin \mathcal{A}_L(\lambda) \end{cases} \tag{4.5}$$

ou tout simplement : $||X^T(Y - X\hat{\beta}^L(\lambda)||_\infty \leq \frac{\lambda}{2}$, avec $||\alpha||_\infty = \sup_k(\alpha_k)$. Dans ce cas l'estimateur est linéaire par morceau et nous pouvons trouver les $p$ valeurs de $\lambda$ qui changent cette linéarité.
Posons : $\varepsilon(\lambda) = X_k^T(Y - X\hat{\beta}^L(\lambda))$ , $\forall\ k \in \{1, \cdots, p\}$. Lorsque $k$ est l'indice d'une variable de coefficient nul, c'est-à-dire $X_j$ est inactive ($j \notin \mathcal{A}_L$), nous pouvons trouver un coefficient $\lambda^{(i)}$ tel que $|\varepsilon(\lambda)| = \frac{\lambda^{(i)}}{2}$. En posant $\lambda = \lambda^{(i)}$, nous avons alors une saturation de la contrainte imposée à la variable $X_k$ et de ce fait elle passe dans l'ensemble $\mathcal{A}_L$, elle devient une variable active parce que son coefficient $\beta_k$ devient non nul. L'algorithme du Lasso est lent dans son exécution et coûteux en mémoire. Il est possible d'approcher les solutions par l'algorithme LARS introduit par Efron [49]. C'est un algorithme d'homotopie qui permet par simple modification de calculer toutes les solutions $\hat{\beta}^L(\lambda)$ du Lasso pour tout réel positif $\lambda$. L'algorithme repose sur la construction des estimateurs $\hat{\beta}^L(\lambda)$ sur la corrélation entre les variables $\{X_1, \cdots, X_P\}$ et la quantité résiduelle. Le LARS construit les estimateurs $X\hat{\beta}^L$ de $X\beta^L$ en un nombre $k$ d'étapes successives. A chaque étape il ajoute une variable de l'ensemble des variables du modèle de telle manière qu'après juste $k$ étapes que les coefficients $\hat{\beta}_L(\lambda)$ soient non nuls. Mais il faut remarquer que lorsqu'il y a une forte corrélation alors l'algorithme du LARS échoue dans la construction des estimateurs $X\hat{\beta}^L$. Dans un groupe de variables corrélées, tout comme le Lasso, il sélectionne de manière arbitraire une seule variable en écartant les

autres du groupe.
Beaucoup de travaux ont montré que l'estimateur Lasso possède des limites en théorie et en pratique. L'estimateur Lasso fait intervenir la matrice Gram.

**La matrice de Gram**

**Définition 14** *On appelle matrice de Gram associée à $X$, la matrice définie par :*

$$\psi^n = \{\psi^n_{j,k}\}_{1\leq j,k\leq p} = \frac{X^T X}{n}$$

$\mathcal{A}$ étant considéré comme l'ensemble des variables actives, la restriction matrice de Gram aux lignes et aux colonnes dont les indices sont des éléments de $\mathcal{A}$ est donnée par :

$$\psi^n_{\mathcal{A},\mathcal{A}} = \frac{X^T_{\mathcal{A}} X_{\mathcal{A}}}{n}$$

On définit également :

$$\psi^n_{\mathcal{A}^c,\mathcal{A}} = \frac{X^T_{\mathcal{A}^c} X_{\mathcal{A}}}{n}$$

Notons que $\psi^n_{\mathcal{A}^c,\mathcal{A}}$ est de dimension $|\mathcal{A}^{\mathcal{C}}| \times |\mathcal{A}|$ où $|A|$ est le cardinal de l'ensemble $A$. L'ensemble des résultats sur l'estimateur $X\hat{\beta}^L$ fait intervenir des restrictions sur la matrice de Gram ce qui impose une faible corrélation entre les variables. Tout ceci limite le champ d'application de l'estimateur Lasso [64]. L'estimateur Lasso ne peut pas prendre en compte certaines informations à priori sur les variables telles que le niveau de corrélation. L'estimateur Lasso possède de bonnes propriétés dans le cadre supervisé, par contre il n'est pas adapté dans le cadre semi-supervisé ou transductif [65]. Les détails sur les résultats théoriques, les limites et critiques sur l'estimateur Lasso ont été largement discutés par M. Hebiri [46]. Les limites théoriques et techniques de l'estimateur Lasso ont été à l'origine de l'extension et de la généralisation de la méthode Lasso. Nous allons présenter quelques unes de ces extensions et généralisations.

### 4.2.3 Propriétés Oracles du Lasso

Soit $\delta$ une procédure de sélection de variables et soit $\beta(\delta)$ l'estimateur des coefficients produit par cette procédure. Selon Fan et Li (2001)[66], la procédure $\delta$ sera dite *Procédure Oracle* si elle possède asymptotiquement les propriétés suivantes :

— Identifier le vrai ensemble de variables actives $\mathcal{A} = \{j : \hat{\beta}_j \neq 0\}$, le support de coefficient $\hat{\beta}$ pour chaque valeur du paramètre de régularisation, on parle de **sélection**.

— Estimer la vraie valeur de $\hat{\beta}$, $\sqrt{n}(\hat{\beta}(\delta)_{\mathcal{A}} - \beta^*_{\mathcal{A}}) \longrightarrow_d \mathbf{N}(0, \Sigma^*_{\mathcal{A}})$
à la vitesse de convergence raisonnable $\sqrt{n}$, on parle **d'estimation**.
— Donner une bonne approximation de $X\hat{\beta}^*$, on parle de **prédiction**.

Des travaux ont été faits sur certains types de données observées et sous certaines conditions l'estimateur Lasso produit dans ces cas possède ces propriétés Oracles.

Bunea et *al* ont étudié les propriétés oracles du problème d'optimisation des moindres carrés avec pénalité $L_1$ dans le cas de la régression non paramétrique avec un design aléatoire [67]. Il a été prouvé que l'estimateur Lasso obtenu satisfait à certaines inégalités oracles et que ces résultats sont également valides en grande dimension aussi dans le cas où la matrice de régression n'est pas définie positive.

Sampson et *al*, ont montré qu'une procédure peut être Oracle, c'est-à-dire produire des estimateurs possédant des propriétés oracles sans être optimale [68]. Ils ont prouvé qu'il existe un certain taux des coefficients nuls, c'est-à-dire des variables non actives que les procédures oracles telles que l'Adaptative Lasso utilisant les paramètres de lissages oracles n'arrivent pas à optimiser. Au delà de ce taux, les procédures peuvent produire des estimateurs avec des propriétés oracles mais l'optimalité n'est pas assurée.

Van de Geer et *al*, par leur travail, ont permis d'établir des inégalités oracles sous certaines conditions sur la matrice de design, des conditions de restriction de Bickel sur les valeurs propres, des conditions de compatibilité faible de Van de Geer et *al* [69].

Kwemou dans le cadre de régression logistique utilisant la procédure lasso ou Group lasso [70], a établi des inégalités oracles sur les estimateurs d'une fonction approchée de manière sparse par combinaison linéaire d'éléments pris dans un dictionnaire de fonctions de base. Ces résultats sont non asymptotiques et sont obtenus sous des conditions de restriction sur les valeurs propres.

### 4.2.4 Extensions, généralisation et variantes du Lasso

Le Lasso présente plusieurs extensions et plusieurs généralisations.

#### 4.2.4.1 Non negative Garrote

La méthode Lasso a été fortement inspirée par l'estimateur *Non negative Garrote (NNG)* de Breiman [71]. Posons $K = (K_1, \cdots, K_p)$, $K_j = X_j\hat{\beta}^{OP}_j$ où $\hat{\beta}^{OP}_j$ est un estimateur choisi arbitrairement pour $\beta$. Posons :

$$\hat{g}(\lambda) = (\hat{g}_1(\lambda), \cdots, \hat{g}_p(\lambda))$$

avec

$$\begin{cases} \hat{g}(\lambda) = Arg\min_{\beta\in\mathbf{R}^p} \frac{1}{2n}||Y - Kg||^2 + \lambda\sum_{j=1}^{p} g_j \\ \\ \text{s.c } \ g_j \geq 0, \quad j \in \{1, \cdots, p\} \end{cases} \tag{4.6}$$

L' estimateur NNG est défini par :

$$\hat{\beta}_j^{NNG} = \hat{g}_j(\lambda)\hat{\beta}_j^{OP}, \quad j \in \{1, \cdots, p\} \tag{4.7}$$

Dans le cas orthogonale, c'est-à-dire $X^TX = I_p$ alors on a :

$$g_j(\lambda) = \left(1 - \frac{\lambda}{(\hat{\beta}_j^{OP})^2}\right)_+, \quad j \in \{1, \cdots, p\} \tag{4.8}$$

Dans ses travaux originaux, Breiman a considéré le cas où : $\hat{\beta}^{OP} = \hat{\beta}^{OLS}$ mais il est toujours possible de donner d'autres valeurs à l'estimateur $\hat{\beta}^{OP}$. Yuan et Lin [72] ont considéré l'estimateur $\hat{\beta}^{OP}$ égale à l'estimateur Ridge, c'est-à-dire

$$\hat{\beta}^{OP} = \hat{\beta}^{R} = (X^TX + \lambda I_p)^{-1}X^TY$$

et ils ont proposé un algorithme de type LARS pour approcher l'estimateur NNG. Ils sont arrivés à prouver que cette méthode est consistante en estimation et sélection de variables avec $p \leq n$

#### 4.2.4.2 La méthode SCAD

Fan et Li ont proposé une fonction pénalité non concave dénommée *SCAD (smoothly clipped absolute deviation)* [66]. Cette pénalité est définie par :

$$[66]P_\lambda^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{si} \quad |\beta_j| \leq \lambda \\ \\ -\left(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right) & \text{si} \quad \lambda < |\beta_j| \leq a\lambda \\ \\ \frac{(a+1)\lambda^2}{2} & \text{si} \quad |\beta_j| > a\lambda \end{cases} \tag{4.9}$$

Ce qui correspond à une pénalité quadratique des splines aux nœuds $\lambda$ et $a\lambda$. On peut constater que cette fonction de pénalité est continue dérivable et dont la dérivée première se met sous la forme :

$$P_\lambda^{'}(\beta) = \lambda\{I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda}I(\beta > \lambda)\} \tag{4.10}$$

avec $\beta > 0$ et $a > 2$.
Cette pénalité est continue, différentiable sur $]-\infty, 0[$ et sur $]0, +\infty[$ mais

elle présente une singularité en 0 et de dérivée nulle en dehors de l'intervalle $[-a\lambda, a\lambda]$. Ceci fait que le SCAD annule les petits coefficients et garde les grands tels qu'ils sont. De ce fait l'ensemble de sparsité du SCAD est raisonnable et les coefficients élevés sont non biaisés. L'estimateur SCAD se met sous la forme :

$$\hat{\beta}_j^{SCAD} = \begin{cases} (|\hat{\beta}_j| - \lambda)_+ sgn(\hat{\beta}_j) & \text{si} \quad |\hat{\beta}_j| \leq 2\lambda; \\ [(a-1)\hat{\beta}_j - sgn(\hat{\beta}_j)a\lambda]/a - 2 & \text{si} \quad 2\lambda < |\hat{\beta}_j| \leq a\lambda; \\ \hat{\beta}_j & \text{si} \quad |\beta_j| > a\lambda \end{cases} \tag{4.11}$$

Cette méthode présente deux paramètres $(\lambda, a)$ qui peuvent être déterminés par la méthode de validation croisée. Il faut savoir que la complexité de l'algorithme fera que les calculs seront difficiles et très coûteux en temps. Après étude sur beaucoup de cas, les auteurs proposent $a = 3.7$ et ainsi les utilisateurs peuvent déterminer $\lambda$ par validation croisée.

#### 4.2.4.3 Elastic net

La méthode *Elastic net* a été introduite par H. ZOU et T. Hastie [73], ceci à cause de certaines insuffisances de la méthode Lasso. Les auteurs sont partis de trois constats :

- En grande dimension $p \gg n$, le Lasso sélectionne au plus $n$ variables et arrive à saturation à cause de la convexité de son problème d'optimisation. Ce qui limite fort bien cette méthode de sélection de variables.
- Si il existe un groupe de variables à l'intérieur duquel il a une forte corrélation, le Lasso à tendance à sélectionner une seule des variables et ne s'occupe pas de celle qui est sélectionnée. Ce fait écarte les autres variables de l'étude.
- En dimension faible $n > p$, si il a une forte corrélation entre les prédicteurs, de manière empirique on peut constater que les performances en prédiction du Lasso sont dominées par celles de la régression Ridge.

En se basant sur ces constats, les auteurs proposent dans un premier temps la méthode dite *Naïve Elastic Net (NEN)*. Cette méthode veut que la variable réponse soit centrée et que les prédicteurs soient *Standardisés*. Ce qui donne :

$$\begin{cases} \sum_{i=1}^n y_i = 0 \\ \sum_{i=1}^n x_{ij} = 0, \quad \text{pour} \;\; j = 1, \cdots, p \\ \sum_{i=1}^n x_{ij}^2 = 1 \end{cases} \tag{4.12}$$

Le problème d'optimisation de la méthode $NEN$ se pésente comme suit :

$$\hat{\beta}^{NEN}(\lambda_1, \lambda_2) = Arg \min_{\beta \in \mathbf{R}^p} \left[ ||Y - X\beta||_2^2 + \lambda_2 \sum_{i=1}^{p} \beta_j^2 + \lambda_1 \sum_{i=1}^{p} |\beta_j| \right] \tag{4.13}$$

Ce qui peut être reécrit simplement :

$$\hat{\beta}^{NEN}(\lambda_1, \lambda_2) = Arg \min_{\beta \in \mathbf{R}^p} \left[ |Y - X\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1 \right] \tag{4.14}$$

En posant $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, le problème (4.14) revient à :

$$\begin{cases} \hat{\beta}^{NEN}(\lambda_1, \lambda_2) = Arg \min_{\beta \in \mathbf{R}^p} |Y - X\beta|^2 \\ \\ \text{s.c } \ (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leq t, \quad t \in \mathbf{R} \end{cases} \tag{4.15}$$

Dans l'équation (4.15), la quantité $(1-\alpha)|\beta|_1 + \alpha|\beta|^2$ est la pénalité de la méthode Naive Elastic Net. On constate aisément que :

— si $\alpha = 0$, la méthode NEN devient la méthode lasso (4.42)

— si $\alpha = 1$, la méthode NEN devient la méthode Ridge

Dans l'equation (4.13) chaque paramètre joue un rôle précis. Le paramètre $\lambda_1$ assure la sparsité et le paramètre $\lambda_2$ permet de mesurer la corrélation entre les variables. Dans le cas orthogonale, on arrive à montrer qu'on peut exprimer les estimateurs NEN, Lasso et Ridge en fonction de l' estimateur OLS. On a :

$$\hat{\beta}^{NEN}(\lambda_1, \lambda_2) = \frac{(|\hat{\beta}^{OLS}| - \lambda_1/2)_+}{1 + \lambda_2} sgn(\hat{\beta}^{OLS}) \tag{4.16}$$

$$\hat{\beta}^{Ridge}(\lambda_2) = \hat{\beta}^{OLS}/(1 + \lambda_2) \tag{4.17}$$

$$\hat{\beta}^{Lasso}(\lambda_1) = (|\hat{\beta}^{OLS}| - \lambda_1/2)_+ sgn(\hat{\beta}^{OLS}) \tag{4.18}$$

où $\hat{\beta}^{OLS} = X^T Y$. Les auteurs se sont rendus compte que le NEN rencontre les mêmes problèmes que le Lasso en sélection de variables. De plus, l'introduction des deux constantes $\lambda_1$ et $\lambda_2$ dans le problème d'optimisation 4.13 n'arrange pas bien les choses. Il faut d'abord tourner l'algorithme pour $\lambda_2$ et ensuite le reprendre pour $\lambda_1$. Ceci ne fera qu'augmenter la variance et le compromis entre le biais et la variance ne peut plus être obtenu. Ainsi, les auteurs proposent une nouvelle méthode appelée *Elastic net* qui s'écrit comme suit :

$$\hat{\beta}^{EN}(\lambda_1) = (1 + \lambda_2)\hat{\beta}^{NEN}(\lambda_1, \lambda_2) \tag{4.19}$$

Il a été prouvé que cet estimateur est meilleur en sélection de variables comparé à la méthode Lasso et la méthode Ridge [73]. Li Qing et *al* ont introduit en 2010 la version bayésienne de la méthode EN [74]. Si on suppose que $(Y|X, \beta) \sim$

$N(X\beta, \sigma^2 I_n)$, l'estimateur *Bayesian Elastic Net (BEN)* est le mode marginale à posteriori de la distribution de $(\beta|Y)$ sachant que sa distribution à priori est $\pi(\beta)$ telle que :

$$\pi(\beta) \propto \exp\{-\lambda_1||\beta||_1 - \lambda_2||\beta||_2^2\} \tag{4.20}$$

où $\exp\{\alpha\}$ est la loi exponetielle de paramètre $\alpha$.
De plus on a : $\pi(\beta|\sigma^2)$ telle que :

$$\pi(\beta|\sigma^2) \propto \exp\{-\frac{1}{2\sigma^2}(\lambda_1||\beta||_1 + \lambda_2||\beta||_2^2)\} \tag{4.21}$$

Ainsi on a :

$$(\beta|\sigma^2) \sim \exp\left\{-\frac{1}{2\sigma^2}(\lambda_1||\beta||_1 + \lambda_2||\beta||_2^2)\right\} \tag{4.22}$$

En prenant $\sigma^2 = 1$, $p = 1$ et en prenant des valeurs judicieuses pour le couple $(\lambda_1, \lambda_2)$ l'estimateur BEN conduit aux estimateurs Lasso, Ridge et Elastic Net. Les travaux de De Mol, De Vito, Rosaco ont montré la consistance en sélection de la méthode EN [75].

