Correction de l'effet du biais d'échantillonnage dans la modélisation de la qualité des habitats écologiques : application au principal vecteur du paludisme en Guyane française

Yi Moua

Université de Guyane
École doctorale 587 "Diversités, santé et développement en Amazonie"

THÈSE

pour obtenir le grade de
Docteur de l'université de Guyane

Correction de l'effet du biais d'échantillonnage dans la modélisation de la qualité des habitats écologiques – Application au principal vecteur du paludisme en Guyane française

par

Yi Moua

soutenue publiquement le 10 février 2017

Jury

M.	**Jean Gaudart**	Professeur, Université d'Aix-Marseille	Rapporteur
M^me^	**Sandra Luque**	Directrice de Recherche, IRSTEA	Rapporteure
M.	**Carlo Costantini**	Directeur de Recherche, IRD	Examinateur
M^me^	**Florence Fournet**	Chargée de Recherche, IRD	Examinatrice
M.	**Sébastien Briolant**	Médecin en chef, IRBA	Co-encadrant
M.	**Emmanuel Roux**	Chargé de Recherche, IRD	Co-encadrant
M^me^	**Frédérique Seyler**	Directrice de Recherche, IRD	Directrice de thèse

Financement de la thèse :

Partenariats de la thèse :

Participartion financière :

— le projet GAPAM-Sentinela et
— le projet DS BIODIVA.

Cette thèse a été réalisée au sein de l'Unité Mixte de Recherche (UMR 228) Espace Dev.

Remerciements

Je tiens à remercier Sandra Luque et Jean Gaudart d'avoir accepté d'être rapporteurs de ce travail de thèse. Je remercie également Florence Fournet et Carlo Costantini d'avoir d'examiné ce travail.
Je voudrais remercier Frédérique d'avoir dirigé cette thèse. Merci à Sébastien pour son aide et ses conseils. Je remercie Emmanuel, sans qui cette thèse n'aurait pas pu être réalisée. Merci, non seulement pour son encadrement génial mais, également pour ses encouragements même si ce n'était pas toujours facile.
Je remercie Isabelle Dusfour et Romain Girod pour le temps qu'ils m'ont consacré ainsi que toutes les discussions autour des anophèles. Je remercie également Benoit de Thoisy pour les nombreux conseils qu'il a apporté. Un grand merci à Antoine Adde pour sa bonne humeur et les fous rires.
Un grand merci à Abdennebi, pour le temps qu'il m'a consacré et pour son aide précieuse dans les démarches administratives bien avant que je ne commence cette thèse. Merci pour ses encouragements tout au long de ce travail.
Un grand merci à mes collègues et amis de la MTD : Cláudio, Christian, Mojdeh, Eva, Alexandre, Savio, Eudes, Zhi Chao, Phuong, Mojdeh, avec qui j'ai partagé non seulement de merveilleux moments mais également des kilos de café. Merci aux chercheurs avec qui j'ai pu échangé. Merci à Sylvie et à Agnès de m'avoir évité l'abîme de l'administration.
Merci à l'équipe de l'IRD de Cayenne : Christophe, Pape, Rosiane, Marie-Claude, Jeannine, Olivier, Justine, Tommy et Max pour ses marinades de crevettes et ses crèmes de cupuaçu qui illuminaient mes journées de thésard. Merci à Youven pour tous les échanges méthodologiques, les astuces culinaires et les encouragements.
Merci aux amis qui m'ont accompagné de près ou de loin, merci à Minh et Pierre, à Suab Nkauj et Rémy ; à mes colloc' : Margaux, Nico et Baptiste pour les soirées Wikipédia. Merci à Brigitte, à Philippe, à David, à Youssef et à Gabrielle, pour leur gentillesse.
Je remercie ma famille pour son écoute et ses encouragements.
Merci à Adrien, d'avoir été présent et de m'avoir soutenu même dans les moments les plus difficiles.
Et ma plus grande reconnaissance va envers mes parents, qui n'ont jamais cessé de m'encourager.

À Pog.

Table des matières

Table des figures **xiii**

Liste des tableaux **xvii**

Introduction générale **1**

1 État de l'art **7**
I. Modèles de distribution d'espèces . . . 9
I. 1. