

UNIVERSITE D'AIX-MARSEILLE

ED 62-SCIENCES DE LA VIE ET DE LA SANTE

UFR DE MEDECINE

UMR 1252-Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale

Thèse présentée pour obtenir le grade universitaire de docteur

**Discipline :** Pathologie humaine

**Spécialité :** Recherche clinique et Santé Publique

Thèse préparée dans le cadre du réseau doctoral en santé publique animé par l'EHESP

**Parcours :** Biostatistique et Sciences de l'Information

**Abdoulaye GUINDO**

Modélisation de l'hétérogénéité spatiale du risque environnemental dans les  
essais de prévention randomisés contre les maladies transmissibles

Soutenue le 16/12/2019 devant le jury composé de :

Pr Roch GIORGI	Aix-Marseille Université	Président
Pr Fati KIRAKOYA	Université Libre de Bruxelles	Examinatrice
Pr Erik-André SAULEAU	Université de Strasbourg	Rapporteur
Pr Mahamadou Aly THERA	FMOS / USTTB, Mali	Rapporteur
Pr Issaka SAGARA	MRTC-OKD, Mali	Co-Directeur de thèse
Pr Jean GAUDART	Aix-Marseille Université	Directeur de thèse

Numéro national de thèse / suffixe local : PC/JLM/SD/2019

<b>Table des matières</b>	1
<b>Résumé (français)</b>	3
<b>Abstract (Anglais)</b>	5
<b>Remerciement</b>	7
<b>Liste des publications et conférences</b>	10
<b>Introduction générale</b>	11
<b>Partie 1 : Hétérogénéité spatiale du risque environnemental dans les essais de prévention : Impact et modélisation</b>	
1.1. Contexte	14
1.2. Méthodes	15
1.2.1. Plan de simulation	16
1.2.1.1. Localisation (répartition spatiale des individus)	16
1.2.1.2. Risque de base	18
1.2.1.3. Source de risque environnemental (gîtes larvaires)	18
1.2.1.4. Facteur traitement et son effet	20
1.2.1.5. Autres facteurs de risque et leurs effets	20
1.2.1.6. Temps d'évènement et temps de censure	21
1.2.2. Evaluations et comparaison des modèles	23
1.2.2.1. Modèle de Cox à risque proportionnels (Cox-PH)	24
1.2.2.2. Modèle Additif Généralisé (GAM)	25
1.2.2.3. Modèle SPDE	26
1.2.2.3.1. Modèle Cox-SPDE	28
1.2.2.3.2. Modèle P-SPDE	28
1.2.3. Indicateurs de performances des modèles	30
1.2.3.1. Biais	31
1.2.3.2. Erreur quadrique moyenne (MSE)	31
1.2.3.3. Taux de couverture (CR)	31
1.2.3.4. Taux de significativité (SR)	32
1.2.4. Exemple d'application	32
1.3. Résultats	32
1.3.1. Modèle de simulation des données (DGM)	32
1.3.2. Impact du risque de base sur la qualité des estimations	32
1.3.3. Estimations de l'effet traitement	33
1.3.4. Estimations de l'effet âge	39
1.3.5. Estimations de l'effet sexe	41
1.4. Discussions	43
1.5. Conclusion	47

## **Partie 2 : Modélisation de l'hétérogénéité spatiale dans le contexte d'évènement récurrents : Application aux essais de prévention contre le paludisme.**

2.1. Contexte	49
2.2. Méthodes	50
2.2.1. Structuration des données	51
2.2.1.1. Etude de Bougouni (effet de l'azithromycine associé à la CPS)	51
2.2.1.2. Etude de Bandiagara (moyen de protection contre le paludisme)	52
2.2.2. Méthodes statistiques	52
2.2.2.1. Modèle d'Andersen-Gill (AG)	53
2.2.2.2. Modèle de fragilité	53
2.2.2.3. Modèle SPDE	54
2.2.2.3.1. Modèle SPDE avec effet récurrent non structuré (SPDE-NS)	55
2.2.2.3.2. Modèle SPDE avec effet récurrent autorégressif (SPDE-AR)	56
2.2.2.3.3. Modèle SPDE avec effet récurrent sigmoïdal (SPDE-S)	56
2.2.2.4. Indicateurs de performances de modèles	56
2.3. Résultats	57
2.3.1. Evaluation de l'effet de l'azithromycine pour la prévention du paludisme	57
2.3.2. Evaluation des moyens de prévention contre le paludisme à Bandiagara	58
2.4. Discussions	60
2.5. Conclusion	62

## **Partie 3 : Guide d'utilisation des modèles SPDE avec la méthode INLA**

1.1. Bref aperçu théorique sur la méthode INLA	63
1.2. Application pratique au modèle SPDE avec le package R-INLA	63
<b>Conclusion générale et perspectives</b>	77
<b>Bibliographie</b>	79
<b>Liste des acronymes</b>	95
<b>Notations</b>	96
<b>Liste des figures</b>	98
<b>Annexes</b>	99
<b>Intitulés des doctorats AMU</b>	108

## Résumé en français

Dans le contexte des maladies transmissibles (*ex.* le paludisme, le choléra, la dengue, etc.), la proximité avec des individus contagieux ou des environnements favorables à la transmission de la maladie (*ex.* proximité de gîtes larvaires ou d'un site favorable à la survie des vecteurs) augmente le risque d'infection, entraînant ainsi une hétérogénéité spatiale de ce risque. Cependant, lors des essais de prévention ou thérapeutiques, ces risques environnementaux ne sont pas toujours exhaustivement observables (*ex.* dans le contexte du paludisme ou de la dengue, il est difficile de dénombrer totalement tous les gîtes ; de même, dans le contexte du choléra, il est quasiment impossible d'évaluer toutes les sources d'eau potentiellement contagieuses).

L'objectif de cette thèse était de modéliser cette hétérogénéité spatiale du risque environnemental non observé dans un essai de prévention randomisé.

Dans une première partie, nous avons montré que la randomisation seule ne permettait pas d'éliminer le biais dû à l'hétérogénéité spatiale du risque environnemental. A l'aide d'études de simulations, nous avons trouvé que cette hétérogénéité spatiale du risque environnemental, lorsqu'il était important, pouvait conduire à une sous-estimation de l'effet d'un traitement dans un essai de prévention randomisé. Nous avons montré que l'approche SPDE (Stochastic Partial Differential Equations) implémenté avec la méthode INLA (Integrated Nested Laplace approximations) et modélisant cette hétérogénéité spatiale à travers la localisation des individus par un champ gaussien dont la matrice de covariance était la fonction de Matérn, permettait de corriger cette sous-estimation. Ce résultat a été confirmé sur des données d'applications réelles issues d'un essai vaccinal contre le paludisme mené à Bancoumana (Mali) de 2015 à 2017.

Dans la deuxième partie de cette thèse, nous nous sommes intéressés à la modélisation de l'hétérogénéité spatiale du risque environnemental dans le contexte d'évènements récurrents contrairement à la première partie qui se limitait à l'analyse du premier évènement en ignorant tous les éventuels évènements postérieurs.

Nous avons réanalysé les données issues d'un essai de prévention qui avait conclu un effet non significatif d'un ajout de l'azithromycine à la Chimio-prévention du Paludisme Saisonnier (CPS) contre les infections palustres et l'efficacité des moyens de protection contre le paludisme à Bandiagara (Mali). Cette réanalyse avec le modèle SPDE impliquant l'effet d'évènements récurrents a permis de mettre en évidence le bénéfice significatif d'un ajout de l'azithromycine à la CPS contre les infections palustres contrairement modifiant les résultats de l'analyse initiale sans prise en compte de l'effet spatial. En ce qui concerne la réanalyse de l'évaluation des moyens de protection contre le paludisme à Bandiagara, sans modification du résultat final, la prise en compte de l'effet spatial a apporté un léger gain en puissance avec une réduction de l'intervalle de crédibilité un peu plus étroit.

Nous avons enfin élaboré un guide d'utilisation du modèle bayésien SPDE, utilisant la méthode d'estimation INLA, qui implémente toutes approches présentées dans cette thèse afin de faciliter la mise en œuvre des modèles prenant en compte l'hétérogénéité spatiale.

**Mots clés :** hétérogénéité spatiale, risques environnementaux, essais de prévention randomisés, évènements récurrents, SPDE, INLA.

## **Abstract in English**

In the context of communicable diseases (*e.g.* malaria, cholera, dengue fever, etc.), proximity to contagious individuals or environments favoring disease transmission (*e.g.* breeding sites or area favoring to vector survival) increases infection risk, thus causing spatial heterogeneity of this risk. However, in prevention or clinical trials, these environmental aspects are not always fully observable (*e.g.* in the context of malaria or dengue, it is very difficult to count completely all breeding sites; similarly, in the cholera context, it is almost impossible to assess all potentially contagious water sources).

The aim of this thesis was to model this spatial heterogeneity of environmental risk not observed in a prevention trial.

In the first section, we have shown that randomization alone did not eliminate the bias due to the spatial heterogeneity of the environmental risk. Using simulation studies, we found that this spatial heterogeneity, if significant, could induce an underestimation of the treatment effect in a randomized prevention trial. We have shown that the stochastic partial differential equations (SPDE) approach implemented with integrated nested Laplace approximations (INLA) method and modelling this spatial heterogeneity through the location of individuals by a Gaussian field whose covariance matrix was the Matèrn function, allowed to correct this underestimation. This result was confirmed by an application on the collected data from a malaria vaccine trial conducted in Bancoumana, Mali, from 2015 to 2017.

In the second section of this thesis, we focused on modelling the spatial heterogeneity of environmental risk in the context of recurrent events unlike the first section that was limited to the analysis of the first event ignoring all the eventual subsequent events. We re-analyzed the data from a prevention trial that concluded to non-significant effect of azithromycin in addition to seasonal malaria chemoprevention (SMC) against clinical malaria and a study evaluating the efficacy of malaria protection measures in Bandiagara (Mali). This reanalysis with the SPDE

approach involving the effect of recurrent events revealed the significant benefit of azithromycin in addition to SMC against clinical malaria. With regard to the reanalysis, taking into account the spatial effect brought a slight gain in power with a slightly narrower reduction in the credibility interval.

Finally, we have elaborated a guide for using the Bayesian SPDE model with the INLA estimation method, which implements all the approaches presented in this thesis in order to facilitate the implementation of models taking into account spatial heterogeneity.

**Keywords:** spatial heterogeneity, environmental risks, randomized prevention trials, recurrent events, SPDE, INLA.

## **Remerciements**

Je voudrais tout d'abord exprimer toute ma reconnaissance et ma gratitude à mes directeurs de thèse le Pr Jean GAUDART et Pr Issaka SAGARA qui ont accepté volontiers de m'accompagner dans cette thèse. Chers Professeurs, chers Maîtres, vous m'avez guidé et conseillé pendant ces 4 années de thèse avec patience, beaucoup de bienveillance et surtout de grandes rigueurs scientifiques, sans vous ce travail n'aurait jamais pu être mené à bout. Vos conseils et accompagnements constants m'ont permis d'étendre mes limites et de surpasser les doutes, je ne saurais jamais vous remercier assez de votre disponibilité constante pour éclairer et orienter nos pas durant cette thèse avec toutes les protections nécessaires. Sachez que j'ai pris note de tous vos conseils et critiques constructifs qui m'accompagneront pour le restant de ma carrière. Que le tout puissant vous protège, exhausse vos vœux et vous garde longtemps.

Ensuite, je remercie chaleureusement les membres du jury qui ont accepté volontiers de participer au parachèvement de ce travail. Chers Pr Mahamadou Ali THERA et Pr Erik-André SAULEAU, je vous remercie pour votre rapport de qualité. Je vous en serais éternellement reconnaissant pour les efforts consentis afin d'améliorer ce manuscrit. Au Pr Fati KIRAKOYA, toutes mes reconnaissances pour avoir examiné ce travail et contribué à son amélioration.

Je remercie spécifiquement le Pr Roch GIORGI qui m'a accepté au sein de son département et présidé le jury lors de la soutenance de cette thèse. Nous sommes reconnaissants pour tous vos efforts consentis afin que nous ne manquions de rien dans ce département depuis le début de cette thèse pour l'avancement de nos travaux. Votre rigueur dans le travail en toute courtoisie et votre charisme ont été formateurs pour nous. Recevez ici toutes nos gratitude.

Toutes mes reconnaissances vont également à l'égard de mon comité de suivi de thèse, le Pr Renaud PIARROUX et Dr Patrizia CARRIERI qui nous ont accompagné toute la durée de cette thèse avec un suivi régulier et des conseils.

Mes remerciements vont à l'égard de tous nos tuteurs du SESSTIM, particulièrement à notre maître Jean Charles DUFOUR qui a toujours su répondre à nos appels à l'aide avec de précieux conseils. Trouvez ici toute ma sincère gratitude et reconnaissance. Je remercie également tout le personnel administratif du SESSTIM pour le climat convivial et familial avec lequel nous avons été accueilli et accompagné durant cette thèse. Mention particulière à Mr Laurent MEYER, Mr Farid DJENAD, Mme Geneviève NOIR, Mme Lamia DAFFRI, Mme Zohra GORINE pour ne citer que ceux-ci.

Un grand merci à tous mes co-doctorants et stagiaires du SESSTIM que nous avons connu durant ce travail avec mention spéciale pour Sokhna DIENG qui a toujours été là tant dans les moments de joies que dans les moments de stress. Merci et reconnaissance également à mes aînés doctorants et post-doctorants du SESSTIM notamment Juste Aristide GOUGOUNGA, Nathalie GRAFFEO, Kankoe SALLAH, Pierre MICHEL et Boukary OUEDRAOGO qui n'ont cessé de nous apporter leurs conseils et expériences tout au long de cette thèse. Mes chaleureux remerciements également à Dr Jordi LANDIER et Mady CISSOKO qui nous ont rejoint au sein de l'équipe mais dont la compagnie m'a été d'un apport significatif, trouvez ici toutes mes reconnaissances.

Mes sincères remerciements à nos maîtres et aînés du MRTC-OKD qui n'ont ménagé aucun effort pour l'accomplissement de ce travail. Notamment la « Génération F5 » de feu « Prof Ogo » en formation à Marseille dont je me garde de citer les noms pour éviter le risque d'omission. Merci pour le réconfort fraternel que j'ai pu bénéficier auprès de vous durant cette thèse. Merci également à tout le personnel administratif et financier du MSC pour leur accompagnement constant.

Un remerciement particulier à l'Agence Française de Développement qui a pris la lourde charge financière de cette thèse durant quatre bonnes années, c'est le lieu de vous dire un grand merci pour la qualité de la prestation. Merci également à la direction et aux doctorants du réseau doctoral en santé publique de l'EHESP qui m'ont permis de bénéficier d'une formation interdisciplinaire courant cette thèse, qui me suivra pour la suite de ma carrière professionnelle.

Enfin, je remercie ma famille et mes ami(e)s auprès de qui je prends appuis pour me réconforter et qui ont accepté toutes mes absences prolongées et répétitives tout en rendant un vibrant hommage à la mémoire de notre Papa Feu Issa GUINDO, que la miséricorde de Dieu soit sur lui. Mention spéciale à mes grands frères Djibril et Seydou qui ont toujours été un appui important pour moi dans les moments de doutes, c'est le lieu de vous dire que j'en suis très reconnaissant et vous en remercie. Un grand merci également à mon grand-frère Philippe DARA et sa famille depuis Paris pour les soutiens fraternels et conseils depuis nos premiers pas en France. A mes amis de tous les temps, le trio formé depuis les bas âges et qui a survécu au périple de la vie universitaire en passant par le lycée jusqu'à aujourd'hui, trouvez ici tous mes respects et considérations, merci pour la fidèle compagnie. A Mme GUINDO Aïcha et à notre fille chérie Kadidia Témélou, je dis un grand merci particulier, vous avez été ma source de motivation. Je n'ai jamais manqué de votre amour et de vos soutiens sous toutes ses formes durant cette thèse, trouvez ici ma profonde reconnaissance et tout mon amour.

Bref, merci à tous ceux et toutes celles qui de près ou de loin, nous ont soutenu et participé à l'accomplissement de cette thèse, retrouvez ici toute ma reconnaissance.

Je ne saurai terminer cette partie sans rendre un vibrant hommage à la mémoire notre grand maître, feu « Prof Ogo » arraché brutalement à notre affection le 09 Juin 2018, un jour noir pour nous et pour l'ensemble du monde de la science. Reposez en paix cher maître, qu'Allah vous récompense de votre bienveillance. Je me souviens de votre phrase au début de cette thèse et qui a été une grande source de motivation pour moi : « Je te connais travailleur et sérieux, je t'envoie chez Pr Gaudart, c'est mon étudiant mais j'ai beaucoup d'estime pour lui, saches profiter au maximum de ses connaissances afin de faire profiter tes autres collègues à ton tour ». J'ai été profondément affligé par votre disparition prématurée, mon bonheur aurait été maximal si vous étiez là aujourd'hui pour juger ce travail, j'espère avoir atteint les objectifs.

## Publications en premier auteur dans des revues internationales à comité de lecture

1. **Guindo A**, Sagara I, Ouédraogo B, Dicko A, Sallah K, Doumbo O, Gaudart J. *Modélisation de l'hétérogénéité spatiale de l'exposition : essais cliniques dans le contexte du paludisme*. Rev DÉpidémiologie Santé Publique. 2018;66:S134. doi:10.1016/j.respe.2018.03.342. **Abstract form**
2. **Guindo A**, Sagara I, Ouedraogo B, Sallah K, Assadou MH, Healy S, Duffy P, Doumbo O, Dicko A, Giorgi R, Gaudart J. *Spatial heterogeneity of exposure in randomized prevention trials: consequences and modeling*. *BMC Med Res Methodol*. 2019;19(1):149. doi:[10.1186/s12874-019-0759-z](https://doi.org/10.1186/s12874-019-0759-z)
3. **Guindo A**, Sagara I, S Dieng, Coulibaly D, Rouamba T, Cairns M, Dicko A, Giorgi R, Gaudart J. *Modeling spatial heterogeneity of environmental risk in the context of recurrent events: a reanalysis of the additional protective effect of azithromycin in seasonal malaria chemoprevention*. *International Journal of Environmental Malaria Journal* (**soumission en cours**).

## Autres publications

1. Rouamba T, Nakanabo-Diallo S, Derra K, Rouamba E, Kazienga A, Inoue Y, Ouédraogo EK, Waongo M, Dieng S, **Guindo A**, Ouedraogo B, Sallah K, Barro S, Yaka P, Kirakoya-Samadoulougou F, Tinto H, Gaudart J. *Socioeconomic and environmental factors associated with malaria hotspots in the Nanoro demographic surveillance area, Burkina Faso*. *BMC Public Health*. 2019;19(1):249. doi:[10.1186/s12889-019-6565-z](https://doi.org/10.1186/s12889-019-6565-z).
2. Dieng S, Ba EH, Cissé B, Sallah K, **Guindo A**, Ouedraogo B, Piarroux M, Rebaudet S, Piarroux R, Landier J, Sokhna C, Gaudart J. *Spatio-temporal variation of malaria hotspots in central Senegal, 2008-2012*. *BMC Infectious Diseases* (**soumis, juin 2019**)
3. Maiga H, Barger B, **Guindo A**, Sagara I, Traore OB, Tekete M, Dara A, Traore ZI, Diarra M, Coumare D, Kodio A, Toure OB, Doumbo OK, Djimde A. *Impact of Three Years of Intermittent Preventive Treatment Using Artemisinin-based Combination Therapy on Malaria Morbidity in Malian Schoolchildren*. *Malaria Journal* (**soumis, avril 2019**).

## Conférence internationale à comité de lecture

**Guindo A**, Sagara I, Ouédraogo B, Dicko A, Sallah K, Doumbo O, Gaudart J. *Modélisation de l'hétérogénéité spatiale de l'exposition : essais cliniques dans le contexte du paludisme*. 12<sup>ième</sup> conférence francophone d'épidémiologie clinique - EPICLIN. Mai 2018, Nice, France.  
**Communication orale**

## Autres présentations en lien avec la thèse

1. **Guindo A**, Sagara I, Ouédraogo B, Dicko A, Sallah K, Doumbo O et Gaudart J. *“Modélisation de l'interaction spatiale dans les essais thérapeutiques : Cas du paludisme”*. Journal Club Officiel du MRTC – OKD. Octobre 2017, Bamako, Mali. **Communication orale**
2. **Guindo A**, Sagara I, Ouédraogo B, Dicko A, Sallah K, Doumbo O et Gaudart J. *Modélisation de l'hétérogénéité dans les essais de préventions contre le paludisme*. 8<sup>ième</sup> rencontre scientifique du réseau doctorale en santé publique. Mars 2018, Bordeaux, France. **Communication orale**
3. **Guindo A**, Sagara I and Gaudart J. *Spatial interaction modeling in survival analysis: Example of the breeding site effect on Malaria clinical trials*. 25<sup>ième</sup> rencontre annuelle de l'école doctorale sciences de la vie et de la santé. Juin 2017, Marseille, France. **Poster**

## **Introduction générale**

Dans le contexte des maladies transmissibles telles que les maladies à transmission vectorielle (*ex.* paludisme, dengue, etc.) ou toutes autres maladies liées à l'environnement ou aux contacts rapprochés (*ex.* cholera, grippe, bilharziose, etc.), la localisation des individus influence le risque de contagion. En effet, la proximité avec des individus contagieux ou avec un environnement favorable à la transmission de la maladie augmente le risque d'infection, entraînant ainsi une hétérogénéité spatiale de celui-ci (*ex.* proximité de gîtes larvaires ou un site favorable à la survie des moustiques dans le contexte du paludisme) [1].

De nombreux chercheurs conduisant des essais de prévention considèrent que la randomisation était suffisante pour assurer la comparabilité des différents bras d'intervention des essais et par conséquent ils ne tiennent généralement pas compte de cette hétérogénéité spatiale du risque environnemental [2–5]. En effet, le modèle de Cox à risques proportionnels (Cox-PH, Cox Proportional Hazard model) est bien l'approche multivariée la plus utilisée pour l'évaluation des essais de prévention dans un contexte d'analyses de survie mais il n'est pas toujours bien adapté quand il s'agit de prendre en compte l'hétérogénéité spatiale du risque.

Dans certaines études, cette hétérogénéité spatiale est prise en compte en utilisant des modèles mixtes [6–8], qui associent un effet aléatoire à une échelle spatiale spécifique (pays, région, district, aire d'influence d'un centre de santé, village, etc.) en supposant que le risque est homogène au sein de chaque unité, à l'échelle considérée et qu'aucune interaction spatiale ne se produit entre les unités à cette échelle (indépendance spatiale entre les unités). Cependant, dans le contexte des maladies comme le paludisme, une hétérogénéité spatiale de l'incidence a été observée à de petites échelles (villages, maisons) [9], ce qui pourrait compromettre les hypothèses qui sous-tendent l'application des modèles mixtes.

En plus, il est souvent difficile d'évaluer cette hétérogénéité spatiale de façon exhaustive. En effet, dans le contexte où la transmission de la maladie est liée à des facteurs environnementaux, ces facteurs spécifiques ne sont pas toujours observables ou du moins leurs observations demandent un travail de terrain approfondi et souvent coûteux. Par exemple, les gîtes larvaires peuvent être nombreux, insoupçonnés et de petites tailles comme pour les espèces *Anopheles* et *Aedes* (vecteur de transmission du paludisme et de la dengue respectivement) [10–14]. De même, dans le contexte du choléra, il est presque impossible d'évaluer toutes les sources d'eau potentiellement contagieuses (puits, marres, pompes d'eaux, etc.).

Par contre, il est de plus en plus fréquent (et facile) de collecter les coordonnées géographiques des maisons des personnes enrôlées dans les études, notamment pour des besoins de cartographies de la pathologie étudiée. Cette source d'information peut être utilisée en l'absence d'information précise sur la localisation des facteurs de risques environnementaux.

La motivation principale de cette thèse était de modéliser l'hétérogénéité spatiale du risque dans un contexte d'analyses de survie, sans pourtant mesurer forcément les facteurs environnementaux conduisant à cette hétérogénéité, en utilisant les coordonnées géographiques des individus ou de leurs ménages (probables lieux de contamination).

Dans un premier temps, nous nous sommes limités à l'évaluation de l'impact de l'hétérogénéité spatiale et sa modélisation dans un contexte d'analyse de survie classique, c'est-à-dire, nous nous sommes intéressés uniquement au temps survie de la première infection en ignorant tous les événements postérieurs. A partir d'une étude de simulation, nous avons déterminé les conditions dans lesquelles l'hétérogénéité spatiale pouvait biaiser l'évaluation des essais de prévention et comparé différentes approches de modélisation permettant de corriger ce biais, en particulier SPDE (Stochastic Partial Differential Equation).

Enfin, nous avons utilisé des données réelles issues d'un essai vaccinal contre le paludisme mené à Bancoumana (Mali) de 2015 à 2017 pour appuyer les résultats des données simulées.

Dans un second temps, nous nous sommes intéressés à la modélisation de l'hétérogénéité spatiale dans un contexte d'analyse d'évènements récurrents lors d'un essai de prévention. En effet, nous avons déterminé si la prise en compte de l'hétérogénéité spatiale dans ce contexte d'épisodes récurrents permettait d'améliorer la précision des estimations dans l'évaluation d'un essai visant à évaluer l'effet de l'ajout de l'azithromycine à la Chimio-prévention du Paludisme Saisonnier (CPS) sur les accès palustres comparativement aux modèles classiques d'analyses d'évènements récurrents (modèle d'Andersen-Gill et modèle de fragilité).

Enfin, nous avons élaboré un guide pratique d'utilisation des modèles SPDE avec la méthode bayésienne INLA facilitant la compréhension et la mise en œuvre de nos recommandations.

## Partie 1

### Hétérogénéité spatiale du risque environnemental dans les essais de prévention : Impact et modélisation.

#### 1.1. Contexte

Les modèles de survie sont souvent utilisés pour évaluer les essais de prévention randomisés notamment parce qu'ils permettent de prendre en compte la durée de suivi, les différents types de censures, l'occurrence de l'évènement d'intérêt, etc. Dans ce contexte, le modèle de Cox-PH (Cox Proportional Hazards model) est l'approche multivariée la plus utilisée malgré qu'il ne soit pas toujours correctement appliqué pour la prise en compte de l'hétérogénéité spatiale du risque environnemental. À cet égard, différentes méthodes ont été proposées permettant de prendre en compte l'hétérogénéité spatiale dans le contexte d'analyses de survie.

Le modèle GAM (Generalized Additive Model), initialement développé pour modéliser les relations non-linéaires, est de plus en plus utilisé pour modéliser l'hétérogénéité spatiale en utilisant une fonction spline bivariée sur les coordonnées géographiques comme des covariables [15–18].

Dans ces derniers temps, les modèles SPDE (Stochastic Partial Differential Equations) sont couramment utilisés pour modéliser explicitement la variation spatiale du risque d'infection à partir de la localisation exacte (les coordonnées géographiques) des individus ou de leurs habitations [19]. Ainsi, *Lindgren et al.* ont proposé un modèle SPDE utilisant un champ gaussien défini par la fonction de covariance de Matérn qui possède de bonnes propriétés de calcul [20, 21]. De plus, le modèle SPDE implémenté à l'aide de l'algorithme INLA (Integrated Nested Laplace Approximation) est actuellement utilisé dans un éventail de contextes toujours plus large [22–26].

Cette partie de la thèse avait pour objectif, dans un premier temps, de déterminer à l'aide d'une étude de simulation, dans quelles conditions et dans quelle mesure l'hétérogénéité spatiale du risque environnemental pouvait biaiser les résultats d'un essai de prévention randomisé.

Et dans un second temps, de comparer la performance des différents modèles spatiaux pour l'estimation de l'effet d'un traitement expérimental en utilisant les données simulées.

Par la suite, ces modèles ont été appliqués aux données réelles issues d'un essai vaccinal contre le paludisme mené à Bancoumana (Mali) entre 2015 et 2017 pour confirmer les résultats obtenus par l'étude de simulation.

## **1.2. Méthodes**

Nous avons simulé 432 scénarios (50 échantillons de 1000 individus chacun). Ces scénarios ont été d'abord analysés avec un modèle de Cox-PH non spatialisé, puis quatre modèles de survie différents prenant en compte l'hétérogénéité spatiale, y compris le modèle DGM (Data Generating Model) qui a servi à générer les données.

Pour chaque facteur de risque (âge, sexe et traitement), la performance des modèles a été évaluée par la quantification du biais de l'estimation de l'effet associé, l'erreur quadratique moyenne (MSE, Mean Squared Error), le taux de couverture (CR, Coverage Rate) et le taux de signification (SR, Significance Rate).

Pour les 32 scénarios les plus pertinents (niveau fort et faible de l'effet gîte, niveau fort et faible de la densité de gîtes, niveau fort et faible de la densité de populations associés à un niveau de l'effet traitement fort, moyen et faible), nous avons simulé 500 échantillons de 1000 individus chacun pour valider les résultats obtenus avec les 50 échantillons précédents (*Figure A.1*).

### 1.2.1. Plan de simulation

Le temps d'évènement a été simulé à l'aide du modèle de Cox-PH classique en tenant compte des différents facteurs de risque (y compris le facteur environnemental) en utilisant une loi de Weibull.

Le temps de censure a été simulé en utilisant la loi exponentielle (un cas particulier de la loi de Weibull avec le paramètre de forme valant 1).

Nous partons de l'hypothèse qu'étant donné un vecteur de covariables  $X$ , la fonction de risque instantané  $\lambda_i(t, X)$  pour un individu  $i$  au temps  $t$  était défini par l'équation suivante :

$$\lambda_i(t, X) = \lambda_i(t, X^{(i)}) + \lambda_i(t, X^{(-i)})$$

où  $\lambda_i(t, X^{(i)})$  est l'effet fixe dépendant de  $i$  et  $\lambda_i(t, X^{(-i)})$  est l'effet spatial dépendant du voisinage de l'individu  $i$ .

Pour cela, nous avons simulé un essai de prévention randomisé et contrôlé à 2 bras de traitement avec une taille de  $n = 1\,000$  individus dans un carré  $\Omega$  de  $400\text{ km}^2$  de superficie.

Avant la simulation de la randomisation dans les bras de traitement, différentes situations de risque (scénarios) ont été simulés à partir de la localisation (répartition spatiale des individus), le risque de base et l'hétérogénéité du risque environnemental (localisation et densité de gîtes) (nous y reviendrons plus en détail par la suite). En plus de l'effet traitement, deux autres facteurs non spatialisés ont été simulés : l'un continu avec un effet significatif (facteur âge) et l'autre binaire avec un effet nul (facteur sexe).

La taille des échantillons et le nombre de simulations nécessaires ont été calculé avec une précision de 0.01 et une variance de 0.09 suivant les recommandations standards de simulation [27, 28].

#### 1.2.1.1. Localisation (répartition spatiale des individus)

Pour simuler la distribution spatiale des individus dans la zone de l'étude  $\Omega$ , nous avons utilisé un

Processus de Poisson Inhomogène (PPI) dans lequel la densité de la population dépendait de la localisation. Nous avons considéré trois points de concentration  $L_1, L_2$  et  $L_3$  autour desquels la population était plus dense, et nous avons introduit un paramètre de concentration  $\tau$  permettant de contrôler de concentration de la population autour de ces points. Par la suite, ce paramètre est appelé « densité de population » pour faciliter les termes.

La localisation de ces trois points a été simulé en utilisant un Processus Ponctuel Homogène (PPH) de sorte à respecter une certaine distance minimum entre eux. En effet, afin de respecter une couverture uniforme de la zone d'étude pour éviter qu'il ait des creux gênant, la distance entre ces points de concentration devait être supérieure à 4% de l'étendu de la zone de l'étude, sinon, le tirage était repris à nouveau.

Pour chaque point de concentration, la densité  $\lambda(.)$  du PPI a été définie comme suit :

$$\forall s \in \Omega, \quad \lambda(s) = \begin{cases} C(\lambda_0) \times e^{-\tau \times d(L1,s)} & \text{si } d(L1,s) \leq \min(d(L2,s), d(L3,s)) \\ C(\lambda_0) \times e^{-\tau \times d(L2,s)} & \text{si } d(L2,s) \leq \min(d(L1,s), d(L3,s)) \\ C(\lambda_0) \times e^{-\tau \times d(L3,s)} & \text{si } d(L3,s) \leq \min(d(L1,s), d(L2,s)) \end{cases}$$

où  $s$  était un point géo-localisé dans la zone d'étude (représentant la localisation d'un individu),  $\lambda_0$  la densité moyenne de population dans la zone,  $C(.)$  la fonction de normalisation nécessaire pour atteindre l'effectif fixé  $n$  et  $d(.,.)$  la distance euclidienne.