#### 4.2.4.4 Fused Lasso

ZHOU et *al* ont proposé en 2005 la méthode du *Fused Lasso* [76]. L'estimateur Fused Lasso se met sous la forme :

$$\hat{\beta}^{FL}(\lambda_1, \lambda_2) = Arg \min_{\beta \in \mathbf{R}^p} \left[ ||Y - X\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2 \sum_{i=1}^{p} |\beta_j - \beta_{j-1}| \right] \tag{4.23}$$

$\lambda_1$ et $\lambda_2$ sont les paramètres de régularisation. La quantité $\lambda_2 \sum_{i=1}^{p} |\beta_j - \beta_{j-1}|$ est appelée pénalité de fusion.
Les travaux de Rinaldo imposant des contraintes supplémentaires aux variables par block ont permis à la version améliorée du Fused Lasso d'être consistante en sélection [77].

#### 4.2.4.5 Group Lasso

La sélection de variables par la méthode Lasso donnant la possiblité de regrouper certaines variables en groupe et de leur appliquer une même pénalité a été proposée par Yuan et Li [78]. Elle utilise une contrainte de type $L_2$ mais la pénalité est aplliquée à chaque groupe de variables fixé. Le problème d'optimisation se présente comme suit :

$$\hat{\beta}^{GL}(\lambda) = Arg \min_{\beta \in \mathbf{R}^p} \left( ||y - 1 \sum_{l=1}^{L} X_l \beta_l||_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l}||\beta_l||_2 \right) \tag{4.24}$$

où $\sqrt{p_l}$ est le nombre de variables explicatives du groupe et $||\ .\ ||_2$ est la norme euclidienne. Cette pénalité fonctionne de la même manière que le Lasso mais au niveau de chaque groupe. Si la longueur de chacun des groupes est 1 alors l'estimateur Group Lasso est confondu à l'estimateur Lasso. Cette méthode a été étendue au modèle de régression logistique par Meir et *al* [79].
Il faut remarquer que cet algorithme ne produit pas de sparsité à l'intérieur d'un groupe. Si dans un groupe de variables, une seule variable a son coefficient non nul alors toutes les autres variables ont leurs coefficients non nuls également. De même si une seule variable a son coefficient nul alors toutes les autres variables ont leurs coefficients nuls également. De plus à l'intérieur de chaque groupe, la matrice $X_l$ doit être orthogonale, ce qui n'est pas toujours évident.
Cette méthode a été améliorée par Friedman et *al* en 2010 par le *sparse group Lasso* [80].
Il permet de sélectionner des groupes et de sélectionner des variables à l'intérieur de chaque groupe en mettant un terme de pénalité sur les groupes et un autre sur chacune des variables. Il peut être vu comme une combinaison entre le *Group Lasso* et le Lasso. Et dans ce cas la condition d'orthogonalité est nécessaire sur les matrices de groupe. L'estimateur du *Sparse grouped Lasso* s'ecrit :

$$\hat{\beta}^{SGL}(\lambda) = Arg \min_{\beta \in \mathbf{R}^p} \left( ||y - \sum_{l=1}^{L} X_l \beta_l||_2^2 + \lambda_1 \sum_{l=1}^{L} \sqrt{p_l} ||\beta_l||_2 + \lambda_2 \sum_{l=1}^{L} ||\beta||_1 \right) \tag{4.25}$$

où $\beta = (\beta_1, \cdots, \beta_l, \cdots, \beta_L)$ avec $\beta_l$ le vecteur des coefficients du groupe $l$. Pour des raisons de commodité on pourra écrire simplement :

$$\hat{\beta}^{SGL}(\lambda) = Arg \min_{\beta \in \mathbf{R}^p} \left( ||y - \sum_{l=1}^{L} X_l \beta_l||_2^2 + \lambda_1 \sum_{l=1}^{L} ||\beta_l||_2 + \lambda_2 \sum_{l=1}^{L} ||\beta||_1 \right) \tag{4.26}$$

La résolution du problème (4.24) conduit à la résolution des équations de la forme :

$$- X_l^T (y - \sum_{l=1}^{L} X_l \beta_l) + \lambda s_l = 0; l = 1, \cdots, L \tag{4.27}$$

avec

$$\begin{cases} s_l = \beta_l / ||\beta_l|| & \text{si } \beta_l \neq 0 \\ ||s_l||_2 < 1 & \text{sinon} \end{cases} \tag{4.28}$$

Dans ce cas les estimateurs Group Lasso sont :

$$\hat{\beta}_l = \begin{cases} 0 & \text{si } ||X_l^T(y - \sum_{k\neq l} X_k\hat{\beta}_k)|| < \lambda \\ \\ (X_l^T X_l + \lambda/||\hat{\beta}_l||)^{-1} X_l^T(y - \sum_{k\neq l} X_k\hat{\beta}_k) & \text{sinon} \end{cases} \tag{4.29}$$

On pose $r_l = (y - \sum_{k\neq l} X_k\hat{\beta}_k)$ pour simplifier la notation. Dans le cas d'orthogonalité c'est-à-dire $X_l^T X_l = I$ avec $s_l = X_l^T r_l$, on obtient tout simplement :

$$\hat{\beta}_l = (1 - \lambda||s_l||)s_l$$

Ce qui conduit à un algorithme et une procédure de descente de coordonnées par block.

#### 4.2.4.6 Adaptative Lasso

La version adaptative du Lasso a été introduite par H. Zou [81]. Il se définit comme suit :

$$\hat{\beta}^{(n),AdL}(\lambda_n) = Arg\min_{\beta\in\mathbf{R}^p}\left[||Y - X\beta||_n^2 + \lambda_n\sum_{j=1}^{p}\omega_j|\beta_j|\right] \tag{4.30}$$

où $\omega_j$ est un vecteur de poids inconnu mais dépendant des données. L'objectif à atteindre par l'auteur est d'obtenir un estimateur qui possède les mêmes propriétés en sélection, en estimation et en prédiction que le Lasso. Mieux encore cet estimateur doit posséder des propriétés oracles. Pour obtenir un tel estimateur, on procède ainsi :

1. On choisit $\hat{\beta}$ un estimateur consistant et toutes les composantes sont non nulles, c'est-à-dire
$$\hat{\beta}_i \neq 0 \quad \forall i \in \{1, \cdots, p\}$$
Par exemple $\hat{\beta} = \hat{\beta}^{(OLS)}$ ou $\hat{\beta} = \hat{\beta}^{(ER)}$
2. On choisi $\gamma > 0$ et on définit le vecteur des poids par :
$$\hat{\omega} = \frac{1}{|\hat{\beta}|^\gamma}$$
3. L'estimateur Adaptative lasso s'écrit :
$$\hat{\beta}^{(n),AdL}(\lambda_n) = Arg\min_{\beta\in\mathbf{R}^p}\left[||Y - X\beta||_n^2 + \lambda_n\sum_{j=1}^{p}\frac{1}{|\hat{\beta}_j|^\gamma}|\beta_j|\right] \tag{4.31}$$

Le poids $\hat{\omega}_j$ utilisé ici permet de contrôler la pénalisation appliquée au coefficient $\beta_j$. L'auteur arrive à prouver que pour $\lambda_n$ bien choisi, l'estimateur $\hat{\beta}^{(n),AdL}$ possède les propriétés oracles sous certaines conditions.
Supposons que :
$\lambda_n/\sqrt{n} \longrightarrow 0$ et $\lambda_n n^{(\gamma-1)} \longrightarrow \infty$. alors on a :

- La méthode Adaptative Lasso est consistante en sélection de variables : $\lim_{n\longrightarrow\infty} P(\mathcal{A}^* = \mathcal{A})^* = 1$
- L'estimateur de l'Adaptative Lasso converge normalement vers le bon estimateur : $\sqrt{n}(\hat{\beta}^{(n),AdL}_{\mathcal{A}} - \hat{\beta}^*_{\mathcal{A}}) \longrightarrow_d \mathcal{N}(0, \sigma^2 \times C_0)$ où $C_0$ est une matrice de taille $p_0 \times p_0$

avec
$\mathcal{A}^* = \{j : \beta^*_j \neq 0\}$, $\mathcal{A}_n = \{j : \hat{\beta}^{(n),AdL} \neq 0\}$, $\mathbf{E}[Y|X] = X\beta^*$, $|\mathcal{A}^*| = p_0$.
Le problème de l'estimateur Adaptative Lasso dans les équations (4.30 et 4.31) est un problème d'optimalité essentiellement convexe avec une contrainte de type $L_1$ donc peut être résolu de manière efficace par l'algorithme du Lasso.

#### 4.2.4.7 Dantzig selector

Cette méthode a été introduite par Candes et Tao [82]. On suppose que : $Y = X\beta + \varepsilon$,
$\varepsilon$ est $i.i.d$, $\varepsilon \sim N(0, \sigma^2)$ et que les données sont en grande dimension, c'est-à-dire $p \gg n$. Cet estimateur noté $\hat{\beta}^{DS}(\lambda)$ est défini par :

$$\hat{\beta}^{DS}(\lambda) = \begin{cases} Arg\min_{\hat{\beta}\in\mathbf{R}^p} ||\hat{\beta}||_1 \\ \\ \text{s.c} \quad ||X'(Y - X\beta)||_\infty \leq (1 + t^{-1})\sqrt{2\log(p)}.\sigma \end{cases} \tag{4.32}$$

Avec $t \in \mathbf{R}^+$. Cet estimateur possède certaines propriétés *oracles*.
Si $X$ obéit au principe de l'incertitude uniforme avec pour norme par colonne l'unité et que le vecteur $\beta$ est suffisamment sparse. Il a été prouvé qu'avec une grande probabilité on a :

$$||\hat{\beta} - \beta||_2^2 \leq C^2.\log(p).\left(\sigma^2 + \sum_i \min(\beta_i^2, \sigma^2)\right) \tag{4.33}$$

Il faut remarquer que ce résultat sur $\hat{\beta}$ est non-asymptotique et une valeur peut être donnée à la constante $C$. Dans le cas où le paramètre $\beta$ est sparse de longueur $S$, c'est-à-dire *S-sparse* alors on prouve de manière non asymptotique

avec une grande probabilité que :

$$||\hat{\beta} - \beta||_2^2 \leq C_2^2.\lambda_p^2.\left(\sigma^2 + \sum_i \min(\beta_i^2, \sigma^2)\right) \tag{4.34}$$

avec

$$\alpha_{2S} + \theta_{S,2S} < 1 - t, \lambda_p := (\sqrt{1+a} + t^{-1})\sqrt{2\log p}$$

et la constante $C_2$ dépend uniquement de $\alpha_{2S}$ et $\theta_{S,2S}$.
L'estimateur $DS$ possède des similitudes avec l'estimateur Lasso. Dans certains cas, ils sont équivalents en pratique. James et *al* ont considéré certains cas où l'équivalence entre les deux estimateurs est obtenue et ils ont aussi prouvé la similitude dans le cas de corrélations particulières entre les variables [83].

#### 4.2.4.8 LAD-Lasso

Le *LAD-Lasso (Least Absolute Deviation-Lasso)* est une extension du Lasso introduite par Wang H. Li G. et Jiang G. [84]. Cet estimateur se base sur une perte de type $L_1$ au lieu de la perte quadratique utilisée généralement. Cette méthode est efficace dans le cas où les erreurs $\varepsilon$ du modèle gaussien ont une distribution à queue lourde ou dans le cas où il y a des observations aberrantes. Cet estimateur se définit comme suit :

$$\hat{\beta}^{LD}(\lambda) = Arg \min_{\beta \in \mathbf{R}^p} \frac{1}{n} \sum_{i=1}^{n} |Y_i - X_i\beta| + \sum_{j=1}^{p} \lambda_i |\beta_j| \tag{4.35}$$

Cet estimateur est efficace dans le cas où $p \leq n$.
Définissons la perte de Huber par :

$$l(v) = \begin{cases} v^2 & \text{si} \quad |v| \leq t \\ \\ 2t|v| - t^2 & \text{sinon} \end{cases}$$

Rosset et Zhu [85] ont utilisé cette perte combinée avec la pénalité Lasso pour un $t$ fixé : Cet estimateur se définit comme :

$$\hat{\beta}^{LDH}(\lambda) = Arg \min_{\beta \in \mathbf{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(y - x_i\beta) + \lambda||\beta||_1 \tag{4.36}$$

Van der Geer [86, 87] et Koltchinskii [88] ont utilisé cette perte et en grande dimension et ont obtenu de bons résultats. ils ont également produit des résultats sur l'erreur de prédiction au sens de la norme $L_2$ sous des hypothèses sur la matrice de Gram [46].

#### 4.2.4.9 BoLasso

L'estimateur *BoLasso (Bootstrapped Lasso)* a été introduit par Francis Bach [4, 5]. Il est consistant en sélection. On suppose que le nombre de paramètres à estimer est $p$, $n$ le nombre d'observations. avec $p \leq n$. Soit $\lambda_0$ un réel strictement positif fixé, et soit $\lambda_n$ le paramètre de régularisation est tel que $\sqrt{n}\lambda_n = \lambda_0$. L'auteur du BoLasso a constaté que le Lasso sélectionne les variables pertinentes $X_i$ expliquant la variable $Y$ avec une forte probabilité (tendant vers 1 ) à une grande vitesse (exponentielle) [4]. Les variables non pertinentes sont sélectionnées avec une probabilité strictement comprise entre 0 et 1. La méthode BoLasso se présente comme suit :

**Algorithme 4.1** BoLasso [4, 5]

1: Importation des données $(X, Y) \in \mathbf{R}^{n\times(p+1)}$
2: Choix du nombre $m$ de réplications bootstrap
3: Choix du paramètre $\lambda_0$
4: Génération de $m$-échantillons bootstrap à partir des données
5: Pour chaque échantillon bootstrap, on détermine l'estimateur Lasso $\hat{\beta}^{L,k}$, On estime l'ensemble de sparsité $\mathcal{A}^k = \{j, \hat{\beta}_j^{L,k} \neq 0\}$, où $k$ est le numéro de l'échantillon bootstrap
6: On détermine $\mathcal{A}_{BoLasso} = \bigcap_{k\in\{1,\cdots,m\}} \mathcal{A}^k$ qui est le support de l'estimateur BoLasso.
7: L'estimateur Bolasso sera défini comme l'estimateur des moindres carrés non pénalisés sur $\mathcal{A}_{BoLasso}$. Posons $\mathcal{A}_{BoLasso} = \mathcal{A}_{BL}$, on a :

$$\hat{\beta}^{BoLasso} = Arg \min_{\beta\in\mathbf{R}^p} \frac{1}{2n}||Y - X_{\mathcal{A}_{BL}}\beta||_2^2 \tag{4.37}$$

La complexité en calcul de l'algorithme est $\mathcal{O}(m(p^3 3 + p^2 n))$. En réalité pour chaque valeur du paramètre de régularisation, l'algorithme estime l'ensemble $\mathcal{A}_{BoLasso}$ sur l'ensemble de sparsité $\mathcal{A}_L$. Il parvient à prouver que :
Pour $\sqrt{n}\lambda_n = \lambda_0$, $\lambda_0 > 0$ et pour tout $m > 1$, la probabilité pour que l'algorithme Bolasso ne sélectionne pas exactement le modèle correct c'est-à-dire $\mathbf{P}(\mathcal{A}_{BoLasso} \neq \mathcal{A}_L)$ est telle que :

$$\mathbf{P}(\mathcal{A}_{BoLasso} \neq \mathcal{A}_L) \leq mA_1 e^{-A_2 n} + A_3 \frac{\log(n)}{\sqrt{n}} + A_4 \frac{\log(m)}{m}$$

où $A_1$, $A_2$, $A_3$, $A_4$ sont des constantes strictement positives.
Ce qui peut aussi se traduire par le fait qu'avec une probabilité élevée, on a $\mathcal{A}_L \subseteq \mathcal{A}_{BoLasso}$. Aussi, lorsque $p$ est tel que $p^6 \leq C_n$, où $C_n$ est une constante strictement positive avec des conditions supplémentaires sur le nombre d'échantillons bootstrap, l'estimateur Bolasso est consistant en sélection. Remarquons

que même en grande dimension, cet estimateur sélectionne de manière efficace le bon ensemble de sparsité [5]. L'algorithme proposé par l'auteur ne subit pas trop de modification lorsque le paramètre de régularisation change de valeur mais il faut qu'il soit faible pour que le nombre de variables pertinentes sélectionnées à chaque itération soit important [46].