Afin d'étudier l'impact de la densité de population autour des trois de concentrations, trois situations différentes de densité de populations  $\tau$  ont été simulé :

$$\tau = 0.2 \text{ (faible densité de population)}$$

$$\tau = 0.5 \text{ (densité moyenne de population)}$$

$$\tau = 0.8 \text{ (forte densité de population)}$$

### 1.2.1.2. Risque de base

Le risque de base  $\lambda_0(\cdot)$  était supposé être constant pour tous les individus. Il a été simulé suivant une loi de Weibull de paramètre de forme égale à 1, garantissant ainsi une valeur constante dans le temps.

Trois niveaux de risque de base correspondant à trois faciès épidémiologiques observés ont été utilisés pour le paramètre d'échelle  $\gamma$  en prenant comme exemple, la prévalence du paludisme au Mali [29]:

Faible prévalence :  $\gamma = 6\%$  (*ex.*, situation de Bamako),

Prévalence médiane :  $\gamma = 37\%$  (*ex.*, situation de Ségou),

Forte prévalence :  $\gamma = 60\%$  (*ex.*, situation de Mopti).

### 1.2.1.3. Source de risque environnemental (gîtes larvaires)

L'hétérogénéité spatiale du risque environnemental a été simulé selon la localisation des gîtes larvaires, la variation de l'effet des gîtes (défini en Risque Relatif  $RR_g$ ), la variation de la densité de gîtes et le rayon d'influence des gîtes considéré comme constante. Le rayon d'influence d'un gîte larvaire a été interprété comme la distance moyenne parcourue par un moustique pour prendre un repas de sang. Il a été fixé à  $r = 600\text{ m}$  [30] et était constant pour tous les gîtes.

Un individu est considéré comme exposé à un gîte si la distance euclidienne entre cet individu et l'épicentre du gîte est plus petite que le rayon d'influence du gîte.

En considérant que la densité de gîtes larvaires ne dépendait pas de sa localisation, nous avons simulé la distribution spatiale des gîtes selon un processus ponctuel homogène marqué (PPHM) où la marque était le rayon d'influence  $r$  du gîte. Ainsi, un individu pouvait être exposé à zéro, un ou plusieurs gîtes larvaires augmentant progressivement le risque environnemental.

La densité de gîtes larvaires était utilisée pour quantifier le pourcentage de la surface exposé au risque environnemental (c'est-à-dire la surface couverte par le rayon d'influence d'au moins un gîte) par rapport à la surface totale de l'étude (400 km<sup>2</sup>). Cette densité pouvait être interprété comme la probabilité d'être exposé à au moins un gîte.

Concrètement,  $A_T$  étant l'aire totale de la zone d'étude,  $n_g$  le nombre de gîtes et  $\pi r^2$  l'aire couverte par un gîte de rayon  $r$ , alors l'aire  $A_g$  couverte par l'ensemble des gîtes présents dans la zone d'étude était définie par :

$$A_g = n_g \pi r^2$$

Ainsi, la densité de gîtes était définie par :

$$D_g = \frac{A_g}{A_T} = \frac{n_g \pi r^2}{A_T}$$

Pour simuler différents scénarios, la densité de gîtes a été fixée à trois niveaux (0.25, 0.5 et 0.75).

La productivité des gîtes et la protection contre les piqûres des moustiques pouvant être très variables, l'effet gîte (défini en Risque Relatif  $RR_g$ ) a également été simulé selon quatre situations :

- Effet gîte très faible ( $RR_g = 1.05$ ),
- Effet gîte faible ( $RR_g = 1.20$ ),
- Effet gîte fort ( $RR_g = 1.5$ ),
- Effet gîte très fort ( $RR_g = 3$ ).

#### 1.2.1.4. Facteur traitement et son effet

L'attribution au bras expérimental et au bras de référence a été simulée suivant une loi de Bernoulli équilibrée  $\mathcal{B}(n, 0.5)$  en utilisant une randomisation individuelle sans tenir compte de la localisation des individus ou des gîtes, ni des autres facteurs (âge et sexe).

Par la suite, quatre niveaux d'effet traitement (défini en Risque Relatif  $RR_t$ ) ont été simulés:

- Effet traitement très faible ( $RR_t = 0.95$ ),
- Effet traitement faible ( $RR_t = 0.8$ ),
- Effet traitement fort ( $RR_t = 0.6$ ),
- Effet traitement très fort ( $RR_t = 0.25$ ).

En effet, l'efficacité des essais de prévention, notamment contre le paludisme, est très variable, mais dans de nombreux cas, elle est comprise dans le panel de risque relatif allant de 0.25 à 0.95 [31–36].

#### 1.2.1.5. Autres facteurs de risque et leurs effets

Bien que l'objectif de notre étude ait été d'évaluer l'impact de l'hétérogénéité spatiale du risque environnemental sur l'effet d'un traitement expérimental, nous avons inclus deux facteurs supplémentaires dans notre analyse : l'un associé à la maladie (âge), l'autre non (sexe). Ces deux facteurs étaient indépendants de la localisation des individus et des gîtes et leur effet était constant dans le temps.

Ainsi, une variable binaire, identifiée comme le facteur sexe, a donc été simulée selon une loi binomiale équilibrée avec un effet nul ( $RR_s = 1$ ).

$$\text{Sexe} \sim \mathcal{B}(n, p = 0.5)$$

Une autre variable continue associée à la maladie, identifiée comme le facteur âge, a été simulée selon une loi uniforme par morceau à partir de la répartition de la population par tranche d'âge au Mali [37]. L'effet protecteur de cette variable (défini comme le Risque Relatif  $RR_a$ ) était issue de la littérature [29] était fixe et constant dans le temps ( $RR_a = 0.84$ ) :

$$Age1 \sim \mathcal{U} (47.3\%n, [0.25, 14])$$

$$Age2 \sim \mathcal{U} (19.2\%n, ]14, 24])$$

$$Age3 \sim \mathcal{U} (26.8\% n, ]24, 54])$$

$$Age4 \sim \mathcal{U} (37.6\% n, ]54, 64])$$

$$Age5 \sim \mathcal{U} (29.4\% n, ]64, 75])$$

#### 1.2.1.6. Temps d'évènement et temps de censure

Le temps d'évènement  $T_{event}$  sans censure a été simulé selon un modèle Cox-PH classique qui tenait compte des différents facteurs de risque détaillés ci-dessus, notamment le risque de base, le traitement, l'âge, le sexe et le risque environnemental (les gîtes). La fonction de risque instantané de l'individu  $i$  a été calculée comme suit :

$$\lambda_i(t|X) = \lambda_0 \exp(\beta_1 Age_i + \beta_2 Sex_i + \beta_3 Treatment_i + \beta_4 Gite_i) \quad [\text{equation 1}]$$

$\lambda_0$  était le risque de base considéré constant dans le temps,  $Gite_i$  était le nombre de gîtes auxquels l'individu  $i$  avait été exposé et  $(\beta_1, \beta_2, \beta_3, \beta_4)$  étaient respectivement les paramètres associés aux covariables traitement, âge, sexe et gîte.

Le temps de censure  $T_{cens}$  a été simulé selon une loi exponentielle  $\mathcal{E}(\alpha)$  indépendamment des facteurs de risque (fixes ou spatiaux). Le paramètre  $\alpha$  de cette loi exponentielle était défini par l'inverse du 3<sup>ième</sup> quartile du temps de suivi maximal de l'étude.

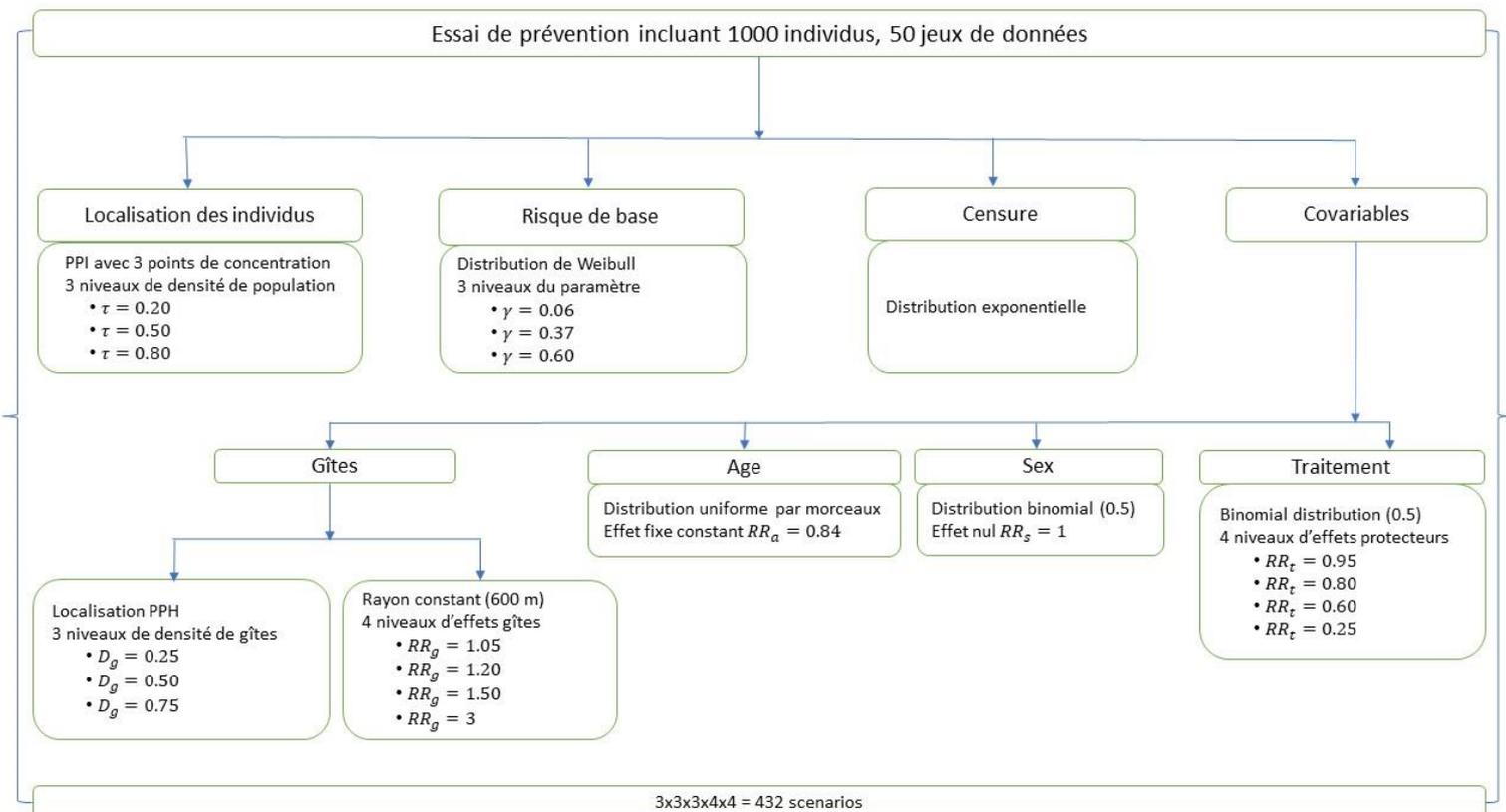
$$\alpha = \frac{1}{Q(T_{event}, 0.75)}$$

Ainsi, le temps d'observation  $T$  a été défini comme étant le minimum entre le temps de censure et le temps d'évènement :

$$T = \min(T_{event}, T_{cens})$$

La combinaison de tous ces paramètres de simulation a permis de constituer 432 scénarios. Chaque scénario était constitué de 50 jeux de données de 1000 individus chacun (Figure 1) permettant de garantir une précision de 0.01 et une variance 0.09. La taille des échantillons de 1000 individus ont permis d'assurer une puissance supérieure à 85% dans le cas où l'effet traitement ( $RR_t$ ) était inférieur à 0.8. Pour un effet traitement à  $RR_t = 0.95$ , la puissance était d'environ 65%.

**Figure 1:** Plan de simulation des différents scénarios.

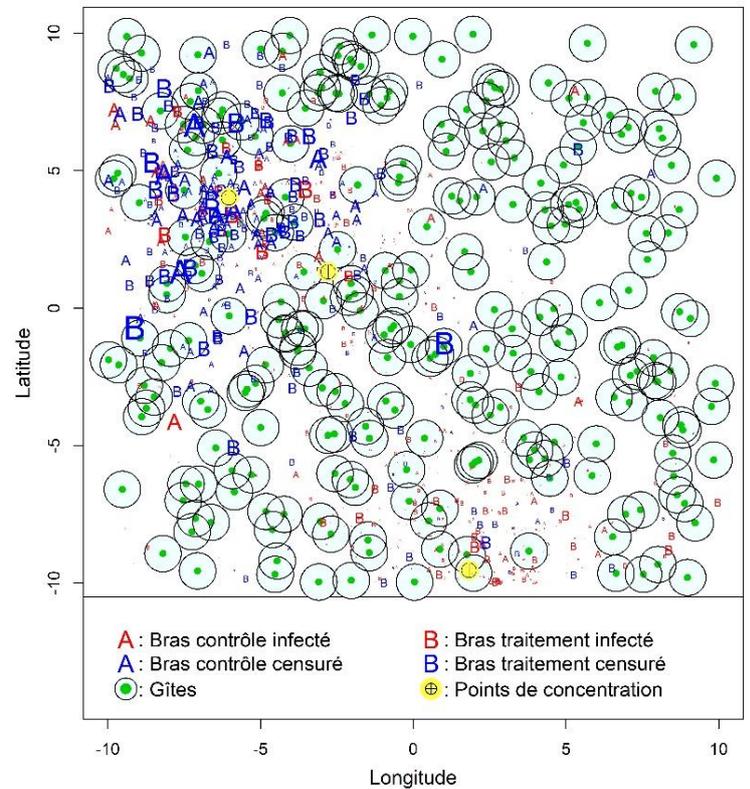
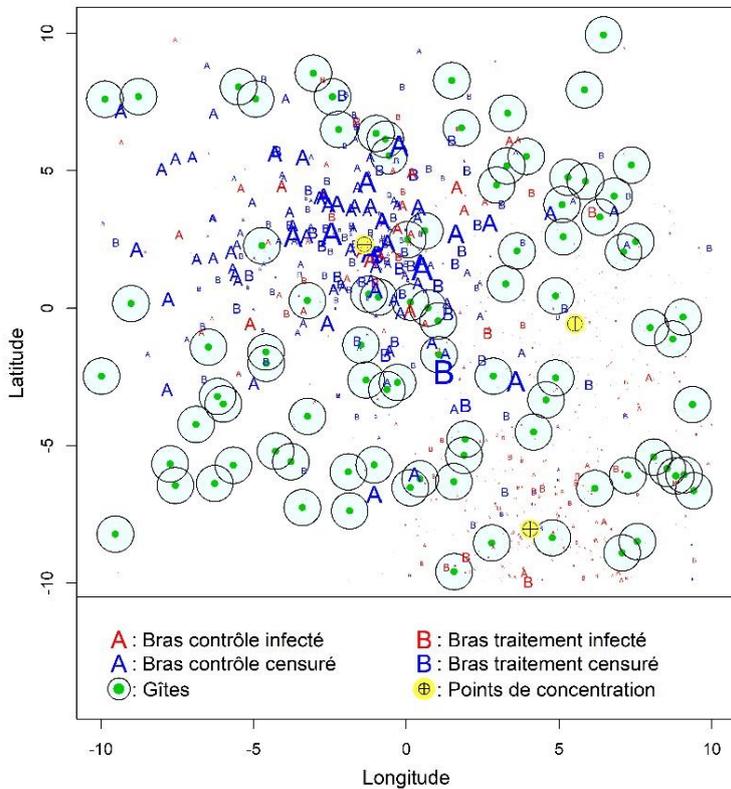


Un exemple de la structure des données simulées est représenté dans la **Figure 2**.

**Figure 2:** Structure des données simulées (la taille des points est proportionnelle au temps de survie)

RRg=1.20, Dg=0.25, RRt=0.25, Dens.Pop=0.50, R.base=0.37

RRg=3, Dg=0.75, RRt=0.25, Dens.Pop=0.50, R.base=0.37



*La taille des points A et B représente la durée du suivi pour les individus (plus le point est volumineux, plus l'individu dure longtemps dans l'étude).*

### 1.2.2. Evaluation et comparaison des modèles

La popularité des modèles de survie dans le domaine de la biomédecine, notamment dans l'évaluation des essais randomisés, nous amène à nous intéresser à ce type de modèle dans cette thèse. L'intérêt de l'analyse de survie par rapport aux méthodes classiques vient du fait qu'on modélise non seulement l'occurrence de l'évènement étudié mais aussi le temps de cette occurrence, de ce fait, la durée du suivi est prise en compte ainsi que les censures. Ce qui n'est pas le cas pour les modèles multivariés classiques tels que les modèles de régression logistique ou de Poisson qui modélisent uniquement l'occurrence de l'évènement étudié sans tenir compte ni du temps de suivi ni des individus censurés.

Pour chaque scénario, cinq modèles de survie ont été comparé sur chacun des 50 jeux de données simulées, qui sont :

Le modèle DGM qui a servi à générer les données. C'est un modèle hypothétique où l'on suppose qu'on a observé et localisé la totalité des gîtes dans la zone d'étude. Ainsi avait été comptabilisé pour chaque individu, le nombre de gîte auxquels il était exposé et intégrer cela comme variable dans un modèle de Cox classique.

Un modèle Cox-PH classique non spatial, dont les paramètres ont été estimés par la méthode de maximum de vraisemblance en utilisant une équation similaire à l'équation 1 mais sans tenir compte du risque environnemental [38, 39] ;

Un modèle GAM, également à risques proportionnels, modélisant l'effet spatial avec une fonction spline bivariée [18] (pour plus de détail, voir dessous) ;

Deux modèles SPDE modélisant l'hétérogénéité spatiale à l'aide d'un champ gaussien dont la fonction de covariance était définie par la fonction de covariance de Matérn : l'une modélisait la durée de survie suivant une loi de Weibull (Cox-SPDE) et l'autre modélisait le nombre d'évènement suivant une loi de Poisson (P-SPDE).

Les 2 modèles SPDE (Cox-SPDE et P-SPDE) ont été estimés, selon une approche bayésienne, avec l'algorithme INLA pour optimiser les temps de calculs et d'autres propriétés [40, 41].

#### **1.2.2.1. Modèle de Cox à risques proportionnels (Cox-PH)**

Le modèle Cox-PH est le modèle multivarié le plus classique adapté aux données de survie. Malgré sa popularité dans le domaine de la biomédecine, ce modèle ne prend pas en compte l'hétérogénéité spatiale du risque.

De façon générale, il s'écrit comme suit :

$$\lambda(t, X) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)$$

$\lambda(t, X)$  est la fonction de risque instantané à l'instant  $t$ ,  $\lambda_0(t)$  est le risque de base (parfois constant dans le temps),  $X = (X_1, X_2, \dots, X_n)$  est le vecteur de covariables à effets fixes et  $\beta = (\beta_1, \beta_2, \dots, \beta_n)$  est le vecteur des paramètres liés à chaque covariable.

Ce modèle à effets fixes est utilisé pour analyser le temps de survenue d'un événement d'intérêt avec possibilité d'ajustement sur des covariables [42, 43]. Comme habituellement, les paramètres de ce modèle ont été estimés par la méthode de maximum de vraisemblance.

#### **1.2.2.2. Modèle Additif Généralisé (GAM)**

Le modèle GAM a été conçu au départ pour étudier des relations non linéaires [44], le plus souvent à l'aide de fonctions splines. Dans notre cas particulier, nous avons utilisé une fonction spline bivariée des coordonnées géographiques des individus (latitude et longitude) ou du moins du lieu probable de leur contamination (habitation) pour prendre en compte l'hétérogénéité spatiale du risque environnemental [18, 45]. Dans notre contexte, nous avons supposé une relation linéaire pour les covariables (comme dans la simulation), les propriétés du modèle GAM a été mis à contribution uniquement pour modéliser l'effet spatial.

Le modèle s'écrit comme suit

$$\lambda(t, X) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + f(long, lat))$$

$f$  était une fonction spline en plaque mince bivariée modélisant l'effet spatial,  $long$  était la longitude et  $lat$  la latitude.

Concrètement, nous avons utilisé la méthode de spline de régression en plaque mince (thin plate spline) à cause de ses propriétés notamment son isotropie et sa capacité à prendre un nombre quelconque de covariables de lissages (longitude et latitude). L'autre intérêt de cette méthode de spline est qu'elle n'est pas sensible au nombre de nœuds. De plus, le résultat du lissage ne change pas par rotation du système de coordonnées à cause de l'isotropie. D'autres méthodes de spline pouvaient être utilisées telles le spline cubique qui est plus sensible au nombre nœud. Simon Wood propose aussi la méthode de produit tensoriel lorsque les covariables de lissage sont d'unités différentes, par exemple lorsqu'on veut faire un lissage en fonction de l'espace et du temps [18]. L'utilisation des coordonnées géographiques comme variables de lissage de spline s'explique par le fait que l'on ne dispose pas de la localisation précise des sources d'infection dans notre contexte. Cependant, au Mali, les vecteurs principaux sont issus des complexes *Anopheles gambiae* et *Anopheles fuscinegus*. Le comportement est majoritairement anthropophage, endophile et endophage. Elles piquent majoritairement à la tombée de la nuit et avant le lever du soleil. Bien que des variations puissent être observées, le lieu principal de la contamination est, le plus souvent, le lieu d'habitation. Dans ce contexte, prendre les coordonnées géographiques des habitations est justifié pour modéliser l'effet spatial.

### 1.2.2.3. Modèle SPDE

Le modèle SPDE de type bayésien initialement conçu pour modéliser les structures spatiales à l'aide d'un champ gaussien [46]. Il s'écrit globalement comme suit :

$$Y | \beta, X, Z \sim \mathbb{P}(Y | \mu, \phi)$$

$$Z \sim GF(0, \Sigma)$$

$$\beta, \mu, \phi \sim N(0, 0.001)$$

$\mathbb{P}$  est la loi de  $Y$  (variable dépendante),  $\beta$  est le vecteur des effets associés aux covariables  $X$  (effets fixes),  $Z$  est un champ gaussien ( $GF$ ) dont la fonction de covariance est la fonction de Matérn  $\Sigma$  (effet spatial aléatoire) et  $\mu = E(Y)$  et  $\phi$  sont les paramètres de la loi de  $Y$ ,  $\mu = E(Y) = h(\beta X + Z)$ , où  $h$  est la fonction de lien canonique.

La loi à priori des paramètres ou hyperparamètre est la loi normale centrée d'écart type 0.001. Selon certains auteurs, l'approximation de Laplace utilisée dans la méthode INLA, permet de diminuer l'impact des lois a priori. Cependant, une analyse de sensibilité était faite pour déterminer l'impact du choix de la loi a priori, sur les scénarios avec le plus d'effet environnemental (forte densité de gîtes avec effet fort et effet traitement moyen)

Dans ce travail, la fonction de Matérn a été utilisée pour modéliser la matrice de covariance  $\Sigma$  permettant ainsi de contrôler la régularité du champ gaussien et la dépendance spatiale entre deux individus  $i$  et  $j$  en fonction de la distance euclidienne  $d_{i,j}$  qui les sépare. L'expression mathématique de la fonction de Matérn est donnée par la formule suivante :

$$\Sigma_{ij}(\theta_1, \theta_2) = \sigma_X \text{Cov}(X_i, X_j) = \sigma_X \frac{2^{1-\theta_1}}{\Gamma(\theta_1)} (\theta_2 d_{i,j})^{\theta_1} K_{\theta_1}(\theta_2 d_{i,j})$$

$\sigma_X$  est l'écart-type marginal de la variable  $X$  à étudier,  $\Gamma$  était une fonction gamma,  $K_{\theta_1}$  était une fonction de Bessel et  $\theta_1$  et  $\theta_2$  étaient les paramètres de décroissance de la fonction de Matérn permettant de contrôler la régularité et la trajectoire de cette fonction.

D'autres fonctions de covariance spatiale comme la fonction de covariance exponentielle ou gaussienne pouvaient être utilisées mais ces fonctions sont en réalité des cas particuliers de la fonction de Matérn [47]. Par exemple, la fonction de covariance exponentielle est obtenue facilement en prenant  $\theta_1 = 1/2$  dans la formule ci-dessus de la fonction de Matérn. Pour rester dans cette généralité, la covariance de Matérn a été préférée.

La matrice de covariance de Matèrn telle que défini ici avec des paramètres constants signifie que le champ gaussien utilisé est isotrope. En dehors de l'isotropie, le maillage par triangulation utilisé dans la méthode INLA, permet de s'adapter à toute forme de zone, cela est en partie une réponse à une problématique de forme particulière de zone. Mais d'une façon générale, il pourrait être important de considéré un champ gaussien anisotrope parce que selon l'expérience de terrain, il n'y a pas de barrière à la dispersion des vecteurs.

Dans notre cas particulier, nous avons appliqué le modèle SPDE de deux façons différentes selon la loi de variable dépendante : Nous avons appliqué un modèle spatial de Cox-PH (Cox-SPDE) modélisant directement le temps de survie, et un modèle de Poisson (P-SPDE) discrétisant le temps de survie et modéliser le nombre d'événements dans chaque intervalle de temps (distribution de Poisson).

#### **1.2.2.3.1. Modèle de Cox-SPDE**

Le modèle Cox-SPDE n'est autre que le modèle de Cox classique auquel est ajouté un champ gaussien pour modéliser la structure spatiale [41, 46].

Dans notre cas particulier, la variable dépendante ( $Y$ ) était le temps de survie ( $T$ ) suivant une loi de Weibull dont la fonction de risque instantané est donnée par :

$$\lambda(t, X) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + Z)$$

où  $X = (X_1, X_2, \dots, X_n)$  est le vecteur des covariables et  $\beta = (\beta_1, \beta_2, \dots, \beta_n)$  est le vecteur des effets associés aux covariables, les autres éléments du modèles étant définis plus haut.

#### **1.2.2.3.2. Modèle P-SPDE**

Le modèle P-SPDE est un modèle exponentiel par intervalle obtenu par un processus de comptage [46, 48, 49].

Ainsi, pour un nombre d'évènements  $D$  ( $\delta_{ik}$  étant l'indicatrice de l'évènement chez l'individu  $i = 1, 2, \dots, n$  dans l'intervalle de temps  $k = 1, 2, \dots, D$  et  $T_{ik}$  étant le temps de survie chez l'individu  $i$  dans le  $k^{ième}$  intervalle), en admettant que  $\delta$  suit une loi de Poisson de paramètre  $\pi$  (nombre moyen d'évènements par intervalle de temps).

Dans notre cas particulier, la structure spatiale était modélisée par un champ gaussien, comme décrit précédemment et la variable dépendante était le nombre d'évènement par intervalle de temps et non plus le temps de survie.

$$\delta \sim \text{Poisson}(\pi)$$

$$\log(\pi) = \lambda_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \log(T) + Z$$

$$\beta, \pi \sim N(0, 0.001)$$

où  $X = (X_1, X_2, \dots, X_n)$  est le vecteur des covariables et  $\beta = (\beta_1, \beta_2, \dots, \beta_n)$  est le vecteur des effets associés aux covariables, les autres éléments du modèles étant définis plus haut.

Les modèles à effets mixtes ou les modèles de type CAR [50–52] ont également une certaine capacité à modéliser la variation spatiale d'une exposition en introduisant une fragilité liée à l'appartenance à une zone donnée mais ces modèles n'apparaissent pas dans ce travail. En effet, ces modèles impliquent une subdivision de la zone de l'étude en unité spatiale préexistante ou construite (région, village, centre de santé, etc.) à l'intérieur de laquelle l'exposition serait supposé homogène. Cela suppose que l'hétérogénéité spatiale soit réellement observée avec les frontières d'homogénéité, alors que cette thèse s'adresse à une hétérogénéité spatiale non observée (comme les gîtes dans le cas du paludisme) où il est difficile (voir impossible) de définir les limites d'homogénéité du risque étant donné que les sources ne sont pas ou difficilement observables.

Il existe d'autres approches géoadditives [53, 54] modélisant également la variation spatiale d'une exposition à partir de la localisation exacte en utilisant une matrice de corrélation spatiale mais nous considérons que ces approches sont de même type que le SPDE qui utilise également une fonction de corrélation spatiale (l'échelle spatiale peut être différente d'une méthode à une autre).

### 1.2.3. Indicateurs de performance des modèles

Dans notre étude de simulation, le paramètre  $\beta$  représentait la vraie valeur de l'effet de chaque covariable estimé par  $\hat{\beta}$ , avec un écart-type  $\overline{Sd(\hat{\beta})}$ . Il faut souligner que pour les modèles bayésiens, l'estimation ponctuelle  $\hat{\beta}$  de la distribution a posteriori de l'effet des covariables était la moyenne. Les intervalles de crédibilités étaient calculés en utilisant les quantiles de la loi marginale a posteriori. En effet, les bornes de l'intervalle de crédibilité dans ce document correspondent à 0.025-quantile et 0.975-quantile, ce qui correspond à l'intervalle de confiance à 5% dans le contexte fréquentiste.

L'estimation est faite à partir des  $K$  jeux de données pour chaque scénario avec un intervalle de confiance (intervalle de crédibilité pour les modèles bayésiens) à 95%  $IC(\beta)$ .  $\hat{\beta}_{i=1,K}$  était l'effet estimé sur le  $i^{ième}$  jeu de données des  $K$  jeux de données d'un scénario spécifique.

$$\overline{Sd(\hat{\beta})} = \frac{1}{K} \sum_{i=1}^K Sd(\hat{\beta}_i)$$

$$\text{où } Sd(\hat{\beta}_i) = \sqrt{Var(\hat{\beta}_i)} = \frac{1}{K} \sqrt{(\hat{\beta}_i - \beta)^2}$$

Pour comparer les différents modèles, quatre indicateurs de performances ont été utilisé comme recommandé dans les références [27, 28] : le biais estimé  $B(\hat{\beta})$ , l'erreur quadratique moyenne (MSE), le taux de couverture (CR) et le taux significativité (SR) défini dessous.

Dans la suite de ce document, l'intervalle de confiance fréquentiste et son homologue bayésien, l'intervalle de crédibilité pour les modèles bayésiens (modèle SPDE) seront confondu pour faciliter la lecture du document.

### 1.2.3.1. Biais

Le biais  $B(\hat{\beta})$  était un indicateur de performance d'un estimateur exprimant l'écart entre l'effet réel et l'effet estimé :

$$B(\hat{\beta}) = \bar{\hat{\beta}} - \beta$$

$\bar{\hat{\beta}} = 1/K \sum_{i=1}^K \hat{\beta}_i$  était l'effet moyen estimé pour un scenario sur la base des 50 jeux de données.

### 1.2.3.2. Erreur quadratique moyenne (MSE, Mean Square Error)

L'erreur quadratique moyenne (MSE) a permis d'évaluer la stabilité et la précision des estimations.

Elle était donnée par la formule mathématique suivante :

$$MSE(\hat{\beta}) = (B(\hat{\beta}))^2 + (\overline{Sd(\hat{\beta})})^2$$

### 1.2.3.3. Taux de couverture (CR, Coverage Rate)

Le CR évaluait le nombre de fois où l'effet réel était couvert par l'intervalle de confiance (intervalle de crédibilité pour les modèles bayésiens) à 95%. Il était donné par la formule mathématique suivante :

$$CR(\beta) = \frac{1}{K} \sum_{i=1}^K \mathbb{1}_{\{\beta \in IC(\hat{\beta})\}}$$

#### 1.2.3.4. Taux de significativité (SR, Significance Rate)

Le SR exprimait la proportion du nombre de fois où l'effet estimé était significatif, autrement dit, le nombre de fois où l'intervalle de confiance des paramètres ne contenait pas la valeur zéro. Il était donné par la formule mathématique suivante :

$$SR(\beta) = \frac{1}{K} \sum_{i=1}^K \mathbb{1}_{\{0 \notin IC(\hat{\beta})\}}$$

#### 1.2.4. Exemple d'application

Les quatre modèles (Cox-PH, GAM, Cox-SPDE et P-SPDE) ont été appliqués, suivant les spécifications précédentes, aux données d'un essai vaccinal visant à tester un candidat vaccin bloquant la transmission du paludisme : le Pfs230 [55]. Cet essai randomisé contrôlé a été mené de 2015 à 2016 dans la localité de Bancoumana, au Mali (pour plus de détails cf. [56]). Dans le cadre de notre travail, l'évènement d'intérêt était l'accès clinique palustre. L'âge, le sexe et les coordonnées géographiques (GPS) des participants ont aussi été recueillis. La fraction préventive du vaccin a été estimée en prenant une prévalence estimée de 77%.

### 1.3. Résultats

#### 1.3.1. Modèle de simulation des données (DGM, Data-Generating Model)

Le DGM est le modèle qui avait servi à simuler les données en tenant compte du risque environnemental. Comme prévu pour ce modèle, le biais et le MSE étaient presque nuls et présentaient presque toujours le CR souhaité, quel que soit le scénario (*Annexe I*).