#### 4.2.4.10 Le smooth-Lasso

La méthode de sélection pénalisée *Smooth-Lasso* a été introduite par M. Hebiri [46, 64] en 2009. Cet estimateur se définit comme suit :

$$\hat{\beta}^{SL}(\lambda) = Arg \min_{\beta \in \mathbf{R}^p} \left[ ||Y - X\beta||_n^2 + pen(\beta) \right] \tag{4.38}$$

où $X = (x_1^T, \cdots, x_n^T)^T$, $Y = (y_1, \cdots, y_n)$ et $pen : \mathbf{R}^p \longrightarrow \mathbf{R}$ est une fonction positive convexe appelée pénalité. Pour tout vecteur $u = (u_1, \cdots, u_n)^T$, on suppose que $||u||_n^2 = n^{-1} \sum_{n=1}^n |u_i|^2$. L'auteur a proposé comme fonction pénalité

$$pen(\beta) = \lambda_1 |\beta| + \lambda_2 \sum_{j=2}^{p} (\beta_j, \beta_{j-1})^2.$$

Cet estimateur se présente comme la combinaison du $L_2$-*Fusion* introduite par Land et Breiman [89] et de la pénalité $L_1$ du Lasso. L'auteur démontre que cet estimateur possède les propriétés de sparsité, prend en compte la corrélation entre les variables ou prédicteurs successifs et s'applique bien dans le cas en grande dimension. Pour un nombre de variables $p$ fixé, l'auteur établit la normalité asymptotique, la consistance en sélection de variables. Il en ressort des résultats théoriques et ses applications que le *S-Lasso* possède de nouvelles propriétés en sélection de variables comparée aux méthodes concurrentes. Dans le cas de fortes corrélations entre les variables explicatives, les résultats montrent que le *S-Lasso* domine la méthode Elastic-Net. Pour simplifier les calculs, l'auteur propose que les prédicteurs soient standardisés, c'est-à-dire

$$n^{-1} \sum_{i=1}^{n} x_{i,j}^2 = 1 \text{ et } n^{-1} \sum_{i=1}^{n} x_{i,j} = 0 \tag{4.39}$$

et il faut que la variable d'intérêt soit centrée, c'est-à-dire

$$\sum_{i=1}^{n} y_i = 0 \tag{4.40}$$

#### 4.2.4.11 Autres méthodes

Nous pouvons aussi citer d'autres méthodes de sélection de variables semblables au Lasso ceci sans détailler ces méthodes. Entre autres, on a :

— le *Compressive sensing* David L. Donoho en 2004 [90]
— Le *Graphical Lasso* de Friedman. F, Hastie. T, Tibshirani. R en 2007 [91]
— Le *Near isotonic regularization* de [92], cette méthode s'apparente comme le Fused Lasso dimuniée de la pénalité $L_1$.
— le *Matrix complétion* Emmanuel J. Candès et Benjamin Recht [93]
— le *Multivariate Method* D. M. Witten, R. Tibshirani,T. Hastie [94]

## 4.3 Modèle linéaire généralisé avec pénalisation $L_1$ (GLM-Lasso)

Lorsque l'étude porte sur des observations non continues et qu'on décide de sélectionner des variables par la méthode Lasso, la procédure qui correspond le mieux à cette approche est l'estimation des paramètres avec un modèle Linéaire Généralisé combinée avec une pénalisation de type $L_1$. Dans ce cas on utilise les équations d'optimisation 4.41 et 4.42

$$\hat{\beta}(t) = \begin{cases} arg\max_{\beta \in \mathbf{R}^p} \left[ ||Y - X\hat{\beta}||^2 \right] \\ \\ \text{s.c } \ ||\beta||_1 \leq t, \end{cases} \tag{4.41}$$

.

$$\hat{\beta}(\lambda) = Arg \max_{\beta \in \mathbf{R}^p} \left[ ||Y - X\hat{\beta}||^2 - \lambda ||\beta||_1 \right] \tag{4.42}$$

Ici le risque quadratique sera remplacé par la log-vraisemblance des paramètres du modèle. Ainsi nous avons :

$$\hat{\beta}(t) = \begin{cases} arg\max_{\beta \in \mathbf{R}^p} \left[ l(\beta) \right] \\ \\ \text{s.c } \ ||\beta||_1 \leq t, \end{cases} \tag{4.43}$$

La seconde définition de détermination des $\hat{\beta}$ en terme d'optimisation de la log-vraisemblance pénalisée s'écrit sous la forme :

$$\hat{\beta}(\lambda) = Arg \max_{\beta \in \mathbf{R}^p} \left[ l(\beta) - \lambda ||\beta||_1 \right] \tag{4.44}$$

Pour une vraisemblance donnée fixée, les deux définitions sont équivalentes. L'équation 4.44 peut être construite comme la version basée sur le multiplicateur de Lagrange du problème d'optimisation dans l'équation 4.43. L'équation 4.44 possède en plus une interprétation Bayésienne, c'est le mode de la distribution postérieure des coefficients $\beta$ avec la condition que à priori chaque coefficient suit de manière indépendante une distribution double exponentielle (distribution de Laplace) pour une même valeur du paramètre $\lambda$. Cette approche a été utilisée dans le cas du modèle à risque proportionnel de Cox [95]. La méthode d'estimation des paramètres se base sur l'algorithme du gradient complet (*full gradient*). Notons que l'algorithme proposé dans ce contexte est souple dans la mesure où elle peut automatiquement passer à la méthode de Newton-Raphson lorsque la convergence vers la solution optimale tend à se faire avec beaucoup d'itérations, ceci permet d'éviter la convergence lente de l'algorithme du *gradient ascent* proposé par Kim et Kim [96].
Posons : $\beta = (\beta_1, \cdots, \beta_p)^T$ la fonction cible et

$$l_{pen}(\beta) = l(\beta) - \lambda||\beta||_1 = l(\beta) - \lambda \sum_{i=1}^{p} |\beta_i| = l(\beta) + pen(\beta) \tag{4.45}$$

la fonction $\beta \mapsto pen(\beta)$ est concave, continue mais seulement différentiable aux points $\beta_j \neq 0$ pour tout $j$. De ce fait elle se comporte moins bien que la fonction $\beta \mapsto l(\beta)$ qui est continue, concave et différentiable. Ainsi la fonction $\beta \mapsto l_{pen}(\beta)$ qui est la somme de deux fonctions concaves est elle aussi concave mais n'est pas strictement concave. La stricte concavité de la vraisemblance faciliterait beaucoup les choses. Elle se produit uniquement aux points où elle présente un sommet plat. Ces points singuliers sont obtenus lorsque aucune variable du modèle n'est combinaison linéaire d'une ou de plusieurs variables du modèle. Dans ce cas la vraisemblance pénalisée peut présenter une concavité faible dans un petit voisinage pour de petites valeurs du paramètre $\lambda$. Nous pouvons remarquer que la log-vraisemblance pénalisée n'est pas différentiable en tout point $\beta$ à cause de la pénalisation. Mais pour la résolution du problème d'optimisation dans l'équation 4.44, nous pouvons définir les dérivées directionnelles du premier et du second ordre, la matrice Hessienne pour la log-vraisemblance et la log-vraisemblance pénalisée. Pour tout coefficient $\beta$ et pour tout vecteur directionnel $v \in \mathbf{R}^p$ on définie la dérivée directionnelle de $l_{pen}$ par :

$$l'_{pen}(\beta, v) = \lim_{t \to 0} \frac{1}{t} \left[ l_{pen}(\beta + tv) - l_{pen}(\beta) \right] \tag{4.46}$$

Posons

$$v_{opt} = Arg \max_{||v||=1} (l'_{pen}(\beta, v)) \tag{4.47}$$

Ainsi pour tout $\beta$, on peut définir le gradient tel que :

$$g(\beta) = \begin{cases} l'_{pen}(\beta, v_{opt}).v_{opt} & \text{si} \quad l'_{pen}(\beta, v_{opt}) \geq 0 \\ \\ 0_p & \text{sinon} \end{cases} \tag{4.48}$$

où $0_p$ est $p$-vecteur composé de zéros. La concavité de la vraisemblance pénalisée $l_{pen}$ fait qu'il existe un seul $\beta$ pour lequel la dérivée directionnelle atteint son maximum avec $l'_{pen}(\beta, v_{opt}).v_{opt} < 0$ bien que l'ensemble des points pour lesquels $l'_{pen}(\beta, v_{opt}).v_{opt} = 0$ peut être contigu à condition que la fonction cible n'est pas strictement concave à l'optimum.
Notons aussi que le gradient $g(\beta) = (g_1(\beta), \cdots, g_p(\beta))^T$ peut bien être calculé utilisant le gradient de la log-vraisemblance $h(\beta) = \frac{\partial l(\beta)}{\partial \beta} = (h_1(\beta), \cdots, h_p(\beta))^T$ comme :

$$g_i(\beta) = \begin{cases} h_i(\beta) - \lambda sgn(\beta_i) & \text{si} \quad \beta_i \neq 0 \\ \\ h_i(\beta) - \lambda sgn[h_i(\beta)] & \text{si} \quad \beta_i = 0 \quad \text{et } |h_i(\beta)| > \lambda \\ 0_p & \text{sinon} \end{cases} \tag{4.49}$$

Nous pouvons alors remarquer que le gradient est discontinu en tout point où la log-vraisemblance pénalisée n'est pas différentiable, en tout point où $\beta_i = 0$ pour certaines valeurs de $i$. Par analogie, on définit aussi la dérivée directionnelle d'ordre deux telle que :

$$l''_{pen}(\beta, v) = \lim_{t \to 0} \frac{1}{t} \left[ l'_{pen}(\beta + tv) - l'_{pen}(\beta) \right] \tag{4.50}$$

même si la matrice Hessienne n'est pas définie. La dérivée directionnelle de second ordre de la vraisemblance pénalisée est donnée par :

$$l''_{pen}(\beta, v) = v^T \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} v \tag{4.51}$$

Dans la pratique, il n'est jamais nécessaire d'évaluer la matrice Hessienne de $l(\beta)$ de dimension $p \times p$ avant de calculer la dérivée directionnelle de second ordre puisque la direction d'intérêt $v$ est la direction du gradient.
Dans le cas du modèle à risque proportionnel de Cox [96] ainsi que dans un GLM avec une fonction de lien canonique, la matrice Hessienne se définit comme suit :

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = X^T W X \tag{4.52}$$

où $X$ est une $n \times p$-matrice de design et $W$ est une $n \times n$ matrice de poids. Notons que cette structure de la matrice Hessienne évite sa totale construction et permet d'évaluer

$$l''_{pen}(\beta, v) = (Xv)^T W(Xv) \tag{4.53}$$

comme dans un GLM. La discontinuité du gradient comme décrit plus haut divise le domaine de la vraisemblance pénalisée $l_{pen}$ en $3^p$ sous-domaines et dans chacun de ces sous domaines, le gradient est continu. Ceci parce que chaque composante du gradient divise le domaine de $l_{pen}$ en trois sous-domaines et le gradient possède $p$ composantes. L'algorithme du gradient ascendant se base sur une suite d'approximation de Taylor. A chaque étape l'algorithme détermine approximativement et de manière locale $l_{pen}$ à partir de $\beta$ dans la direction du gradient par un développement de Taylor de second ordre. On a :

$$l_{pen}(\beta, tg(\beta)) \approx l_{pen}(\beta) + t\ l'_{pen}(\beta, g(\beta)) + \frac{1}{2} t\ l''_{pen}(\beta, g(\beta)) + t^2\ \varepsilon(\beta, g(\beta) \tag{4.54}$$

avec

$$\lim_{t \to 0} (t^2\ \varepsilon(\beta, g(\beta)) = 0 \tag{4.55}$$

Et on retient simplement que :

$$l_{pen}(\beta, tg(\beta)) \approx l_{pen}(\beta) + t\ l'_{pen}(\beta, g(\beta)) + \frac{1}{2} t\ l''_{pen}(\beta, g(\beta)) \tag{4.56}$$

Cette approximation de $l_{pen}$ n'est valable que dans un domaine de continuité du gradient, c'est-à-dire pour $t \in ]0,\ t_{seuil}[$ avec :

$$t_{seuil} = \min_i \left\{ -\frac{\beta_i}{g_i(\beta)} \quad : \quad sgn(\beta_i) = -sgn[g_i(\beta)] \neq 0 \right\} \tag{4.57}$$

L'optimum de l'approximation de Taylor est obtenu à :

$$t_{opt} = -\frac{l'_{pen}(\beta, g(\beta))}{l''_{pen}(\beta, g(\beta))} \tag{4.58}$$

ce qui donne $t_{opt} < t_{seuil}$ sinon l'optimum serait obtenu à $t_{seuil}$. L'approximation directionnelle de Taylor à une étape est obtenue à partir de l'optimum obtenu à l'étape précedente et la convergence est acquise si $g(\beta) = 0$.

Une fois que le gradient et la matrice Hessienne sont connus alors le calcul de $\beta^{(i+1)}$ utilisant $\beta^{(i)}$ devient aisé. Par contre l'algorithme peut nécessiter plus d'étapes avant sa convergence. Mais cette lente convergence peut être évitée car l'algorithme est capable de continuer les calculs par la méthode de Newton-Raphson qui converge beaucoup plus rapidement.

**Algorithme 4.2** Gradient ascendant

1: Initialisation $\beta^{(0)}$
2: pour $i \in \{1, 2, \cdots\}$

$$\beta^{(i+1)} = \beta^{(i)} + \min\{t_{opt}, t_{seuil}\} g(\beta^{(i)}) \tag{4.59}$$

l'étape (2) est répétée jusqu'à convergence.

### Méthode de Newton-Raphson

Nous supposons que la fonction cible est concave et deux fois différentiable, on suppose que l'algorithme du gradient ascendant est comme une série d'optimisation, la contrainte sur chaque élément de cette série étant dans le sous-domaine de continuité du gradient. Nous savons qu'il y en a $3^p$ sous-domaines de ce genre chacun d'eux étant défini par :

$$sgn(\beta) = (sgn(\beta_1), \cdots, sgn(\beta_p))^T \tag{4.60}$$

et dans chaque sous-domaine le gradient comme fonction de $\beta$ est continu. Posons

$$\begin{aligned} sgn(\beta_+) &= \lim_{\varepsilon \to 0} [sgn(\beta + \varepsilon g(\beta))], \\ J &= \{j \text{ tel que } sgn(\beta^+) \neq 0\} \\ m &= Card(J) \end{aligned}$$

Il est clair que $m \leq p$ et $J$ n'est rien d'autre que l'ensemble des variables actives ou l'ensemble de sparsité. La meilleure approximation est obtenue si $t_{opt} < t_{seuil}$.
Posons :

$$\begin{aligned} \tilde{\beta} &= (\beta_{J_1}, \cdots, \beta_{J_m})^T \\ g(\tilde{\beta}) &= (g_{J_1}(\beta), \cdots, g_{J_m}(\beta))^T \\ \tilde{H}_{k,l}(\beta) &= \frac{\partial^2 l(\beta)}{\partial \beta_{J_k} \partial \beta_{J_l}} \quad k = 1, \cdots, m, \quad l = 1, \cdots, m \end{aligned}$$

où $g(\tilde{\beta})$ est le gradient dans le domaine de contrainte et $\tilde{H}$ la matrice Hessienne de l'optimisation sous contrainte.
A chaque étape, cet algorithme donne :

$$\tilde{\beta}^{(i+1)} = \tilde{\beta}^{(i)} - [\tilde{H}(\beta^{(i)})]^{-1} \tilde{g}(\beta^{(i)}) \tag{4.61}$$

De ce fait on peut poser $\beta_{NR}^{(i+1)}$ le vecteur de Newton-Raphson de dimension $p$, le vecteur obtenu en prenant $\tilde{\beta}^{(i+1)}$ et en prenant zéro pour les variables non

actives. Cette considération est faite uniquement dans le sous-domaine courant de contrainte. Ce qui fait que $\beta_{NR}^{(i+1)}$ est juste si $\text{sgn}(\beta_{NR}^{(i+1)}) = \text{sgn}(\beta_{+}^{(i)})$

**Algorithme 4.3** Newton-Raphson

1: Initialisation $\tilde{\beta}^{(0)}$

2: Evaluation de :

$$\begin{aligned} sgn(\beta_+) &= \lim_{\varepsilon \to 0} \left[ sgn(\beta + \varepsilon g(\beta)) \right], \\ J &= \{ j \ : \ sgn(\beta_j^+) \neq 0 \} \\ m &= Card(J) \end{aligned}$$

3: Pour $t_{opt} < t_{seuil}$

$$\begin{aligned} \tilde{\beta} &= (\beta_{J_1}, \cdots, \beta_{J_m})^T \\ g(\tilde{\beta}) &= (g_{J_1}(\beta), \cdots, g_{J_m}(\beta))^T \\ \tilde{H}_{k,l}(\beta) &= \frac{\partial^2 l(\beta)}{\partial \beta_{J_k} \partial \beta_{J_l}} \quad k = 1, \cdots, m, \ \ l = 1, \cdots, m \end{aligned}$$

4:

$$\tilde{\beta}^{(i+1)} = \tilde{\beta}^{(i)} - [\tilde{H}(\beta^{(i)})]^{-1} \tilde{g}(\beta^{(i)}) \tag{4.62}$$

L'étape (4) est répétée jusqu'à convergence.

**Gradient ascendant pour la vraisemblance pénalisée avec incorporation de la méthode Newton-Raphson**

**Algorithme 4.4** Gradient ascendant & Newton-Raphson

1: Initialisation $\beta^{(0)}$

2: Pour $i \geq 1$

$$\beta^{(i+1)} = \begin{cases} \beta^{(i)} + t_{seuil} \times g(\beta^{(i)}) & \text{si} \quad t_{opt} \geq t_{seuil} \\ \beta_{NR}^{(i+1)} & \text{si} \quad t_{opt} \leq t_{seuil} \\ & \text{et} \quad \text{sgn}(\beta_{NR}^{(i+1)}) = \text{sgn}(\beta_{+}^{(i)}) \\ \beta^{(i)} + t_{opt} \times g(\beta^{(i)}) & \text{sinon} \end{cases} \tag{4.63}$$

3: L'étape (2) est répétée jusqu'à convergence.

Cet algorithme est aussi applicable dans le cas où tous les paramètres ne sont pas pénalisés, c'est-à-dire chaque paramètre possède une pénalisation

propre. Dans ce cas, il suffit juste de reécrire l'équation 4.45 sous la forme :

$$l_{pen}(\beta) = l(\beta) - \sum_{i=1}^{p} \lambda_i |\beta_i| \tag{4.64}$$

et si un paramètre n'est pas pénalisé alors il prend $\lambda_i = 0$. Cet algorithme prend également en compte le cas où la pénalité est une combinaison de types $L_1$ et $L_2$. Le problème d'optimisation dans l'équation 4.44 se reécrit sous la forme :

$$\hat{\beta}(\lambda) = Arg \max_{\beta \in \mathbf{R}^p} \left[ l(\beta) - \lambda_1 ||\beta||_1 - \lambda_2 ||\beta||_2^2 \right] \tag{4.65}$$

Notons que la pénalisation de type $L_2$ est deux fois différentiable, concave et continue. Elle se comporte mieux que la pénalisation $L_1$. Mais son inconvénient principal est qu'elle ne facilite pas les choses dans la sélection de variables. Elle tend à augmenter l'ensemble de sparsité si elle est relativement plus large que la pénalité $L_1$. Ceci oblige l'algorithme à procéder à l'inversion de grande matrice à chaque étape dans l'algorithme de Newton Raphson. Mais dans le cas du modèle à risque proportionnel de Cox, ceci peut être éviter en faisant de nouvelles paramétrisations [95].

## 4.4 Méthode de double validation croisée stratifiée

### 4.4.1 Validation Croisée

La validation croisée est d'un principe simple, efficace et largement utilisée pour estimer une erreur moyennant un surplus de calcul. L'idée est d'itérer l'estimation de l'erreur sur plusieurs échantillons de validation puis d'en calculer la moyenne. C'est indispensable pour réduire la variance et ainsi améliorer la précision lorsque la taille de l'échantillon initial est trop réduite pour en extraire des échantillons de validation et test de taille suffisante. Soit $\Upsilon$ une fonction d'indexation définie par :

$$\Upsilon : \{1, \ldots, n\} \longrightarrow \{1, \ldots, K\}, \Upsilon(i) = k \tag{4.66}$$

qui à chaque observation $i$ associe uniformément et de manière aléatoire sa classe. L'estimation par validation croisée de l'erreur de prévision est :

$$\hat{R_{cv}} = \frac{1}{n} \sum_{i=1}^{n} l[y_i, \hat{f}^{(-\Upsilon(i))}(xi)] \tag{4.67}$$

où

$$f^{(-\Upsilon(i))} = f^{(-k)}$$

désigne l'estimation de $f$ sans la prise en compte de la $k$ième partie de l'échantillon.

— Si $K \neq 1$, on parle de *K-fold* cross-validation 1.6.2.
— Si $K = 1$, on parle de Leave-one-out (*loo*) 1.6.3

Par défaut dans les logiciels de calcul comme R, la valeur de $K$ est $K = 10$. Dans les expériences en général, l'estimation de l'erreur de prévision par validation se calcule en fonction d'un paramètre. Ainsi pour optimiser le choix d'un modèle au sein d'une famille paramétrée, une bonne approche serait de minimiser cette erreur pour les différentes valeurs du paramètre considéré. Une estimation $\hat{f}$ de $f$ est définie par $\hat{\theta}$ et :

$$\hat{\theta} = Arg \min_{\theta} \hat{R}_{cv}(\theta) \tag{4.68}$$

Dans le cas de régressions linéaires, les valeurs ajustées sont fonctions linéaires des observations. $\hat{y} = \mathbf{H}y$ avec $\mathbf{H} = (h_{i,j})_{n\times n}$ est la *hat-matrix*. En régression linéaire multiple, $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}$. On trouve des formes de $\mathbf{H}$ très proches de celle-ci dans les régressions pénalisées (Bridge, Lasso, section 4.2). Pour ces genres estimateurs linéaires, l'estimation (loo) de l'erreur quadratique par validation croisée (PRESS) se met sous la forme :

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}^{(-i)}(x_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(x_i)}{1 - h_{ii}}\right]^2$$

Dans ce processus, une seule estimation de $\hat{f}$ est requise mais le calcul de la matrice diagonale $\mathbf{H}$ peut prendre beaucoup de temps lorsque $n$ ou $p$ sont grands.

### 4.4.2 Validation Croisée stratifiée

La validation croisée stratifiée dans le cadre de ce travail est une double CV qui consiste à faire une seconde validation croisée à l'intérieur de la première. Elle est nommée LOLO-DCV (*Leave-One-Level-Out Double Cross-Validation*). En clair lorsque pour la première CV l'ensemble des données a été scindé en deux parties, ensemble d'apprentissage ($E_A$) ensemble de test ($E_T$), on fait une validation croisée complète sur $E_A$ afin de déterminer le modèle optimal de prédiction sur $E_T$. Les strates de la seconde CV sont les ensembles d'apprentissage à chaque étape de la première CV.

Dans le cadre de notre travail, cette validation croisée tient compte de la structure des données. Ainsi les blocs utilisés dans la CV ne sont pas obtenus

**Algorithme 4.5** Validation croisée à deux niveaux (LOLO-DCV)

1. A chaque étape du premier niveau de la validation-croisée
   (a) Les données sont divisées en $N$-blocs
   (b) Les blocs sont regroupés en deux parties : $E_A$ et $E_T$, $E_A$ : l'ensemble d'apprentissage qui contient les observations de $(N-1)$-blocs, $E_T$ : l'ensemble de test, contenant les observations du dernier bloc.
   (c) On met de côté $E_T$
   (d) Deuxième niveau de validation croisée.
      i. Les données de $E_A$ sont divisées en $(N-1)$-blocs
      ii. Les blocs sont regroupés en deux parties : $E_{A_2}$ et $E_{T_2}$, $E_{A_2}$ : l'ensemble d'apprentissage qui contient les observations de $(N-2)$-blocs de $E_A$, $E_{T_2}$ : l'ensemble de test, contenant les observations du dernier bloc de $E_A$.
      iii. On met de côté $E_{T_2}$
      iv. On reprend le processus 1(d)ii $(N-1)$ fois afin que chacun des $(N-1)$ blocs soit $E_{T_2}$
   (e) On reprend le processus 1b $N$ fois afin que chacun des $N$ blocs soit $E_T$

de manière aléatoire, ils sont déterministes. Un bloc est l'ensemble de toutes les observations d'une maison de capture. Cette méthode de constitution des blocs nous permet de rester cohérent avec l'objectif final qui est de faire des prédictions optimales dans des zones dont aucune information n'a été utilisée dans l'apprentissage.

## 4.5 Algorithme combiné : GLM-Lasso et double validation croisée stratifiée

Cette méthode développée dans ce travail consiste à faire de la sélection de variables et de la prédiction en combinant le GLM-Lasso et la double validation croisée stratifiée. Elle présente comme suit.

**Algorithme 4.6** Algorithme combiné GLMM-Lasso et LOLO-DCV

1. Utilisation du GLM-Lasso au second niveau de LOLO-DCV pour sélectionner les variables optimales selon l'ensemble $E_T$
2. Utilisation d'une méthode spécifique pour sélectionner le sous ensemble optimal de variables
3. Prédiction avec le GLM par simple CV utilisant les blocs de l'étape 1a du LOLO-DCV 4.5

### 4.5.1 Interactions entre les variables

Dans ce travail, les algorithmes développés apprennent automatiquement sur toutes les variables et interactions du second ordre et produit un ensemble optimal de variables pour la prédiction. Si $p$ est le nombre de covariables originales, le nombre total de covariables avec les interactions du second ordre est $N_{cov}$, $\mathcal{V}_{\mathcal{O}} = (V_1, \ldots, V_p)$ est le vecteur des variables de départ (originales ou recodées). L'ensemble des covariables d'interactions du second ordre est défini par : $\mathcal{I}_{\mathcal{V}_{\mathcal{O}}} = \{V_i : V_j, 1 \leq i, j \leq p, \quad i \neq j\}$. Les interactions du second ordre sont disponibles pour les variables : numérique croisée avec numérique, numérique croisée avec non-numérique et non-numérique croisée avec non-numérique. Ainsi, le nombre d'interactions du second ordre est $p(p-1)/2$ et $N_{cov} = p + p(p-1)/2$. Supposons que le nombre total des observations est $N_{obs}$.

#### 4.5.1.1 Variable numérique croisée variable numérique

$V_k$ and $V_l$ sont deux variables numériques. La variable d'interaction obtenu de $V_k$ et $V_l$ est notée $V_k : V_l$ et définie par :

$$(V_k : V_l)_i = (V_k)_i \times (V_l)_i, \ 1 \leq i \leq N_{obs}$$

#### 4.5.1.2 Variable numérique croisée variable non-numérique

$V_k$ est une variable numérique et $V_l$ une variable non-numérique avec $d_l$ modalités. $V_l$ est considérée comme une variable numérique avec $d_l$-dimension. Elle peut être remplacée par les indicatrices de ses modalités. Supposons que les modalités sont $V_{lq}, \quad 1 \leq q \leq d_l$. L'indicatrice $I_{lq}$ associatée à $V_{lq}$ est définie par :

$$(I_{lq})_i = \begin{cases} 1 & \text{si } \ (V_l)_i = V_{lq} \\ 0 & \text{sinon} \end{cases}$$

$V_l$ peut être remplacée par $\{I_{lq}, \quad 1 \leq q \leq d_l\}$. La variable d'interaction obtenue est $V_k : V_l$ avec $d_l$-dimension peut être remplacée par $\{V_k : I_{lq}, \quad 1 \leq q \leq d_l\}$.

Chaque $V_k : I_{lq}$ est définie par :

$$(V_k : I_{lq})_i = \begin{cases} (V_k)_i & \text{si } (I_{lq})_i = 1 \\ 0 & \text{sinon} \end{cases}$$

#### 4.5.1.3 Variable non-numérique croisée variable non-numérique

$V_k$ et $V_l$ deux variables non-numériques avec $d_k$ et $d_l$ modalités respectivement. La variable d'interaction obtenue de $V_k$ et $V_l$ est $V_k : V_l$. $V_k : V_l$ est $d_k \times d_l$-dimension, peut être remplacée par $\{I_{kp} : I_{lq}, \quad 1 \leq p \leq d_k \quad \text{et} \quad 1 \leq q \leq d_l\}$. Chaque $I_{kp} : I_{lq}$ est définie par :

$$(I_{kp} : I_{lq})_i = \begin{cases} 1 & \text{si } (I_{kp})_i = (I_{lq})_i = 1 \\ 0 & \text{sinon} \end{cases}$$

#### 4.5.1.4 Identifiabilité des variables

Pour l'identifiabilité des variables incluant celles d'interaction, un vecteur $\mathcal{H}$ de nombres entiers est automatiquement généré, $\mathcal{H} = \{h_1, h_2, \ldots, h_{N_{cov}}\}$. Si $\mathcal{V}$ est l'ensemble de toutes les covariables incluant celles d'interaction alors $\mathcal{V} = \{V_1, V_2, \ldots, V_{N_{cov}}\}$. $\mathcal{H}$ et $\mathcal{V}$ sont deux vecteurs avec la même longueur $N_{cov}$. La composante $h_s$ de $\mathcal{H}$ est la dimension de la covariable $V_s$, $1 \leq s \leq N_{cov}$. Dans le processus de sélection, même si une variable non-numérique $V_s$ est remplacée par les indicatrices de ses modalités, les indicatrices sont automatiquement identifiées et groupées conformément aux composantes $h_s$ de $\mathcal{H}$ correspondant à cette variable.

#### 4.5.1.5 Variables fréquentes

Soit $\mathcal{V} = \{V_1, V_2, \ldots, V_{N_{cov}}\}$ l'ensemble de toutes les variables y compris les interactions du second ordre. Pour chaque valeur $\lambda$ du paramètre de pénalisation, le vecteur des coefficients des covariables est noté $\beta(\lambda)$. On peut alors déterminer la présence ou l'absence de chaque variable. Pour tout $\lambda$, définissons la fonction "Présence" de variable par :

$$\begin{cases} \mathcal{P}_k(V_r) = 1 & \text{si} \quad \beta_r(\lambda) \neq \Theta \\ \mathcal{P}_k(V_r) = 0 & \text{sinon} \end{cases} \tag{4.69}$$

où $\beta_r(\lambda)$, $1 \leq s \leq N_{cov}$ est le vecteur $V_r$ et $\Theta$ le vecteur nul. La longueur de $\beta_r(\lambda)$ est fonction de la composante $h_r$ de $\mathcal{H}$. Pour un seuil $s$, $1 \leq s \leq 100$,

l'ensemble des variables fréquentes (FV) est

$$\mathrm{FV}(\lambda) = \left\{ V_r, \frac{100}{N_f} \times \sum_{k=1}^{N_f} \mathcal{P}_k(V_r) \geq s \right\} \tag{4.70}$$

## 4.6 Méthodologie de construction d'une fonction de prévision par combinaison du GLM-Lasso et de la double validation croisée

Dans cette partie du travail, le processus de construction d'une fonction de prévision se présente comme suit :

1. On introduit dans le modèle toutes les variables explicatives.
2. Toutes interactions de second ordre entre toutes les variables sont générées automatiquement.
3. On fixe un niveau dans la hiérarchisation. Les données sont divisées en k blocks, où k est le nombre total de parties à ce niveau dans la hiérarchisation des données. Les blocs ne sont pas aléatoires dans le k-fold CV, ils sont bien déterministes comme le LOO CV.
4. (a) On utilise une validation croisée *Leave one out* dans le quel on considère un bloc comme l'individu à mettre de côté. C'est une validation croisée combinée k-folds et *Leave one out*, c'est la validation croisée Leave One Level Out (LOLO) 4.5.
   (b) Les blocs sont séparés en deux parties $E_A$ contenant $(k-1)$ blocs et $E_T$ contenant le dernier bloc.
   (c) Le GLMM-Lasso génère un vecteur $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_m)$, tel que, pour $\lambda_0$ toutes les variables sont présentes dans le modèle et pour $\lambda_m$ aucune variable n'est présente dans le modèle. Le seul prédicteur est la constante $\beta_0$
   (d) On réalise une validation complète sur $E_A$ pour déterminer le ($\lambda$ qui minimise l'erreur de prédiction $l(y, \phi(x)) = \sum (y - \phi(x))^2$
   (e) Avec $\lambda$, On apprend sur $E_A$ et on prédit sur $E_T$. On détermine la présence ou l'absence $\mathcal{P}_k(V_r)$ 4.69 de chacune des variables.
   (f) Pour un seuil fixé, on détermine le sous ensemble des variables fréquentes $FV(\lambda)$ 4.70
   (g) Avec les éléments de $FV(\lambda)$, on prédit chaque observation par validation croisée *Leave one out*. On détermine le risque quadratique en prédiction pour ce modèle.

5. On répète l'étape 4 k fois en variant $E_A$ et $A_T$ jusqu'à obtenir k valeurs de $\lambda$.
6. On choisira comme valeur optimale $\lambda_{opt}$ de $\lambda$ celui qui minimise la perte quadratique en prédiction.
7. Le modèle optimal pour la prévision sera celui obtenu par $\lambda_{opt}$.

Chapitre 5

# Prédiction de l'exposition palustre local utilisant un algorithme basé sur le GLM-Lasso et une double validation croisée

Dans ce travail, nous proposons une méthode d'apprentissage machine automatique pour la sélection de variables en combinant Lasso, GLM et une validation croisée à deux niveaux dans le contexte épidémiologique [97]. L'un des objectifs de cette approche est de surmonter les prétraitements des experts en médecine et en épidémiologie sur les données collectées. L'approche utilise toutes les variables explicatives disponibles sans Traitement et génère automatiquement toutes les interactions du second ordre entre elles. Le Lasso fait simultanément de la sélection et de l'estimation et est robuste pour la sélection de variables en grande dimension. Dans certaines études, le nombre d'observations est faible. La méthode classique de reéchantillonnage utilisée est la validation croisée. Il est bien connu que la validation croisée peut conduire à un sur-apprentissage et une solution alternative est le *percentil-cv* [98]. Pour éviter le sur-apprentissage, nous proposons la validation croisée stratifiée à deux niveaux (DCV). La variable d'intérêt, le nombre d'anophèle caractéristique principale du risque palustre est une variable de comptage. Il est aussi connu que les estimateurs Lasso sont biaisés. Une combinaison de GLM et de Lasso (GLM-Lasso) est mise en œuvre basée sur la DCV. Le GLM simple est utilisé pour débiaiser les estimateurs Lasso compte tenu de la nature de la variable d'intérêt. Pour la prédiction du risque palustre, deux stratégies de sélection de variables basées sur le GLM-Lasso et la DCV : LDLM, LDLS, sont implémentées. Ces stratégies utilisent la déviance du modèle. Chaque stratégie est appliquée sur deux groupes de covariables ( groupe 1 : originales, groupe 2 : recodées ). Plusieurs algorithmes implémentés dans ce travail sont basés sur les tra-

vaux [99, 100, 73]. Les résultats obtenus sont comparés à ceux obtenus par la méthode (B-GLM), section 3.

## 5.1 Matériels et méthode

### 5.1.1 Matériels

La zone d'étude, la méthode de collecte et d'identification des moustiques, les données environnementales et celles liées aux comportements, la description des variables sont entièrement détaillées dans le travail [3]. Les traitements opérés sur les variables originales sont détaillés dans les tableaux 5.7 et 5.8. En général les experts en épidémiologie et en médecine choisissent certaines interactions compte tenu de leurs connaissances et expérience. Pour éviter cette manière de faire, nous avons généré automatiquement toutes les interactions du second ordre dans l'ensemble des variables explicatives utilisées dans le modèle. Ceci entraîne une croissance exponentielle du nombre de variables et les méthodes classiques de sélection de variables échoueraient dans ce cas. L'algorithme apprend automatiquement sur toutes les variables et interactions du second ordre et produit un ensemble optimal de variables pour la prédiction.

### 5.1.2 Méthode statistique de travail

Les études de cohorte en général génèrent de grandes bases de données contenant des dizaines de variables. Dans le processus d'analyse des données, les experts en médecine et en épidémiologie utilisent leurs connaissances empiriques sur les phénomènes pour réaliser des pré-traitements sur les variables. Ces pré-traitements consistent à recoder certaines variables et à choisir certaines interactions du second ordre entre les variables. Ils utilisent ensuite des méthodes classiques de sélection de variables telles que *wrapper (forward, backward, stepwise, etc.), embedded, filter et ranking* pour la sélection de variables. Le but de la méthode *wrapper* est de sélectionner un sous ensemble de variables avec une faible erreur de prédiction. Son algorithme a été amélioré par le *structural wrapper* pour obtenir une suite de sous-ensembles emboîtés de variables pour l'optimalité. En pratique, les méthodes classiques de sélection de variables sont pratiquement inefficaces en grande dimension parce que le nombre de sous-ensembles de variables est $(2^p)$, où $p$ est le nombre de variables.

Les analyses statistiques sont conduites en trois étapes. Dans la première, la sélection de variables est réalisée par la méthode GLM-Lasso à travers la validation croisée à deux niveaux. Dans la seconde étape, les variables sélectionnées

sont débiaisées par un simple GLM et utilisées pour la prédiction du nombre d'anophèles. A la dernière étape, les résultats sont comparés à la méthode de référence pour savoir laquelle des deux est la meilleure pour la sélection et la prédiction.