#### 1.3.2. Impact du risque de base sur la qualité des estimations

Le niveau de risque de base n'avait que très peu d'impact sur les indicateurs de performances des différents modèles, indépendamment des autres paramètres (*Figure A.2*).

Par exemple, avec un faible effet gîte (risque relatif  $RR_g = 1.05$ ), une faible densité de gîtes ( $D_g = 0.25$ ), un effet traitement modéré (risque relatif  $RR_t = 0.6$ ) et une faible densité de population ( $\tau = 0.2$ ), les indicateurs de performances des modèles étaient identiques pour un risque de base de 0.06 et un risque de base de 0.6, comme par exemple pour le modèle de Cox-PH (biais =  $-9.37 \cdot 10^{-3}$  et  $-8.14 \cdot 10^{-3}$ , respectivement) et le modèle Cox-SPDE (biais =  $-26.57 \cdot 10^{-3}$  et  $-21.38 \cdot 10^{-3}$ , respectivement).

De même, pour l'estimation de l'effet âge et de l'effet sexe, les indicateurs de performances des modèles n'étaient impactés par le changement de valeur du risque de base, quel que soit le scénario.

Pour la suite, seuls les résultats concernant le risque de base de 0.37 seraient donnés sachant qu'un changement de cette valeur n'a que très peu d'impact.

### 1.3.3. Estimations de l'effet traitement

Lorsque l'effet gîte était faible ( $RR_g \leq 1.2$ ), la densité de gîtes avait peu d'impact sur le biais et le MSE de l'effet traitement, qui par ailleurs, étaient faible pour tous les modèles, quel que soit le scénario.

En effet, avec un faible effet gîte  $RR_b \leq 1.2$ , le biais de l'effet traitement variait de -0.034 à 0.045 pour tous les modèles quel que soit le scénario hormis le modèle Cox-SPDE (G1 et G2 des Figures 3 et A.3). Dans ces conditions de faible effet gîte, le modèle Cox-SPDE tendait à surestimer légèrement l'effet traitement lorsque la densité de population était faible ( $\tau = 0.2$ ), d'autant plus que l'effet traitement était fort (**Tableau 1**). Par exemple, avec  $RR_t = 0.25$ ,  $RR_g = 1.05$ ,  $D_g = 0.25$  et  $\tau = 0.2$ , le modèle Cox-SPDE surestimait l'effet traitement avec un biais maximum atteignant -0.071.

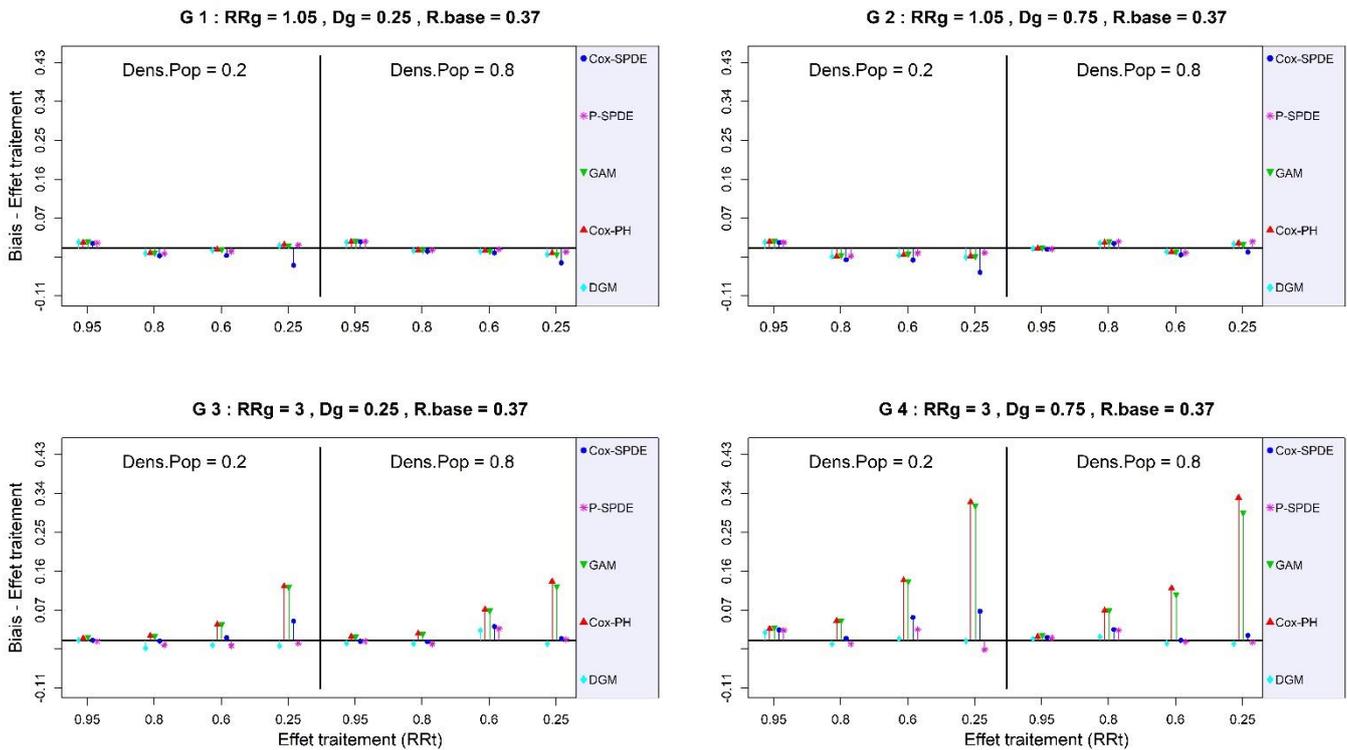
En revanche, avec une forte densité de population ( $\tau = 0.8$ , les autres paramètres demeurants inchangés), le biais de l'effet du traitement était de -0.033. Cependant, même dans des conditions de faible densité de population, cette surestimation était corrigée lorsque l'effet traitement était faible ( $RR_t \geq 0.80$ ). Par exemple, avec  $RR_g = 1.05$ ,  $D_g = 0.25$  et  $\tau = 0.2$ , le biais de l'effet traitement pour le modèle Cox-SPDE variait de -0.011 à 0.012 pour  $RR_t = 0.95$ .

Le SR (*resp.* CR) de l'effet traitement était similaire pour tous les modèles quel que soit le scénario associé au faible effet gîte. Il variait de 2% à 24% (*resp.* de 86% à 100 %) pour  $RR_t = 0.95$ , et atteignait 100% (*resp.* de 82% à 100%) pour  $RR_t = 0.25$ . Comme prévu, il y avait un manque de puissance lorsque l'effet traitement était faible ( $RR_t = 0.95$ ) (G1 et G2 de la Figures 5 et 6). Cependant, lorsque l'effet gîte était fort ( $RR_g \geq 1.51$ ), l'augmentation de la densité de gîtes provoquait une augmentation importante du biais de l'effet traitement pour les modèles Cox-PH et GAM. Le biais de l'effet traitement était encore plus important lorsque cet effet était fort (G3 et G4 de *Figures 3 et 4*).

Par exemple, pour le modèle Cox-PH, avec  $RR_g = 3$ ,  $D_g = 0.25$  et  $\tau = 0.2$ , le biais de l'effet traitement pouvait atteindre un maximum de 0.139 pour le plus haut niveau de l'effet traitement ( $RR_t = 0.95$ ). Avec ces mêmes paramètres, le biais de l'effet traitement atteignait un maximum de 0.131 pour le modèle GAM. Quant aux deux modèles SPDE, le biais de l'effet traitement se limitait à un maximum de 0.054 pour le modèle Cox-SPDE et 0.010 pour le modèle P-SPDE.

Pour une densité de gîtes plus forte ( $D_g = 0.75$ , les autres paramètres restants inchangés), le biais de l'effet traitement atteignait un maximum de 0.342 pour le plus haut niveau de l'effet traitement ( $RR_t = 0.25$ ) pour le modèle Cox-PH. Le biais de l'effet traitement atteignait un maximum de 0.33 pour le modèle GAM. Quant aux deux modèles SPDE, le biais de l'effet traitement restait toujours à un maximum de 0.09 pour le modèle Cox-SPDE et 0.014 pour le modèle P-SPDE.

**Figure 3:** Biases de l'effet traitement avec un risque de base de 0.37



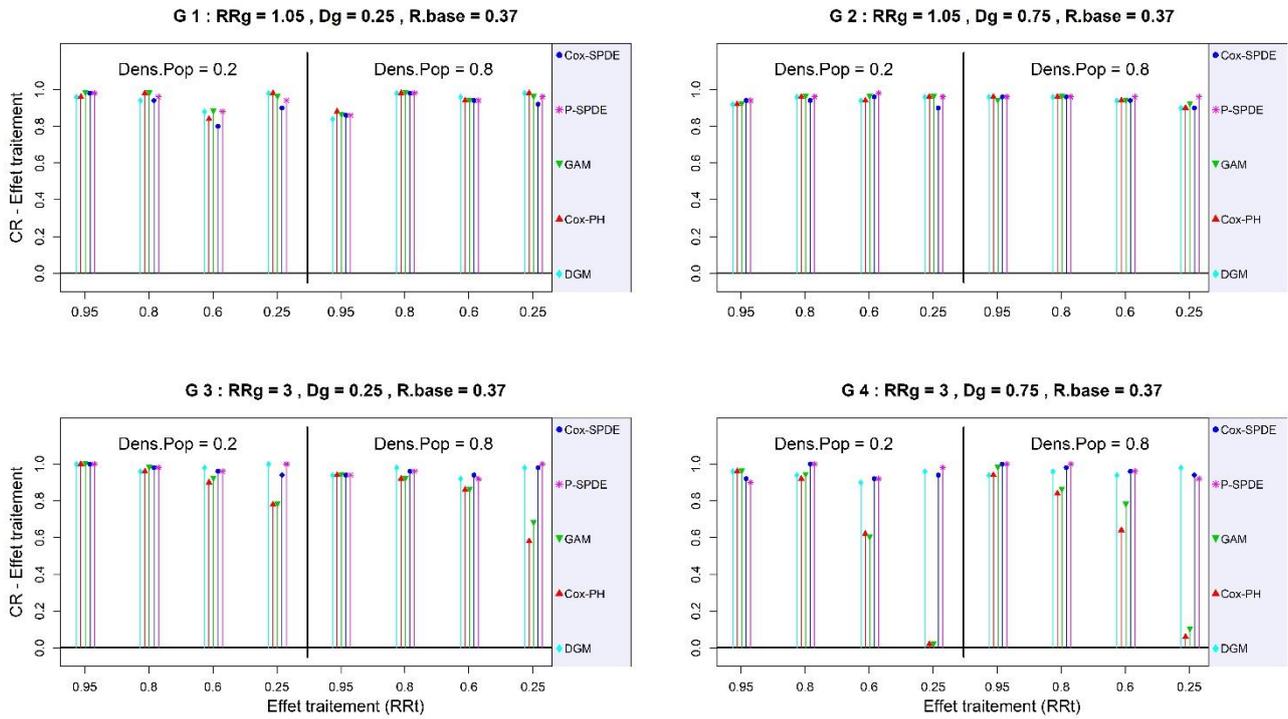
*DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base*

Dans ces conditions de fort effet gîte ( $RR_g \geq 1.51$ ), le CR de l'effet traitement était négativement impacté par une augmentation de la densité de gîtes pour les modèles Cox-PH et GAM (G3 et G4 de la Figure 4 et 5). Cet impact était encore plus important lorsque l'effet traitement était fort.

Par exemple, avec un fort effet gîte  $RR_g = 3$ , une faible densité de gîtes  $D_g = 0.25$  et une forte densité de population  $\tau = 0.8$ , le CR pour le niveau le plus bas de l'effet traitement ( $RR_t = 0.95$ ) était à un maximum de 94% pour le modèle Cox-PH, 92% pour le modèle GAM. Alors que pour le niveau le plus haut de l'effet traitement ( $RR_t = 0.25$ ), le CR était à un maximum de 8% pour le modèle de Cox-PH et 14% pour le modèle GAM. Alors que le CR restait toujours élevé quel que soit le niveau de l'effet traitement pour les deux modèles SPDE (un minimum de 96% pour le modèle Cox-SPDE et 98% pour le modèle P-SPDE).

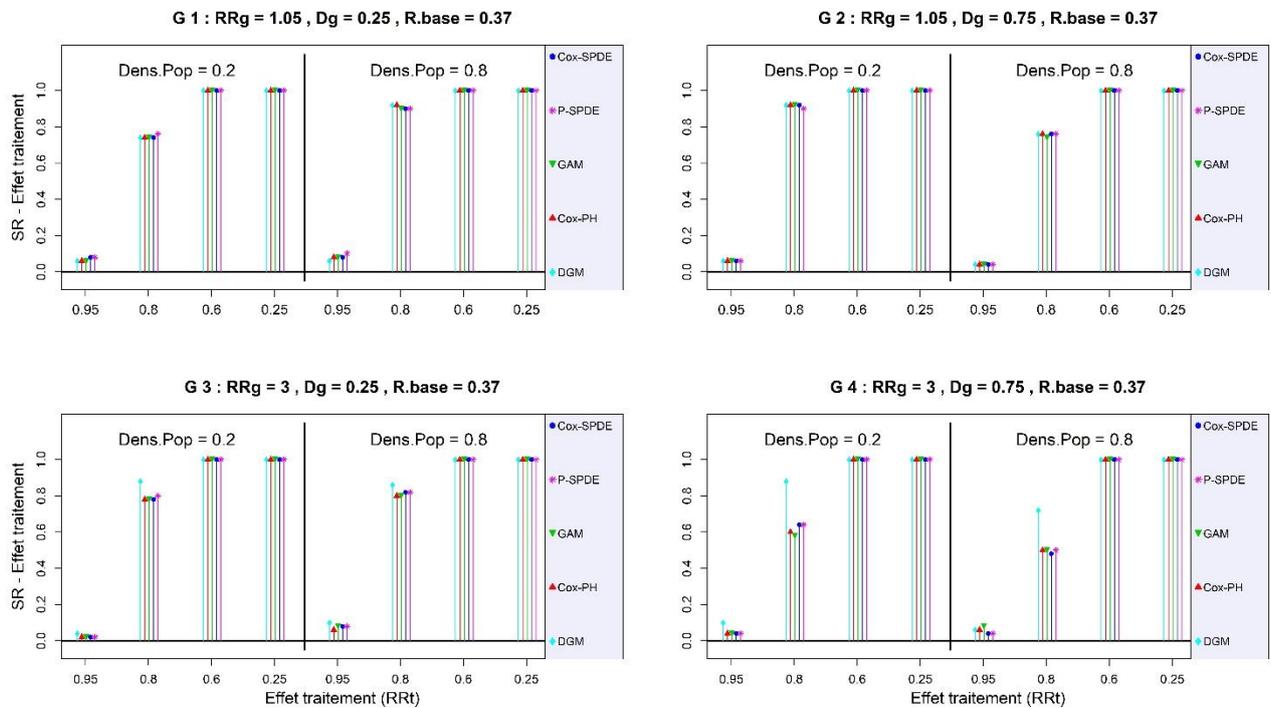
Le SR de l'effet traitement n'était pas impacté par le niveau fort de l'effet gîte. Il était semblable pour tous les modèles, mais variait d'un scénario à l'autre selon le niveau de l'effet traitement (avec un manque de puissance attendu pour le faible effet traitement  $RR_t = 0.95$ ). En effet, pour le niveau le plus élevés de l'effet traitement ( $RR_t = 0.25$ ), le SR de l'effet traitement était compris entre 98% et 100% pour tous les modèles (*Tableau 1*).

**Figure 4:** Taux de couverture de l'effet traitement avec un risque de base de 0.37



DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base.

**Figure 5:** Taux de significativité de l'effet traitement avec un risqué de base de 0.37



DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base.

Tableau 1 : Indicateurs de performance des modèles pour l'estimation de l'effet traitement (risque de base 0.37)

RRg	RRt	Modèles	Densité de gîtes															
			0.25								0.75							
			Densité de population								Densité de population							
			0.2				0.8				0.2				0.8			
			Biais	MSE	CR	SR	Biais	MES	CR	CS	Biais	MSE	CR	SR	Biais	MSE	CR	SR
1.5	0.95	Cox-PH	0.014	0.006	0.96	0.06	0.015	0.006	0.88	0.08	0.016	0.006	0.92	0.06	<0.001	0.0058	0.96	0.04
		GAM	0.014	0.006	0.98	0.06	0.015	0.006	0.86	0.08	0.015	0.006	0.92	0.06	<0.001	0.0058	0.94	0.04
		Cox-SPDE	0.012	0.0063	0.98	0.08	0.016	0.0062	0.86	0.08	0.014	0.0063	0.94	0.06	-0.002	0.0059	0.96	0.04
		P-SPDE	0.013	0.0061	0.98	0.08	0.016	0.0061	0.86	0.1	0.014	0.0061	0.94	0.06	-0.002	0.0059	0.96	0.04
	0.80	Cox-PH	-0.011	0.006	0.98	0.74	-0.005	0.0058	0.98	0.92	-0.018	0.0061	0.96	0.92	0.013	0.006	0.96	0.76
		GAM	-0.012	0.006	0.98	0.74	-0.006	0.0059	0.98	0.9	-0.018	0.0062	0.96	0.92	0.014	0.006	0.96	0.74
		Cox-SPDE	-0.017	0.0065	0.94	0.74	-0.007	0.006	0.98	0.9	-0.026	0.0069	0.94	0.92	0.011	0.0061	0.96	0.76
		P-SPDE	-0.012	0.0061	0.96	0.76	-0.004	0.0059	0.98	0.9	-0.018	0.0062	0.96	0.9	0.016	0.0061	0.96	0.76
	0.60	Cox-PH	-0.003	0.006	0.84	1.00	-0.005	0.006	0.94	1.00	-0.015	0.0062	0.94	1.00	-0.008	0.0061	0.94	1.00
		GAM	-0.004	0.006	0.88	1.00	-0.008	0.0061	0.94	1.00	-0.015	0.0063	0.96	1.00	-0.01	0.0061	0.94	1.00
		Cox-SPDE	-0.017	0.0065	0.8	1.00	-0.011	0.0062	0.94	1.00	-0.027	0.007	0.96	1.00	-0.015	0.0064	0.94	1.00
		P-SPDE	-0.007	0.006	0.88	1.00	-0.002	0.0059	0.94	1.00	-0.011	0.006	0.98	1.00	-0.011	0.006	0.96	1.00
	0.25	Cox-PH	0.009	0.0075	0.98	1.00	-0.011	0.0076	0.98	1.00	-0.018	0.0078	0.96	1.00	0.012	0.0076	0.9	1.00
		GAM	0.005	0.0075	0.96	1.00	-0.016	0.0078	0.96	1.00	-0.02	0.0079	0.96	1.00	0.007	0.0075	0.92	1.00
		Cox-SPDE	-0.039	0.0091	0.9	1.00	-0.033	0.0085	0.92	1.00	-0.055	0.0107	0.9	1.00	-0.008	0.0074	0.9	1.00
		P-SPDE	0.008	0.0061	0.94	1.00	-0.008	0.0061	0.96	1.00	-0.01	0.0061	0.96	1.00	0.016	0.0063	0.96	1.00
3	0.95	Cox-PH	0.005	0.0058	1.00	0.02	0.009	0.0058	0.94	0.06	0.027	0.0065	0.96	0.04	0.009	0.0058	0.94	0.06
		GAM	0.005	0.0058	1.00	0.02	0.007	0.0059	0.94	0.08	0.027	0.0066	0.96	0.04	0.01	0.006	0.98	0.08
		Cox-SPDE	<0.001	0.0068	1.00	0.02	-0.002	0.007	0.94	0.08	0.024	0.0094	0.92	0.04	0.007	0.0085	1.00	0.04
		P-SPDE	-0.002	0.0072	1.00	0.02	-0.002	0.007	0.94	0.08	0.023	0.0101	0.9	0.04	0.006	0.0087	1.00	0.04
	0.80	Cox-PH	0.01	0.006	0.96	0.78	0.016	0.0061	0.92	0.8	0.045	0.0079	0.92	0.6	0.069	0.0106	0.84	0.5
		GAM	0.008	0.006	0.98	0.78	0.013	0.0061	0.92	0.8	0.044	0.0078	0.94	0.58	0.068	0.0106	0.86	0.5
		Cox-SPDE	-0.002	0.0068	0.98	0.78	-0.003	0.007	0.96	0.82	0.004	0.009	1.00	0.64	0.025	0.0091	0.98	0.48
		P-SPDE	-0.011	0.0072	0.98	0.8	-0.008	0.0072	0.96	0.82	-0.008	0.0096	1.00	0.64	0.023	0.0091	1.00	0.5
	0.60	Cox-PH	0.036	0.0073	0.9	1.00	0.072	0.0111	0.86	1.00	0.139	0.0252	0.62	1.00	0.121	0.0204	0.64	1.00
		GAM	0.035	0.0072	0.92	1.00	0.067	0.0105	0.86	1.00	0.135	0.0241	0.6	1.00	0.104	0.0169	0.78	1.00
		Cox-SPDE	0.006	0.0071	0.96	1.00	0.032	0.0084	0.94	1.00	0.053	0.012	0.92	1.00	<0.001	0.0089	0.96	1.00
		P-SPDE	-0.012	0.0072	0.96	1.00	0.027	0.0078	0.92	1.00	0.025	0.0101	0.92	1.00	-0.003	0.0086	0.96	1.00
	0.25	Cox-PH	0.125	0.0228	0.78	1.00	0.136	0.0256	0.58	1.00	0.32	0.1091	0.02	1.00	0.33	0.1155	0.06	1.00
		GAM	0.121	0.0218	0.78	1.00	0.123	0.0222	0.68	1.00	0.309	0.1025	0.02	1.00	0.293	0.0928	0.1	1.00
		Cox-SPDE	0.045	0.0116	0.94	1.00	0.004	0.0098	0.98	1.00	0.067	0.0177	0.94	1.00	0.011	0.0118	0.94	1.00
		P-SPDE	-0.006	0.0074	1.00	1.00	0.002	0.0072	1.00	1.00	-0.021	0.0104	0.98	1.00	-0.004	0.0088	0.92	1.00

MSE : Erreur Quadratique Moyenne ; CR : Taux de couverture ; SR : Taux de significativité ; Cox-PH : Modèle de Cox à risques proportionnels ; GAM : Modèle Additif Généralisé ; Cox-SPDE : Modèle Cox-Equation aux Dérivées Partielles Stochastiques ; P-SPDE : Modèle Poisson-Equation aux Dérivées Partielles Stochastiques ; RRt : risque relatif associé au traitement ; RRg : Risque relatif associé au gîte. Les valeurs en GRAS (fond gris) représentent les cas les plus extrêmes (faible densité de gîte avec effet faible pour un faible effet traitement et forte densité de gîte avec effet fort pour un fort effet traitement) et la sensibilité du Cox-SPDE à la densité de population lorsque l'effet gîte était faible.

## Application

Dans le cadre de l'essai vaccinal Pfs230, aucun des modèles n'a montré un effet significatif du vaccin sur les accès cliniques du paludisme.

Le modèle Cox-SPDE estimait une fraction préventive légèrement plus importante (29.26% [-10.93 ; 51.21%]), suivi du modèle P-SPDE (26.95% [-15.25% ; 49.90%]). En revanche, le modèle Cox-PH non spatial estimait une fraction préventive légèrement plus faible (24.64% [-16.56% ; 47.66%]), tout comme le modèle GAM (22.33% [-22.25 ; 46.82]).

### 1.3.4. Estimations de l'effet âge

Lorsque l'effet gîte était faible ( $RR_g = 1.05$  ou  $1.2$ ), la densité de gîtes avait peu d'impact sur le biais de l'effet âge, quel que soit le modèle. Par exemple, quel que soit le scénario associé à un faible effet gîte, le biais de l'effet âge variait autour de  $-0.0035$  pour tous les modèles hormis le modèle Cox-SPDE (G 1 et G2 de la *Figures A.4* et *A.5*). Tout comme dans le cas de l'effet traitement, le modèle Cox-SPDE tendait à surestimer légèrement l'effet âge, d'autant plus que la densité de population autour des points de concentration était faible. Ainsi, pour une densité de population  $\tau = 0.2$ , le biais était d'environ  $-0.007$  alors qu'il variait autour  $-0.003$  pour une densité de population  $\tau = 0.8$  (*Tableau 2*).

Le SR de l'effet âge était identique à 100% pour tous les modèles, quel que soit le scénario. En revanche, dans cette condition de faible effet gîte, le CR de l'effet âge était légèrement plus bas pour le modèle Cox-SPDE lorsque la densité de population était faible.

En effet, avec un faible effet gîte ( $RR_g = 1.05$ ), une faible densité de gîtes ( $D_g = 0.25$ ) et une densité de population  $\tau = 0.2$ , le CR de l'effet âge variait de 86% à 100% pour le modèle Cox-PH non spatial, de 88% à 100% pour le modèle GAM, de 90% à 100% pour le modèle P-SPDE alors qu'il variait de 62% à 92% pour le modèle Cox-SPDE.

A contrario, lorsque l'effet gîte était fort ( $RR_g = 1.51$  ou  $3$ ), le biais de l'effet âge augmentait de façon majeure avec l'augmentation de la densité de gîtes  $D_g$  pour les modèles Cox-PH et GAM (G3 et G4 dans les **Figures A.4** et **A.5**). Par exemple, pour le modèle Cox-PH, avec un fort effet gîte ( $RR_g = 3$ ) et une densité de population faible ( $\tau = 0.2$ ), le biais de l'effet âge variait autour de  $0.017$  pour une faible densité de gîtes ( $D_g = 0.25$ ) et variait autour de  $0.041$  pour une forte densité de gîtes ( $D_g = 0.75$ ). Le modèle GAM, bien que spatial, avait un biais sur l'effet âge similaire au modèle de Cox-PH non spatial dans tous les scénarios. Pour le modèle Cox-SPDE, le biais de l'effet âge était légèrement impacté par la densité de population  $\tau$  qui par ailleurs était plus faible lorsque la densité de population était forte. Par exemple, avec un fort effet gîte ( $RR_g = 3$ ) et une forte densité de gîtes ( $D_g = 0.75$ ), le biais de l'effet âge variait autour de  $0.01$  pour une faible densité de population ( $\tau = 0.2$ ) et autour de  $0.001$  pour une forte densité de population ( $\tau = 0.8$ ). Pour le modèle P-SPDE, le biais de l'effet âge était faible ( $\sim 0$ ) quel que soit le scénario. Il faut noter que le biais de l'effet âge était similaire entre le modèle P-SPDE et le modèle Cox-SPDE, d'autant plus que la densité de population  $\tau$  était forte.

Dans ce cas de fort effet gîte, le CR de l'effet âge était faible pour les modèles Cox-PH et GAM (G3 et G4 dans la Figure A.6). Par exemple, pour le modèle Cox-PH (*resp.* modèle GAM), pour un effet gîte fort ( $RR_g = 3$ ), le CR de l'effet âge était égal quasiment nul (*resp.*  $< 2\%$ ) lorsque la densité de gîtes était forte ( $D_g = 0.75$ ). Dans ces mêmes conditions, pour les deux modèles SPDE, le CR de l'effet âge était élevé avec un minimum de  $70\%$  pour le modèle Cox-SPDE et un minimum de  $82\%$  pour le modèle P-SPDE. Cependant, le SR de l'effet âge était de  $100\%$  pour tous les modèles, quel que soit le scénario (**Figure A.7**). Il est important de souligner que pour ces jeux de données simulées, le niveau de l'effet traitement n'avait aucun impact sur les estimations de l'effet âge, quel que soit le scénario.

Tableau 2 : Indicateurs de performance des modèles pour l'estimation de l'effet âge (risque de base 0.37)

RRg	RRt	Modèles	Densité de gîtes															
			0.25								0.75							
			Densité de population								Densité de population							
			0.2				0.8				0.2				0.8			
Biais	MSE	CR	SR	Biais	MSE	CR	SR	Biais	MSE	CR	SR	Biais	MSE	CR	SR			
1.5	0.95	Cox-PH	0.002	< 0.0001	0.96	1.00	< 0.001	< 0.0001	0.9	1.00	-0.002	< 0.0001	0.98	1.00	-0.001	< 0.0001	0.98	1.00
		GAM	0.001	< 0.0001	0.94	1.00	-0.001	0.0001	0.9	1.00	-0.002	< 0.0001	0.96	1.00	-0.002	0.0001	0.96	1.00
		Cox-SPDE	-0.004	< 0.0001	0.76	1.00	-0.003	< 0.0001	0.86	1.00	-0.006	0.0001	0.72	1.00	-0.003	< 0.0001	0.88	1.00
		P-SPDE	< 0.001	< 0.0001	0.94	1.00	-0.001	< 0.0001	0.96	1.00	-0.001	< 0.0001	0.96	1.00	-0.001	< 0.0001	0.92	1.00
	0.80	Cox-PH	< 0.001	< 0.0001	0.96	1.00	-0.001	< 0.0001	0.96	1.00	< 0.001	< 0.0001	0.96	1.00	-0.002	< 0.0001	0.84	1.00
		GAM	< 0.001	< 0.0001	0.96	1.00	-0.002	0.0001	0.96	1.00	-0.001	< 0.0001	0.98	1.00	-0.003	0.0001	0.88	1.00
		Cox-SPDE	-0.004	0.0001	0.76	1.00	-0.003	< 0.0001	0.86	1.00	-0.006	0.0001	0.68	1.00	-0.005	0.0001	0.72	1.00
		P-SPDE	-0.001	< 0.0001	0.96	1.00	-0.001	< 0.0001	0.92	1.00	< 0.001	< 0.0001	0.94	1.00	-0.002	< 0.0001	0.92	1.00
	0.60	Cox-PH	0.001	< 0.0001	0.9	1.00	-0.001	< 0.0001	0.96	1.00	-0.002	< 0.0001	0.96	1.00	0.001	< 0.0001	0.96	1.00
		GAM	< 0.001	< 0.0001	0.92	1.00	-0.002	0.0001	0.96	1.00	-0.002	< 0.0001	0.94	1.00	< 0.001	0.0001	0.96	1.00
		Cox-SPDE	-0.003	< 0.0001	0.84	1.00	-0.003	< 0.0001	0.86	1.00	-0.007	0.0001	0.66	1.00	-0.002	< 0.0001	0.88	1.00
		P-SPDE	< 0.001	< 0.0001	0.94	1.00	-0.001	< 0.0001	0.94	1.00	-0.002	< 0.0001	0.98	1.00	< 0.001	< 0.0001	0.94	1.00
	0.25	Cox-PH	< 0.001	< 0.0001	0.98	1.00	-0.001	< 0.0001	0.98	1.00	-0.001	< 0.0001	0.94	1.00	< 0.001	< 0.0001	0.96	1.00
		GAM	< 0.001	< 0.0001	0.98	1.00	-0.003	0.0001	0.9	1.00	-0.002	< 0.0001	0.92	1.00	< 0.001	< 0.0001	0.94	1.00
		Cox-SPDE	-0.006	0.0001	0.66	1.00	-0.004	< 0.0001	0.8	1.00	-0.006	0.0001	0.76	1.00	-0.003	< 0.0001	0.8	1.00
		P-SPDE	< 0.001	< 0.0001	0.96	1.00	-0.001	< 0.0001	0.96	1.00	< 0.001	< 0.0001	0.94	1.00	< 0.001	< 0.0001	1.00	1.00
3	0.95	Cox-PH	0.018	0.0004	0.16	1.00	0.016	0.0003	0.32	1.00	0.041	0.0017	0.00	1.00	0.042	0.0018	0.00	1.00
		GAM	0.017	0.0003	0.2	1.00	0.014	0.0002	0.5	1.00	0.04	0.0016	0.00	1.00	0.037	0.0014	0.00	1.00
		Cox-SPDE	0.007	0.0001	0.74	1.00	< 0.001	0.0001	0.98	1.00	0.01	0.0002	0.84	1.00	0.003	0.0001	0.86	1.00
		P-SPDE	-0.001	< 0.0001	0.92	1.00	-0.001	< 0.0001	0.96	1.00	< 0.001	< 0.0001	0.94	1.00	< 0.001	< 0.0001	0.96	1.00
	0.80	Cox-PH	0.018	0.0004	0.18	1.00	0.019	0.0004	0.18	1.00	0.042	0.0018	0.00	1.00	0.041	0.0017	0.00	1.00
		GAM	0.017	0.0003	0.26	1.00	0.018	0.0004	0.32	1.00	0.04	0.0017	0.00	1.00	0.036	0.0014	0.00	1.00
		Cox-SPDE	0.007	0.0001	0.86	1.00	0.005	0.0001	0.9	1.00	0.011	0.0002	0.8	1.00	0.002	0.0001	0.98	1.00
		P-SPDE	0.001	< 0.0001	0.92	1.00	0.001	< 0.0001	0.96	1.00	0.002	< 0.0001	0.9	1.00	< 0.001	< 0.0001	0.96	1.00
	0.60	Cox-PH	0.017	0.0003	0.26	1.00	0.016	0.0003	0.3	1.00	0.042	0.0018	0.00	1.00	0.04	0.0017	0.00	1.00
		GAM	0.016	0.0003	0.28	1.00	0.015	0.0003	0.52	1.00	0.041	0.0017	0.00	1.00	0.037	0.0014	0.00	1.00
		Cox-SPDE	0.007	0.0001	0.8	1.00	0.001	0.0001	0.94	1.00	0.012	0.0002	0.76	1.00	0.001	0.0001	0.96	1.00
		P-SPDE	0.001	< 0.0001	0.98	1.00	-0.001	< 0.0001	0.98	1.00	0.002	< 0.0001	0.94	1.00	-0.001	< 0.0001	0.96	1.00
	0.25	Cox-PH	0.018	0.0004	0.2	1.00	0.016	0.0003	0.28	1.00	0.042	0.0018	0.00	1.00	0.043	0.0019	0.00	1.00
		GAM	0.017	0.0003	0.2	1.00	0.014	0.0003	0.44	1.00	0.041	0.0017	0.00	1.00	0.038	0.0015	0.00	1.00
		Cox-SPDE	0.007	0.0001	0.8	1.00	-0.001	0.0001	0.94	1.00	0.011	0.0002	0.78	1.00	0.003	0.0001	0.92	1.00
		P-SPDE	0.001	< 0.0001	0.96	1.00	-0.001	< 0.0001	0.92	1.00	< 0.001	< 0.0001	0.98	1.00	0.001	< 0.0001	0.94	1.00

MSE : Erreur Quadratique Moyenne ; CR : Taux de couverture ; SR : Taux de significativité ; Cox-PH : Modèle de Cox à risques proportionnels ; GAM : Modèle Additif Généralisé ; Cox-SPDE : Modèle Cox-Equation aux Dérivées Partielles Stochastiques ; P-SPDE : Modèle Poisson-Equation aux Dérivées Partielles Stochastiques ; RR<sub>t</sub> : risque relatif associé au traitement ; RR<sub>g</sub> : Risque relatif associé au gîte. Les valeurs en GRAS (fond gris) représentent les cas les plus extrêmes (faible densité de gîte avec effet faible pour un faible effet traitement et forte densité de gîte avec effet fort pour un fort effet traitement) et la sensibilité du Cox-SPDE à la densité de population lorsque l'effet gîte était faible.

### 1.3.5. Estimations de l'effet sexe

Quel que soit le scénario, le biais, le MSE, le CR et les SR de l'effet du sexe étaient à peu près identiques pour tous les modèles (*Tableau 3*), avec très peu de variations d'un scénario à l'autre (*Figure A.8 à A.11*).

*Tableau 3* : Indicateurs de performance des modèles pour l'estimation de l'effet sexe (risque de base 0.37)

RRg	RRt	Modèles	Densité de gîtes															
			0.25								0.75							
			Densité de population								Densité de population							
			0.2				0.8				0.2				0.8			
Biais	MSE	CR	SR	Biais	MSE	CR	SR	Biais	MSE	CR	SR	Biais	MSE	CR	SR			
1.5	0.95	Cox-PH	-0.009	0.0058	1.00	0.00	-0.022	0.0063	0.88	0.12	-0.013	0.006	0.96	0.04	0.004	0.0058	0.92	0.08
3	0.95	GAM	-0.009	0.0059	1.00	0.00	-0.021	0.0063	0.88	0.12	-0.013	0.006	0.96	0.04	0.005	0.0058	0.92	0.08
		Cox-SPDE	-0.009	0.0062	1.00	0.00	-0.024	0.0066	0.88	0.12	-0.015	0.0064	0.96	0.04	0.006	0.006	0.96	0.04
		P-SPDE	-0.008	0.006	1.00	0.00	-0.024	0.0064	0.88	0.12	-0.014	0.0061	0.96	0.04	0.005	0.0059	0.96	0.04
		Cox-PH	-0.01	0.0059	0.96	0.04	<0.001	0.0058	0.96	0.04	0.012	0.0059	0.96	0.04	0.011	0.0059	0.92	0.08
	0.80	GAM	-0.01	0.006	0.98	0.02	<0.001	0.0058	0.98	0.02	0.013	0.006	0.96	0.04	0.012	0.006	0.92	0.08
		Cox-SPDE	-0.01	0.0062	0.96	0.04	0.001	0.0059	0.98	0.02	0.014	0.0063	0.96	0.04	0.013	0.0062	0.94	0.06
		P-SPDE	-0.011	0.0061	0.96	0.04	<0.001	0.0059	0.98	0.02	0.013	0.0061	0.96	0.04	0.012	0.006	0.94	0.06
		Cox-PH	-0.024	0.0063	0.98	0.02	0.004	0.0058	0.98	0.02	0.002	0.0058	0.92	0.08	<0.001	0.0058	0.94	0.06
	0.60	GAM	-0.024	0.0064	0.98	0.02	0.003	0.0058	0.98	0.02	0.003	0.0058	0.92	0.08	0.002	0.0058	0.94	0.06
		Cox-SPDE	-0.023	0.0066	0.96	0.04	0.004	0.0059	0.98	0.02	0.001	0.0061	0.96	0.04	0.003	0.006	0.94	0.06
		P-SPDE	-0.023	0.0064	0.96	0.04	0.005	0.0059	0.98	0.02	0.001	0.0058	0.98	0.02	0.003	0.0059	0.94	0.06
		Cox-PH	-0.004	0.0058	0.96	0.04	0.007	0.0058	0.98	0.02	-0.006	0.0058	0.94	0.06	-0.001	0.0058	0.96	0.04
	0.25	GAM	-0.003	0.0058	0.96	0.04	0.006	0.0058	0.98	0.02	-0.007	0.0059	0.94	0.06	-0.001	0.0058	0.96	0.04
		Cox-SPDE	-0.005	0.0062	0.96	0.04	0.008	0.006	0.98	0.02	-0.008	0.0061	0.94	0.06	<0.001	0.0059	0.94	0.06
		P-SPDE	-0.005	0.0059	0.96	0.04	0.008	0.0059	0.98	0.02	-0.007	0.0059	0.94	0.06	<0.001	0.0059	0.96	0.04
		Cox-PH	0.002	0.0058	0.98	0.02	-0.019	0.0061	0.92	0.08	-0.026	0.0064	0.92	0.08	0.003	0.0058	0.96	0.04
	0.95	GAM	0.002	0.0058	0.98	0.02	-0.019	0.0062	0.9	0.1	-0.028	0.0066	0.92	0.08	0.004	0.006	0.96	0.04
		Cox-SPDE	0.002	0.0067	0.96	0.04	-0.022	0.0075	0.9	0.1	-0.039	0.0103	0.96	0.04	0.013	0.0087	0.94	0.06
		P-SPDE	0.003	0.0072	0.96	0.04	-0.022	0.0074	0.9	0.1	-0.042	0.0113	0.94	0.06	0.014	0.0089	0.94	0.06
		Cox-PH	0.004	0.0058	0.94	0.06	-0.008	0.0058	0.94	0.06	-0.006	0.0058	0.96	0.04	-0.004	0.0058	0.96	0.04
0.80	GAM	0.005	0.0059	0.94	0.06	-0.008	0.0059	0.94	0.06	-0.009	0.006	0.94	0.06	-0.001	0.006	0.94	0.06	
	Cox-SPDE	0.005	0.0068	0.92	0.08	-0.009	0.007	0.98	0.02	-0.01	0.0089	0.96	0.04	0.002	0.0085	0.96	0.04	
	P-SPDE	0.006	0.0072	0.92	0.08	-0.009	0.0072	0.96	0.04	-0.01	0.0096	0.96	0.04	0.003	0.0086	0.96	0.04	
	Cox-PH	-0.016	0.0061	0.98	0.02	0.002	0.0058	0.98	0.02	-0.01	0.0058	0.96	0.04	-0.007	0.0058	0.9	0.1	
0.60	GAM	-0.018	0.0062	0.98	0.02	0.003	0.0059	0.96	0.04	-0.01	0.0059	0.98	0.02	-0.004	0.006	0.92	0.08	
	Cox-SPDE	-0.018	0.007	0.98	0.02	0.001	0.007	0.94	0.06	-0.014	0.0088	0.96	0.04	-0.001	0.0085	0.96	0.04	
	P-SPDE	-0.018	0.0074	0.98	0.02	0.001	0.0071	0.92	0.08	-0.013	0.0095	0.92	0.08	-0.001	0.0086	0.96	0.04	
	Cox-PH	-0.02	0.0062	0.9	0.1	-0.003	0.0058	0.98	0.02	-0.001	0.0058	0.92	0.08	-0.006	0.0058	0.92	0.08	
0.25	GAM	-0.018	0.0062	0.92	0.08	-0.001	0.0059	0.98	0.02	<0.001	0.0059	0.9	0.1	0.002	0.006	0.92	0.08	
	Cox-SPDE	-0.015	0.007	0.9	0.1	-0.003	0.0071	0.96	0.04	0.007	0.0089	0.96	0.04	-0.002	0.0085	0.92	0.08	
	P-SPDE	-0.015	0.0074	0.9	0.1	-0.004	0.0071	0.98	0.02	0.009	0.0098	0.96	0.04	-0.002	0.0086	0.9	0.1	

*MSE* : Erreur Quadratique Moyenne ; *CR* : Taux de couverture ; *SR* : Taux de significativité ; *Cox-PH* : Modèle de Cox à risques proportionnels ; *GAM* : Modèle Additif Généralisé ; *Cox-SPDE* : Modèle Cox-Equation aux Dérivées Partielles Stochastiques ; *P-SPDE* : Modèle Poisson-Equation aux Dérivées Partielles Stochastiques ; *RRt* : risque relatif associé au traitement ; *RRg* : Risque relatif associé au gîte.

## 1.4. Discussion

L'objectif de ce travail était de mettre en évidence l'impact de l'hétérogénéité spatiale du risque environnemental, souvent non-mesurée, sur les résultats d'un essai de prévention à partir d'une étude de simulation selon différents scénarios. Nous avons montré que malgré la randomisation, cette hétérogénéité spatiale si elle n'est pas bien prise en compte, pouvait conduire à une sous-estimation de l'effet traitement avec un CR pouvant être très faible dans certaines situations. Dans cette étude de simulation, les conditions amenant aux résultats les plus biaisés étaient un fort effet du risque environnemental (effet gîte dans notre application) associé à une forte densité du risque environnemental (densité de gîtes dans notre application). Ce résultat était attendu, puisque le risque environnemental varie en fonction de la localisation des individus et en fonction de facteurs environnementaux [57]. Les modèles prenant en compte la localisation des individus comme « proxy » de la variation du risque environnemental, notamment le modèle P-SPDE permettait de corriger ce biais.

Les limites de l'effet gîte (exprimé en Risque Relatif  $RR_g$ ) ont été fixées entre 1.05 et 3 et celles de la densité de gîtes  $D_g$  (exprimant la probabilité d'être exposé à au moins un gîte) entre 0.25 et 0.75. Le biais augmentait linéairement et était important à partir d'un  $RR_g = 1.5$ , même pour une faible densité de gîtes ( $D_g = 0.25$ ). Par contre, les scénarios présentant une forte densité de gîtes ( $D_g = 0.75$ ) et un faible effet gîte ( $RR_g = 1.05, 1.20$ ) ne montraient pas de biais nets.

De même, l'effet traitement avait un impact sur les estimations malgré la randomisation. En effet, le biais augmentait linéairement avec l'effet traitement, sauf lorsque l'effet gîte était faible ( $RR_g \leq 1.2$ ). Avec un fort effet traitement ( $RR_t \leq 0.6$ ), un fort effet gîte ( $RR_g \geq 1.5$ ) et une forte densité de gîtes ( $D_g = 0.75$ ), la sous-estimation était maximale et le CR était quasiment nul.

Ni le risque de base ni la densité de population n'ont eu un impact important sur la qualité des estimations. Le panel de valeurs choisies pour le risque de base était dans les limites observées dans notre contexte applicatif (paludisme).

Cependant, l'absence d'impact du risque de base était attendue, puisque ce risque s'appliquait uniformément à l'ensemble de la zone d'étude, n'influençant donc pas les estimations.

Dans notre contexte applicatif, le risque environnemental dépendait peu de la densité de population, mais principalement de la densité de gîtes, la maladie étant transmise par un vecteur dont le gîte larvaire est la source. Nos résultats auraient sans doute été différents si la pathologie étudiée avait été fortement dépendante de la densité de population (par exemple le choléra).

Comme attendus, aucun effet significatif n'a pu être mis en évidence dans le cadre de l'essai vaccinal utilisé en exemple applicatif. L'absence d'effet significatif du vaccin sur les accès palustres pourrait expliquer la faible différence des estimations entre les différents modèles. Cependant, les estimations des modèles allaient dans le sens que les résultats de nos simulations, puisque les effets estimés par les modèles Cox-PH et GAM étaient plus faibles que celles des modèles Cox-SPDE et P-SPDE.

Dans la littérature, l'hétérogénéité spatiale est généralement prise en compte en utilisant les modèles mixtes classiques dans les analyses de survie. Bien que ces modèles n'aient pas été élaborés explicitement pour tenir compte de l'hétérogénéité spatiale, ils peuvent saisir une partie de cette hétérogénéité spatiale en agrégeant les individus présentant le même profil de risque. L'un de ces modèles est le modèle de régression additif structuré (Structured Additive Regression, STAR) [58–61], qui tient compte à la fois de la structure de contiguïté et de la corrélation spatiale entre zones. Dans ce modèle, la zone d'étude est divisée en plusieurs subdivisions prédéfinies pour représenter l'hétérogénéité spatiale du risque environnemental, qui est alors supposé homogène dans chaque subdivision [7, 62].

Cependant, comme l'indique notre étude, deux personnes géographiquement proches peuvent présenter des risques différents selon la proximité de la source de l'exposition aux risques, ce qui signifie que l'hypothèse d'homogénéité n'est pas toujours tenable. C'est notamment le cas lorsque des subdivisions administratives sont utilisées ou lorsque la maladie étudiée est fortement liée à l'environnement telle que le paludisme, pour laquelle l'hétérogénéité spatiale à petite échelle du risque a été décrite [9]. Une subdivision plus fine tendant à obtenir l'homogénéité spatiale dans chaque subdivision induit un nombre excessif de subdivisions à inclure dans le modèle mixte entraînant ainsi une inflation du nombre de paramètres à estimer et l'estimation de la variance des trop petites subdivisions devient instable. En d'autres termes, cette approche implique un choix entre l'hypothèse d'homogénéité et le nombre de paramètres à inclure dans les modèles et la stabilité des estimations, et les résultats obtenus dépendent du choix de la forme et de la taille des subdivisions. Dans le contexte des maladies à transmission vectorielle comme le paludisme, la construction de zones à risque homogènes nécessiterait l'observation de chaque gîte larvaire, ce qui est difficilement réalisable [10].

Alors que l'objectif de notre étude n'était pas de tenir compte d'une structure spatiale particulière observée (à laquelle un modèle STAR pourrait être appliqué), mais de mettre en évidence l'impact de l'hétérogénéité spatiale sur les résultats de l'étude et de proposer différentes approches pour modéliser cette hétérogénéité lorsque les facteurs de risque spatiaux ne sont pas précisément mesurés. Bien que peu d'études aient tenu compte de l'hétérogénéité spatiale du risque environnemental, beaucoup ont examiné la variation temporelle du risque.

Une méthode utilisée pour étudier cette variation est le modèle de Cox-PH, dans lequel la durée de l'étude est stratifiée en intervalles de temps afin d'obtenir des périodes de risque homogène [63, 64]. Cependant, des approches plus continues (en particulier celles utilisant des fonctions spline) ont également été utilisées [65]. Un nombre croissant d'études utilisent le modèle GAM pour prendre en compte l'hétérogénéité spatiale du risque [44, 66].

Dans cette approche, l'hétérogénéité spatiale est modélisée par une fonction spline bivariée des coordonnées géographiques des individus [18]. Cependant, dans notre étude, le modèle GAM bien que prenant en compte une composante spatiale, montrait peu de différences en termes de qualité des estimations, avec le modèle Cox-PH non spatial. Cela peut s'expliquer par le fait que le modèle GAM estime le risque local en agrégeant les données individuelles [45, 65], alors qu'en fait, les approches de survie (comme celle que nous avons utilisée) exigent que ces données soient séparées.

De plus, même lorsque les véritables sources spatiales du risque environnemental (gîtes pour le paludisme) ne sont pas mesurées, il est connu que leur localisation diffère de celle des individus. Cette imprécision, ainsi que la méthode d'estimation des fonctions splines bivariées, peuvent expliquer les mauvais résultats obtenus dans notre étude par le modèle GAM [41, 46].

Dans cette étude de simulation, les modèles Cox-SPDE et P-SPDE modélisant l'effet spatial à l'aide d'un champ gaussien (effet spatial aléatoire) avec une fonction de covariance de Matérn avaient les meilleures performances. Le modèle P-SPDE demeurait stable et suivait de près le modèle DGM quel que soit le scénario. Cependant, le modèle Cox-SPDE était un peu sensible à la densité de population, surestimant légèrement les paramètres lorsque la densité de population était faible. Enfin, puisque la fonction Matérn est décroissante en fonction la distance, l'impact de l'effet spatial sur le temps de survie devenait très faible.

La différence de performance entre les deux modèles SPDE pour les faibles densités de population peut donc s'expliquer par le fait que le modèle P-SPDE modélisait le nombre d'événements plutôt que le temps de survie. Il convient toutefois de noter que les deux modèles ont donné de meilleurs résultats que le modèle Cox-PH classique.

Dans cette thèse, nous avons considéré un champs gaussien isotrope. En effet, dans le contexte de terrain à l'origine de notre question de recherche, le risque est plutôt spatialement isotrope en

l'absence de forte pente. Cependant, dans d'autres contextes ou pour d'autres pathologies, l'anisotropie pourrait être prise en compte en redéfinissant les coefficients de la covariance de Matérn [67]. C'est-à-dire, on permettra à ces paramètres de varier selon la localisation, donc au lieu d'avoir  $\theta_1$  et  $\theta_2$  comme paramètres spatialement constants, on aurait  $\theta_1(s)$  et  $\theta_2(s)$  qui dépendront de la localisation.

## 1.5. Conclusion

Notre étude a montré que le biais dû à l'hétérogénéité spatiale du risque environnemental n'était pas adéquatement éliminé par la randomisation. La sous-estimation de l'effet du traitement mise en évidence dans notre étude (avec des CR presque nuls dans certaines situations) peut expliquer pourquoi certains traitements ou stratégies notamment contre le paludisme finissent par être rejetés. La modélisation à travers la localisation spatiale peut réduire les biais dus à cette hétérogénéité spatiale du risque environnemental, cette dernière étant parfois difficile à mesurer. Pour ce faire, les modèles SPDE qui modélisent l'hétérogénéité spatiale avec un champ gaussien semblaient être les plus appropriés.

Ce premier travail de cette thèse a été publié dans le journal *BMC Medical Research Methodology* (couverture ci-dessous).

Cependant, dans cette partie, seule la survenue du premier événement était modélisée, les autres événements ultérieurs n'étaient pas pris en compte. Pourtant, le phénomène d'événements récurrents est très fréquent dans le domaine biomédical et peut également biaiser l'évaluation d'un essai de prévention. Dans le contexte du paludisme, un individu peut avoir plusieurs épisodes palustres durant une étude et l'immunité s'acquiert et/ou se perd en fonction du nombre d'épisodes. Cette problématique fait l'objet de la deuxième partie de cette thèse.

# Spatial heterogeneity of environmental risk in randomized prevention trials: consequences and modeling

Abdoulaye Guindo<sup>1,2\*</sup>, Issaka Sagara<sup>1,2</sup>, Boukary Ouedraogo<sup>1</sup>, Kankoe Sallah<sup>1,4</sup>, Mahamadoun Hamady Assadou<sup>2</sup>, Sara Healy<sup>5</sup>, Patrick Duffy<sup>5</sup>, Ogobara K. Doumbo<sup>1^</sup>, Alassane Dicko<sup>2</sup>, Roch Giorgi<sup>6</sup> and Jean Gaudart<sup>6</sup>

## Abstract

**Background:** In the context of environmentally influenced communicable diseases, proximity to environmental sources results in spatial heterogeneity of risk, which is sometimes difficult to measure in the field. Most prevention trials use randomization to achieve comparability between groups, thus failing to account for heterogeneity. This study aimed to determine under what conditions spatial heterogeneity biases the results of randomized prevention trials, and to compare different approaches to modeling this heterogeneity.

**Methods:** Using the example of a malaria prevention trial, simulations were performed to quantify the impact of spatial heterogeneity and to compare different models.

Simulated scenarios combined variation in baseline risk, a continuous protective factor (age), a non-related factor (sex), and a binary protective factor (preventive treatment). Simulated spatial heterogeneity scenarios combined variation in breeding site density and effect, location, and population density.

The performances of the following five statistical models were assessed: a non-spatial Cox Proportional Hazard (Cox-PH) model and four models accounting for spatial heterogeneity—i.e., a Data-Generating Model, a Generalized Additive Model (GAM), and two Stochastic Partial Differential Equation (SPDE) models, one modeling survival time and the other the number of events. Using a Bayesian approach, we estimated the SPDE models with an Integrated Nested Laplace Approximation algorithm.

For each factor (age, sex, treatment), model performances were assessed by quantifying parameter estimation biases, mean square errors, confidence interval coverage rates (CRS), and significance rates. The four models were applied to data from a malaria transmission blocking vaccine candidate.

**Results:** The level of baseline risk did not affect our estimates. However, with a high breeding site density and a strong breeding site effect, the Cox-PH and GAM models underestimated the age and treatment effects (but not the sex effect) with a low CR.

When population density was low, the Cox-SPDE model slightly overestimated the effect of related factors (age, treatment). The two SPDE models corrected the impact of spatial heterogeneity, thus providing the best estimates.

**Conclusion:** Our results show that when spatial heterogeneity is important but not measured, randomization alone cannot achieve comparability between groups. In such cases, prevention trials should model spatial heterogeneity with an adapted method.

Trial registration: The dataset used for the application example was extracted from Vaccine Trial #NCT02334462 (ClinicalTrials.gov registry).

**Keywords:** Randomized prevention trials, Spatial heterogeneity, Stochastic Partial Differential Equation, Integrated Nested Laplace Approximation, Environmental factors

\* Correspondence: [abdoulaye.guindo@etu.univ-amu.fr](mailto:abdoulaye.guindo@etu.univ-amu.fr)

<sup>^</sup>Deceased

Aix Marseille Univ, NSERM, RD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'information Médicale, Marseille, France <sup>2</sup>Malaria Research and Training Center-Ogobara K Doumbo, FMOS-FAPH, Mali-NIAID-CER, Université des Sciences, des Techniques et des Technologies de Bamako, Mali

Full list of author information is available at the end of the article



© The Author(s). 2019 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0

BMC International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

## **Partie II**

### **Modélisation de l'hétérogénéité spatiale dans le contexte d'évènements récurrents :**

#### **Application aux essais de prévention contre le paludisme.**

##### **2.1. Contexte**

Dans le contexte de la modélisation de l'hétérogénéité spatiale du risque dans un cadre d'analyses de survie, les auteurs s'intéressent souvent uniquement au premier évènement, ignorant ainsi les éventuels évènements ultérieurs. Cependant, le phénomène d'évènements récurrents est très fréquent dans le domaine biomédical où l'évènement étudié se répète chez le même individu à des instants différents. Ces évènements peuvent être corrélés entre eux [68, 69], soit parce que les mêmes facteurs de risques persistent soit parce que l'évènement modifie la probabilité de l'évènement suivant. Par exemple, dans le contexte du paludisme, les individus sont soumis aux mêmes facteurs de risque environnementaux. Par ailleurs, ils acquièrent une prémunition (immunité clinique partielle et labile) au fil du temps à cause des multiples infections [70–72]. Cette prémunition se perd si l'exposition disparaît [73, 74]. De ce fait, les épisodes successifs d'accès palustres sont corrélés entre eux. Or, certains auteurs ont montré que cette corrélation entre les évènements successifs chez un même individu pouvait constituer un biais en l'absence de prise en compte. Les extensions du modèle de Cox notamment le modèle d'Andersen-Gill et le modèle de fragilité étaient mieux adaptés pour prendre en compte cette corrélation intra-individu dans un contexte de survie [75]. Cependant, ces modèles ne prennent pas en compte spécifiquement l'hétérogénéité spatiale du risque.

Dans la littérature, les deux problématiques ont toujours été résolues séparément lors de l'analyse des essais de prévention. De nombreux travaux modélisaient uniquement les évènements récurrents [75–78] ou uniquement l'hétérogénéité spatiale d'une exposition [7, 50, 62, 79–81] mais à notre connaissance, aucun auteur ne combinait ces deux problématiques dans une seule analyse.

Les approches notamment bayésiennes peuvent servir à modéliser conjointement l'effet spatial à l'aide d'un champ gaussien et la corrélation intra-individuelle en utilisant une fragilité individuelle prenant en compte l'ordre des événements [82–84].

L'objectif de travail dans un premier temps était de déterminer si la prise en compte de l'hétérogénéité spatiale dans le contexte d'analyse d'événements récurrents permettait d'améliorer la précision des estimations et dans un deuxième temps comparer différentes approches. Nous avons réanalysé un essai de prévention en évaluant l'effet d'un ajout de l'azithromycine à la Chimio-prévention du Paludisme Saisonnier (CPS) sur les accès palustres, à Bougouni, Mali. Différentes approches prenant en compte simultanément l'effet spatial et événements récurrents étaient comparé aux modèles classiques d'analyses d'événements récurrents (modèle d'Andersen-Gill et modèle de fragilité).

De même, l'efficacité des moyens de prévention contre le paludisme dans la ville de Bandiagara (Mali) a été également évaluée afin de comparer l'estimation des différentes approches.

## **2.2. Méthodes**

Ce travail est une réanalyse des données issues de l'essai de prévention de Bougouni visant à tester l'effet d'un ajout de l'azithromycine à la CPS sur la mortalité et la morbidité globale chez les enfants âgés de 3 à 59 mois. Les résultats principaux ont été déjà publiés [85]. Cette étude a été menée dans le district sanitaire de Bougouni (Mali) et de Houndé (Burkina) de juin 2014 à juin 2017. Dans le présent travail, nous n'avons évalué que les données du Mali.

En ce qui concerne les données de la ville Bandiagara, l'étude a été menée de juin 2009 à janvier 2015, incluant environ 300 enfants. Une partie de ces données a déjà été publié [86–88].

## **2.2.1. Structuration des données**

### **2.2.1.1. Etude de Bougouni (effet de l'azithromycine sur le paludisme)**

Les enfants de l'étude, suivi pendant trois ans, étaient repartis entre deux bras de traitement, l'un recevant la CPS accompagné du placebo et l'autre recevant la CPS accompagné de l'azithromycine pendant la période de haute transmission du paludisme avec un mois d'intervalle (Août, septembre, octobre et novembre). Lorsque qu'un enfant dépassait les 59 mois, il ne recevait plus les produits de l'étude mais était suivi et pris en charge en cas de maladie. L'étude étant randomisée et double aveugle, les agents se contentaient d'administrer les produits aux enfants selon leur bras de traitement sans savoir lequel contenaient de l'azithromycine. Pour plus de détails sur le design et la collecte des données, l'article principale de cette étude est déjà publié et accessible [85].

Seuls les enfants ayant reçu les quatre doses prévues par an pendant les trois années de suivis ont été pris en compte dans ce travail (analyse Per Protocole). En effet, certains enfants du bras expérimental sensé avoir les doses d'azithromycine, n'ont rien reçu (enfant malade, absent, sous traitement antipaludique au moment du passage, etc.), ils n'ont pas été impliqués dans ce travail.

Les accès palustres étaient détectés soit au moment de l'administration des produits de l'études ou soit courant le suivi de morbidité (TDR positifs et quelques fois confirmé par gouttes épaisses). En effet, tous les enfants de l'étude étaient pris en charge gratuitement pour n'importe quelle pathologie dans tous les centres de santé partenaires de l'étude ou alors les frais de soins étaient remboursés. Cela a permis de détecter le plus grand nombre de cas de paludisme chez les enfants inclus dans l'étude.

En plus du facteur traitement, nous avons inclut dans l'analyse le facteur âge (mois). L'âge a été calculé au moment de l'évènement, cela a permis de prendre en compte l'évolution de l'âge des enfants dans le temps au cours des trois années de suivis. En effet, le risque palustre diminue avec l'âge [75, 86] dans les zones de forte endémie.

A chaque évènement (infection clinique du paludisme), les enfants bénéficiaient d'un traitement par l'artemether lumefantrine, ce qui leurs confère une protection de deux à trois semaines contre une nouvelle infection. Ce délai a été pris en compte lors du calcul du temps de survie d'une infection palustre en laissant quatorze jours comme temps de non-risque après chaque évènement.

### **2.2.1.2. Etude de Bandiagara (moyens de protection contre le paludisme)**

Les données de la cohorte de Bandiagara ont été structurées identiquement à celles de Bougouni. L'évènement d'intérêt étaient l'infection du paludisme, symptomatique ou asymptomatique. En effet, des visites mensuelles étaient programmées pour détecter l'infection parasitaire (Goutte épaisse et TDR). Parallèlement, une permanence des agents de santé était assurée dans les centres de santé pour accueillir les enfants enrôlés dans cette étude en cas maladies, en dehors de ces visites programmées.

Les moyens de protection étaient définis par l'utilisation des moustiquaires, de serpentins, de fumigation ou d'autres. Une variable binaire a donc été créée selon l'utilisation ou non d'au moins un de ces moyens de protection.

Dans cette analyse, les enfants manquant l'une de ces informations ont été exclus (localisation exacte, moyens de protection, âge, confirmation ou non de paludisme (GE ou TDR) à chaque visite, date de visite et date d'inclusion). Sur environ 300 enfants de l'étude principale, 280 enfants étaient inclus dans cette analyse.

### **2.2.2. Modèles statistiques**

Cinq approches modélisant les évènements récurrents ont été comparées : Deux approches non spatialisées (modèle d'Andersen-Gill et le modèle de fragilité) et 3 approches spatialisées de type SPDE (modèle SPDE avec évènements récurrents à corrélation non structurées, modèle SPDE

avec évènements récurrents à corrélation autorégressives, modèle SPDE avec évènements récurrents à corrélation sigmoïdale).

### 2.2.2.1. Modèle d'Andersen-Gill (Modèle AG)

Le modèle d'Andersen-Gill est l'un des modèles d'évènements récurrents des plus populaires. Introduit par Andersen P. K. et Gill R. D. en 1982 [89], il est issu d'une extension du modèle de Cox [90–93].

Le modèle s'écrit globalement comme suit :

$$\lambda_{ij}(t|\beta, X) = \mathbb{1}_{ij}(t)\lambda_0(t) \exp(\beta X_{ij})$$

$\lambda_{ij}(\cdot)$  est la fonction de risque instantané pour le  $j^{ième}$  évènement de l'individu  $i$ ,  $\lambda_0(\cdot)$  est le risque de base commun à tous les individus à chaque évènement,  $\mathbb{1}_{ij}(\cdot)$  l'indicateur de  $j^{ième}$  évènement chez l'individu  $i$  (1 il y a évènement et 0 sinon),  $X_{ij}$  est l'observation des covariables  $X$  au  $j^{ième}$  évènement de l'individu  $i$  et  $\beta$  est l'effet associé aux  $X_{ij}$ .

### 2.2.2.2. Modèle de fragilité

Développé pour prendre en compte un effet aléatoire latent ou pour les analyses des données groupées, le modèle de fragilité est également utilisé dans la modélisation d'évènements récurrents [94–96]. Il nécessite l'ajoute d'un effet aléatoire individuel pour prendre en compte la corrélation intra-individu. En fait, ce modèle suppose que le risque de base dépend de l'ordre de l'évènement, de ce fait, il définit un risque de base pour chaque évènement selon son ordre dans la succession. Théoriquement, le modèle s'écrit comme suit :

$$\lambda_{ij}(t|\alpha, \beta, X) = \lambda_0(t) \exp(\beta X_{ij} + \alpha_i)$$

$\lambda_{ij}(\cdot)$  est la fonction de risque instantané pour le  $j^{ième}$  évènement de l'individu  $i$ ,  $\lambda_0(\cdot)$  est le risque de base commun à tous les individus à chaque évènement,  $\alpha_i$  est l'effet aléatoire lié à

l'individu  $i$ ,  $X_{ij}$  est l'observation des covariables  $X$  au  $j^{\text{ième}}$  évènement de l'individu  $i$  et  $\beta$  est l'effet associé aux  $X_{ij}$ .