#### 5.1.2.1 Modèle de travail

Les analyses statistiques sont basées sur le GLM et les processus sur les données par la méthode Lasso. Cette approche est appelée le GLM-Lasso [27, 27]. La variable cible, le nombre d'anophèles capturés conditionnellement aux données environnementales et climatiques suit une loi de Poisson. Les lois de Poisson constituent une famille exponentielle de dispersion dont la fonction densité de probabilité est :

$$\begin{aligned} \mathcal{P}(y|\mu) &= e^{-\mu}\frac{\mu^y}{y!} \\ \mathcal{P}(y|\mu) &= \frac{1}{y!}\exp\{y\theta - e^{\theta}\} \end{aligned} \tag{5.1}$$

avec $\theta = \log(\mu)$. Sa fonction de variance unité est $\mu$ et la déviance associée est définie par :

$$\begin{aligned} d(y|\mu) &= -2\int_y^{\mu}\frac{y-u}{u}du \\ d(y|\mu) &= -2\left\{(y - y\log(y)) - (\mu - y\log(\mu))\right\} \end{aligned} \tag{5.2}$$

Cette fonction est convexe, son minimum est nul et est obtenu à $\mu = y$. Ce qui implique que $d(y|\mu)$ est positive. La fonction densité de probabilité peut être définie utilisant la déviance par :

$$\mathcal{P}(y|\mu) = \frac{y^y e^{-y}}{y!}\exp\left\{-\frac{1}{2}d(y|\mu)\right\} \tag{5.3}$$

Conformément à l'équation 5.3, minimiser la déviance équivaut à maximiser la vraisemblance. Pour chaque observation $i$, l'équation 5.3 est définie par :

$$\mathcal{P}(y_i|\mu(x_i,\beta)) = \frac{y_i^{y_i} e^{-y_i}}{y_i!}\exp\left\{-\frac{1}{2}d(y_i|\mu(x_i,\beta)\right\} \tag{5.4}$$

Le modèle GLM sous forme matricielle se met sous la forme :

$$g[E(Y|X,\beta)] = X\beta \tag{5.5}$$

où la distribution de $Y$ conditionnellement à $(X = x)$ est une distribution de Poisson de paramètre $E(Y|X = x,\beta)$, $X$ est une matrice de dimension

$n \times (p+1)$ des covariables, $n$ est le nombre d'observations, $p$ est le nombre de covariables. $\beta$ est un $(p+1)$-vecteur des paramètres fixes y compris l'intercept, $Y$ est le vecteur de la variable cible.

$$(Y = y_i | X = x_i) \sim \mathbb{P}(\mu_i); \tag{5.6}$$

où $x_i\beta = \log(\mu_i)$ et $\mathbb{P}(\mu_i)$ est une distribution de Poisson de paramètre $\mu_i$. Ainsi

$$\mathbb{P}(Y = y_i | X = x_i) = \frac{e^{(x_i\beta)^{y_i}}}{(y_i)!} \times e^{-e^{x_i\beta}} \tag{5.7}$$

où $y_i$ est un entier positif, $x_i$ un vecteur $(x_{i1}, \ldots, x_{ip})$ de nombres réels. Si $\mathcal{D} = \{(Y = y_i, X = x_i), 1 \leq i \leq n\}$, la vraisemblance des $n$ observations peut être définie par :

$$L_{GLM}(\beta | \mathcal{D}) = \prod_{i=1}^{n} \frac{e^{(x_i\beta)^{y_i}}}{(y_i)!} \times e^{-e^{x_i\beta}} \tag{5.8}$$

et la log-vraisemblance est :

$$\begin{aligned} \mathcal{L}_{GLM}(\beta | \mathcal{D}) &= \log\left(\prod_{i=1}^{n} \frac{e^{(x_i\beta)^{y_i}}}{(y_i)!} \times e^{-e^{x_i\beta}}\right) \\ \mathcal{L}_{GLM}(\beta | \mathcal{D}) &= -\sum_{i=1}^{n} \log((y_i)!) + \sum_{i=1}^{n} y_i(x_i\beta) - e^{(x_i\beta)} \end{aligned} \tag{5.9}$$

Minimiser la déviance sous contrainte $\sum_i |\beta_j| < t$ ce qui est équivalent à $\lambda \sum_j |\beta_j| < 1$, est réduit à minimiser sans contrainte sur le vecteur $\beta$ des paramètres de la fonction de régression $\sum_i d(y_i|\mu(x_i, \beta) + \lambda \sum_j |\beta_j|$

$$\begin{aligned} \sum_i d(y_i|\mu(x_i, \beta) + \lambda \sum_j |\beta_j| &= -2\sum_j (y_i - y_i \log(y_i)) - (\mu(x_i, \beta) - y \log(\mu(x_i, \beta)) \\ &\quad +\lambda \sum_j |\beta_j| \\ &= +2(\sum_j (\mu(x_i, \beta) - y \log(\mu(x_i, \beta)) + \frac{1}{2}\lambda \sum_j |\beta_j|) \\ &\quad -2\sum_j (y_i - y_i \log(y_i)) \end{aligned} \tag{5.10}$$

La quantité $\sum_j (y_i - y_i \log(y_i))$ est indépendante du paramètre $\mu$ du modèle. Ainsi minimiser $\sum_i d(y_i|\mu(x_i, \beta) + \lambda \sum_j |\beta_j|$ se réduit à minimiser $(\sum_j (\mu(x_i, \beta) - y \log(\mu(x_i, \beta)) + \frac{1}{2}\lambda \sum_j |\beta_j|)$. Utilisant $\lambda$ en lieu et place de $\frac{1}{2}\lambda$, on peut aussi minimiser $(\sum_j (\mu(x_i, \beta) - y \log(\mu(x_i, \beta)) + \lambda \sum_j |\beta_j|)$. Si

$$Q = \sum_j (\mu(x_i, \beta) - y \log(\mu(x_i, \beta)) + \lambda \sum_j |\beta_j| \tag{5.11}$$

alors

$$\begin{aligned} Q &= -\left(\sum_j(-\mu(x_i,\beta)+y\log(\mu(x_i,\beta))-\lambda\sum_j|\beta_j|\right) \\ &= -\left(\sum_j y\log(\mu(x_i,\beta))-\mu(x_i,\beta)+\log((y_i)!)-\lambda\sum_j|\beta_j|\right)+\sum_j\log((y_i)!) \\ Q &= -(L_{GLM}(\beta|\ \mathcal{D})-\lambda\sum_j|\beta_j|)+\sum_j\log((y_i)! \end{aligned} \tag{5.12}$$

minimiser la quantité Q est la même chose que maximiser $L_{GLM}(\beta|\ \mathcal{D}) - \lambda\sum_j|\beta_j|$. Ainsi

$$\begin{aligned} \mathcal{L}_{pen}(\beta(\lambda)|\ \mathcal{D}) &= L_{GLM}(\beta|\ \mathcal{D})-\lambda\sum_j|\beta_j| \\ \mathcal{L}_{pen}(\beta(\lambda)|\ \mathcal{D}) &= -\sum_{i=1}^n\log((y_i)!)+\sum_{i=1}^n y_i(x_i\beta)-e^{(x_i\beta)}-\lambda\sum_j|\beta_j| \end{aligned} \tag{5.13}$$

Selon l' équation 5.13, la méthode GLM-Lasso est une méthode régularisante qui consiste à pénaliser la vraisemblance du modèle GLM en ajoutant un terme de pénalité

$$\mathcal{P}(\lambda)=\lambda\sum_{i=1}^p|\beta_i|,\quad \text{with}\ \ \lambda\geq 0 \tag{5.14}$$

$$\begin{aligned} \mathcal{L}_{pen}(\beta(\lambda)|\ \mathcal{D}) &= -\sum_{i=1}^n\log((y_i)!)+\sum_{i=1}^n y_i(x_i\beta)-e^{(x_i\beta)}-\mathcal{P}(\lambda) \\ \mathcal{L}_{pen}(\beta(\lambda)|\ \mathcal{D}) &= L_{GLM}(\beta|\ \mathcal{D})-\mathcal{P}(\lambda) \end{aligned} \tag{5.15}$$

Les coefficients du GLM-Lasso sont donnés par :

$$\hat{\beta}(\lambda)=Arg\max_\beta\left[L_{GLM}(\beta|\ \mathcal{D})-\mathcal{P}(\lambda)\right] \tag{5.16}$$

Le choix du paramètre de régularisation lambda est donné en minimisant une fonction de score. Dans la pratique, cette équation ne possède pas une solution numérique exacte. On peut utiliser la combinaison de l'approximation de Laplace, la méthode de Newton-Raphson ou la méthode du score de Fisher pour résoudre le problème. Cette procédure est utilisée à chaque étape de l'apprentissage. La déviance peut être définie comme suit :

$$Deviance(\beta|\mathcal{D})=\sum_{i=1}^n d(y_i|\mu(x_i,\beta)) \tag{5.17}$$

où

$$\frac{1}{2}d(y_i|\mu(x_i,\beta) = (y_i\log(y_i) - y_i) - (y_i\log(\mu(x_i|\beta)) - \mu(x_i|\beta)) \tag{5.18}$$

et $d(y_i|\mu(x_i,\beta)$ est la contribution des observations $(y_i, x_i)$ à la deviance. Ainsi

$$\begin{aligned}
\frac{1}{2}\sum_i d(y_i|\mu(x_i,\beta) &= \sum_i (y_i\log(y_i) - y_i) - (y_i\log(\mu(x_i|\beta)) - \mu(x_i|\beta)) \\
&= \sum_i (y_i\log(y_i) - y_i - \log(y_i!)) \\
&\quad - \sum_i (y_i\log(\mu(x_i|\beta)) - \mu(x_i|\beta) - \log(y_i!)) \\
\frac{1}{2}\sum_i d(y_i|\mu(x_i,\beta) &= \mathcal{L}(\mathcal{M}(sat)) - \mathcal{L}(\mathcal{M}(\beta)) \\
Deviance(\mathcal{M}(\beta)) &= 2\left(\mathcal{L}(\mathcal{M}(sat)) - \mathcal{L}(\mathcal{M}(\beta))\right)
\end{aligned} \tag{5.19}$$

où $\mathcal{M}(sat)$ est le modèle "saturé" et $\mathcal{M}(\beta)$ est le modèle de la régression de Poisson. Il est clair que : $Deviance(\mathcal{M}(Sat)) = 0$. La déviance du modèle "null" est notée $\mathcal{M}(Null)$ (le modèle contenant uniquement le terme constant) est défini par :

$$\sum_{i=1}^{n}(y_i\log(y_i) - y_i) - (y_i\log(\bar{y}) - \bar{y}) \tag{5.20}$$

alors

$$\begin{aligned}
Deviance(\mathcal{M}(\beta)) &= 2\left(\mathcal{L}(\mathcal{M}(sat)) - \mathcal{L}(\mathcal{M}(Null)) + \mathcal{L}(\mathcal{M}(Null)) - \mathcal{L}(\mathcal{M}(\beta))\right) \\
Deviance(\mathcal{M}(\beta)) &= Deviance(\mathcal{M}(Null)) - 2(L(\mathcal{M}(\beta)) - L(\mathcal{M}(Null))) \\
Deviance(\mathcal{M}(\beta)) &= Deviance(\mathcal{M}(Null)) - ResidDev(\mathcal{M}(\beta)) \\
Deviance(\mathcal{M}(\beta)) &= Deviance(\mathcal{M}(Null))\left(1 - \frac{ResidDev(\mathcal{M}(\beta))}{Deviance(\mathcal{M}(Null))}\right)
\end{aligned} \tag{5.21}$$

Ainsi nous avons :

$$\frac{Deviance(\mathcal{M}(\beta))}{Deviance(\mathcal{M}(Null))} = 1 - \frac{Deviance\ Residual(\mathcal{M}(\beta))}{Deviance(\mathcal{M}(Null))} \tag{5.22}$$

où $ResidDev(\mathcal{M}(\beta)) = 2(L(\mathcal{M}(\beta)) - L(\mathcal{M}(Null)))$ est la déviance résiduelle et $\frac{Deviance(\mathcal{M}(\beta))}{Deviance(\mathcal{M}(Null))}$ est le rapport des déviances. C'est la proportion de la déviance du modèle nul expliquée par le modèle $\mathcal{M}(\beta)$). La déviance résiduelle est positive si

$$\hat{\beta}(\lambda) = Arg\max_{\beta}\left[L(\mathcal{M}(\beta)) - \lambda\sum_{j=1}^{p}|\beta_j|\right] \tag{5.23}$$

Supposons que :

$$R = \frac{Deviance(\mathcal{M}(\beta))}{Deviance(\mathcal{M}(Null))} \text{ et } r = \frac{ResidDev(\mathcal{M}(\beta))}{Deviance(\mathcal{M}(Null))}$$

L'équation 5.22 devient :

$$R = 1 - r \tag{5.24}$$

La minimisation de la déviance suivant chacune des valeurs du paramètre $\lambda$ de pénalisation conduit à un modèle de paramètres $\hat{\beta}(\lambda)$ noté $\mathcal{M}(\hat{\beta}(\lambda))$. L'objectif principal du GLM-Lasso est de produire un modèle minimisant le rapport $R$ ou en maximisant le rapport $r$. L' équation 5.24 donne

$$R = \frac{Deviance(\mathcal{M}(\hat{\beta}(\lambda_k)))}{Deviance((\hat{\beta}(\lambda_{max}))} = 1 - r \tag{5.25}$$

ainsi

$$Deviance(\mathcal{M}((\hat{\beta}(\lambda_k))) = (1 - r) \times Deviance(\mathcal{M}((\hat{\beta}(\lambda_{max}))) \tag{5.26}$$

La valeur optimale $\lambda.min$ de $\lambda$ qui minimise la fonction $Deviance$ est :

$$\lambda.min = Arg \min_{\lambda_k}[Deviance(\mathcal{M}((\hat{\beta}(\lambda_k)))] \tag{5.27}$$

La valeur $\lambda.1se$ de $\lambda$ définie par Hastie et *al* qui minimise la déviance plus sa déviation standard [99, 101, 102] est :

$$\lambda.1se = Arg \min_{\lambda_k}[Deviance(\mathcal{M}(\hat{\beta}(\lambda_k))) + Std(Deviance(\mathcal{M}(\hat{\beta}(\lambda_k))))]. \tag{5.28}$$

#### 5.1.2.2 Algorithme (LOLO-DCV) appliqué aux données du paludisme

Dans le cadre de notre travail, un bloc est une maison de capture. Les observations d'un bloc sont toutes les observations d'une maison. Au second niveau de LOLO-DCV, les deux paramètres de régularisation $\lambda.min_k$ et $\lambda.1se_k$ sont déterminés. La méthode spécifique à utiliser pour déterminer le sous ensemble optimal de variables au premier niveau de LOLO-DCV est la présence $\mathcal{P}(X_i)$ de chacune des variables utilisant $\lambda.min_k$ et $\lambda.1se_k$. On a l'algorithme suivant :

**Algorithme 5.1** LOLO-DCV appliqué aux données

1. Les données sont disivées en $N_f$-blocs
2. A chaque premier niveau $k$
   (a) Les blocs sont regroupés en deux lots : $A_k$ et $E_k$, $A_k$ : l'ensemble d'apprentissage contenant les observations de $(N_f - 1)$-blocs, $E_k$ : l'ensemble de test, contenant les observations du dernier bloc.
   (b) Mise de côté $E_k$
   (c) Le second niveau de validation croisée
      i. Une validation croisée complète est exécutée sur $A_k$
      ii. Les deux paramètres de régularisation $\lambda.min_k$ et $\lambda.1se_k$ sont obtenus.
      iii. Les coefficients des variables actives, c'est-à-dire les variables à coefficients non nul associés à ces paramètres sont débiaisés
      iv. Les prédictions sont déterminées via un simple GLM sur $E_k$
      v. La présence $\mathcal{P}(X_i)$ de chacune des variables est déterminée utilisant $\lambda.min_k$ et $\lambda.1se_k$ sur $A_k$
3. L'étape 2c est répétée jusqu'à ce que les prédictions soient déterminées pour toutes les observations.

#### 5.1.2.3 Critères de qualité

Les critères de qualité utilisés dans cette étude sont : la moyenne des prédictions, le risque quadratique des prédictions, le risque absolu des prédictions et la déviance du modèle.

#### 5.1.2.4 Sélection par la méthode des variables fréquentes

Soit $\mathcal{V} = \{V_1, V_2, \ldots, V_{N_{cov}}\}$ l'ensemble de toutes les variables y compris les interactions du second ordre. D'après l'algorithme LOLO-DCV 4.5, à chaque premier niveau $k, 1 \leq k \leq N_f$, le second niveau de validation croisée donne deux valeurs de lambda : $\lambda.min_k$ et $\lambda.1se_k$ équation. 5.27 et 5.28. $\lambda.min_k$ et $\lambda.1se_k$ engendrent deux vecteurs $\beta(\lambda.min_k)$ et $\beta(\lambda.1se_k)$ de coefficients des covariables. Se basant sur ceci, on peut déterminer la présence ou l'absence de chaque variable, équation 4.69 Pour un seuil $s$, $1 \leq s \leq 100$, l'ensemble des variables fréquentes (FV) est donné par l'équation 4.70

#### 5.1.2.5 Stratégies de sélection de variables

Dans ce travail, deux stratégies de sélection de variables ont été implémentées et comparées à la méthode de référence. Chaque stratégie de sélection de variables est appliquée à deux groupes de variables, les variables originales et les

variables recodées. La première LDLM basée sur LOLO-DCV utilisant $\lambda.min$ de l'équation. 5.27. La deuxième stratégie, LDLS est base sur LOLO-DCV utilisant $\lambda.1se$ de l'équation. 5.28. A la fin du processus, LDLM et LDLS sélectionnent les meilleurs sous-ensembles de variables qui sont utilisés pour la prédiction. La différence entre ces deux stratégies est la valeur du paramètre lambda dans les équations. 5.27 et 5.28 qui a servi comme base dans le processus de sélection. Le minimum $s$ de l'équation (4.70) prend des valeurs différentes pour mesurer la souplesse de l'algorithme en sélection et en prédiction. Pour cela, $s \in \{75, 80, 90, 95, 100\}$. Si le pourcentage de présence est égal à ce minimum, cette covariable est considérée comme présente et peut appartenir à l'ensemble des variables fréquentes. Les sous-ensembles correspondants obtenus sont utilisés pour prédiction.