### 2.2.2.3. Modèle SPDE.

Le modèle SPDE est de plus en plus utilisé pour modéliser les structures spatiales, en utilisant la méthode INLA pour l'estimation [82–84, 97, 98].

Cette approche permet de modéliser l'hétérogénéité spatiale à partir, notamment, d'un champ gaussien dont la matrice de covariance est la fonction de Matérn [79].

C'est un modèle de type bayésien qui s'écrit globalement comme suit :

$$T | \beta, X, Z \sim \mathcal{W}(T | \mu, \phi)$$

$$Z \sim GF(0, \Sigma(\theta_1, \theta_2))$$

$$\beta, \mu, \phi, \theta_1, \theta_2 \sim N(0, 0.001)$$

$T$  est le temps de survie de loi  $\mathcal{W}$ ,  $X$  est le vecteur des covariables (effets fixes),  $Z$  est un champ gaussien ( $GF$ ) dont la fonction de covariance  $\Sigma$  (effet spatial aléatoire) et  $(\mu, \phi)$  est respectivement le paramètre d'échelle et de forme de la loi de Weibull (loi de la survie  $T$ ).

Dans notre, nous ajoutons à modèle ci-dessus, une partie permettant de prendre en compte l'effet des évènements récurrents. On obtient donc un modèle plus complet qui s'écrit comme suit

$$T | \beta, X, Z, V \sim \mathcal{W}(T | \mu, \phi)$$

$$Z \sim GF(0, \Sigma)$$

$$V \sim \mathcal{L}(\omega)$$

$$\beta, \mu, \phi, \theta_1, \theta_2, \omega \sim N(0, 0.001)$$

$V$  est l'ordre dans la succession des évènements chez les individus,  $\mathcal{L}$  est sa loi de paramètre  $\omega$ , les autres éléments du modèles étant définis plus haut.

De façon plus détaillé, en notant  $\lambda_{ij}(t|\cdot)$  la fonction de risque instantané associé à l'individu  $i$  au  $j^{\text{ème}}$  évènement, le modèle ci-dessus s'écrit explicitement comme suit :

$$T | \beta, X, Z, V \sim \mathcal{W}(T | \mu, \phi)$$

$$\lambda_{ij}(t | X, Z, V) = \lambda_0(t) \exp(\beta X + Z_i + V_{ij})$$

$\lambda_0(\cdot)$  est le risque de base ;  $X$  est le vecteur des covariables ;  $\beta$  est le vecteur des effets associés aux covariables ;  $Z$  est l'effet spatial aléatoire (champ gaussien) ;  $V$  est la fragilité individuelle lié la corrélation intra-individuelle.

Pour bien mettre en évidence l'impact de l'effet spatial, différentes modélisations d'évènements récurrents combinées au modèle SPDE avait été utilisé afin identifier la plus appropriée : une approche supposant l'existence d'une corrélation entre les évènements successifs chez un individu sans prendre en compte sa nature (non structuré), une deuxième approche supposant une corrélation du type autorégressive d'ordre 1 entre les évènements successifs. La différence entre ces 2 approches ne se situe pas au niveau de la modélisation de l'hétérogénéité spatiale, qui par ailleurs est le SPDE pour les 2, mais plutôt au niveau de la modélisation d'évènements récurrents.

#### **2.2.2.3.1. Modèle SPDE avec effets récurrents non structuré (SPDE-NS)**

Cette approche suppose qu'il y a une corrélation entre les évènements successifs chez un même individu mais sans modélisation de sa nature, et sans tenir compte de l'ordre de survenue. Dans ce cas, les estimations sont seulement ajustées en tenant compte de la corrélation intra-individuelle considéré comme un paramètre de nuisance supposé constant dans le temps. C'est-à-dire, le terme de fragilité  $V_{ij} = \psi$ ,  $\psi$  est un paramètre de nuisance.

### 2.2.2.3.2. Modèle SPDE avec effet récurrent autorégressif (SPDE-AR)

Cette approche, avec une hypothèse plus forte que la première, suppose que les évènements successifs sont corrélés entre eux avec une structure autorégressive d'ordre 1. Elle prend en compte l'ordre des évènements contrairement à la première approche en supposant une liaison linéaire entre les épisodes successifs [99, 100]. Cette fragilité individuelle autorégressive était définie dans ce cas par un paramètre  $\omega$  tel que  $V_{ij} = \omega V_{i(j-1)}$

### 2.2.2.3.3. Modèle SPDE avec effets récurrent sigmoïdal (SPDE-S)

Dans cette approche, comme la précédente, on suppose qu'il y a une corrélation entre les évènements successifs dont la structure à la forme d'une sigmoïde contrôlée par trois paramètres [101]. Il y a donc une structure liant les événements successifs chez un même individu, une structure définie par une fonction sigmoïdale à trois paramètres  $\omega = (\omega_1, \omega_2, \omega_3)$  telle que :

$$S(i) = cov(V_{ij}, V_{ij+1}) = \frac{\omega_1}{1 + \exp(\omega_2 V_{ij})} - \omega_3$$

$\omega_1$  est le paramètre contrôlant la portée de la fonction de covariance  $S$ ,  $\omega_2$  est le paramètre contrôlant sa pente et  $\omega_3$  est le paramètre contrôlant sa symétrie.

## 2.2.3. Indicateurs de performances de modèles

Les modèles ont été comparés selon différents critères de performances notamment l'erreur standard, la largeur de l'intervalle de confiance (intervalle de crédibilité pour les modèles bayésiens) à 95%, la valeur du log vraisemblance, l'AIC et son équivalent bayésien le DIC [102].

## 2.3. Résultats

### 2.3.1. Evaluation de l'effet l'azithromycine pour la prévention du paludisme à Bougouni (Mali)

Au total 1730 enfants de 3 à 59 mois ont été inclus dans ce travail, ce qui correspond au nombre d'enfants qui ont effectivement participé à l'étude pendant les trois années de suivi et qui ont pris toutes les doses de médicament, contenant l'azithromycine ou non selon leur bras d'appartenance. Certains enfants (89) avaient dépassé l'âge au cours de l'étude mais ont été suivi et inclus dans l'analyse, l'âge étant pris en compte. La taille des deux bras de traitement était équilibrée avec respectivement 864 enfants dans le bras d'intervention et 866 dans le bras de contrôle. L'effet de l'azithromycine (exprimé en Risque Relatif, RR) n'était pas significatif au risque de 5% pour les modèles non spatialisés (0.909 [0.825 ; 1.002] pour le modèle AG et 0.914 [0.828 ; 1.010] pour le modèle de fragilité).

Par contre, pour les modèles SPDE cet effet, bien que faible, était significatif au même risque de 5% avec un RR [IC] de 0.910 [0.853 ; 0.995], 0.921 [0.853 ; 0.995] pour le modèle SPDE-AR et 0.919 [0.851 ; 0.993] pour le modèle SPDE-NS). L'erreur standard était autour de 0.039 pour tous ces modèles. Tous les modèles ont montré un effet significatif de l'âge (en mois) autour de 0.97. Tous les modèles avaient la même erreur standard concernant l'effet âge (0.002) (**Tableau 4**).

**Tableau 4** : Estimations de l'effet de l'azithromycine et l'âge sur le paludisme avec l'erreur standard associée.

Modèles	Azytromicine		Age (en mois)	
	RR [95% IC]	SD	RR [95% IC]	SD
Model AG	0.909 [0.825 ; 1.002]	0.038	0.970 [0.967 ; 0.974]	0.002
Model Fragilité	0.914 [0.828 ; 1.010]	0.039	0.960 [0.956 ; 0.964]	0.002
Model SPDE-NS	<b>0.919 [0.851 ; 0.993]</b>	0.039	0.974 [0.971 ; 0.977]	0.002
Model SPDE-AR	<b>0.921 [0.853 ; 0.995]</b>	0.039	0.974 [0.971 ; 0.977]	0.002
Model SPDE-S	<b>0.910 [0.844, 0.983]</b>	0.039	0.974 [0.972 ; 0.978]	0.002

**Model AG**: modèle d'Andersen-Gill, **Model Fragilité**: Modèle de fragilité, **Model SPDE-NS**: Modèle SPDE à récurrence non structuré, **Model SPDE-AR**: Modèle SPDE à recurrence Autorégressive, **Model SPDE-S**: Modèle SPDE à récurrence sigmoïdale, **SD**: Standard Deviation, **RR**: Risque Relatif, **IC** : Intervalle de confiance (ou de crédibilité pour les 2 derniers modèles), les valeurs en **GRAS** (fond gris) montre la significativité de l'azithromycine pour les modèles spatiaux.

La valeur de la log-vraisemblance (log-vraisemblance marginale pour les modèles bayésiens) et la valeur de l'AIC (ou DIC pour les modèles bayésiens) étaient très différentes entre les modèles SPDE et les modèles non spatialisés, avec des valeurs plus favorables pour les modèles SPDE. De même, la log-vraisemblance pour les modèles non spatialisés était très faible par rapport à son homologue bayésien (log-vraisemblance marginale) pour les deux modèles SPDE (*Tableau 5*)

**Tableau 5:** Indicateurs de performance de modèles pour l'évaluation de l'effet de l'azithromycine contre le paludisme

Modèles	AIC / DIC	Log-vraisemblance / marginale
<b>Model AG</b>	39 916.808	-19 956.404
<b>Model Fragilité</b>	39 576.384	-19 117.442
<b>Model SPDE-SN</b>	6 622.852	-3376.696
<b>Model SPDE-AR</b>	6 619.048	-3 374.027
<b>Model SPDE-S</b>	6690.454	-3364.514

*Model AG* : modèle d'Andersen-Gill, *Model Fragilité* : Modèle de fragilité, *Model SPDE-NS* : Modèle SPDE à récurrence non structuré, *Model SPDE-AR* : Modèle SPDE à récurrence Autorégressive, *Model SPDE-S* : Modèle SPDE à récurrence sigmoïdale, *SD* : Standard Deviation, *AIC* : Akaike Information Criterion, *DIC* : Deviance Information Criterion, **fond bleu** : modèles spatiaux, **fond blanc** : modèles non spatiaux.

### 2.3.2. Evaluation des moyens de prévention contre le paludisme à Bandiagara (Mali)

Pour l'évaluation des moyens de protection contre le paludisme, au total 280 enfants ont été suivis de 2009 à 2014, correspondant au nombre d'enfants géoréférencés pour lesquels nous savons avec précision la charge parasitaire à chaque visite, ainsi le moyen de protection utilisé. L'âge moyen des enfants était de 54 mois avec un intervalle de confiance à 95% de [51 ; 57].

Les modèles non spatialisés ont montré un effet (RR [IC]) significatif des moyens de protection utilisés à Bandiagara au risque de 5% (0.689 [0.525 ; 0.904] pour le modèle AG et 0.729 [0.554 ; 0.961] pour le modèle de fragilité). De même, le modèle SPDE-S a montré un effet significatif avec un risque relatif 0.77 [0.60 ; 0.99]. Ainsi, le modèle SPDE-AR a montré également un effet (RR [IC]) significatif mais légèrement plus faible que les modèles non spatialisés (0.762 [0.593 ;

0.991) pendant que le SPDE-NS, bien que spatialisé, montrait un effet non significatif des moyens de protection à Bandiagara (0.772 [0.602 ; 1.003]). Tous les modèles présentaient une erreur standard semblable autour de 0.130 (**Tableau 6**).

Pour le facteur âge, tous les modèles présentaient des estimations très proches autour d'un risque relatif de 0.99 avec une erreur standard identique pour tous (0.001), sauf le modèle de fragilité qui estimait un effet âge avec un risque relatif de 0.96 et une erreur standard de 0.002.

**Tableau 6** : Estimations de l'effet des moyens de protection et l'âge sur le paludisme avec l'erreur standard associée

Modèles	Moyen de protection		Age (en mois)	
	RR [95% IC]	SD	RR [95% IC]	SD
Model AG	0.689 [0.525 ; 0.904]	0.127	0.992 [0.989 ; 0.995]	0.001
Model Fragilité	0.729 [0.554 ; 0.961]	0.138	0.961 [0.956 ; 0.965]	0.002
Model SPDE-NS	<b>0.772 [0.602 ; 1.003]</b>	0.130	0.990 [0.988 ; 0.992]	0.001
Model SPDE-AR	<b>0.762 [0.593 ; 0.991]</b>	0.131	0.989 [0.987 ; 0.991]	0.001
Model SPDE-S	<b>0.770 [0.600 ; 0.994]</b>	<b>0.129</b>	<b>0.992[0.991 ; 0.993]</b>	<b>0.002</b>

**Model AG**: modèle d'Andersen-Gill, **Model Fragilité**: Modèle de fragilité, **Model SPDE-NS**: Modèle SPDE à récurrence non structuré, **Model SPDE-AR**: Modèle SPDE à récurrence Autorégressive, **Model SPDE-S**: Modèle SPDE à récurrence sigmoïdale, **SD**: Standard Deviation, **RR**: Risque Relatif, **IC** : Intervalle de confiance (ou de crédibilité pour les 2 derniers modèles), les valeurs en **GRAS** (fond gris) montre la sous-estimation de l'effet de la protection pour les modèles spatiaux.

Les modèles non spatialisés présentaient un AIC largement plus élevé et une vraisemblance largement plus basse que les modèles spatialisés (**Tableau 7**).

**Tableau 7**: Indicateurs de performance de modèles pour l'évaluation de l'efficacité des moyens de protection contre les moustiques.

Modèles	AIC (DIC)	Log-vraisemblance (marginale)
<b>Model AG</b>	22 149.8	-11 072.9
<b>Model Fragilité</b>	21 533.1	-10 537.7
<b>Model SPDE-SN</b>	2 206.5	-1 171.3
<b>Model SPDE-AR</b>	2 196.9	-1 160.8
<b>Model SPDE-S</b>	2254.436	-1151.746

**Model AG**: modèle d'Andersen-Gill, **Model Fragilité**: Modèle de fragilité, **Model SPDE-NS**: Modèle SPDE à récurrence non structuré, **Model SPDE-AR**: Modèle SPDE à récurrence Autorégressive, **Model SPDE-S**: Modèle SPDE à récurrence sigmoïdale, **SD**: Standard Deviation, **AIC**: Akaike Information Criterion, **DIC**: Deviance Information Criterion, fond bleu : modèles spatiaux, fond blanc : modèles non spatiaux.

## 2.4. Discussion

L'objectif ce travail était de réanalyser 2 études de prévention en tenant compte conjointement de l'hétérogénéité spatiale du risque et de la récurrence des événements. Les deux modèles SPDE ont montré des performances similaires bien en termes d'intervalle de crédibilité qu'en termes de DIC même si, très légèrement, le modèle SPDE-AR semblait plus adéquat. Cela peut s'expliquer peut-être par la façon de modéliser les événements récurrents (le premier impliquant une structure précise d'auto-régression dans la succession des événements tandis que le second suppose juste une corrélation sans modélisation de sa nature).

Le modèle de fragilité était sensiblement intermédiaire entre le modèle AG non spatialisé et les modèles SPDE spatiaux en termes d'intervalle de confiance et d'estimation de l'effet de l'azithromycine mais pas pour l'effet âge. Il est à noter que le modèle de fragilité implique également un terme de fragilité individuel, terme pouvant capturer une partie de l'effet spatial individuel non modélisé. Cette grande différence entre l'AIC des modèles non spatialisés et son homologue DIC des modèles spatialisés pourrait s'expliquer par le fait du nombre élevé de paramètre des modèles spatiaux. Cette même raison pourrait expliquer également la différence entre la vraisemblance des modèles non spatialisés et la vraisemblance marginale des modèles spatiaux.

Dans la littérature, nous n'avons pas connaissance d'une étude qui associait ces deux problématiques dans en une seule analyse. Nous avons montré à travers l'analyse de l'efficacité de l'azithromycine sur le paludisme clinique que la prise en compte de l'effet spatial pouvait entraîner un gain de puissance dans les estimations. Ainsi, nous avons montré que l'azithromycine associé à la CPS avait un effet bénéfique significatif sur le paludisme clinique, ce qui n'était pas le cas pour les analyses non spatialisées. Les modèles d'événements récurrents non spatialisés l'effet de l'azithromycine (0.91 [0.825 ; 1.01]) avec moins de puissance mais restaient à la limite

de la significativité. La prise en compte de la spatialisation en plus de la récurrence des infections estimait un effet de l'azithromycine similaire mais la puissance (0.92 [0.85 ; 0.99]) par rapport à l'analyse initiale. Il faut noter que l'analyse initiale impliquant plus d'enfants (Mali et Burkina, au moins 9800 enfants pour chaque pays) avait trouvé l'effet de l'azithromycine légèrement moins élevé tout en restant à la limite de significativité (0.97 [0.93 ; 1.00]) [85]. Mais les résultats de notre travail n'étaient pas directement comparable à l'analyse initiale à cause de critère de sélection des enfants, en effet, l'analyse initiale impliquait plus d'enfants avec moins de restriction (9800 enfants dans chaque pays), alors que ce travail impliquait uniquement les enfants du Mali (1700 enfants qui ont reçus les 12 doses courant les 3 années de suivis ont été inclus dans cette analyse). La puissance de l'analyse initiale aurait dû ainsi être plus grande. La modélisation de l'effet spatial et de la récurrence des évènements a dû, tout de même, permit d'avoir une puissance suffisante.

Cependant, cet effet significatif doit être interpréter avec prudence : le risque relatif estimé restait toujours proche de 1, la fraction préventive (1-RR, à vérifier) étant de 9%. Cette faible fraction préventive est compréhensible car, dans cet essai, l'azithromycine s'ajoute à l'AQ+SP (à vérifier) qui a déjà fait preuve de son efficacité dans la prévention du paludisme. Evidemment, d'autres études à plus petites échelles (grandes zones géographiques) sont nécessaires avant l'inclusion d'azithromycine dans une politique préventive.

De façon plus générale, dans la littérature, certains auteurs trouvaient une lente action prolongée des antibiotiques, y compris l'azithromycine contre les parasites du paludisme [103, 104]. Ce qui se confirmait avec certaines études qui ont montré un effet bénéfique de l'azithromycine sur le paludisme notamment clinique [105–110]. Par contre, d'autres auteurs, suggérant une étude plus poussé étaient également optimistes pour l'utilisation future de l'azithromycine en association avec les anti-palustres existants [85, 111–114]. On peut cependant craindre l'apparition d'une résistance bactérienne à cet antibiotique en cas d'utilisation répétées et prolongées en population.

Ce gain en puissance venant de la modélisation de l'hétérogénéité spatiale n'avait pas été observé pour les données de l'évaluation des moyens de protection contre le paludisme à Bandiagara. Nous avons observé une réduction de l'estimation de l'efficacité et une baisse de puissance pour le modèle SPDE-NS pour qui l'effet recherché n'était pas significatif. Cela pourrait s'expliquer par le fait que le recrutement des enfants de l'étude n'avait pas été fait en tenant compte de leur localisation. La répartition géographique des enfants était ainsi hétérogène, avec des zones creuses sans aucun enfant recruté. De ce fait, la structuration spatiale du risque a pu être sous-estimée par manque d'observation. Il y a aussi le manque de puissance (la faible taille de l'échantillon, 280 enfants) qui pourrait expliquer le fait que les modèles spatiaux ne modifiaient pas les résultats. Peut-être aussi que le risque environnemental à Bandiagara était homogène, auquel cas, la modélisation spatiale modifierait peu les estimations.

## **2.5. Conclusion**

Ce travail a permis de montrer que dans un essai de prévention randomisé où la maladie étudiée pouvait infecter plusieurs fois la même personne et sa transmission est liée à l'environnement, il est plus convenable d'associer une analyse spatiale à une analyse d'évènements récurrents. En effet, la modélisation spatiale impliquait au moins un gain en puissance, c'est dire, une augmentation de chance de détecter des produits potentiellement efficaces, notamment dans le contexte du paludisme, avec l'apparition inquiétante, en Asie, de résistances parasitaires contre les produits existants. Dans ce travail, la prise en compte l'hétérogénéité spatiale du risque n'a pas mis en évidence une modification de l'estimation ponctuelle de l'efficacité du traitement, mais un gain en puissance. Ce qui a permis de mettre en évidence une significativité de l'azithromycine contre le paludisme clinique que les modèles classiques non spatialisés d'évènement récurrents n'ont pas pu montrer. Par contre, dans l'étude de Bandiagara, la modélisation spatiale n'a pas mis en évidence une modification des estimations, pouvant être due, au moins en partie, aux modalités d'inclusion des enfants. Il est donc recommandé de prévoir un recrutement spatialement homogène pour

couvrir convenablement la zone de l'étude lors des essais de prévention contre les maladies liées à l'environnement notamment le paludisme.

## **Partie III**

### **Guide d'utilisation de la méthode INLA**

#### **3.1. Bref aperçu théorique sur la méthode INLA**

Grâce aux logiciels statistiques libres dotés de langage de programmation tels que le logiciel R ou encore le logiciel Python avec l'avènement des machines de grandes capacités de calculs intensifs de plus en plus performantes, les approches bayésiennes ont connu une rapide avancée depuis les deux dernières décennies. L'une des méthodes les plus populaires de l'estimation bayésienne est la méthode de Monte Carlo par Chaîne de Markov (MCMC) à cause de sa flexibilité et de sa précision avec de bonnes propriétés de convergences [115]. La méthode MCMC est très générale et peut être appliquée à une large gamme de modèles et implémentée dans plusieurs logiciels statistiques, par exemple les logiciels WinGUGS [116, 117], R [118–120], etc.

Cependant, dans la pratique, cette précision appréciable de la méthode MCMC est à moduler avec la complexité des modèles et le temps d'exécution jusqu'à la convergence. La gestion du temps d'exécution des modèles représente particulièrement un obstacle lorsqu'il s'agit de grandes bases de données complexes. En plus, cette méthode comporte une étape importante du choix de valeurs initiales des paramètres à estimer, qui peuvent retarder la convergence si elles étaient mal choisies.

C'est dans ce contexte que la méthode d'approximation de Laplace emboîtée par intégration (INLA, Integrated Nested Laplace Approximation) est apparue comme une alternative permettant d'estimer ces modèles dans un temps beaucoup moindre [121]. C'est une méthode très générale applicable à presque tous les modèles bayésiens pourvu qu'ils soient théoriquement bien formulés. L'un des avantages de la méthode INLA est l'absence de valeurs initiales nécessaires, qui peut parfois, entraver la convergence de l'algorithme MCMC. On peut relever

également que la gamme de lois disponibles est plus riche pour le MCMC que pour le INLA qui est encore en construction mais l'estimation est généralement plus précise pour cette dernière.

Son nom en dit long sur la performance de cette méthode :

- Approximation de Laplace : parce que l'on utilise l'**Approximation de Laplace** pour déterminer les distributions inconnues marginales cibles.
- Emboîtée (**Nested**) : Parce que les approximations de Laplace sont **Emboîtées** les unes dans les autres (la  $n^{\text{ième}}$  valeur est nécessaire pour l'estimation de la  $(n+1)^{\text{ième}}$  valeur)
- Intégration : Parce que les distributions marginales a posteriori recherchées sont obtenues par **Intégration** (numérique) de l'approximation de Laplace.

Notons  $g(x)$  la fonction de distribution marginale a posteriori que l'on cherche à déterminer.

L'idée principale est d'approximer  $\log(g(x))$  en utilisant une fonction quadratique de son développement de Taylor autour de la mode  $\hat{x}$  :

- Selon le développement de Taylor d'ordre 2 de  $\log(g(x))$  autour de  $\hat{x}$ , nous obtenons :

$$\log g(x) \approx \log g(\hat{x}) + \frac{\partial \log g(\hat{x})}{\partial x} (x - \hat{x}) + \frac{1}{2} \frac{\partial^2 \log g(\hat{x})}{\partial x^2} (x - \hat{x})^2$$

Puisque la dérivée d'une constante est nulle, nous avons :

$$\frac{\partial \log g(\hat{x})}{\partial x} = 0 \quad \text{car } \log g(\hat{x}) \text{ est une constante par rapport à } x$$

Donc, l'expression ci-dessus se réduit à :

$$\log(g(x)) = \log g(\hat{x}) + \frac{1}{2} \frac{\partial^2 \log g(\hat{x})}{\partial x^2} (x - \hat{x})^2$$

En résolvant l'équation  $\frac{\partial \log g(\hat{x})}{\partial x} = 0$ , nous obtenons facilement :

$$\hat{\sigma}^2 = - \frac{1}{\frac{\partial^2 \log g(\hat{x})}{\partial x^2}}$$

On a alors,  $\log g(x) \approx \log g(\hat{x}) + \frac{1}{2\hat{\sigma}^2} (x - \hat{x})^2$

En passant l'intégration de part et d'autre de l'expression ci-dessus, nous avons :

$$\int g(x) = \int \exp(\log g(x)) \approx \text{Const} \times \int \exp \left\{ -\frac{1}{2\hat{\sigma}^2} (x - \hat{x})^2 \right\}$$

Ainsi, en vertu de l'approximation de Laplace, nous obtenons l'approximation suivante:

$$g(x) \approx N(\hat{x}, \hat{\sigma}^2)$$

Par conséquent, la distribution marginale *a posteriori* de n'importe quelle variable peut être approximé à l'aide de la loi normale en utilisant la méthode INLA.

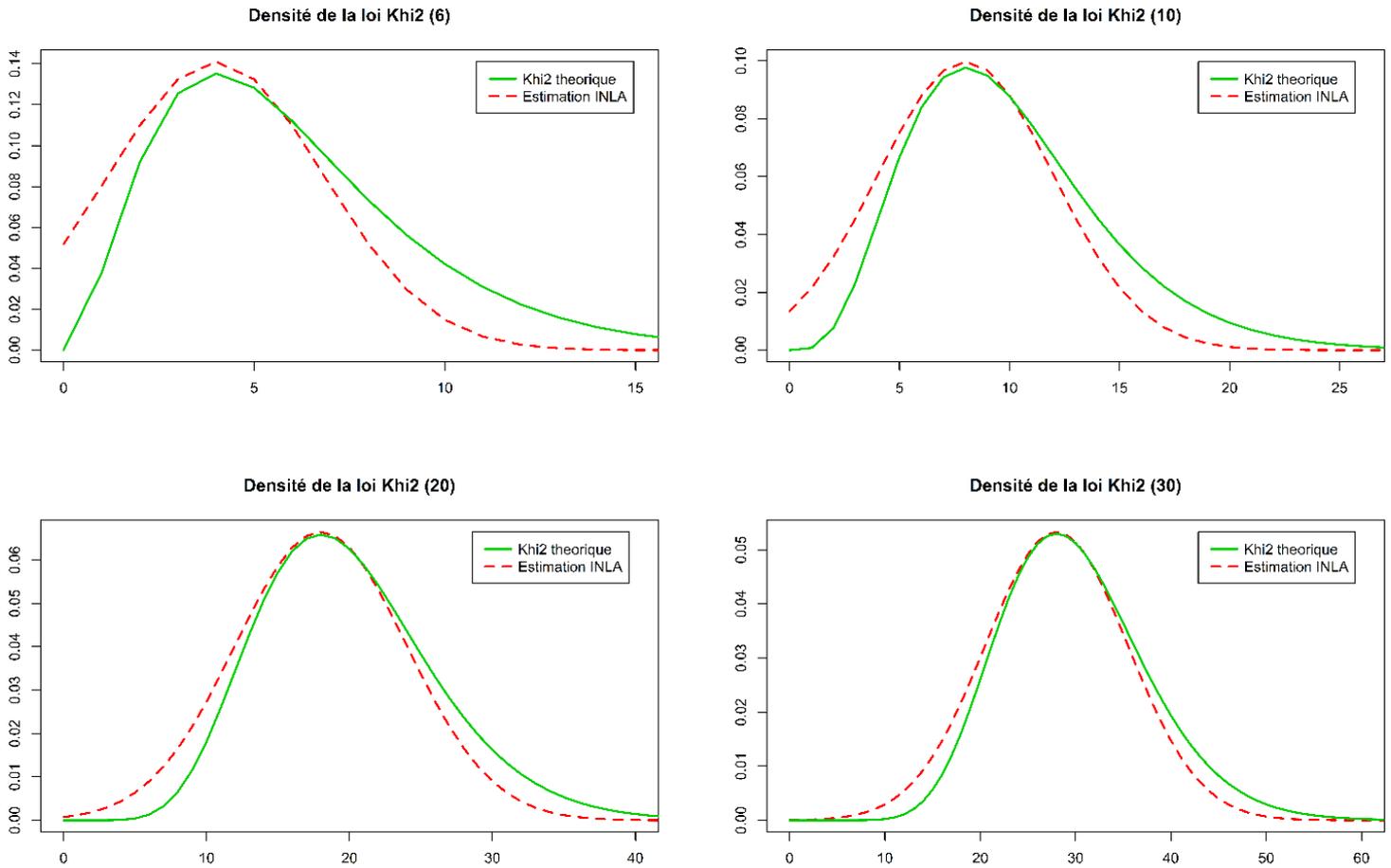
Un exemple concret pour mieux comprendre ces formules théoriques en considérant une

distribution de  $\chi_k^2$  à  $k$  degrés de liberté :  $p(x) = \frac{g(x)}{c} = \frac{1}{c} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$ , alors, nous avons :

- $l(x) = \log(g(x)) = \left(\frac{k}{2} - 1\right) \log(x) - \frac{x}{2}$
- $l'(x) = \left(\frac{k}{2} - 1\right) x^{-1} - \frac{1}{2}$
- $l''(x) = -\left(\frac{k}{2} - 1\right) x^{-2}$
- Alors, en résolvant l'équation  $l'(x) = 0$ , nous obtenons :  $\hat{x} = k - 2$
- L'évaluation  $-\frac{1}{l''(\hat{x})}$  en  $\hat{x}$  donne :  $\hat{\sigma}^2 = 2(k - 2)$
- Par conséquent, en vertu de l'approximation de Laplace, nous pouvons approximer une distribution de  $\chi^2(k)$  à  $k$  degrés de liberté par une distribution gaussienne  $N(k - 2, 2(k - 2))$

Cet exemple théorique est représenté dans le graphique ci-dessous et montre que la méthode INLA est bien adapté pour estimer les distributions marginales *a posteriori* (**Figure 6**).

**Figure 6 :** Estimation de la distribution de Khi2 par une distribution gaussienne selon la méthode INLA



*L'approximation de la loi de Khi2 par la méthode INLA selon le degré de liberté (6,10,20 et 30). Courbe verte est la vraie densité de la loi de Khi2 et la courbe rouge (en pointillé) est l'approximation de celle-ci par la méthode INLA. Plus le degré de liberté augmente, plus l'approximation est meilleure (les deux courbes se rapprochent).*

### 3.2. Application pratique au modèle SPDE avec le package R-INLA

Supposons que l'on cherche à évaluer l'efficacité d'un traitement contre le maladie, ajusté sur 2 cofacteurs, maladies dont la transmission est spatialisée.

Le modèle de survie bayésien s'écrit alors :

$$T | \beta, X, Z \sim \mathcal{W}(T | \mu, \phi)$$

$$Z \sim GF(0, \Sigma)$$

$$\beta, \mu, \phi \sim N(0, 0.001)$$

$$\lambda(t, X) = \lambda_0(t) \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + Z)$$

$T$  est le temps de survie de l'infection du paludisme clinique suivant une loi de Weibull de paramètres  $\mu$  et  $\phi$  à estimer,  $\beta = (\beta_1, \beta_2, \beta_3)$  est l'effet à estimer du vecteur des covariables  $X_1$ ,  $X_2$  et  $X_3$  (traitement, âge et sexe),  $Z$  est un champ gaussien ( $GF$ ) dont la fonction de covariance est la fonction de Matèrn  $\Sigma$  (effet spatial aléatoire) et  $\beta_0$  est l'intercept.

A l'aide de données simulées, exécutons pas à pas ce modèle à l'aide de la méthode INLA implémenté dans le package INLA. Télécharger la base de données [DataBase](#).

Importer cette base de données dans R et observer sa structure de base de données permettant de mieux comprendre la suite de l'exercice.

```
DataBase = read.csv(« chemin\\DataBase.txt », header = TRUE, sep = "\t")
```

```
head(DataBase, 20)
```

	identifiant	xcoord	ycoord	times	status	age	sexe	traitement
1	ML0001	4.047070	-9.827182	0.217094945	1	8.31	1	1
2	ML0002	6.364235	-11.145389	0.898059089	1	5.80	1	1
3	ML0003	1.919225	-10.124872	9.711163189	1	11.64	1	1
4	ML0004	2.793248	-9.188998	0.005411331	1	9.81	1	0
5	ML0005	5.539924	-10.411047	0.597485079	1	1.04	0	1
6	ML0006	6.718432	-10.398433	5.436690019	1	8.07	1	0
7	ML0007	3.152560	-8.432556	0.741785810	1	5.23	1	0
8	ML0008	4.813428	-8.954703	0.145891726	1	4.74	0	1
9	ML0009	5.783661	-11.364728	0.524752488	1	8.44	0	0
10	ML0010	4.902436	-8.560995	7.576496047	1	11.78	0	1
11	ML0011	5.268172	-10.540383	6.108958728	1	4.76	1	1
12	ML0012	7.073133	-11.949233	15.479956004	0	10.22	0	1
13	ML0013	-0.548998	-11.256232	7.490271011	1	13.53	1	1
14	ML0014	5.430873	-9.826511	0.305666721	1	11.33	1	1
15	ML0015	5.998813	-9.661879	0.014394935	1	1.64	1	1
16	ML0016	3.566877	-9.845713	1.756534017	1	3.30	1	0
17	ML0017	6.968825	-10.407774	0.475881381	1	5.32	0	0
18	ML0018	5.784216	-11.447215	4.403236553	1	8.25	0	1
19	ML0019	4.594270	-9.263435	0.304751643	1	7.13	1	0
20	ML0020	7.862039	-10.159541	13.352251432	1	13.10	1	0

Les huit variables sont :

- ✓ identifiant : identifiant (unique) des individus,
  - ✓ (xcoord, ycoord) : coordonnées géographiques des individus (un individu par localisation dans cet exercice mais il était possible avoir plusieurs individus à la même localisation, par exemple un ménage par localisation).
  - ✓ times : le temps de participation, *i.e.* l'écart de temps entre le début de l'étude jusqu'à l'évènement (première infection du paludisme) ou la fin de l'étude s'il n'y a pas eu d'infection.
  - ✓ status : l'indicateur de l'infection (1 s'il y a eu infection et 0 s'il y a eu censure)
  - ✓ age : l'âge des individus (en année)
  - ✓ sexe : le sexe des individus (0 pour femme et 1 pour homme)
  - ✓ traitement : bras de traitement des individus (0 pour traitement de référence et 1 pour traitement expérimental).
- Installer le package INLA à partir du site web de [r-inla.org](http://r-inla.org).

```
install.packages("INLA", repos = c(getOption("repos"), INLA = "https://inla.r-inla-download.org/R/stable"), dep = TRUE)  
library(INLA)
```

Le package INLA est en pleine amélioration, donc des modifications peuvent être apporté à tout moment. Il est donc important de penser à faire des mises à jour fréquemment afin de bénéficier de ces améliorations et des nouvelles fonctionnalités (et de conserver les anciennes versions pour reprendre d'anciennes analyses).

```
inla.upgrade(testing = TRUE)
```

Commençons par l'aspect spatial du modèle. Nous déterminons d'abord l'espace dans lequel l'étude a été menée. L'un des intérêts de l'approche INLA dans la modélisation spatiale est de permettre la prise en compte des effets de bords directement dans le cadrage de la zone de l'étude.

En effet, il y a deux zones : une première couche qui quadrille de façon précise la localisation des individus et une deuxième couche qui quadrille globalement la zone de l'étude et couvre la première couche dans l'objectif de prendre en compte les effets de bord.

- Construction des limites convexes pour un ensemble des individus et la zone de l'étude.

```
location <- cbind(DataBase$xcoord, DataBase$ycoord)
boundary1 <- inla.nonconvex.hull(location, convex = 0.9, resolution = 100)
boundary2 <- inla.nonconvex.hull(location, convex = 2.5, resolution = 100)
```

L'option « *convex* » (ou « *concave* ») détermine la précision que l'on cherche dans la délimitation de la zone de l'étude. Les plus petites valeurs sont plus précises mais c'est à moduler avec le temps d'exécution du modèle plus tard. L'option « *resolution* » détermine la précision des calculs internes de la triangulation (une plus grande valeur implique une plus grande précision). Un message d'avertissement s'affiche lorsque la résolution fixée n'est pas suffisante et qu'elle devrait être augmenté jusqu'à atteindre une valeur minimale requise. Il y a un outil d'aide pour cette étape.

```
help(inla.nonconvex.hull)
```

- Maillage par triangulation de la localisation des individus pour délimiter la zone de l'étude. Il y a plusieurs fonctions permettant de faire du maillage par triangulation mais dans ce cas, nous avons utilisé la fonction « *inla.mesh.2d* ». La qualité de la triangulation et du maillage peut être contrôlée séparément à l'intérieur de la zone de l'étude et à l'extérieure de la zone permettant de contrôler les éventuels effets de bords mais devrait être moins raffiné pour limiter la dimension du modèle.

```
mesh <- inla.mesh.2d(location, boundary = list(boundary1, boundary2), max.edge = c(0.6, 1.8))
```

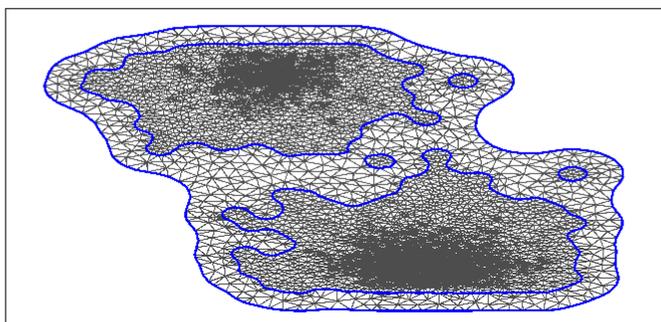
L'option « *max.edge* » détermine la longueur maximale des triangles respectivement à l'intérieur et à l'extérieur de la zone de l'étude. Les plus petites valeurs impliquent une meilleure triangulation mais le temps d'exécution du modèle serait plus long (**Figure 7**). Il existe d'autres options permettant de raffiner la qualité de la triangulation. Voir l'outil d'aide pour plus de détails.

*help(inla.mesh.2d)*

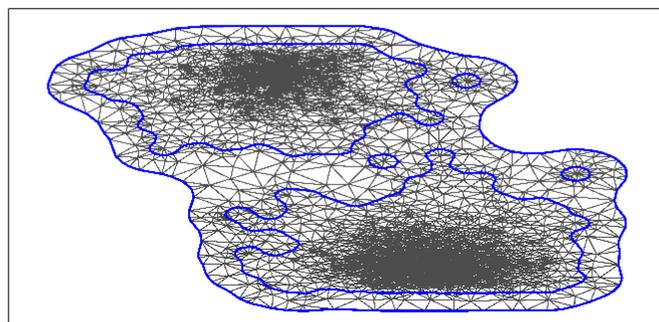
Visualisation du maillage par triangulation avec différentes combinaisons de valeurs de « *convex* » et « *max.edge* » avec une la fonction *plot(mesh)*

**Figure 7:** Maillage par triangulation selon différentes combinaisons de valeurs des paramètres.

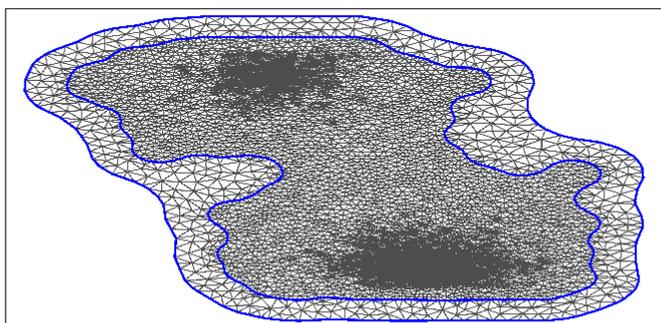
**convex=(0.7,2.5), max.edge=c (0.6, 1.2)**



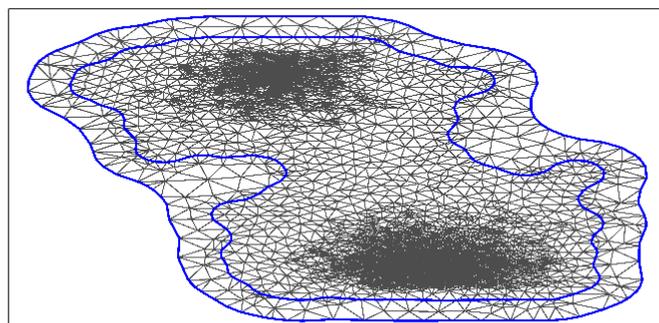
**convex=(0.7, 2.5), max.edge=c (1.2, 2.2)**



**convex=(1.4,3.5), max.edge=c (0.6, 1.2)**



**convex=(1.4, 3.5), max.edge=c (1.2, 2.2)**



Maillage par triangulation selon différentes combinaison de 2 valeurs de la convexité du contour de la zone d'étude (*convex*) et 2 valeurs de la taille des triangles (*max.edge*). Les valeurs en couples pour indiquer la valeur du contour intérieur et extérieur de la zone de l'étude. En haut à gauche: faible convexité (0.7, 2.5) et triangles de petite taille (0.6, 1.2), en haut à droite: faible convexité (0.7, 0.25) et triangles de taille moyenne (1.2, 2.2), en bas à gauche : convexité moyenne (1.4, 3.5) et triangles de petite taille (0.6, 1.2), en bas à droite: convexité moyenne (1.4, 3.5) et triangles de taille moyenne (1.2, 2.2)

Après le maillage de la zone de l'étude par triangulation, il faut faire une projection des données observées dans ce maillage.

- Création de la matrice de projection

```
A <- inla.spde.make.A(mesh, location)
```

- Construction du modèle de SPDE avec la fonction de Matèrn comme matrice de covariance

```
SPDE <- inla.spde2.matern(mesh, alpha = 2)
```

Nous avons choisi de faire un modèle SPDE à deux dimensions (longitude et latitude) mais il est possible de faire le modèle SPDE à plus deux dimensions (altitude par exemple). Il y a plusieurs autres options dans cette fonction donnant une large gamme de choix autres que la fonction de Matèrn.

```
help(inla.spde2.matern)
```

- Regroupement de tous les éléments constants le modèle

Cette fonction permet de rassembler la variable réponse, l'intercept, l'effet spatial et les différents facteurs dans un seul objet INLA pour faciliter l'exécution du modèle plus tard.

```
Stack <- inla.stack(data = list(times = DataBase$times/52, status = DataBase$status),  
                  A = list(A, 1), effect = list(list(spatial = 1:SPDE$n.spde),  
                  data.frame(a0 = 1, DataBase[,c("identifiant", "age", "sexe", "traitement"))]))
```

Il est fréquent qu'un problème d'échelles se pose dans ces approches, ce qui conduit à un échec de l'estimation du modèle. C'est le cas dans cet exercice où le temps de survie a été simulé en semaine mais nous l'avons transformé en année (diviser par 52 car une année fait 52 semaines) pour l'exécution le modèle. D'autres mises à échelles sont possibles.

La constante  $a_0 = 1$  a été ajoutée pour un problème de dimension à cause de la matrice de distance qui possède la valeur 0 sur sa diagonale. Donc nous ajoutons aux données une colonne  $a_0 = 1$

mais dans l'exécution du modèle, cet élément serait considéré comme l'intercept et n'influence pas l'estimation des paramètres. La variable notée « *spatial* » représente l'effet spatial issu de la construction du modèle SPDE.

➤ Formulation du modèle SPDE avec tous les facteurs

C'est la principale partie de la modélisation. Il s'écrit de la même façon que dans le cas d'un simple modèle linéaire généralisé. La fonction « *inla.surv* » est utilisé pour créer l'objet de type survie constituant la variable réponse.

```
formula <- inla.surv(times, status) ~ 0 + a0 + age + sexe + traitement + f(spatial, model = SPDE)
```

L'élément intrigant de cette partie est la présence du 0 dans la formule du modèle. En effet, le 0 est mis pour signifier que l'élément  $a_0$  qu'on avait ajouté pour contourner le problème de dimension est l'intercept et non un facteur. En absence de ce 0,  $a_0$  serait considéré comme facteur et biaiserait les estimations. Les facteurs sont ajoutés simplement comme dans un modèle linéaire. La partie spatiale du modèle est ajouté à partir d'une fonction comme un effet aléatoire latent en spécifiant qu'on a utilisé le modèle SPDE qu'on avait construit plus haut.

Dans cet exemple, nous avons construit le modèle SPDE prenant en compte l'effet spatial aléatoire mais la méthode INLA possède tout une batterie d'effets aléatoires latents aussi bien spatiaux que temporels. Par exemple lorsqu'on a une variable temporelle avec une saisonnalité, il va falloir ajouter à la formule du modèle une partie temporelle : « *f(times, model = "seasonal")* ».

La modélisation d'évènements récurrents est faite également à partir de cette fonction en créant une variable indiquant l'ordre des événements chez un même individu. Si par exemple, la variable indiquant l'ordre des événements est nommée « *order* » et qu'il y a une corrélation autorégressive supposée entre les événements (d'autres formes de corrélation existent), il faut ajouter à la formule du modèle « *f(order, model = "ar1")* ».

La liste des effets aléatoires latents pris en compte par la méthode INLA peut être consulté à partir de la fonction R comme suit :

```
inla.list.models("latent")
```

➤ L'ajustement du modèle

```
ModelSPDE <- inla(formula,  
  family = "weibullsurv",  
  data = inla.stack.data(Stack),  
  control.predictor = list(A = inla.stack.A(Stack)),  
  control.compute = list(dic = TRUE, mlik = TRUE, waic = TRUE),  
  control.inla = list(strategy = "laplace", h = 1e-15),  
  verbose = TRUE)
```

La formule du modèle a été déjà définie plus haut et contient tous les facteurs à effets fixes et aléatoires. Comme dans les fonctions classiques de régression, la famille (« *family* ») représente le type de modèle, fonction de la variable dépendante.

Dans cet exercice, nous avons comme variable dépendante, le temps de survie suivant une loi de Weibull comme signifié dans la famille (*family = "weibullsurv"*). Si le temps de survie suivait une loi exponentielle, nous aurions (*family = "exponentialsurv"*). De même, s'il fallait modéliser un processus de comptage, nous aurions pu utiliser la loi de Poisson (*family = "poisson"*) ou la loi binomiale négative (*family = "nbinomial"*).

Une large gamme de distribution est implémentée dans le package R-INLA, la liste complète est disponible avec la fonction suivante :

```
inla.list.models("likelihood")
```

L'élément « *data = Stack* » indique la base de données utilisée. Puisqu'on avait regroupé plus haut, toutes les données dans un objet « *Stack* », il suffit de signifier cela ici.

Dans cet exercice, nous avons besoin de calculer la vraisemblance marginale, le DIC et le log vraisemblance marginal de notre modèle comme indiqué dans *control.compute = list(dic = TRUE, mlik = TRUE)*. Mais il existe d'autres mesures de performance des modèles comme le WAIC, CPO que l'on pouvait calculer en les ajoutant dans la fonction ci-dessus.

```
control.compute = list(dic = TRUE, mlik = TRUE, waic = TRUE, cpo = TRUE)
```

La fonction « *control.inla()* » permet de raffiner les calculs internes du modèle. Cette fonction possède différentes options permettant de contrôler la qualité des estimations. Par exemple, l'option « *h* » indiquant longueur des pas pour les calculs de gradient dans l'estimation des hyper paramètres. Sa valeur par défaut est 0.01 mais une valeur plus petite permet un calcul plus précis. L'option « *strategy* » indique l'approximation utilisée. La valeur par défaut est « *strategy = "gaussian"* » qui est moins couteuse en temps mais elle implique un biais important lorsque la distribution est loin de la loi gaussienne. La stratégie « *strategy = "laplace"* » convient à toutes les distributions mais le temps d'exécution est plus long. D'autres options existent pour raffiner les calculs, se référer à l'outil d'aide.

```
help(control.inla)
```

L'option « *verbose = TRUE* » permet de suivre l'exécution du modèle pas à pas en ayant une vue sur les calculs internes. Sa valeur par défaut est « *verbose = FALSE* » mais l'exécution du modèle ne sera alors pas affichée à l'écran pour voir les anomalies, comme par exemple les matrices non-inversibles qui n'arrêtent pas l'exécution mais risquent de biaiser les estimations. Dans ces cas, une diminution de la valeur « *h* » pourrait souvent résoudre ces problèmes d'inversibilité des matrices ou des valeurs propres non positives.

➤ Sortie de la méthode INLA

Comme toutes les autres méthodes, la sortie de la méthode INLA est consultable à travers la fonction « *summary* » qui affiche tous les éléments essentiels du modèle.

```
summary(ModelSPDE)
```

```
Time used:
  Pre = 10.4, Running = 3402, Post = 4.2, Total = 3416
Fixed effects:
      mean    sd 0.025quant 0.5quant 0.975quant  mode kld
a0      4.607 0.328      3.963   4.607   5.250  4.608  0
age     -0.174 0.002     -0.179  -0.174  -0.170 -0.174  0
sexe    0.057 0.039     -0.020   0.057   0.135  0.057  0
traitement -0.500 0.039     -0.577  -0.500  -0.423 -0.500  0

Random effects:
Name      Model
  spatial SPDE2 model

Model hyperparameters:
      mean    sd 0.025quant 0.5quant 0.975quant  mode
alpha parameter for weibullsurv  1.000 0.00      1.000   1.000   1.000  1.000
Thetal for spatial              -0.548 0.00     -0.548  -0.548  -0.548 -0.548
Theta2 for spatial              -0.718 0.00     -0.718  -0.718  -0.718 -0.718

Expected number of effective parameters(stdev): 148.36(0.014)
Number of equivalent replicates : 26.89

Deviance Information Criterion (DIC) .....: 6515.06
Deviance Information Criterion (DIC, saturated) ....: NaN
Effective number of parameters .....: 140.46

Marginal log-Likelihood: -3026.64
Posterior marginals for the linear predictor and
the fitted values are computed
```

Ci-dessus la sortie du modèle. D'abord, nous avons « *Time used* » qui correspondant au temps total d'exécution du modèle (en secondes). Dans cet exercice, le modèle a pris 3416 secondes (soit 57 mn) pour converger.

Ensuite, nous avons les effets fixes associés à chaque facteur dans la colonne « *mean* » (les coefficients de l'âge (-0.174), du sexe (0.057) et du traitement (-0.500). Suivi de la colonne « *sd* » qui est la valeur de l'erreur standard associé à l'estimation de coefficient correspondant.

Les colonnes « *0.025quant* » et « *0.975quant* » représente respectivement le 1<sup>er</sup> et le 3<sup>ème</sup> quantile, ce sont aussi les bornes de l'intervalle de crédibilité à 95%. Dans cet exercice, les facteurs âge et traitement sont significatifs (-0.179 et -0.170 pour l'âge, -0.577 et -0.423 pour le traitement) parce que les bornes de l'intervalle de crédibilité sont de même signe.

Par contre, le facteur sexe n'est pas significatif parce que les bornes de l'intervalle de crédibilité (-0.020 et 0.135) sont de signe opposé (ça signifie que l'intervalle de crédibilité contient la valeur zéro). La colonne « *mode* », la moins sollicité mais donne des informations sur la distribution des paramètres. Une mode proche de la moyenne « *mean* » signifie que la distribution du paramètre associé est symétrique. Dans cet exercice, la mode est identique à la moyenne pour tous les facteurs, donc la distribution des paramètres (les coefficients) est symétrique, gaussienne, ce qui signifie que les estimations sont de bonnes qualités.

Nous avons également les informations sur les effets aléatoires du modèle et les hyper paramètres aux termes bayésiens qui s'interprètent exactement comme les paramètres.

Les hyperparamètres « *Theta1 of spatial* » et « *Theta2 of spatial* » représentent les deux paramètres de la fonction de covariance de Matérn. Lorsqu'ils sont significatifs, ça veut dire qu'il y a un effet spatial significativement non nul.

Enfin, il y a les indicateurs de performance du modèle notamment le DIC et le log vraisemblance marginale qui sont importants lorsqu'il faut comparer deux modèles. Le meilleur modèle est celui qui possède le plus petit DIC et la plus grande vraisemblance marginale.

Il y a d'autres éléments dans la sortie du modèle que l'on peut voir à travers la fonction « *summary* ».

```
names(ModelSPDE)
```

```
names(summary(ModelSPDE))
```

## **Conclusions générales et perceptives**

Dans cette thèse nous avons montrés que la randomisation n'était pas suffisante pour assurer la comparabilité des différents groupes de traitement dans un essai de prévention. A cet effet, l'utilisation des méthodes classiques non adaptées au risque spatialement hétérogène pourrait conduire à un biais dans l'évaluation d'un essai prévention randomisé. Les modèles SPDE modélisant l'hétérogénéité spatiale au travers la localisation des personnes (les coordonnées géographiques) permettaient de corriger ce biais. Dans ces conditions, nous suggérons aux investigateurs des essais de préventions et/ou cliniques notamment contre les maladies transmissibles, de prendre les coordonnées géographiques des participants depuis le début de l'étude afin de les utiliser au moment de l'analyse à travers la localisation pour de corriger le biais dû à l'hétérogénéité spatiale du risque environnemental.

De même, nous avons montré dans la deuxième partie de cette thèse que lorsque la maladie étudiée pouvait infecter plusieurs fois le même individu courant la durée d'un essai de prévention randomisé, il serait plus adapté d'associer la modélisation de l'hétérogénéité spatiale à la prise en compte des évènements récurrents. Par exemple, contrairement à la conclusion de l'étude principale, nous avons mis en évidence le bénéfice significatif d'un ajout de l'azithromycine à la CPS comme prévention du paludisme en associant une modélisation de l'effet spatiale et des évènements récurrents. La réanalyse de l'étude de Bandiagara n'a montré pas modifié les résultats probablement à cause du recrutement spatialement hétérogène des individus dans la zone de l'étude.

Cependant, l'une des hypothèses fortes de ce travail de thèse était l'immobilité des personnes. En effet, nous avons considéré que les individus étaient immobiles, et le risque associé a été mesuré par les coordonnées GPS de leur maison. En cas de mobilité, l'estimation de l'effet spatial serait alors biaisée [122, 123]. Toutes fois, il est à souligner que cette mobilité serait plus pertinente lorsqu'elle conduit à un niveau de risque environnemental différent, notamment pour une période

de mobilité plus longue. Dans le contexte du paludisme, les déplacements nocturnes correspondant au moment des repas sanguins des moustiques ou un déplacement vers un faciès épidémiologique très différents (une ou plusieurs nuits), pourront biaiser les estimations selon notre approche. Lorsqu'un individu se déplace dans la journée en absence du vecteur et retourne à son lieu d'habitation localisé dans la même journée, l'impact de cette mobilité serait moindre. Cela dépend bien sûr des habitudes trophiques des vecteurs, parfois changeantes.

L'autre hypothèse forte de cette thèse était que l'hétérogénéité spatiale n'évoluait pas dans le temps. Par exemple, dans le contexte du paludisme, l'hétérogénéité de la répartition des gîtes pourrait être plus prononcée pendant l'hivernage qu'en saison sèche, donc la zone la plus dense en gîte pendant l'hivernage peut être différente en saison sèche. En plus, la variation dans le temps de la structure spatiale de l'incidence a été observée, même dans une même saison de transmission palustre, notamment à Bandiagara [86–88].

A propos de la mobilité des personnes, les modèles SPDE et les modèles de mobilité humaine défini dans [122, 123] peuvent être combiné en redéfinissant la matrice de covariance du champ gaussien utilisé dans ce travail avec un coefficient supplémentaire représentant par exemple la viscosité de la zone de l'étude. Cela nécessiterait des données de mobilité notamment de téléphonie mobile ou d'autres objets connectés. A cet effet, une collaboration avec les opérateurs de téléphones mobiles pourrait aider à l'obtention des données de mobilité humaine grâce aux réseaux téléphoniques auxquels les téléphones des usagers se connectent par proximités. Cependant, il pourrait être utile également de prendre en compte les habitudes trophiques, parfois changeantes, du vecteur et le comportement humain.

Concernant la variation temporelle de la structure spatiale, il faudrait introduire une troisième dimension en plus de la latitude et de la longitude dans la définition du champ gaussien pour le modèle SPDE. Une troisième dimension qui représenterait le temps, permettant de définir la distance entre les individus en fonction du temps et par conséquent la structure spatiale également.

## **Bibliographie**

1. Hiscox A, Homan T, Vreugdenhil C, Otieno B, Kibet A, Mweresa CK, et al. Spatial heterogeneity of malaria vectors and malaria transmission risk estimated using odour-baited mosquito traps. *Malaria Journal*. 2014;13 Suppl 1:P41.
2. Tine RCK, Ndour CT, Faye B, Cairns M, Sylla K, Ndiaye M, et al. Feasibility, safety and effectiveness of combining home based malaria management and seasonal malaria chemoprevention in children less than 10 years in Senegal: a cluster-randomised trial. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. 2014;108:13–21.
3. Barry A, Issiaka D, Traore T, Mahamar A, Diarra B, Sagara I, et al. Optimal mode for delivery of seasonal malaria chemoprevention in Ouelessebougou, Mali: A cluster randomized trial. *PLOS ONE*. 2018;13:e0193296.
4. Tine RC, Faye B, Ndour CT, Ndiaye JL, Ndiaye M, Bassene C, et al. Impact of combining intermittent preventive treatment with home management of malaria in children less than 10 years in a rural area of Senegal: a cluster randomized trial. *Malaria Journal*. 2011;10:358.
5. Thera MA, Kone AK, Tangara B, Diarra E, Niare S, Dembele A, et al. School-aged children based seasonal malaria chemoprevention using artesunate-amodiaquine in Mali. *Parasite Epidemiology and Control*. 2018;3:96–105.
6. Getachew Y, Janssen P, Yewhalaw D, Speybroeck N, Duchateau L. Coping with time and space in modelling malaria incidence: a comparison of survival and count regression models. *Statistics in Medicine*. 2013;32:3224–33.
7. Li Y, Ryan L. Modeling spatial survival data using semiparametric frailty models. *Biometrics*. 2002;58:287–97.

8. Gangnon RE, Clayton MK. A hierarchical model for spatially clustered disease rates. *Statistics in Medicine*. 2003;22:3213–28.
9. Sissoko MS, Sauerwein RW, Knight P, Bousema T, Coulibaly M, Samake Y, et al. Spatial patterns of *Plasmodium falciparum* clinical incidence, asymptomatic parasite carriage and *Anopheles* Density in Two Villages in Mali. *The American Journal of Tropical Medicine and Hygiene*. 2015;93:790–7.
10. Overgaard HJ, Olano VA, Jaramillo JF, Matiz MI, Sarmiento D, Stenström TA, et al. A cross-sectional survey of *Aedes aegypti* immature abundance in urban and rural household containers in central Colombia. *Parasites & Vectors*. 2017;10. doi:10.1186/s13071-017-2295-1.
11. Imbahale SS, Paaijmans KP, Mukabana WR, van Lammeren R, Githeko AK, Takken W. A longitudinal study on *Anopheles* mosquito larval abundance in distinct geographical and environmental settings in western Kenya. *Malaria Journal*. 2011;10. doi:10.1186/1475-2875-10-81.
12. Gao Q, Wang F, Lv X, Cao H, Su F, Zhou J, et al. *Aedes albopictus* production in urban stormwater catch basins and manhole chambers of downtown Shanghai, China. *PLOS ONE*. 2018;13:e0201607.
13. Pandey S, Das MK, Dhiman RC. Diversity of breeding habitats of anophelines (Diptera: Culicidae) in Ramgarh district, Jharkhand, India. *J Vector Borne Dis*. 2016;53–4:327–334.
14. Thomas S, Ravishankaran S, Justin JA, Asokan A, Mathai MT, Valecha N, et al. Overhead tank is the potential breeding habitat of *Anopheles stephensi* in an urban transmission setting of Chennai, India. *Malaria Journal*. 2016;15. doi:10.1186/s12936-016-1321-7.

15. Young RL, Weinberg J, Vieira V, Ozonoff A, Webster TF. A power comparison of generalized additive models and the spatial scan statistic in a case-control setting. *International Journal of Health Geographics*. 2010;9:37.
16. Vieira VM, Weinberg JM, Webster TF. Individual-level space-time analyses of emergency department data using generalized additive modeling. *BMC Public Health*. 2012;12. doi:10.1186/1471-2458-12-687.
17. Rebaudet S, Griffiths K, Trazillio M, Lebeau A-G, Abedi AA, Bulit G, et al. Cholera spatial-temporal patterns in Gonaives, Haiti: From contributing factors to targeted recommendations. *Advances in Water Resources*. 2017;108:377–85.
18. Wood SN. *Generalized additive models: an introduction with R*. 2nd edition. London: Chapman and Hall/CRC; 2017. p. 136–71.
19. Tobler WR. A computer movie simulating urban growth in the detroit region. *Economic Geography*. 1970;46:234.
20. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2011;73:423–98.
21. Cameletti M, Lindgren F, Simpson D, Rue H. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*. 2013;97:109–31.
22. Samadoulougou S, Maheu-Giroux M, Kirakoya-Samadoulougou F, De Keukeleire M, Castro MC, Robert A. Multilevel and geo-statistical modeling of malaria risk in children of Burkina Faso. *Parasites & Vectors*. 2014;7:350.

23. Ouédraogo M, Samadoulougou S, Rouamba T, Hien H, Sawadogo JEM, Tinto H, et al. Spatial distribution and determinants of asymptomatic malaria risk among children under 5 years in 24 districts in Burkina Faso. *Malar J.* 2018;17:460.
24. Musenge E, Chirwa TF, Kahn K, Vounatsou P. Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural North East South Africa. *International Journal of Applied Earth Observation and Geoinformation.* 2013;22:86–98.
25. Núñez O, Fernández-Navarro P, Martín-Méndez I, Bel-Lan A, Locutura JF, López-Abente G. Arsenic and chromium topsoil levels and cancer mortality in Spain. *Environmental Science and Pollution Research.* 2016;23:17664–75.
26. Lenaerts E, Mandro M, Mukendi D, Suykerbuyk P, Dolo H, Wonya’Rossi D, et al. High prevalence of epilepsy in onchocerciasis endemic health areas in Democratic Republic of the Congo. *Infectious Diseases of Poverty.* 2018;7. doi:10.1186/s40249-018-0452-1.
27. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine.* 2006;25:4279–92.
28. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine.* 2019;38:2074–102.
29. USAID. The DHS program STATcompiler. 2016. <https://www.statcompiler.com/fr/>. Accessed 1 Jan 2017.
30. Midega JT, Mbogo CM, Mwambi H, Wilson MD, Ojwang G, Mwangangi JM, et al. Estimating dispersal and survival of *Anopheles gambiae* and *Anopheles funestus* along the

Kenyan coast by using mark–release–recapture methods. *Journal of medical entomology*. 2007;44:923–929.

31. Bousema T, Stresman G, Baidjoe AY, Bradley J, Knight P, Stone W, et al. The impact of hotspot-targeted interventions on malaria transmission in Rachuonyo south district in the western Kenyan highlands: A cluster-randomized controlled trial. *PLOS Medicine*. 2016;13:e1001993.

32. Toure OA, Valecha N, Tshefu AK, Thompson R, Krudsood S, Gaye O, et al. A phase 3, double-blind, randomized study of arterolane maleate–piperaquine phosphate vs artemether–lumefantrine for falciparum malaria in adolescent and adult patients in Asia and Africa. *Clinical Infectious Diseases*. 2016;62:964–71.

33. Thera MA, Coulibaly D, Kone AK, Guindo AB, Traore K, Sall AH, et al. Phase 1 randomized controlled trial to evaluate the safety and immunogenicity of recombinant *Pichia pastoris*-expressed *Plasmodium falciparum* apical membrane antigen 1 (PfAMA1-FVO [25-545]) in healthy Malian adults in Bandiagara. *Malaria Journal*. 2016;15. doi:10.1186/s12936-016-1466-4.

34. Sissoko MS, Healy SA, Katile A, Omaswa F, Zaidi I, Gabriel EE, et al. Safety and efficacy of PfSPZ Vaccine against *Plasmodium falciparum* via direct venous inoculation in healthy malaria-exposed adults in Mali: a randomised, double-blind phase 1 trial. *The Lancet Infectious Diseases*. 2017;17:498–509.

35. Dama S, Niangaly H, Djimde M, Sagara I, Guindo CO, Zeguime A, et al. A randomized trial of dihydroartemisinin–piperaquine versus artemether–lumefantrine for treatment of uncomplicated *Plasmodium falciparum* malaria in Mali. *Malaria Journal*. 2018;17. doi:10.1186/s12936-018-2496-x.