### 5.1.3 Résultats et discussion

Le tableau 5.1 présente les résultats sur les critères de qualité pour la méthode de référence BGLM et pour la validation croisée (CV) à un niveau associé au Lasso.

TABLE 5.1 – **Résultats de la méthode de référence BGLM et de la validation croisée à un niveau associé au Lasso**.
*CV $\lambda.min$ signifie validation croisée utilisant Lambda.min et CV $\lambda.1se$ signifie validation croisée utilisant Lambda.1se.*

| Variables | Méthode | Moyenne | Risque quadratique | Risque absolu |
|---|---|---|---|---|
| - | B-GLM | 3.75 | 62.29 | 3.88 |
| Originales | CV $\lambda.min$ | 3.75 | 624.65 | 5.95 |
| | CV $\lambda.1se$ | 3.86 | 190.81 | 5.16 |
| Recodées | CV $\lambda.min$ | 3.96 | 130.27 | 5.15 |
| | CV $\lambda.1se$ | 3.74 | 134.60 | 5.72 |

Pour les deux types de Lambda et pour les deux groupes de variables, les risques quadratique et absolu pour les prédictions par simple validation croisée sont plus élevés que ces mêmes risques pour les prédictions par la méthode de référence B-GLM, tableau 5.1. La méthode Lasso combinée avec une simple validation croisée est moins performante que la méthode de référence. Ceci implique l'introduction d'un second niveau de validation pour améliorer les résultats.

Les figures 5.1 et 5.2 présentent le processus de sélection de variables pour les stratégies LDLM et LDLS ceci pour différents seuils $s$ fixés, le seuil étant la fréquence de présence minimale que doit atteindre une variable pour être

sélectionnée dans le sous-ensemble optimal, Equation (4.70).

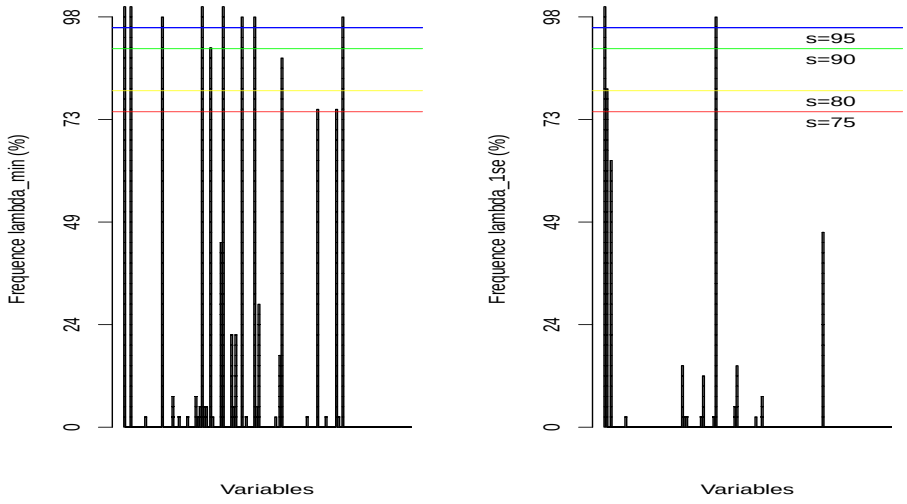



FIGURE 5.1 – **Variables fréquentes parmi les variables originales.** *Sur l'axe des abscisses, les variables et les interactions du second ordre, sur l'axe des ordonnées, les pourcentages de présence des variables. La sous-figure de gauche concerne $\lambda.min$ et celle de droite $\lambda.1se$.*

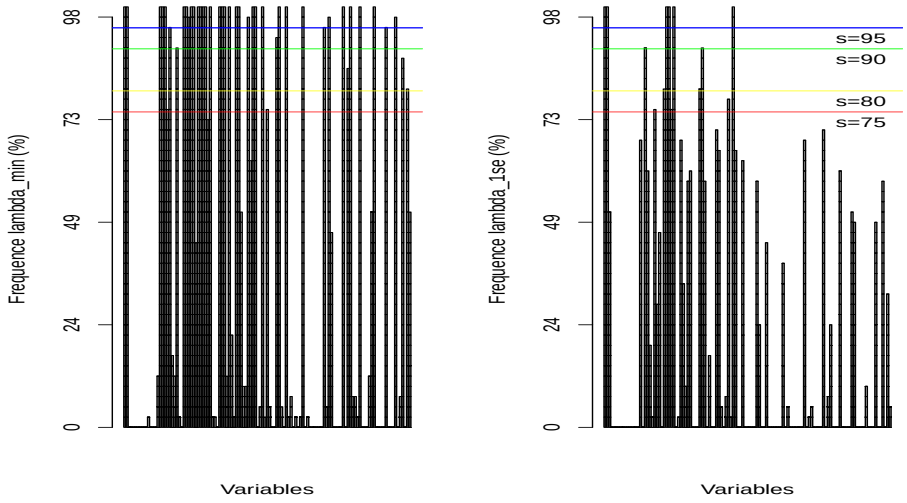



FIGURE 5.2 – **Variables fréquentes parmi les variables recodées.**

Au niveau des figures 5.1 et 5.2, chaque bande verticale représente une variable et la hauteur de la bande est proportionnelle à la fréquence de présence des variables dans les stratégies LDLM et LDLS. Chaque ligne horizontale représente un seuil donné. Ces figures montrent les résultats de sélection de variables originales et recodées. Pour un seuil donné, le nombre de barres verticales qui rencontrent la ligne horizontale correspondant à ce seuil donne le nombre de variables sélectionnées pour constituer le sous ensemble optimal.

Le tableau 5.2 associé à la figure 5.1 présente la fréquence de présence des variables originales stables suivant les différents seuils pour les stratégies LDLM et LDLS.

TABLE 5.2 – **Fréquence de présence des variables originales pour les stratégies LDLM et LDLS**

| **Variable** | **LDLM (%)** | **LDLS (%)** |
|---|---|---|
| Saison | 100 | 100 |
| Qté moyenne de pluie : Ouvertures | 100 | 80 |
| RND10 : Nombre d'habitants | 100 | - |
| RND3 : Végatation | 100 | 95 |
| Saison : Cours d'eau | 95 | - |
| Saison : Type de sol | 95 | - |
| Saison : Village | 95 | - |
| Qté moyenne de pluie : Végetation | 95 | - |
| RND3 : Village | 90 | - |
| Saison : RND3 | 80 | - |
| Saison : Répulsif | 75 | - |
| Saison : Travaux | 75 | - |

*RND10 = Jours de pluie avant la mission*
*RND3 = Jours de pluie pendant la mission*

Les tableaux 5.3 et 5.4 associés respectivement aux figures 5.1 et 5.2 montrent le nombre de variables sélectionnées pour les stratégies LDLM et LDLS lorsque le seuil varie.

TABLE 5.3 – **Nombre de variables originales sélectionnées pour les stratégies LDLM et LDLS**

| **Seuil (%)** | **NV pour LDLM** | **NV LDLS** |
|---|---|---|
| 100 | 4 | 1 |
| 95 | 8 | 2 |
| 90 | 9 | 2 |
| 80 | 10 | 3 |
| 75 | 12 | 3 |

*NV signifie Nombre de variables*

TABLE 5.4 – **Nombre de variables recodées sélectionnées pour les stratégies LDLM et LDLS**

| **Seuil (%)** | **NV pour LDLM** | **NV pour LDLS** |
|---|---|---|
| 100 | 31 | 11 |
| 95 | 39 | 11 |
| 90 | 44 | 16 |
| 80 | 50 | 22 |
| 75 | 52 | 29 |

*NV signifie Nombre de variables*

les tableaux 5.5 et 5.6 montrent les critères de qualité pour la méthode de référence BGLM et pour les stratégies LDLM et LDLS.

TABLE 5.5 – **Critères de qualité pour la méthode B-GLM et les stratégies LDLM et LDLS sur variables originales.**

| Seuil | Méthode | Moyenne | Risque quadratique | Risque absolu |
|---|---|---|---|---|
| - | B-GLM | 3.75 | 62.29 | 3.88 |
| 100 | **LDLM** | **3.74** | **44.26** | **3.30** |
| | LDLS | 3.74 | 54.50 | 3.62 |
| 95 | LDLM | 3.74 | 72.01 | 4.42 |
| | LDLS | 3.74 | 72.03 | 4.40 |
| 90 | LDLM | 3.74 | 72.00 | 4.47 |
| | LDLS | 3.75 | 72.01 | 4.42 |
| 80 | LDLM | 3.75 | 74.00 | 4.71 |
| | LDLS | 3.72 | 73.02 | 4.52 |
| 75 | LDLM | 3.74 | 71.84 | 4.41 |
| | LDLS | 3.74 | 72.00 | 4.31 |

TABLE 5.6 – **Critères de qualité pour la méthode B-GLM et les stratégies LDLM et LDLS sur variables recodées.**

| Seuil | Méthode | Moyenne | Risque quadratique | Risque absolu |
|---|---|---|---|---|
| | B-GLM | 3.75 | 62.29 | 3.88 |
| 100 | LDLM | 3.85 | 82.06 | 4.67 |
| | LDLS | 3.76 | 74.08 | 4.76 |
| 95 | LDLM | 3.84 | 81.06 | 4.61 |
| | LDLS | 3.76 | 74.08 | 4.76 |
| 90 | LDLM | 3.87 | 83.06 | 4.72 |
| | LDLS | 3.75 | 75.07 | 4.86 |
| 80 | LDLM | 3.87 | 84.06 | 4.81 |
| | LDLS | 3.75 | 75.07 | 4.86 |
| 75 | LDLM | 3.89 | 84.05 | 4.79 |
| | LDLS | 3.77 | 75.56 | 4.85 |

Selon les critères de qualité, le meilleur modèle pour l'algorithme Lasso est obtenu par la stratégie LDLM au seuil 100, tableau 5.5, ligne 2. Ainsi, le nombre de variables dans ce modèle optimal final est quatre et ces variables se présentent comme suit : Saison, interaction entre Quantité moyenne de pluie et Ouvertures, interaction entre Jours de pluie avant la mission et Nombre d'habitants, interaction entre Jours de pluie pendant la mission et Végétation. La figure 5.3 montre les prédictions obtenues par la méthode BGLM et celles obtenues par LOLO-DCV comparées aux observations dans huit maisons de captures.

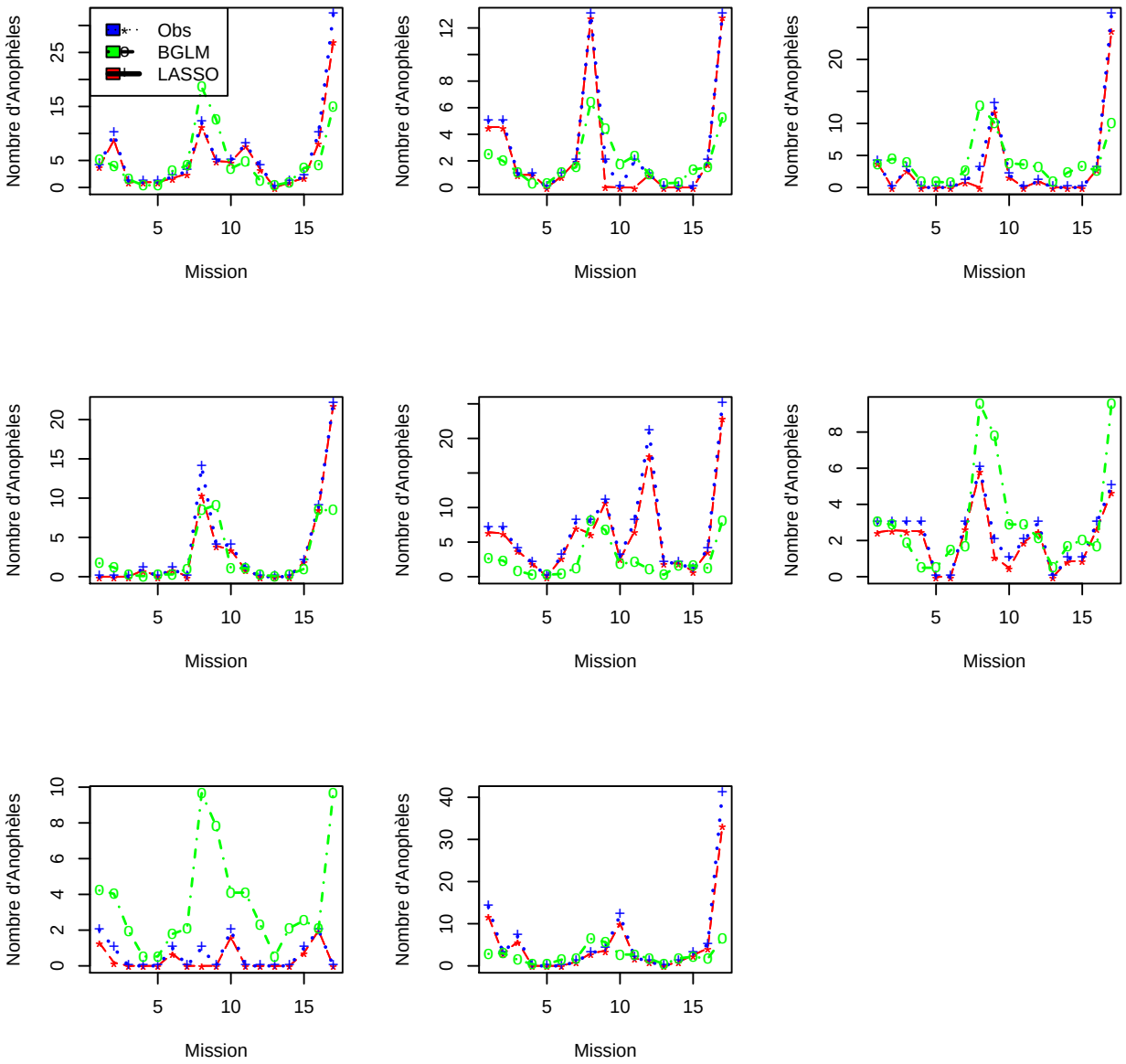



FIGURE 5.3 – **Comparaison des observations et des prédictions BGLM et l'algorithme LOLO-DCV sur huit maisons**.
*Obs=Nombre d'anophèles Observé.*
*La courbe en rouge est celle des observations, la courbe en vert est celle des prédictions du modèle de référence et celle en bleu est celle des prédictions par l'algorithme LOLO-DCV*

La figure 5.4 présente les prédictions obtenues par la méthode BGLM et celles obtenues par LOLO-DCV comparées aux observations dans les neufs villages du projet. La figure 5.6 en annexe, présente les prédictions obtenues

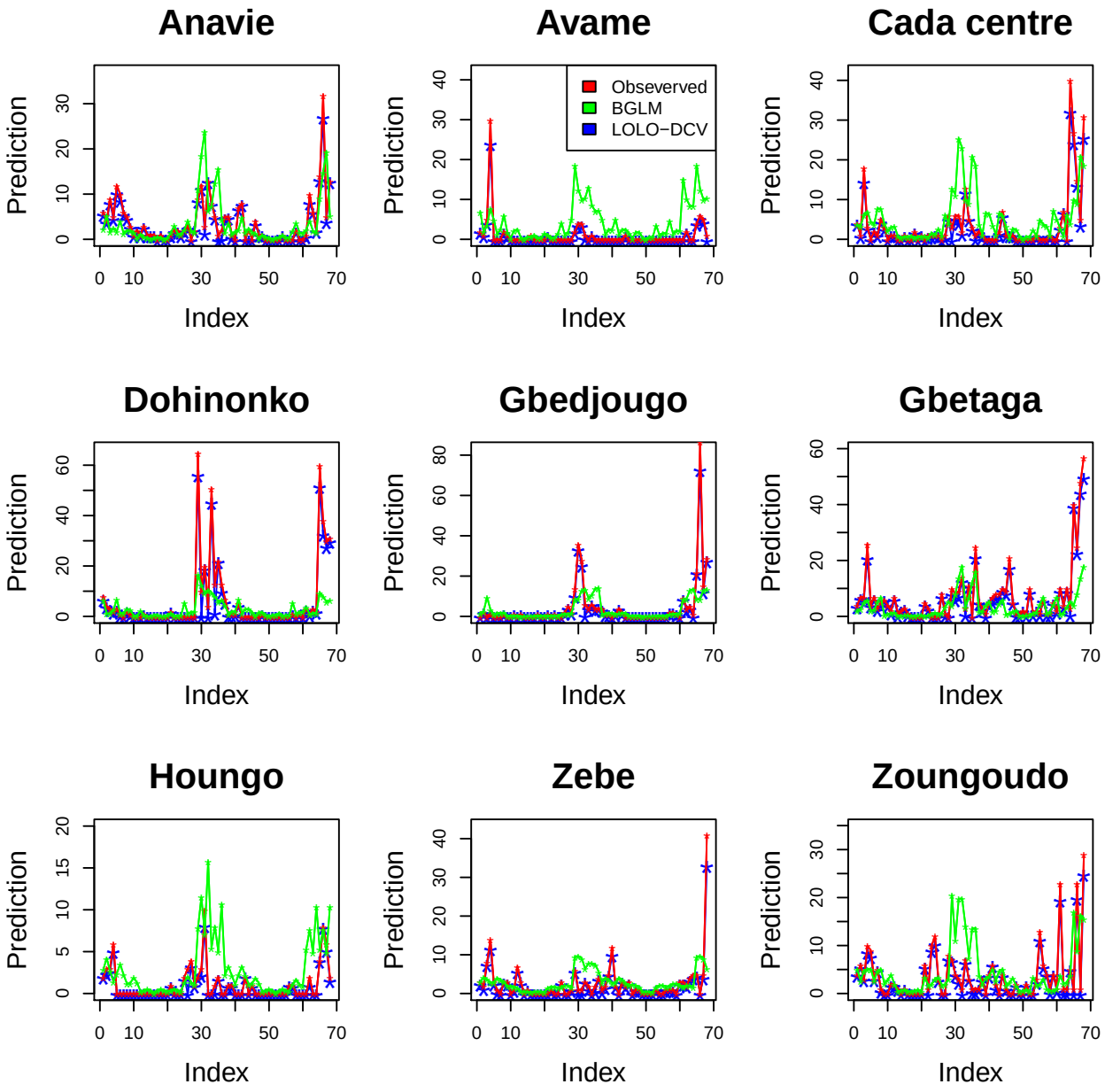



FIGURE 5.4 – **Comparaison des observations et des prédictions BGLM et l'algorithme LOLO-DCV sur les neuf villages**.
*La courbe en rouge est celle des observations, la courbe en vert est celle des prédictions du modèle de référence et celle en bleu est celle des prédictions par l'algorithme LOLO-DCV*

par la méthode BGLM et celles obtenues par LOLO-DCV comparées aux ob-

servations dans 41 maisons de capture du projet.