36. Dicko A, Brown JM, Diawara H, Baber I, Mahamar A, Soumare HM, et al. Primaquine to reduce transmission of *Plasmodium falciparum* malaria in Mali: a single-blind, dose-ranging, adaptive randomised phase 2 trial. *The Lancet Infectious Diseases*. 2016;16:674–84.
37. *Les Humains et la Terre. Atlas des populations et pays du monde*. 2017.  
<https://www.populationdata.net/pays/mali/>. Accessed 1 Mar 2017.
38. Lee ET, Wang JW. *Statistical methods for survival data analysis*. 3rd ed. New York: J. Wiley; 2003. p. 1–165.
39. David C. Modelling survival data. In: *Modelling survival data in medical research*. 2nd edition. London: Chapman and Hall/CRC; 2003. p. 55–109.
40. Rue H, Martino S, Chopin N. Approximate bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2009;71:319–92.
41. Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. Bayesian computing with INLA: a Review. *Annual Review of Statistics and Its Application*. 2017;4:395–421.
42. Lee ET, Wang JW. *Statistical methods for survival data analysis*. 3rd ed. New York: J. Wiley; 2003. p. 1–165.
43. David C. *Modelling survival data in medical research*. 2nd edition. London: Chapman and Hall/CRC; 2003. p. 55–109.
44. Li L, Wu J, Wilhelm M, Ritz B. Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in Southern California. *Atmospheric Environment*. 2012;55:220–8.

45. Venables WN, Dichmont CM. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. *Fisheries Research*. 2004;70:319–37.
46. Lindgren F, Rue H. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*. 2015;63. doi:10.18637/jss.v063.i19.
47. Minasny B, McBratney AlexB. The Matérn function as a general model for soil variograms. *Geoderma*. 2005;128:192–207.
48. Crowther MJ, Riley RD, Staessen JA, Wang J, Gueyffier F, Lambert PC. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Medical Research Methodology*. 2012;12:34.
49. Simpson D, Illian JB, Lindgren F, Sørbye SH, Rue H. Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*. 2016;103:49–70.
50. Diva U, Banerjee S, Dey DK. Modelling spatially correlated survival data for individuals with multiple cancers. *Statistical Modelling: An International Journal*. 2007;7:191–213.
51. Bastos LS, Gamerman D. Dynamic survival models with spatial frailty. *Lifetime Data Anal*. 2006;12:441–60.
52. Banerjee S, Dey DK. Semiparametric Proportional Odds Models for Spatially Correlated Survival Data. *Lifetime Data Anal*. 2005;11:175–91.
53. Sauleau EA, Hennerfeind A, Buemi A, Held L. Age, period and cohort effects in Bayesian smoothing of spatial cancer survival with ge additive models. *Statist Med*. 2007;26:212–29.
54. Hennerfeind A, Brezger A, Fahrmeir L. Ge additive Survival Models: A Supplement. :21.

55. Farrance CE, Rhee A, Jones RM, Musiychuk K, Shamloul M, Sharma S, et al. A plant-produced Pfs230 vaccine candidate blocks transmission of *Plasmodium falciparum*. *Clinical and Vaccine Immunology*. 2011;18:1351–7.
56. Patrick E D. Study of the Safety and Immunogenicity of Pfs230D1M-EPA/Alhydrogel and Pfs25M-EPA/Alhydrogel , a Transmission Blocking Vaccine Against *Plasmodium Falciparum* Malaria, in Adults in the U.S. and Mali. 2015.  
<https://clinicaltrials.gov/ct2/show/study/NCT02334462?term=Pfs230&cond=Malaria%2CFalciparum&cntry=ML>. Accessed 5 Oct 2018.
57. Nissen A, Cook J, Loha E, Lindtjørn B. Proximity to vector breeding site and risk of *Plasmodium vivax* infection: a prospective cohort study in rural Ethiopia. *Malaria Journal*. 2017;16. doi:10.1186/s12936-017-2031-5.
58. Umlauf N, Adler D, Kneib T, Lang S, Zeileis A. Structured additive regression models: An R interface to BayesX. *Journal of Statistical Software*. 2015;63. doi:10.18637/jss.v063.i21.
59. Brezger A, Kneib T, Lang S. BayesX : Analyzing bayesian structured additive regression models. *Journal of Statistical Software*. 2005;14. doi:10.18637/jss.v014.i11.
60. Adebayo SB, Gayawan E, Heumann C, Seiler C. Joint modeling of anaemia and malaria in children under five in Nigeria. *Spatial and Spatio-temporal Epidemiology*. 2016;17:105–15.
61. Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data: A bayesian perspective. *Statistica Sinica*. 2004;14:731–61.
62. Banerjee S, Wall MM, Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*. 2003;4:123–142.

63. Dekker FW, de Mutsert R, van Dijk PC, Zoccali C, Jager KJ. Survival analysis: time-dependent effects and time-varying risk factors. *Kidney International*. 2008;74:994–7.
64. Fisher LD, Lin DY. Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health*. 1999;20:145–57.
65. Kenneth RH. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. english. 1994;13(10):1045–62.
66. Wheeler DC, Waller LA, Cozen W, Ward MH. Spatial–temporal analysis of non-Hodgkin lymphoma risk using multiple residential locations. *Spatial and Spatio-temporal Epidemiology*. 2012;3:163–71.
67. Fuglstad G-A, Lindgren F, Simpson D, Rue H. Exploring a New Class of Non-stationary Spatial Gaussian Random Fields with Varying Local Anisotropy. arXiv:13046949 [stat]. 2014. <http://arxiv.org/abs/1304.6949>. Accessed 13 Dec 2019.
68. Amorim LD, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology*. 2015;44:324–33.
69. Tang Y, Fitzpatrick R. Sample size calculation for the Andersen-Gill model comparing rates of recurrent events. *Statistics in Medicine*. 2019;:sim.8335.
70. Stanley CC, Kazembe LN, Buchwald AG, Mukaka M, Mathanga DP, Hudgens MG, et al. Joint modelling of time-to-clinical malaria and parasite count in a cohort in an endemic area. *J Med Stat Inform*. 2019;7:1.
71. Rambhatla JS, Turner L, Manning L, Laman M, Davis TME, Beeson JG, et al. Acquisition of Antibodies Against Endothelial Protein C Receptor–Binding Domains of *Plasmodium*

*falciparum* Erythrocyte Membrane Protein 1 in Children with Severe Malaria. The Journal of Infectious Diseases. 2019;219:808–18.

72. Farrington L, Vance H, Rek J, Prahl M, Jagannathan P, Katureebe A, et al. Both inflammatory and regulatory cytokine responses to malaria are blunted with increasing age in highly exposed children. Malar J. 2017;16:499.

73. Bediako Y, Ngoi JM, Nyangweso G, Wambua J, Opiyo M, Nduati EW, et al. The effect of declining exposure on T cell-mediated immunity to *Plasmodium falciparum* – an epidemiological “natural experiment.” BMC Med. 2016;14:143.

74. Rono J, Färnert A, Murungi L, Ojal J, Kamuyu G, Guleid F, et al. Multiple clinical episodes of *Plasmodium falciparum* malaria in a low transmission intensity setting: exposure versus immunity. BMC Med. 2015;13:114.

75. Sagara I, Giorgi R, Doumbo OK, Piarroux R, Gaudart J. Modelling recurrent events: comparison of statistical models with continuous and discontinuous risk intervals on recurrent malaria episodes data. Malaria journal. 2014;13:293.

76. Villegas R, Julià O, Ocaña J. Empirical study of correlated survival times for recurrent events with proportional hazards margins and the effect of correlation and censoring. BMC Med Res Methodol. 2013;13:95.

77. Rondeau V, Mazroui Y, Gonzalez JR. frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. Journal of Statistical Software. 2012;47. doi:10.18637/jss.v047.i04.

78. Li Y, Qi L, Sun Y. Semiparametric varying-coefficient regression analysis of recurrent events with applications to treatment switching: Analysis of recurrent events with treatment switching. *Statistics in Medicine*. 2018. doi:10.1002/sim.7856.
79. Guindo A, Sagara I, Ouedraogo B, Sallah K, Assadou MH, Healy S, et al. Spatial heterogeneity of environmental risk in randomized prevention trials: consequences and modeling. *BMC Med Res Methodol*. 2019;19:149.
80. Darmofal D. Bayesian spatial survival models for political event processes. *American Journal of Political Science*. 2009;53:241–257.
81. Grenfell B, Kleczkowski A, Gilligan C, Bolker B. Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases. *Statistical Methods in Medical Research*. 1995;4:160–83.
82. Bakka H, Rue H, Fuglstad G-A, Riebler A, Bolin D, Illian J, et al. Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2018;10:e1443.
83. Lindgren F, Rue H. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*. 2015;63. <http://opus.bath.ac.uk/45256/>. Accessed 24 Jul 2017.
84. Krainski ET, Lindgren F. The R-INLA tutorial: SPDE models. 2013.
85. Chandramohan D, Dicko A, Zongo I, Sagara I, Cairns M, Kuepfer I, et al. Effect of Adding Azithromycin to Seasonal Malaria Chemoprevention. *New England Journal of Medicine*. 2019. doi:10.1056/NEJMoa1811400.
86. Coulibaly D, Travassos MA, Tolo Y, Laurens MB, Kone AK, Traore K, et al. Spatio-Temporal Dynamics of Asymptomatic Malaria: Bridging the Gap Between Annual Malaria Resurgences in a Sahelian Environment. *The American Journal of Tropical Medicine and Hygiene*. 2017;97:1761–9.

87. Coulibaly D, Rebaudet S, Travassos M, Tolo Y, Laurens M, Kone AK, et al. Spatio-temporal analysis of malaria within a transmission season in Bandiagara, Mali. *Malar J.* 2013;12:82.
88. Coulibaly D, Travassos MA, Kone AK, Tolo Y, Laurens MB, Traore K, et al. Stable malaria incidence despite scaling up control strategies in a malaria vaccine-testing site in Mali. *Malar J.* 2014;13:374.
89. Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics.* 1982;10:1100–20.
90. Sokoreli I, Pauws SC, Steyerberg EW, de Vries G-J, Riistama JM, Tesanovic A, et al. Prognostic value of psychosocial factors for first and recurrent hospitalizations and mortality in heart failure patients: insights from the OPERA-HF study: Impact of psychosocial factors on recurrent readmissions and death. *Eur J Heart Fail.* 2018;20:689–96.
91. Xu J, Lam KF, Chen F, Milligan P, Cheung YB. Semiparametric estimation of time-varying intervention effects using recurrent event data: J. XU *ET AL.* *Statist Med.* 2017;36:2682–96.
92. Ozga A-K, Kieser M, Rauch G. A systematic comparison of recurrent event models for application to composite endpoints. *BMC Med Res Methodol.* 2018;18:2.
93. Rauch G, Kieser M, Binder H, Bayes-Genis A, Jahn-Eimermacher A. Time-to-first-event versus recurrent-event analysis: points to consider for selecting a meaningful analysis strategy in clinical trials with composite endpoints. *Clin Res Cardiol.* 2018;107:437–43.
94. Vasudevan A, Choi JW, Feghali GA, Kluger AY, Lander SR, Tecson KM, et al. First and recurrent events after percutaneous coronary intervention: implications for survival analyses. *Scandinavian Cardiovascular Journal.* 2019;;doi:10.1080/14017431.2019.1645349.

95. Tawiah R, McLachlan GJ, Ng SK. Mixture cure models with time-varying and multilevel frailties for recurrent event data. *Stat Methods Med Res.* 2019;;doi:10.1177/0962280219859377.
96. Choi Y-H, Jacqmin-Gadda H, Król A, Parfrey P, Briollais L, Rondeau V. Joint nested frailty models for clustered recurrent and terminal events: An application to colonoscopy screening visits and colorectal cancer risks in Lynch Syndrome families. *Stat Methods Med Res.* 2019;;doi:10.1177/0962280219863076.
97. Illian JB, Sørbye SH, Rue H. A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics.* 2012;6:1499–530.
98. Guindo A, Sagara I, Ouédraogo B, Dicko A, Sallah K, Doumbo O, et al. Modélisation de l'hétérogénéité spatiale de l'exposition : essais cliniques dans le contexte du paludisme. *Revue d'Épidémiologie et de Santé Publique.* 2018;66:S134.
99. Yan G. Séries chronologiques à une et plusieurs variables: synthèse des méthodes classiques et modèles à base de copules. Université du Québec; 2011. <http://depot-e.uqtr.ca/2066/1/030187471.pdf>. Accessed 3 Sep 2019.
100. Proïa F. Autocorrélation et stationnarité dans le processus autorégressif. Université de Bordeaux I; 2013. <https://tel.archives-ouvertes.fr/tel-01128258/document>. Accessed 3 Sep 2019.
101. Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: Mira J, Sandoval F, editors. *From Natural to Artificial Neural Computation.* Berlin, Heidelberg: Springer Berlin Heidelberg; 1995. p. 195–201. doi:10.1007/3-540-59497-3\_175.

102. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. *Stat Comput.* 2014;24:997–1016.
103. White NJ. The assessment of antimalarial drug efficacy. *Trends in Parasitology.* 2002;18:458–64.
104. White NJ, Pukrittayakamee S, Hien TT, Faiz MA, Mokuolu OA, Dondorp AM. Malaria. *The Lancet.* 2014;383:723–35.
105. Arzika AM, Maliki R, Boubacar N, Kane S, Cotter SY, Lebas E, et al. Biannual mass azithromycin distributions and malaria parasitemia in pre-school children in Niger: A cluster-randomized, placebo-controlled trial. *PLoS Med.* 2019;16. doi:10.1371/journal.pmed.1002835.
106. Rosenthal PJ. Azithromycin for Malaria? *The American Journal of Tropical Medicine and Hygiene.* 2016;95:2–4.
107. Abdus-Salam R, Bello F, Fehintola F, Arowojolu A. A comparative study of azithromycin and sulphadoxine-pyrimethamine as prophylaxis against malaria in pregnancy. *Niger Postgrad Med J.* 2016;23:57.
108. Akinyotu O, Bello F, Abdus-Salam R, Arowojolu A. A randomized controlled trial of azithromycin and sulphadoxine–pyrimethamine as prophylaxis against malaria in pregnancy among human immunodeficiency virus–positive women. *Transactions of The Royal Society of Tropical Medicine and Hygiene.* 2019;113:463–70.
109. Desai M, Hill J, Fernandes S, Walker P, Pell C, Gutman J, et al. Prevention of malaria in pregnancy. *The Lancet Infectious Diseases.* 2018;18:e119–32.
110. Moore BR, Benjamin JM, Tobe R, Ome-Kaius M, Yadi G, Kasian B, et al. A Randomized Open-Label Evaluation of the Antimalarial Prophylactic Efficacy of Azithromycin-Piperaquine

versus Sulfadoxine-Pyrimethamine in Pregnant Papua New Guinean Women. *Antimicrob Agents Chemother.* 2019;63:doi:10.1128/AAC.00302-19.

111. Kimani J, Phiri K, Kamiza S, Duparc S, Ayoub A, Rojo R, et al. Efficacy and Safety of Azithromycin-Chloroquine versus Sulfadoxine-Pyrimethamine for Intermittent Preventive Treatment of *Plasmodium falciparum* Malaria Infection in Pregnant Women in Africa: An Open-Label, Randomized Trial. *PLoS ONE.* 2016;11:e0157045.

112. O'Brien KS, Cotter SY, Amza A, Kadri B, Nassirou B, Stoller NE, et al. Mass Azithromycin and Malaria Parasitemia in Niger: Results from a Community-Randomized Trial. *The American Journal of Tropical Medicine and Hygiene.* 2017;97:696–701.

113. Bloch EM, Munoz B, Mrango Z, Weaver J, Mboera LEG, Lietman TM, et al. The impact on malaria of biannual treatment with azithromycin in children age less than 5 years: a prospective study. *Malar J.* 2019;18:284.

114. Shanks GD, Quang HH, Chavchich M, Thanh NX, Edstein MD, Trung TN, et al. In Vivo Efficacy and Tolerability of Artesunate–Azithromycin for the Treatment of *Falciparum* Malaria in Vietnam. *The American Journal of Tropical Medicine and Hygiene.* 2016;95:164–7.

115. Gbedo YG. Les techniques Monte Carlo par chaînes de Markov appliquées à la détermination des distributions de partons. Université Grenoble Alpes; 2016. <https://tel.archives-ouvertes.fr/tel-01743752>. Accessed 22 Aug 2019.

116. Université de Cambridge. MRC Biostatistics Unit. 2009. <https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/>. Accessed 22 Aug 2019.

117. Sturtz S, Ligges U, Gelman A. “R2WinBUGS” : A Package for Running WinBUGS from R. *J Stat Soft.* 2005;12. doi:10.18637/jss.v012.i03.

118. Martin AD, Quinn KM, Park JH. MCMCpack : Markov Chain Monte Carlo in R. J Stat Soft. 2011;42. doi:10.18637/jss.v042.i09.
119. Hadfield JD. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. J Stat Soft. 2010;33. doi:10.18637/jss.v033.i02.
120. Furrer R, Sain SR. spam : A Sparse Matrix R Package with Emphasis on MCMC Methods for Gaussian Markov Random Fields. J Stat Soft. 2010;36. doi:10.18637/jss.v036.i10.
121. Baio G. An introduction to INLA with a comparison to JAGS. 2013.  
<http://www.statistica.it/gianluca/Talks/INLA.pdf>.
122. Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, Sallah K, et al. Using Mobile Phone Data to Predict the Spatial Spread of Cholera. Sci Rep. 2015;5:8923.
123. Sallah K, Giorgi R, Bengtsson L, Lu X, Wetter E, Adrien P, et al. Mathematical models for predicting human mobility in the context of infectious disease spread: introducing the impedance model. Int J Health Geogr. 2017;16:42.

## Liste des acronymes

- CoxPH : Cox Proportional Hazard Model (Modèle de Cox à risques proportionnels)
- GAM : Generalized Additive Model (Modèle Additif Généralisé)
- SPDE : Stochastic Partial Differential Equation (Equation aux Dérivées Partielles Stochastiques)
- GF : Gaussian Field (champ gaussien)
- Cox-SPDE : Cox-Stochastic Partial Differential Equation model (Modèle Cox-Equation aux Dérivées Partielles Stochastiques)
- P-SPDE : Poisson-Stochastic Partial Differential Equation model (Modèle Poisson-Equation aux Dérivées Partielles Stochastiques)
- SPDE-AR : Stochastic Partial Differential Equation-Autoregressive model (Modèle d'Equation aux Dérivées Partielles Stochastiques avec évènements récurrents autorégressifs)
- SPDE-NS : Stochastic Partial Differential Equation-No structured model (Modèle d'Equation aux Dérivées Partielles Stochastiques avec évènements récurrents non structurés)
- AG : Andersen-Gill
- AIC : Akaike Information Criterion (Critère d'Information d' Akaike)
- WAIC : Watanabé-Akaike Information Criterion
- DIC : Deviance Information Criterion (Critère d'Information de la Déviance)
- MSE : Mean Squared Error (Erreur Quadratique Moyenne)
- INLA : Integrated Nested Laplace Approximation
- PPI : Processus Ponctuel de Poisson Inhomogène
- PPH : Processus Ponctuel de Poisson Homogène
- CR : Coverage Rate (Taux de couverture)
- SR : Significance Rate (Taux de significativité)
- DGM : Data-Generating Model (Modèle de génération des données)
- IC : Intervalle de Confiance (Intervalle de Crédibilité pour les méthodes bayésiennes)

## Notations

$\tau$  : densité de la population :

$D_g$  : densité de gîte

$RR_g$  : risque relatif associé aux gîtes

$RR_a$  : risque relatif associé à l'âge

$RR_t$  : risque relatif associé au traitement

$RR_s$  : risque relatif associé au sexe

$\gamma$  : paramètre du risque de base dans la simulation (Weibull)

$n$  : taille de l'échantillon utilisé dans la simulation

$r$  : le rayon des gîtes

$\Omega$  : zone de l'étude dans la simulation

$L_1, L_2, L_3$  : points de concentration de la population

$C(\cdot)$  : constante de normalisation de la densité de population

$d(\cdot, \cdot)$  : distance euclidienne entre deux points d'observation

$\lambda_0$  : densité moyenne de la population dans la zone d'étude

$n_g$  : nombre de gîtes dans la zone de l'étude

$A_g$  : aire de la zone de l'étude couverte par les gîtes

$A_T$  : aire totale de la zone de l'étude pour la simulation

$\mathcal{B}(\cdot, \cdot)$  : loi binomiale

$\mathcal{U}(\cdot, \cdot)$  : loi uniforme

$\mathcal{E}(\cdot)$  : loi exponentielle

$\mathcal{W}(\cdot, \cdot)$  : loi de Weibull

$\alpha$  : paramètre de la loi du temps de censure (exponentielle)

$T_{event}$  : temps d'évènement sans censure

$T_{cens}$  : temps de censure

$T$  : temps d'observation (temps de survie)

$\lambda(.,.)$ : fonction de risqué instantané de la survie

$\lambda_0(.,.)$ : fonction de risqué de base

$f(.,.)$ : fonction spline bivariée

$Y$ : variable réponse quelconque

$X$ : vecteur de covariables

$\beta$ : vecteur des effets associés aux covariables

$Z$ : Effet spatial aléatoire (champ gaussien)

$\Sigma$ : Fonction de covariance de Matèrn

$\theta_1, \theta_2$ : paramètre d'échelle et de forme de la fonction de covariance de Matèrn

$\mathbb{P}$ : loi de variable dépendante (temps de survie, Weibull)

$\mu, \phi$ : paramètre d'échelle et forme de la fonction de la loi du temps de survie (Weibull)

$h$ : est la fonction de lien canonique

$V$ : ordre des évènements

$\mathcal{L}$ : loi de l'ordre dans la succession des évènements chez le même individu

$\omega$ : vecteur de paramètres de la loi de l'ordre des évènements

$\psi$ : paramètre de nuisance lié à la corrélation entre les évènements

## Liste des figures

Figure 1: Plan de simulation des différents scénarios.

Figure 2: Structure des données simulées (la taille des points est proportionnelle au temps de survie)

Figure 3: Biais de l'effet traitement avec un risque de base de 0.37

DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base

Figure 4: Taux de couverture de l'effet traitement avec un risque de base de 0.37

DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base

Figure 5: Taux de significativité de l'effet traitement avec un risqué de base de 0.37

DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg:

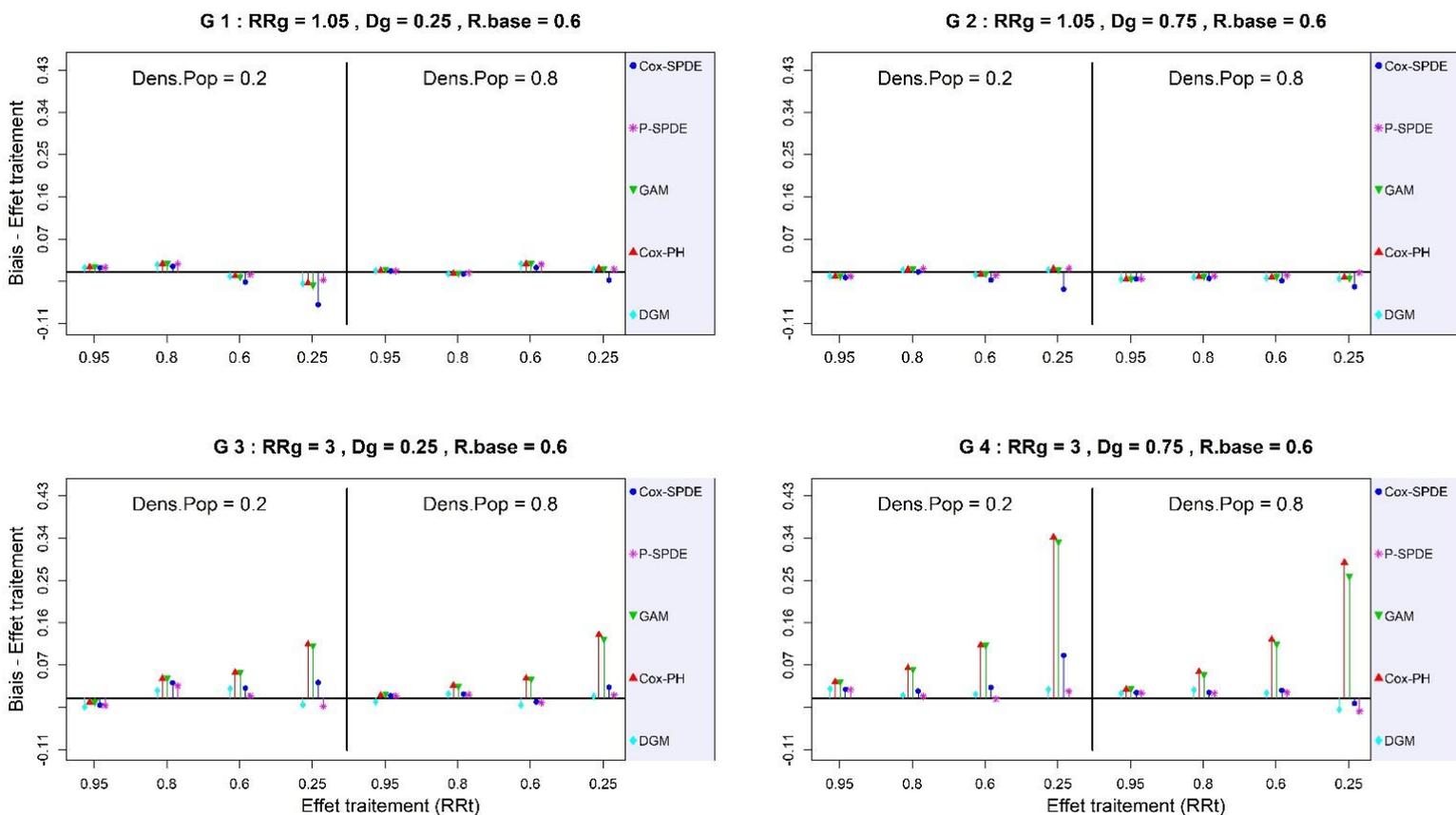
Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base

Figure 6 : Estimation de la distribution de Khi-deux par la méthode INLA

Figure 7 : Maillage par triangulation selon différentes combinaisons de valeurs des paramètres.

## Annexes

Figure A.1: Biais de l'effet traitement avec un risque de base de 0.60 pour 500 simulations



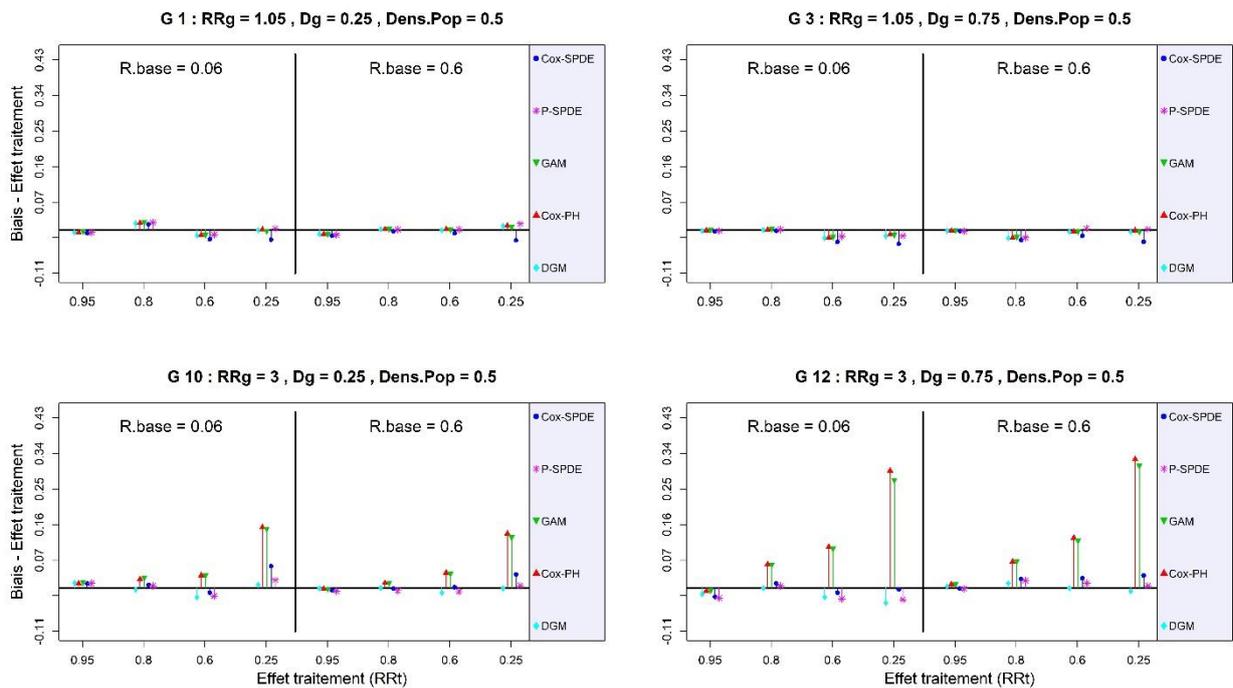
DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base

**Annexe I:** Indicateurs de performance du modèle DGM dans l'estimation de l'effet des différents facteurs avec un risque de base de 0.37.

RRg	RRt	facteurs	Effet gîte																
			0.25								0.75								
			Densité de population								Densité de population								
			0.2				0.8				0.2				0.8				
			Biais	MSE	CR	SR	Biais	MSE	CR	SR	Biais	MSE	CR	SR	Biais	MSE	CR	SR	
1.05	0.95	Age	0.001	0.000	0.98	1.00	0.000	0.000	0.90	1.00	-0.002	0.000	0.98	1.00	-0.001	0.000	0.96	1.00	
		Sexe	-0.010	0.006	1.00	0.00	-0.022	0.006	0.88	0.12	-0.012	0.006	0.96	0.04	0.005	0.006	0.92	0.08	
		Traitement	0.015	0.006	0.96	0.06	0.014	0.006	0.84	0.06	0.015	0.006	0.92	0.06	0.000	0.006	0.96	0.04	
		Gîte	0.011	0.006	0.94	0.12	-0.011	0.006	0.92	0.14	0.006	0.002	1.00	0.26	0.007	0.002	0.94	0.26	
	0.80	Age	0.000	0.000	0.96	1.00	-0.002	0.000	0.96	1.00	-0.001	0.000	0.96	1.00	-0.003	0.000	0.84	1.00	
		Sexe	-0.010	0.006	0.96	0.04	-0.001	0.006	0.96	0.04	0.012	0.006	0.96	0.04	0.012	0.006	0.9	0.10	
		Traitement	-0.012	0.006	0.94	0.74	-0.005	0.006	0.98	0.92	-0.018	0.006	0.96	0.92	0.013	0.006	0.96	0.76	
		Gîte	0.007	0.006	0.92	0.10	0.019	0.007	0.92	0.10	-0.010	0.002	0.98	0.10	-0.004	0.002	0.96	0.16	
	0.60	Age	0.001	0.000	0.90	1.00	-0.001	0.000	0.96	1.00	-0.002	0.000	0.96	1.00	0.000	0.000	0.96	1.00	
		Sexe	-0.024	0.006	0.98	0.02	0.004	0.006	0.96	0.04	0.002	0.006	0.92	0.08	0.001	0.006	0.94	0.06	
		Traitement	-0.004	0.006	0.88	1.00	-0.006	0.006	0.96	1.00	-0.016	0.006	0.94	1.00	-0.008	0.006	0.94	1.00	
		Gîte	0.008	0.006	0.98	0.14	-0.005	0.007	0.96	0.12	-0.004	0.002	0.96	0.16	-0.002	0.002	0.94	0.18	
	0.25	Age	0.000	0.000	0.98	1.00	-0.001	0.000	0.98	1.00	-0.001	0.000	0.94	1.00	0.000	0.000	0.94	1.00	
		Sexe	-0.005	0.006	0.96	0.04	0.007	0.006	0.98	0.02	-0.005	0.006	0.94	0.06	0.001	0.006	0.94	0.06	
		Traitement	0.007	0.007	0.98	1.00	-0.014	0.008	0.98	1.00	-0.020	0.008	0.96	1.00	0.011	0.008	0.9	1.00	
		Gîte	0.013	0.006	0.98	0.08	-0.008	0.007	0.90	0.12	-0.002	0.002	0.94	0.18	-0.005	0.002	0.92	0.24	
	3	0.95	Age	0.001	0.000	0.96	1.00	-0.001	0.000	0.96	1.00	-0.001	0.000	0.94	1.00	0.000	0.000	0.9	1.00
			Sexe	-0.001	0.006	1.00	0.00	-0.011	0.006	0.92	0.08	-0.027	0.007	0.92	0.08	-0.001	0.006	0.94	0.06
			Traitement	0.001	0.006	1.00	0.04	-0.006	0.006	0.94	0.10	0.018	0.006	0.96	0.10	0.004	0.006	0.94	0.06
			Gîte	0.014	0.007	0.88	1.00	-0.012	0.007	0.94	1.00	0.001	0.003	0.94	1.00	0.005	0.003	0.92	1.00
0.80		Age	-0.001	0.000	0.88	1.00	0.001	0.000	0.92	1.00	-0.002	0.000	0.94	1.00	-0.001	0.000	1.00	1.00	
		Sexe	0.006	0.006	0.94	0.06	-0.002	0.006	0.98	0.02	-0.009	0.006	0.94	0.06	0.001	0.006	0.94	0.06	
		Traitement	-0.018	0.006	0.96	0.88	-0.007	0.006	0.98	0.86	-0.008	0.006	0.94	0.88	0.008	0.006	0.96	0.72	
		Gîte	0.017	0.007	0.98	1.00	0.021	0.007	0.96	1.00	0.016	0.003	0.90	1.00	0.008	0.003	0.96	1.00	
0.60		Age	0.000	0.000	0.96	1.00	-0.001	0.000	1.00	1.00	0.000	0.000	0.96	1.00	-0.002	0.000	0.96	1.00	
		Sexe	-0.014	0.006	0.96	0.04	-0.002	0.006	0.96	0.04	-0.007	0.006	0.90	0.10	0.001	0.006	0.96	0.04	
		Traitement	-0.011	0.006	0.98	1.00	0.023	0.007	0.92	1.00	0.004	0.006	0.90	1.00	-0.006	0.006	0.94	1.00	
		Gîte	0.004	0.007	0.96	1.00	0.012	0.007	1.00	1.00	0.006	0.003	0.98	1.00	0.007	0.003	0.96	1.00	
0.25		Age	0.000	0.000	0.98	1.00	-0.002	0.000	0.98	1.00	0.000	0.000	0.98	1.00	-0.001	0.000	0.96	1.00	
		Sexe	-0.016	0.006	0.92	0.08	-0.005	0.006	0.96	0.04	0.009	0.006	0.98	0.02	0.000	0.006	0.96	0.04	
		Traitement	-0.012	0.008	1.00	1.00	-0.007	0.007	0.98	1.00	-0.001	0.007	0.96	1.00	-0.007	0.007	0.98	1.00	
		Gîte	0.003	0.007	0.96	1.00	0.022	0.007	0.98	1.00	0.005	0.003	0.98	1.00	0.008	0.003	0.96	1.00	

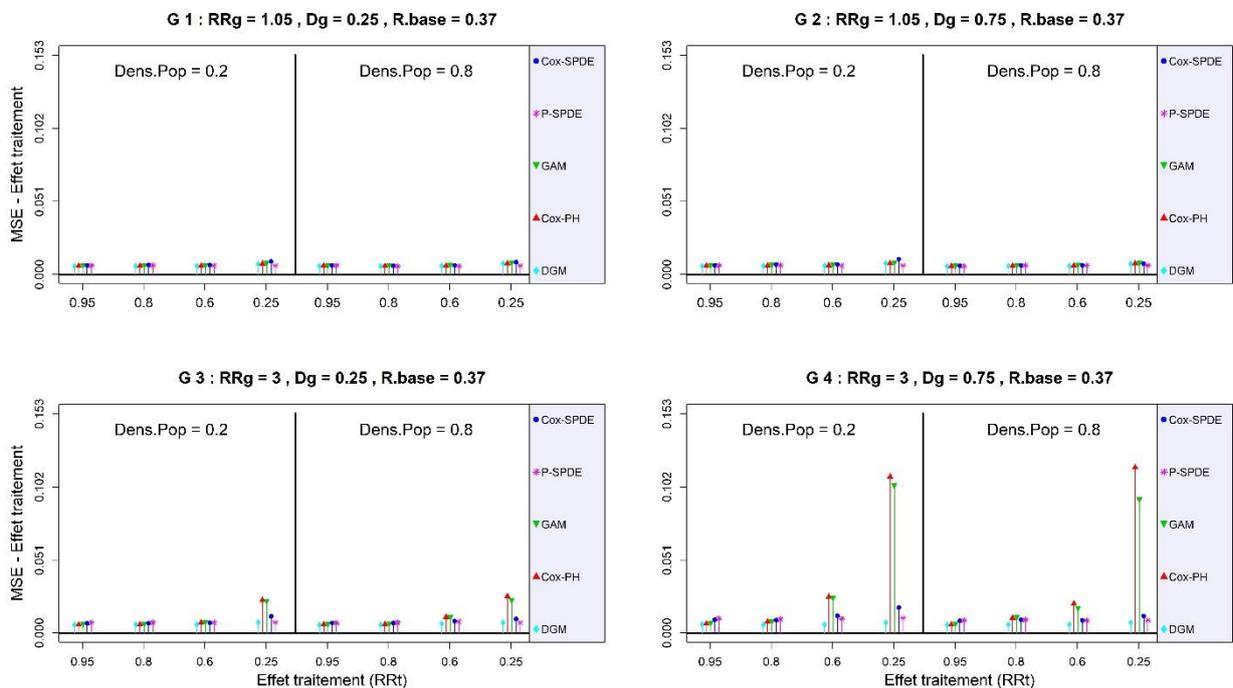
**CR:** Coverage Rate (Taux de couverture), **SR:** Significance Rate (Taux de significativité), **RRt:** Treatment Relative Risk (Risque Relatif associé au traitement), **RRg:** Breeding site Relative Risk (Risque Relatif associé aux gîtes).

**Figure A.2:** Biases de l'effet traitement avec une densité de population de 0.5



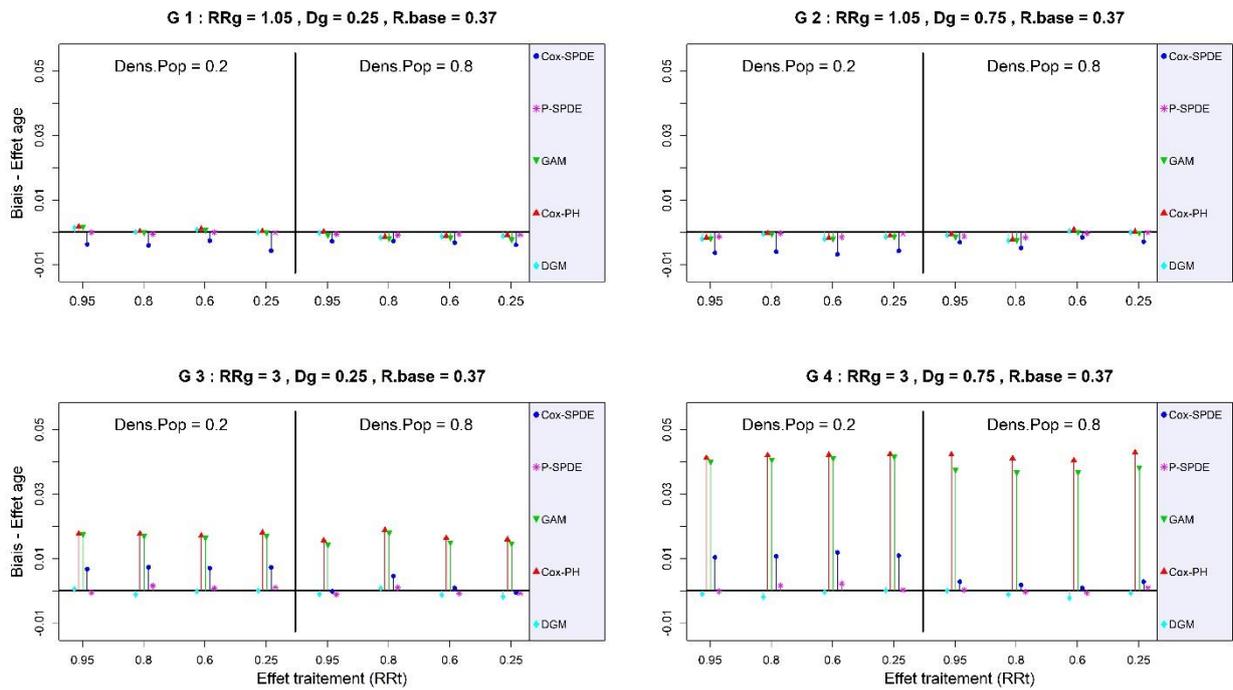
DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base

**Figure A.3:** MSE de l'effet traitement avec un risque de base de 0.37



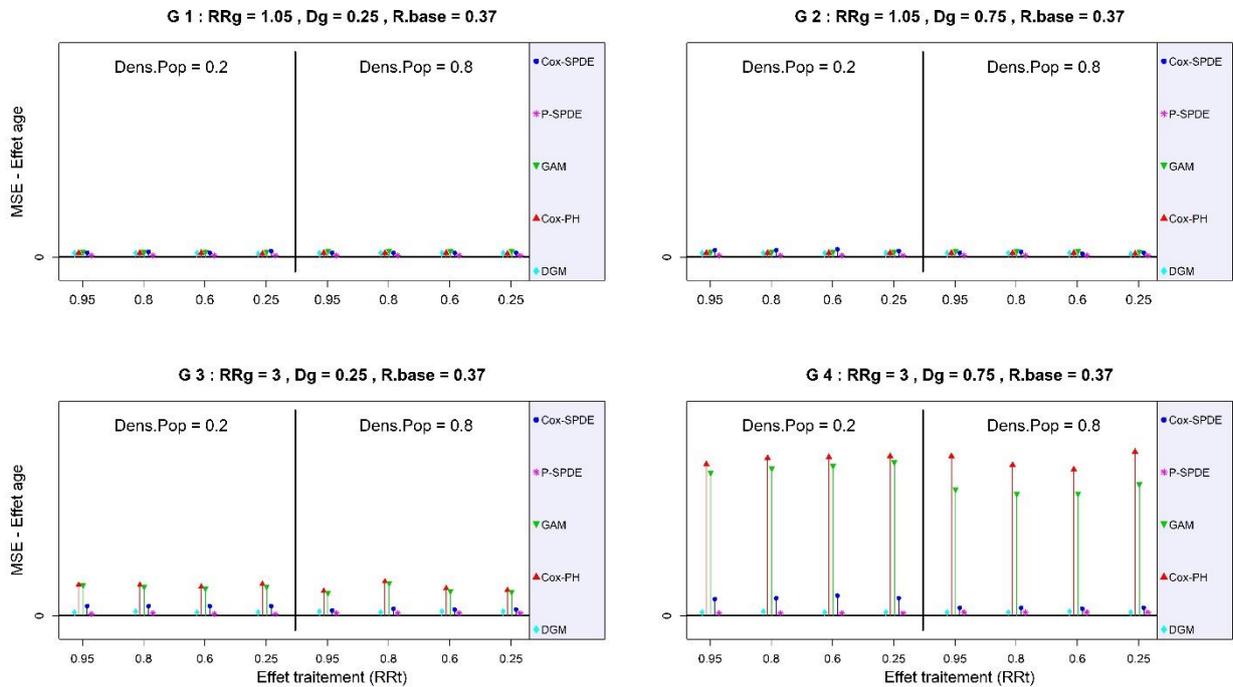
DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base

**Figure A.4:** Biais de l'effet âge avec un risque de base de 0.37



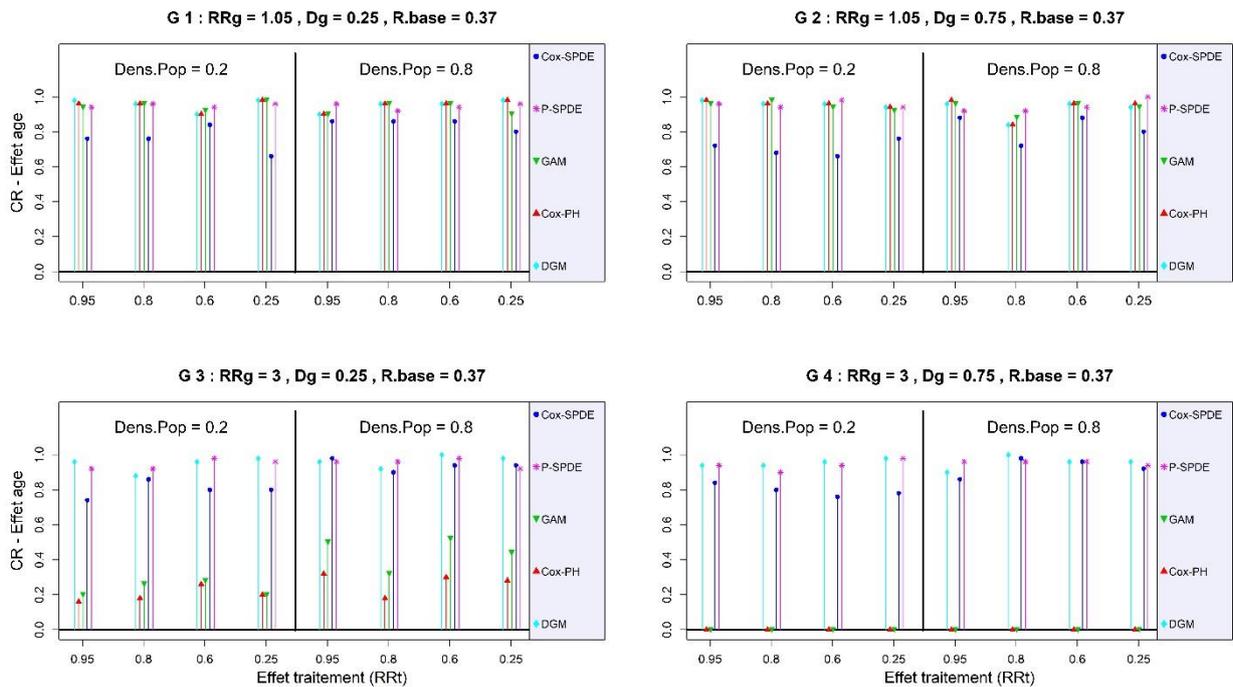
DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base

**Figure A.5:** MSE de l'effet âge avec un risque de base de 0.37



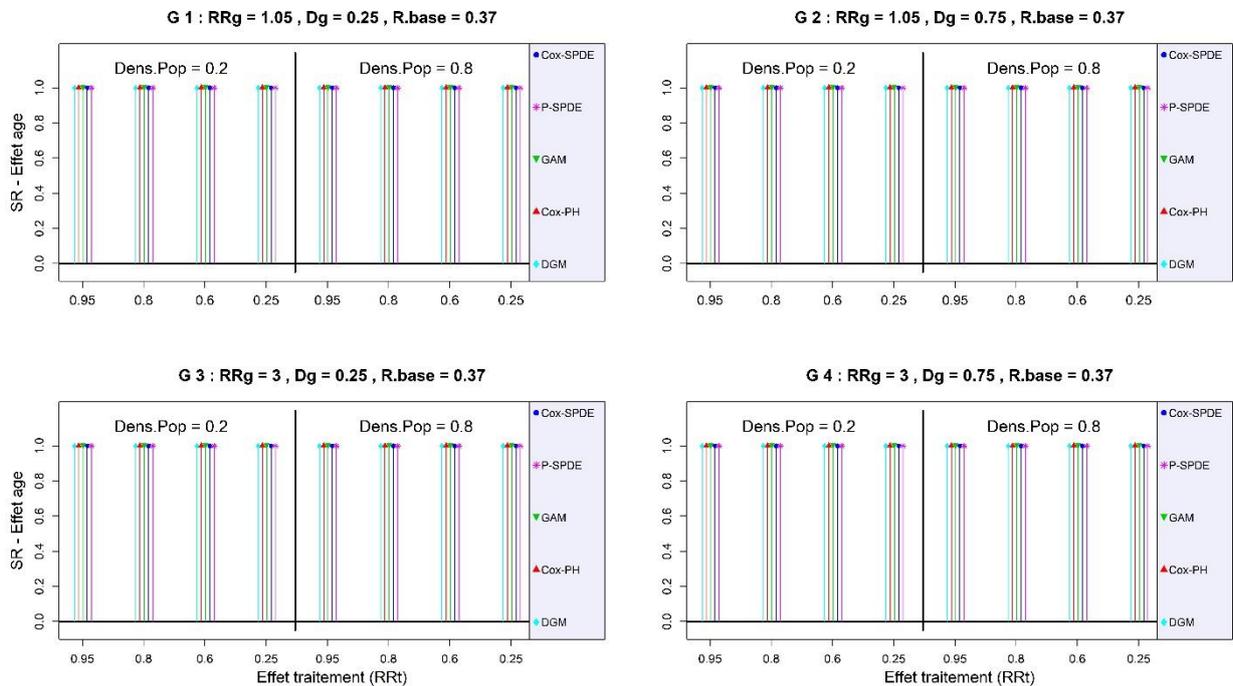
DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base

**Figure A.6:** Taux de couverture de l'effet âge avec un risque de base de 0.37



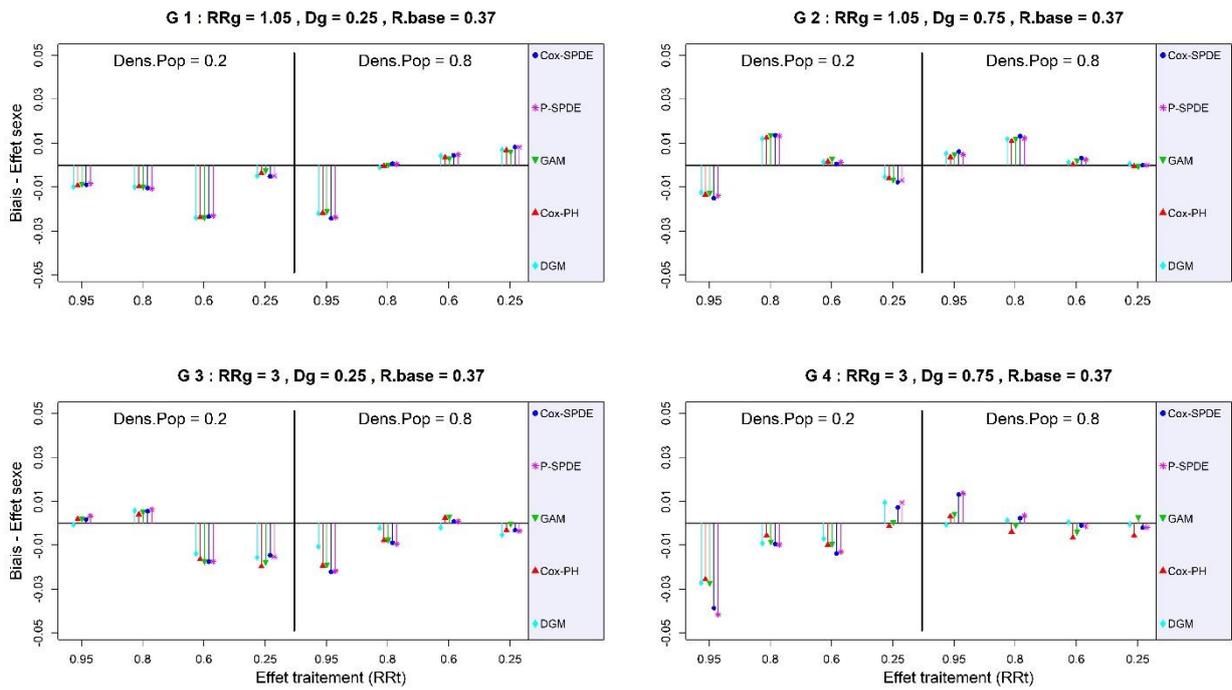
*DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base*

**Figure A.7:** Taux de significativité de l'effet âge avec un risque de base de 0.37



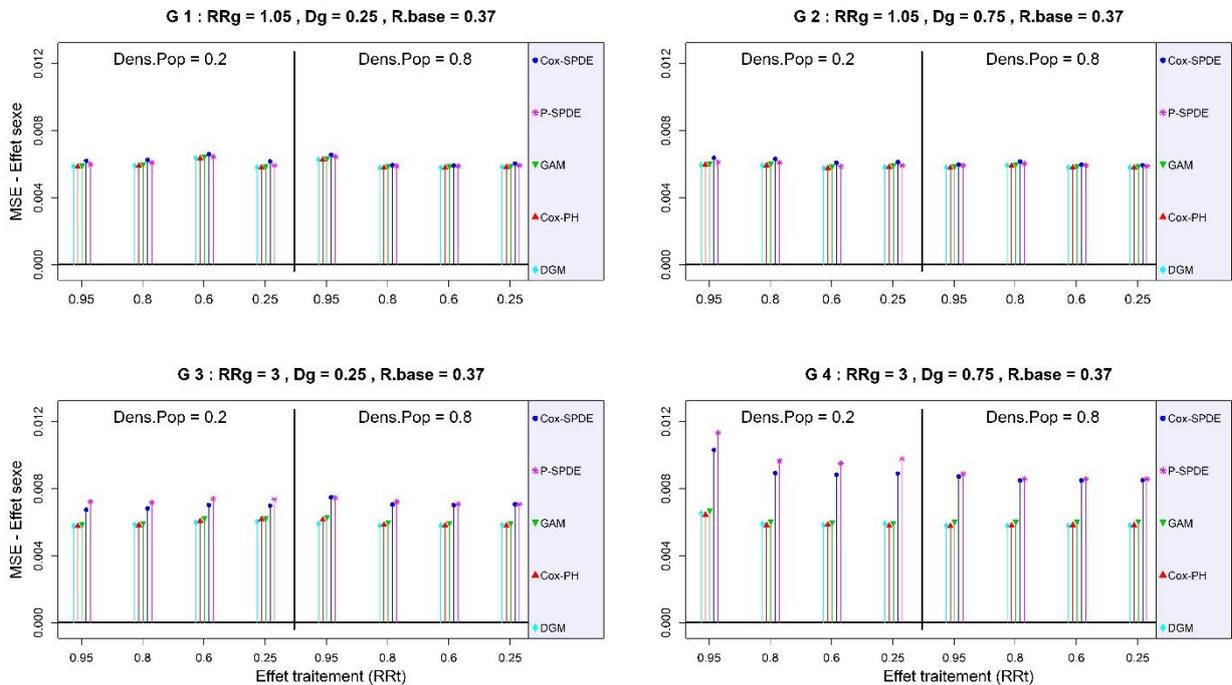
*DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base*

**Figure A.8:** Biais de l'effet sexe avec un risque de base de 0.37



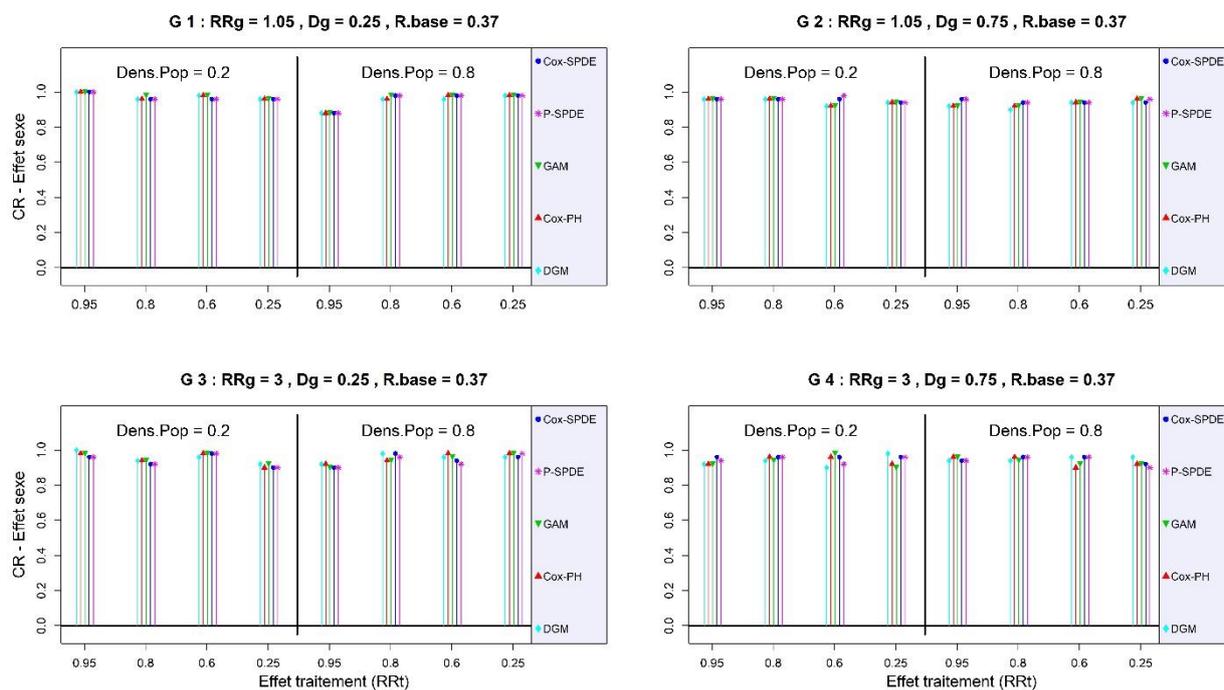
*DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base*

**Figure A.9:** MSE de l'effet sexe avec un risque de base de 0.37



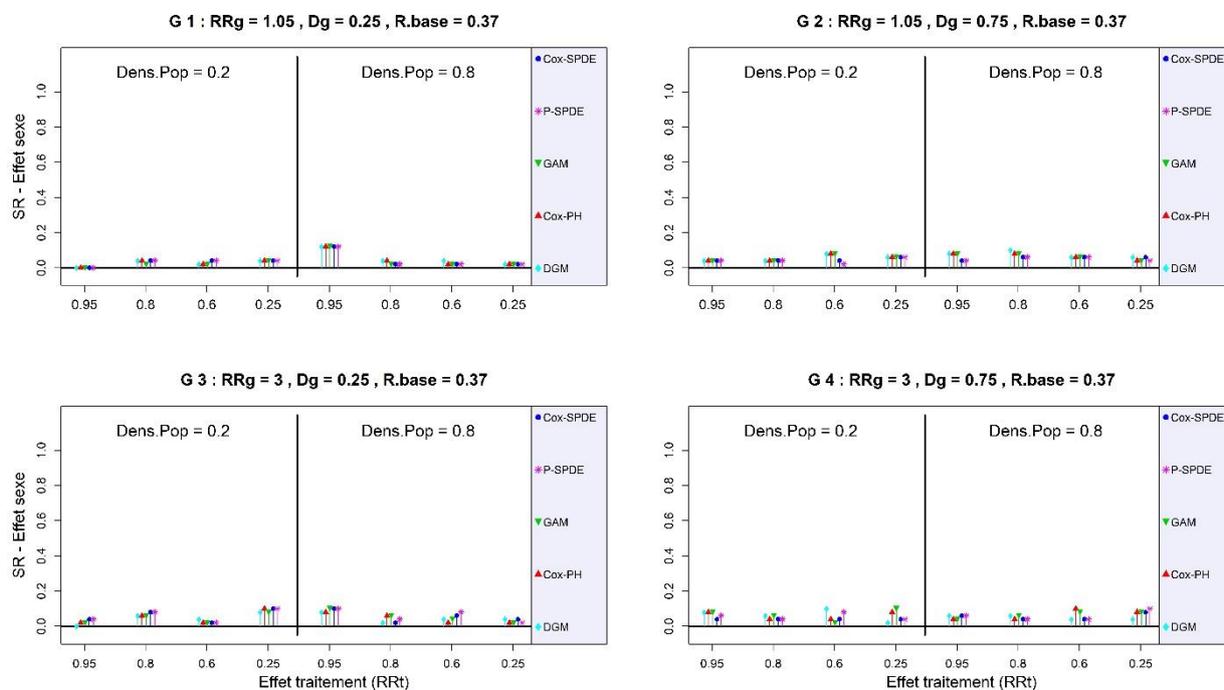
*DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base*

**Figure A.10:** Taux de couverture de l'effet sexe avec un risque de base de 0.37



*DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base*

**Figure A.11:** Taux de significativité de l'effet sexe avec un risque de base de 0.37



*DGM: Data-Generating Model, Cox-PH: Cox Proportional Hazard model, GAM: Generalized Additive Model, Cox-SPDE: Cox Stochastic Partial Differential Equation Model, P-SPDE: Poisson Stochastic Partial Differential Equation, RRg: Risque Relatif associé aux gîtes, Dg: Densité de gîtes, RRt: Risque Relatif associé au traitement, Dens.Pop: Densité de population, R.base: Risque de base*

## **Annexe 2** : Code R pour l'exécution du modèle SPDE avec le INLA

```
DataBase = read.csv("C:\\\\DataBase.txt", sep = "\\t", dec = ".", header = TRUE)

## Pour prendre en compte de possible effet de bords

location <- cbind(DataBase$xcoord, DataBase$ycoord)

boundary1 <- inla.nonconvex.hull(location, convex = 0.7, resolution = 100)
boundary2 <- inla.nonconvex.hull(location, convex = 2.5, resolution = 100)

## Triangulation

mesh <- inla.mesh.2d(location, boundary = list(boundary1, boundary2), max.edge = c(0.6, 1.2))

## Creation de la matrice de projection

A <- inla.spde.make.A(mesh, location)

## Creation du modèle de SPDE

spde <- inla.spde2.matern(mesh, alpha = 2)

## Assemblage de l'intercept, des covariables et de l'effet spatial dans un seul objet

stk <- inla.stack(data = list(times = DataBase$times/52, status = DataBase$status),
                  A = list(A, 1), effect = list( list(spatial = 1:spde$n.spde),
                  data.frame(a0 = 1, DataBase[,c("identifiant", "age", "sexe", "traitement")])))

## Modèle SPDE à PH avec une loi paramétrique (weibull)

# D'abord, la formule du modèle, incluant l'intercept, les covariables et le modèle de SPDE

formula <- inla.surv(times, status) ~ 0 + a0 + age + sexe + traitement + f(spatial, model = spde)
```

*#### Modèle SPDE avec le INLA en utilisant une loi de weibull (modèle paramétrique)*

```
ModelSPDE <- inla(formula,  
  family = "weibullsurv",  
  data = inla.stack.data(stk),  
  control.predictor = list(A=inla.stack.A(stk)),  
  control.compute = list(dic=TRUE,mlik=TRUE),  
  verbose = TRUE,  
  control.inla = list(strategy = "laplace",h=1e-15))
```

*## Sortie du modèle*

```
summary(ModelSPDE)
```

## **Intitulés des doctorats AMU**

Mentions et Spécialités des doctorats votées en CS le 16/10/2012

### **ED 62 – SCIENCES DE LA VIE ET DE LA SANTE**

- Biologie
  - Biochimie structurale
  - Génomique et Bio-informatique
  - Biologie du développement
  - Immunologie
  - Génétique
  - Microbiologie
  - Biologie végétale
- Neurosciences
- Pathologie humaine
  - Oncologie
  - Maladies infectieuses
  - Génétique humaine
  - Conseil en Génétique
  - Pathologie vasculaire et nutrition
  - Ethique
  - Recherche clinique et Santé Publique

### **ED 67 – SCIENCES JURIDIQUES ET POLITIQUES**

- Droit privé
- Droit public

- Histoire du droit
- Droit
- Science politique

#### **ED 184 – MATHEMATIQUES ET INFORMATIQUE**

- Mathématiques
- Informatique
- Automatique

#### **ED 250 – SCIENCES CHIMIQUES DE MARSEILLE**

- Sciences chimiques

#### **ED 251 – SCIENCES DE L'ENVIRONNEMENT**

- Anthropologie biologique
- Ecologie
- Géosciences de l'environnement
- Génie des procédés
- Océanographie
- Chimie de l'environnement

#### **ED 352 – PHYSIQUE ET SCIENCES DE LA MATIERE**

- Astrophysique et Cosmologie
- Biophysique
- Energie, Rayonnement et Plasma
- Instrumentation

- Optique, Photonique et Traitement d'Image
- Physique des Particules et Astroparticules
- Physique Théorique et Mathématique
- Matière Condensée et Nanosciences

### **ED 353 – SCIENCES POUR L'INGENIEUR : MECANIQUE, PHYSIQUE, MICRO ET NANO ELECTRONIQUE**

- Energétique
- Mécanique et Physique des Fluides
- Acoustique
- Mécanique des Solides
- Micro et Nanoélectronique
- Génie Civil et Architecture

### **ED 354 – LANGUES, LETTRES ET ARTS**

- Etudes anglophones
- Etudes germaniques
- Etudes slaves
- Langue et littérature chinoises
- Langue et Littérature françaises
- Littérature générale et comparée
- Arts plastiques et sciences de l'Art
- Musicologie
- Etudes cinématographiques
- Arts du spectacle

### **ED 355 – ESPACES, CULTURES, SOCIETES**

- Géographie
- Urbanisme et Aménagement du territoire
- Préhistoire
- Archéologie
- Histoire de l'Art
- Histoire
- Sciences de l'Antiquité
- Mondes arabe, musulman et sémitique
- Etudes romanes
- Sociologie
- Anthropologie
- Architecture

### **ED 356 – COGNITION, LANGAGE, EDUCATION**

- Philosophie
- Psychologie
- Sciences du Langage
- Sciences de l'Information et de la Communication
- Sciences de l'Education

### **ED 372 – SCIENCES ECONOMIQUES ET DE GESTION**

- Sciences de Gestion
- Sciences Economiques

- Sciences Economiques : AMSE

### **ED 463 – SCIENCES DU MOUVEMENT HUMAIN**

- Sciences du Mouvement Humain
- Biomécanique
- Contrôle Perceptivo-Moteur et Apprentissage
- Physiologie de l'exercice
- Sciences de l'Homme et de la Société