La figure 5.5 montre les quantiles théoriques en fonctions des quantiles empiriques pour la méthode LOLO-DCV et la méthode de référence BGLM. Selon

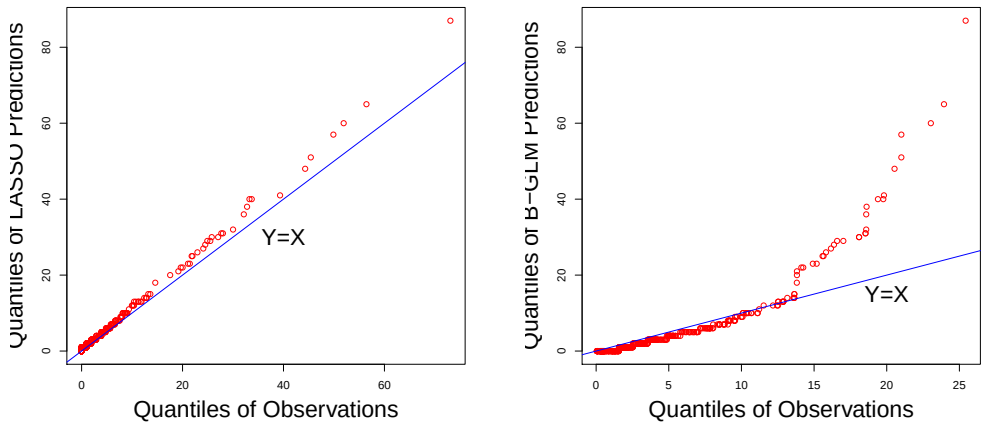



FIGURE 5.5 – **Comparaison des Quantiles des observations en fonction des quantiles des prédictions de l'algorithme LOLO-DCV.**
*La courbe de gauche est celle des prédictions de l'algorithme LOLO-DCV et la courbe de droite est celle des prédictions du modèle de référence. La ligne bleue est la première bissectrice (Y=X).*

les tableaux 5.3 et 5.4, on remarque que le nombre de variables sélectionnées augmente au fur et à mesure que le seuil diminue. Ceci est cohérent parce que plus le seuil diminue, plus on est souple dans les conditions de sélection. On constate également que pour tous les seuils, la stratégie LDLS sélectionne moins de variables que LDLM. LDLS est plus sparse que LDLM. Selon les figures 5.3 et 5.6 présentant les résultats au niveau des maisons de capture, on constate que les prédictions obtenues par l'algorithme LOLO-DCV sont plus proches des valeurs réelles que les prédictions de la méthode de référence BGLM. Ces résultats sont confirmés par la figure 5.5 qui montre que les quantiles des prédictions du LOLO-DCV en fonction des quantiles des observations sont bien ajustés par la première bissectrice (Y=X) tandis que les quantiles des prédictions du B-GLM en fonction des quantiles des observations sont moins ajustés par la première bissectrice. Ceci confirme que l'Algorithme LOLO-DCV est plus performant que la méthode de référence en matière de prédiction.

Pour chacun des groupes, le meilleur sous ensemble de covariables est sélectionné en se basant sur le compromis entre l'application des critères et la sparsité du sous ensemble optimal de covariables selectionné pour la prédiction. Globalement la moyenne en prédiction pour les deux stratégies appliquées aux deux groupes sont proches ou égales à la moyenne des observations.

Les figures 5.1 et 5.2, présentent deux classes de variables, les plus fréquentes et les moins fréquentes. Les tableaux 5.5 et 5.6 montrent que les risques quadratique et absolu augmentent au fur et à mesure que le seuil diminue. Mais il faut remarquer que ces risques sont stables à partir du seuil 95 et ne sont pas très éloignés de ceux obtenus au seuil 100, ceci montre la souplesse de l'algorithme LOLO-DCV.

Pour la sélection de variables dans le groupe 1 avec 136 variables, les méthodes classiques devraient construire $2^{136}$ différents modèles avant de sélectionner le meilleur sous ensemble optimal. La combinaison d'une méthode classique avec la double validation croisée rendrait les calculs très difficiles à cause de la complexité de l'algorithme. La force de LOLO-DCV est la combinaison harmonieuse du Lasso et de la double validation croisée. En un temps relativement court, LOLO-DCV arrive à détecter parmi les variables sélectionnées par le B-GLM, un nombre plus petit de variables interprétables et des interactions du second ordre interprétables entre elles. La distribution des erreurs de prédiction selon les classes d'anophèles montre une variabilité élevée pour B-GLM et faible pour LOLO-DCV. La stabilité de LOLO-DCV est prouvée par le fait que le sous ensemble optimal de variables pour la prédiction est pratiquement le même à chaque étape du premier niveau dans la double validation croisée. Le meilleur sous ensemble de covariables pour la prédiction est sélectionné sur le groupe 1.
Ces résultats montrent que le traitement sur les covariables n'est pas nécessaire. D'autres travaux en biologie, épidémiologie et en médecine ont montré que les résultats des méthodes classiques peuvent être améliorés par les algorithmes d'apprentissage machine [103], [104], [105], [106], [107] [108], [109]. Ces résultats pourraient être améliorés si la possible corrélation dans les données au niveau maison était pris en compte en développant un algorithme combinant le GLMM, le Lasso et la double validation croisée. Les résultats obtenus par cette procédure sont clairement améliorés comparés à ceux obtenus par le B-GLM pris comme méthode de référence. L'amélioration concerne toutes les propriétés telles que la qualité de la sélection et de la prédiction. En plus le temps d'exécution des algorithmes est nettement réduit et seules quelques variables environnementales et climatiques sont associées au risque d'exposition au paludisme avec une amélioration en précision.

## 5.2 Conclusion

La méthode Lasso est une méthode attractive parce qu'elle fait à la fois de la sélection, de l'estimation et de la prédiction. Les estimateurs Lasso sont obtenus par résolution d'un problème d'optimisation des moindres carrés sous contrainte. Ces estimateurs sont linéaires par morceau. Ils possèdent de bonnes

propriétés sous des hypothèses fortes comme celles sur la matrice de Gram. Les propriétes oracles ne sont obtenues que lorsque certaines conditions sont réunies. L'algorithme résolvant le problème Lasso est lent en convergence et les estimateurs ne sont pas toujours consistants surtout dans le cas de colinéarité entre les prédicteurs. Tout ceci montre les limites de l'estimateur du Lasso et permet la mise au point de plusieurs autres méthodes améliorant et généralisant la méthode Lasso. l'algorithme LARS est une amélioration notable pour la méthode Lasso. La forme adaptative du Lasso a permis d'alléger des conditions nécessaires pour obtenir des estimateurs suffisamment sparses, interpretables avec des propriétés oracles. Les différentes variantes du Lasso dépendent de la perte et de la pénalité utilisée dans l'optimisation. Dans le cas où la variable d'intérêt $Y$ n'est pas continue, il convient d'utiliser un modèle GLM combiné avec la méthode Lasso. Ce problème d'optimisation exige des outils un peu plus compliqués. Il faut commencer par l'approximation de Laplace pour l'estimation de la vraisemblance du modèle, le calcul du gradient, la méthode du gradient ascendant pour l'évalution de la plus grande pente, les dérivées directionnelles pour l'approximation de Taylor, la méthode de Newton-Raphson pour résoudre les équations du type $f(x) = 0$, la méthode EM dans la méthode du Score de Fisher, etc.
La variable d'intérêt étant de comptage, nous avons utilisé une combinaison du GLM et du Lasso (GLM-Lasso). Pour éviter le sur-apprentissage une stratégie de double validation croisée intégrée au GLM-Lasso a été développée. Cette combinaison a donné l'algorithme LOLO-DCV. Cette méthode appliquée aux données liées au paludisme a donné de bons résultats comparativement à la méthode B-GLM développée dans le chapitre 4. L'algorithme LOLO-DCV génère automatiquement toutes les interactions du second ordre entre les variables et sélectionne à la fin du processus un sous ensemble de variables cohérentes, pertinentes et interprétables. Le sous ensemble optimal est obtenu sur les variables originales. Tout ceci prouve que les traitements opérés sur les variables ne sont pas nécessaires. Le temps de calcul est réduit par rapport aux méthodes classiques. L'algorithme LOLO-DCV est complètement automatique, rapide, surmonte les traitements sur les variables. Il permet de sélectionner des variables cohérentes, pertinentes et interprétables, et permet de faire de bonnes prédictions. Le temps de calcul a été amélioré ainsi que le sous ensemble de covariables pertinentes obtenu est interprétable.
Ces résultats pourraient être améliorés si la possible corrélation dans les données au niveau des maisons et des villages était prise en compte en développant un algorithme combinant le GLMM, le LASSO et la double validation croisée.

# Conclusion générale et perspectives

Cette thèse nous a permis d'explorer les différentes stratégies de construction de fonction de prévision et de sélection de variables pour cette fonction de prédiction par les méthodes d'apprentissage machine automatique. Les différents algorithmes implémentés et mis en œuvre nous ont aidé à la détermination des facteurs environnementaux pouvant expliquer la variabilité de la densité anophélienne et de pouvoir prédire le risque d'exposition au vecteur palustre au niveau village et maison dans le milieu d'étude. Ces prédictions peuvent être également obtenues dans d'autres milieux d'étude ou les données liées au vecteur palustre ne sont pas disponibles. La formalisation des objectifs des experts en problèmes statistiques nous a permis de travailler d'abord de façon théorique et ensuite de façon concrète sur des données réelles dans un contexte épidémiologique. Les algorithmes développés et appliqués aux données palustres de Tori-Bossito ont permis d'élaborer une nouvelle méthode de validation croisée stratifiée à deux niveaux, une nouvelle approche de sélection de variables et de prédiction en grande dimension. Egalement, les pré-traitements opérés par les experts avant l'analyse des données ont été surmontés dans le cadre de ce travail mais ces pré-traitements pourraient donner de bons résultats si au lieu des quartiles, on pourrait utiliser les déciles, ce qui augmenterait le nombre de classes et la quantité d'information contenue dans les variables recodées. Une autre alternative serait de voir dans quelle mesure les prétraitements pourraient améliorer les résultats obtenus. On pourra procéder à des changements d'échelle sur des variables, transformer des variables numériques continues en variables non numériques ordinales, etc. Nous avons déterminé un nouvel ensemble optimal de facteurs climatiques et environnementaux plus parcimonieux et plus pertinent pour la prévision du risque d'exposition au vecteur palustre qu'est l'anophèle. Les résultats obtenus par la méthode basée sur le Lasso peuvent être améliorés avec la forme adaptative du Lasso mais cette forme n'est pas encore disponible pour les données comme celles que nous avons utilisées. Nous pourrons penser à l'élaboration d'un ensemble d'algorithmes pour réaliser la sélection et la prédiction pour les

données de comptage, hiérarchiques ou emboîtées. Les algorithmes développés dans le chapitre 4 serviront de base à l'élaboration d'un package R offrant de nouvelles stratégies de sélection de variables en grande dimension aux praticiens. Ces packages seront mis sur le CRAN et pourront être exploités par les utilisateurs de R.

# Bibliographie


[1] Bates D (2012). Mixed models in R using the lme4 package Part 5 : Generalized linear mixed models. Departement of Statistics, University of Winconsin-Madison, `Douglas.Bates@R-project.org`.

[2] Bates D (2012). Fitting mixed-effect models using the lme4 package in R . International Meetting of the Psychometric Society, Departement of Statistics, University of Winconsin-Madison, `Douglas.Bates@R-project.org`.

[3] Cottrell G, Kouwaye B, Pierrat C, le Port A, Bouraïma A, et al. (2012) Modeling the Influence of Local Environmental Factors on Malaria Transmission in Benin and Its Implications for Cohort Study. PlosOne 7 : 1.

[4] Bach F (2008). Bolasso : Model Consistent Lasso Estimation through the Bootstrap. *Appearing in Proceedings of 25th* International Conference on Machine Learning, Helsinki, Finland,.

[5] Bach F (2009). Model-Consistent Sparse Estimation through the Bootstrap. *Willow Project-team Laboratoire d'Informatique de l'Ecole Normale Supérieure (CNRS/ENS/INRIA UMR 8548)* 45, rue d'Ulm, 75230 Paris, France francis.bach@mines.org.

[6] Organisation mondiale de la santé (OMS) (2015) Rapport sur le paludisme dans le monde. Rapport annuel : 28.

[7] WHO (2015) World Health Organisation, World malaria REPORT 2015, World global malaria programme. WHO Library Cataloguing-in-Publication Data : 248.

[8] Ministère de la santé (Bénin), Direction de la Programmation et de la prospective (2010) Annuaire des Statistiques sanitaires (ASS) .

[9] Ministère de la santé (Bénin), Programme national de lutte contre le paludisme (2010) Evaluation des activités de lutte contre le paludisme au bénin. Rapport MIS : 171.

[10] Ministère de la santé (Bénin), Direction Nationale de la Santé Publique (2014) Programme national de lutte contre le paludisme. Rapport annuel d'activité : 43.

[11] Organisation Mondiale de la santé (OMS) (2015) Paludisme : Groupes à haut risque. OMS : 243.

[12] Coll O, Menendez C, Botet F, Dayal R (2008) it Treatment and prevention of malaria in pregnancy and newborn,. Infections Med 36(1) : 15-29.

[13] Greenwood M A, Armstrong R J, Byass P, Snow W R, Greenwood M B (1992) Malaria chemoprophylaxis, birth weight and child survival. Trans R Soc Trop Med Hyg 86 : 483-5.

[14] Le Hesran J Y, Cot M, Personne P, Fievet N, Dubois B, et al. (1997) Maternal placental infection with Plasmodium falciparum and malaria morbidity during the first 2 years of life,. Am J Epidemiol 146 : 826-831.

[15] Mutabingwa T K, Bolla M C, Li J L, Domingo G J, Li X, et al. (2005) Maternal malaria and gravidity interact to modify infant susceptibility to malaria,. PLoS Med 2(12) : e407.

[16] Schwarz N G, Adegnika A A, Breitling L P, Gabor J, Agnandji S T, et al. (2008) Placental malaria increases malaria risk in the first 30 months of life,. Clin Infect Dis 47(8) : 117-125.

[17] Abdulla S, Schellenberg J A, Nathan R, Mukasa O, Marchant T, et al. (2001) Impact on malaria morbidity of a programme supplying insecticide treated nets in children aged under 2 years in tanzania : community cross sectional study. BMJ 322(7281) : 270-273.

[18] Yé Y, Louis V R, Simboro S, Sauerborn R (2007) Effect of meteorological factors on clinical malaria risk among children : an assessment using village-based meteo- rological stations and community-based parasitological survey. BMC 7 : 101.

[19] Garcia A, Alle Baba D, Rouget F, Migot-Nabias F, Le Hersan J Y, et al. (2004) Role of environment and behaviour in familial resemblances of plasmodium falciparum infection in a population of senegalese children. Microbes and infections 6 : 68-75.

[20] Li L, Ling B, Guiyun Y (2004) A study of the distribution and abundance of the adult malaria vector in western kenya malaria vector in western kenya. Microbes and infections 6 : 68-75.

[21] Silué K D, Raso G, Yapi A, Vounatsou P, Tanner M, et al. (2008) Spatially-explicit risk profiling of plasmodium falciparum infections at a small scale : a geostatistical modelling approach. Malaria Journal 7 : 111.

[22] Guthmann H, Llanos-Cuentas A, Palacios A, Hall A J (2002) Environmental factors as determinants of malaria risk. a descriptive study on the northern coast of peru. Trop Med Int Health 7 : 518-525.

[23] Mabaso M H L, Craig M, Ross A, Smith T (2007) Environmental predictors of the season- nality of malaria transmission in africa : the challenge. Am J Trop Med Hyg 76(1) : 33-38.

[24] Kouwaye B, Fonton N, Rossi F (2015) Sélection de variables par le glmlasso pour la prédiction du risque palustre. In : 47èmes Journees de Statistique de la SFdS, Lille, France. hal-01196450, Hal.

[25] Kouwaye B, Fonton N, Rossi F (2015) Lasso based feature selection for malaria risk exposure prediction. In : 11th International Conference, MLDM 2015 Hamburg, Germany, July 2015 Poster Proceedings, ibai publishing. Petra Perner (Ed.), Machine Learning and Data Mining in Pattern Recognition.

[26] Kouwaye B, Fonton N, Rossi F (2015) Sélection de variables par le GLM-Lasso pour la prédiction du risque palustre . URL `https://arxiv.org/format/1509.02873`. Working paper or preprint.

[27] Kouwaye B, Fonton N, Rossi F (2015) Lasso based feature selection for malaria risk exposure prediction. URL `https://arxiv.org/format/1511.01284`. Working paper or preprint.

[28] Kouwaye B (2016) Anopheles number prediction on environmental and climatevariables using Lasso and stratified two levels cross validation. URL `https://hal.archives-ouvertes.fr/hal-01336317`. Working paper or preprint.

[29] Vapnik N V (1995) The nature of statistical learning theory. New York, NY, USA : Springer-Verlag New York, Inc.

[30] Ben Ishak A (2007). Sélection de variables par les machines à vecteurs supports pour la discrimination binaire et multiclasse en grande dimension.

[31] Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 43 : 59-69.

[32] Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning *Data mining, Inference, Prediction*. Second edition.

[33] Dohono L D, Johnstone M I (1995) Adapting to unknown smoothess via wavlet shrinkage. AmerStatis Assoc : 1200-1224.

[34] Dohono D L, Johnstone I M (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika : 425-455.

[35] Brigé L (2006) Model selection via testing : an alternative to (penalized) maximum likelihood estimators. Ann Inst H Poincaré Probab Statis 42(3) : 273-325.

[36] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate : a pratical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57(1) : 289-300.

[37] Bauera P, Potschera M B, Hackla P (1998) Model selection by multiple test procedures. Statistics 19(1) : 39-44.

[38] Pötsher M B (1983) Order estimation in ARMA-models by Lagrangian multiplier tets. Ann Statis 11(3) : 872-885.

[39] Bunea F, Wegkamp H M, Auguste A (2006) Consistent variable selection in high dimension regression via multimle testing. J Statis Plann Inference 136(12) : 4349-4364.

[40] Foster P D, George I E (1994) The risk inflation criterion for multiple regression. Ann Statist 22 (1994).

[41] Foster P D, George I E (2000) Calibration and empirical bayes variable selection. Biometrika 87 : 731–747.

[42] Mallows C L (1973) Some comments on Cp. Technometrics 15(4) : 661-675.

[43] Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Second International Symposuim on Information Theory (Tsahkador, 1971) : 267-281.

[44] Schwart G (1978) Estimating the dimension of a model. Ann Statis 6(2)) : 461-464.

[45] Birge L, Massart P (2001) A generalized Cp criterion for gaussian model selection. *Prépublication de Laboratoire* n°647 : 39.

[46] Hebiri M (2009) Quelques questions de sélections de variables autour de l'estimateur Lasso. Ph.D. thesis.

[47] Vapnik N V (1999) Statistical learning theory. Wiley Inter science,.

[48] Kouwaye B (2011) Modelisation du risque spatio temporel d'exposition palustre à Tori-Bossito (Bénin). Master's thesis.

[49] Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. The Annals of statistics 32 : 407–499.

[50] Bates D (2010). Linear mixed model implementation in lme4. Departement of Statistics, University of Winconsin-Madison, `Douglas.Bates@R-project.org`.

[51] Tempelman J R, Gianola D (1993) Marginal maximum likelihood estimation of variance components in Poisson mixed models using Laplacian integration. Genetics Selection Evolution 25 : 305-319.

[52] Breslow E N, Clayton G D (1993) Approximate inference in generalized linear mixed model. Journal of American Statistical Association 88(421) : 9-25.

[53] Pinheiro J, Bates D (1995) Approximations to the log-likelihood function in the nonlinearmixed-effects models. JCGS 4(1) : 12-35.

[54] Bates G (2004). Sparse matrix representations of linear mixed models. R Development Core Team.

[55] Bates D, DebRoy S (2006) Linear mixed models and penalized least squares. JMA 91(1) : 1-17.

[56] Skrondal A, Rabe-Hesketh S (2007) Prediction and diagnostics, in generalized linear mixed models, recent advances in multilevel modelling : Methodology and applications. Royal Satistical Society 33 : 6.

[57] Rabe-Hesketh S, Skrondal A (2008) Multilevel and Longitudinal Modeling Using Stata. Stata Press, 562 pp. 2nd Ed.

[58] Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall/CRC, editor, 436 pp.

[59] Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) : 267–288.

[60] Chen S S, Donoho L D, Saunders A M (1998) Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20 : 33–61.

[61] Dohono L D, Johnstone M I (1995) Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association 90 : 1200–1224.

[62] Donoho L D, Johnstone I, Johnstone M I (1993) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81 : 425–455.

[63] Zou H, Hastie T, Tibshirani R (2004) On the "degrees of freedom" of the lasso. Technical report, Annals of Statistics.

[64] Chesneau C, Hebiri M (2008). Some theorical results on the Gouped Variables Lasso. Hal-00145160, Version 3-3.

[65] Vapnik V (1998) The Nature of Statistical Learning Theory. Springer-Verlag.

[66] Fan J, Li R (2001) Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association 96 : 1348-1360.

[67] Bunea F, Tsybakov A, Wegkamp M (2007) Sparsity oracle inequalities for the lasso. Electronic Journal of Statistics 1 : 169-94.

[68] Sampson N J, Chattrejee N (2010) Oracle is not optimal : Adapting the adaptive lasso. Biostatistics 1 : 1-27.

[69] Van de Geer S, Bühlmann P (2009) On the conditions used to prove oracle results for the lasso. arXiv [MathsST] 0910.0722v1 : 1-33.

[70] Kwemou M (2012) On the conditions used to prove oracle results for the lasso. arXiv [MathsST] 1206.0710v3 : 1-39.

[71] Breiman L (1995) Better subset regression using the nonnegative garrote. Technometrics 37(4) : 373-384.

[72] Yuan M, Lin Y (2007) On the non-negative garrote estimator. J R Stat Soc Ser B Stat Methodol 69(2) : 143-161.

[73] Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society : Series B (Statistical Methodology) 67 : 301–320.

[74] Li Q, Lin N (2010) The bayesian elastic net. Bayesian Analysis 5(1) : 151-170.

[75] De Mol C, De Vito E, Rosasco L (2009) Elastic-net regularization in learning theory. Journal of Complexity .

[76] Tibshirani R, Saunders M (2005) Sparsity and smoothness via the fused lasso. J R Statist Soc B 67 : 91-108.

[77] Rinaldo A (2009) Properties and refinements of the fused lasso. The Annals of Statistics 37 : 2922-2952.

[78] Yuan M, Yuan M, Lin Y, Lin Y (2006) Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68 : 49–67.

[79] Meier L, Van De Geer S, Bühlmann P (2008) The group lasso for logistic regression. Journal of the Royal Statistical Society, Series B 70(1) : 53-71.

[80] Friedman J, Hastie T, Tibshirani R (2010) A note on the group lasso and a sparse group. ArXiv : 10010736v1 [mathST] : 1-8.

[81] Zou H, Hastie T (2006) The adaptative Lasso and its oracle properties. Journal of the American Statistical Association 101 : 1418-1429.

[82] Candes E, Tao T (2007) The Dantzig selector : Statistical estimation when p is much larger than n. Annals of statistics 35 : 2313-2351.

[83] James M G, Radchenko P, Dasso J Lv (2009) Connections between the dantzig selector and lasso. J Royal Statist Soc,Ser B 71 : 127-142.

[84] Wang H, Li G, Jiang G (2007) Robust regression shrinkage and consistent variable selection through the LAD-Lasso. J Bus Econom Statist 25(3) : 347–355.

[85] Rosset S, Zhu J (2007) Piecewise linear regularized solution paths. Ann Statist 35(3) : 1012-1030.

[86] Sara A, Van de Geer S (2008) High-dimensional generalized linear models and the lasso. The Annals of Statistics Vol. 36, No. 2 : 614–645.

[87] Sara A, Van de Geer S (2007) The deterministic Lasso. Technical Report 140. Seminar für Statistik, ETH Züricht.

[88] Koltchinskii V (2009) Sparsity in penalized empirical risk minimization. Ann Inst H Poincaré Probab Statist Volume 45, Number 1 : 7-57.

[89] Land R S, Friedman H J (1994) variable fusion : A new adaptive signal regression methods,. Technical Report 656, Department of Statistics, Stanford University.

[90] Donoho D (2004) Compressed sensing. Information Theory, IEEE Transactions 52 (4) : 1289 - 1306.

[91] Friedman F, Hastie T, Tibshirani R (2007) Sparse inverse covariance estimation with the graphical lasso. The Annals of Statistics .

[92] Tibshirani J R, Hoefling H, Tibshirani R (2007) Sparse inverse covariance estimation with the graphical lasso. The Annals of Statistics .

[93] Candes J E, Recht B (2008) Exact matrix completion via convex optimization. Foundation of Computational Mathematics 9 : 717-772.

[94] Witten M D, Tibshirani R, Hastie T (2009) Compressed sensing. Biostatistics 10(3) : 515-534.

[95] Goeman J J (2010) L1 penalized estimation in the cox proportional hazards model. Biometrical Journal 52 : 70–84.

[96] Kim Y, Kim Y (2004) Gradient Lasso for feacture selection. *In Preceedings of the 21st International Conference of machine learnig*. ACM International Conference Preceedings Series 69 : 473-480.

[97] Kouwaye B, Rossi F, Fonton N, Garcia A, Dossou Gbete S, et al. (2017) Predicting local malaria exposure using a lasso-based two-level cross validation algorithm. PLoS ONE 12(10) : e0187234. https ://-doi.org/10.1371/ journal.pone.0187234 : 14.

[98] Andrew Y Ng (1997) Preventing "overfitting" of cross-validation data. In : Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., ICML '97, pp. 245–253. URL `http://dl.acm.org/citation.cfm?id=645526.657119`.

[99] Friedman J, Hastie T, Simon N, Tibshirani R (2015). Lasso and elastic-net regularized generalized linear models. `http://www.jstatsoft.org/v33/i01/` R CRAN.

[100] Goeman JJ (2010) $L_1$ Penalized Estimation in Cox Proportional Hazards Model. Biometrical Journal 52 : 70-84.

[101] Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 : 1–22.

[102] Hastie J T, Tibshirani R, Friedman J, Jerome H (2009) The elements of statistical learning : data mining, inference, and prediction. Springer series in statistics. New York : Springer. URL `http://opac.inria.fr/record=b1127878`. Autres impressions : 2011 (corr.), 2013 (7e corr.).

[103] Bontempi G (2005) Structural feature selection for wrapper methods. In : ESANN 2005, 13th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 27-29, 2005, Proceedings. pp.

405–410. URL https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2005-97.pdf. :conf/esann/Bontempi05

[104] Kourou K, Exarchos TP, Exarchosa KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 13 : 8-17.

[105] Oermann EK, Rubinsteyn A, Ding D, Mascitelli J, Starke RM, et al. (2016) Using a Machine Learning Approach to Predict Outcomes after Radiosurgery for Cerebral Arteriovenous . Scientific Reports 13 : 12.

[106] Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D (2012) Machine Learning for Personalized Medicine : Predicting Primary Myocardial Infarction from Electronic Health Records. Association for the Advancement of Artificial Intelligence : AI MAGAZINE : 13.

[107] Li S, Oh S (2016) Improving feature selection performance using pairwise pre-evaluation. BMC Bioinformatics : 13.

[108] Wang H, Liu S (2016) An Effective Feature Selection Approach Using the Hybrid Filter Wrapper. International Journal of Hybrid Information Technology 9 : 119-128.

[109] van der Ploeg T, Steyerberg EW (2016) Feature selection and validated predictive performance in the domain of Legionella pneumophila : a comparative study. BMC Research Notes : 7.

TABLE 5.7 – **Description des variables originales**.
*NM=Nombre de modalités*

| | **Nature** | **Nm** | **Modalité** |
|---|---|---|---|
| Répulsif | Non-numérique | 2 | Oui/ Non |
| Moustiquaire | Non-numérique | 2 | Oui/ Non |
| Type de toit | Non-numérique | 2 | Tôle/ Paille |
| Ustensiles | Non-numérique | 2 | Oui/ Non |
| Présence de constructions | Non-numérique | 2 | Oui/ Non |
| Type de sol | Non-numérique | 2 | Humide/ Sec |
| Cours d'eau | Non-numérique | 2 | Oui/ Non |
| Classe majoritaire | Non-numérique | 3 | 1/4/7 |
| Saison | Non-numérique | 4 | 1/2/3/4 |
| Village | Non-numérique | 9 | |
| Maison | Non-numérique | 41 | |
| Jours de pluie avant la mission | Numérique | Discrète | $0/2/\cdots/9$ |
| Jours de pluie pendant la mission | Numérique | Discrète | $0/1/\cdots/3$ |
| Indice de fragmentation | Numérique | Discrète | $26/\cdots/71$ |
| Ouvertures | Numérique | Discrète | $1/\cdots/5$ |
| Nombre d'habitants | Numérique | Discrète | $1/\cdots/8$ |
| Quantité moyenne de pluie | Numérique | Continue | $0/\cdots/82$ |
| Végetation | Numérique | Continue | $115.2/\cdots/$ 159.5 |
| Nombre total de moustiques | Numérique | Discrète | $0/\cdots/481$ |
| Nombre total d'anophèles | Numérique | Discrète | $0/\cdots/87$ |
| Nombre total d'anophèles infectés | Numérique | Discrète | $0/\cdots/9$ |

TABLE 5.8 – **Description des variables recodées.**
*Les variables avec une étoile ont été recodées. NM=Nombre de modalités*

| | **Nature** | **Nm** | **Modalité** |
|---|---|---|---|
| Répulsif | Non-numérique | 2 | Oui/ Non |
| Moustiquaire | Non-numérique | 2 | Oui/ Non |
| Type de toit | Non-numérique | 2 | Tôle/ Paille |
| Utensils | Non-numérique | 2 | Oui/ Non |
| Présence de constructions | Non-numérique | 2 | Oui/ Non |
| Type de sol | Non-numérique | 2 | Humide/ Sec |
| Cours d'eau | Non-numérique | 2 | Oui/ Non |
| Classe majoritaire * | Non-numérique | 3 | 1/2/3 |
| Saison | Non-numérique | 4 | 1/2/3/4 |
| Village* | Non-numérique | 9 | |
| Maison * | Non-numérique | 41 | |
| Jours de pluie avant la mission * | Non-numérique | 3 | Quartile |
| Jours de pluie pendant la mission | Numérique | Discrète | $0/1/\cdots/3$ |
| Indice de fragmentation * | Non-numérique | 4 | Quartile |
| Ouvertures* | Non-numérique | 4 | Quartile |
| Nombre d'habitants * | Non-numérique | 3 | Quartile |
| Quantité moyenne de pluie * | Non-numérique | 4 | Quartile |
| Végetation * | Non-numérique | 4 | Quartile |
| Nombre total de moustiques | Numérique | Discrète | $0/\cdots/481$ |
| Nombre total d'anophèles | Numérique | Discrète | $0/\cdots/87$ |
| Nombre total d'anophèles infectés | Numérique | Discrète | $0/\cdots/9$ |

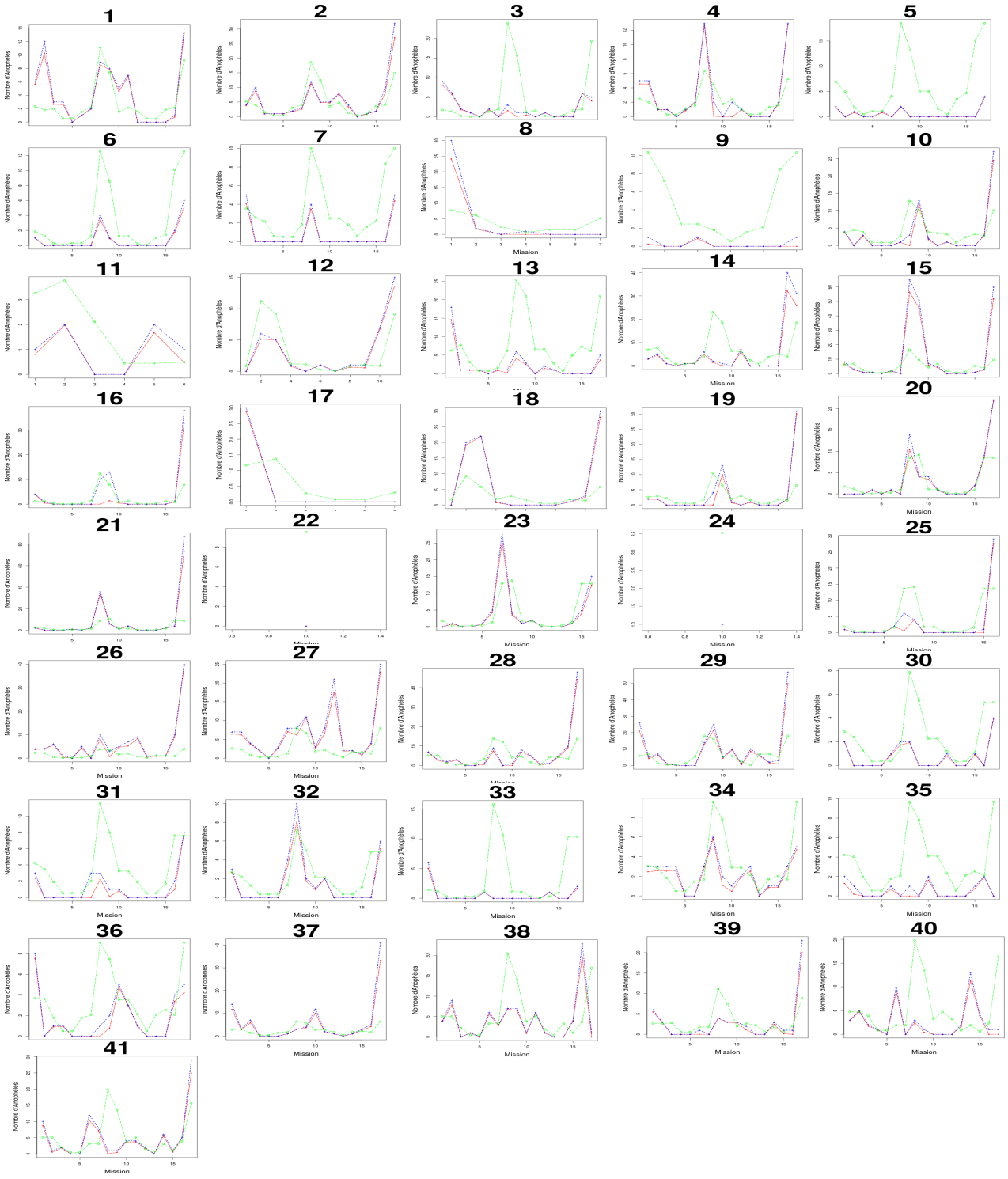



FIGURE 5.6 – **Comparaison des observations et des prédictions BGLM et l'algorithme LOLO-DCV sur les 41 maisons.**

*La courbe en rouge est celle des observations, la courbe en vert est celle des prédictions du modèle de référence et celle en bleu est celle des prédictions par le l'algorithme LOLO-DCV*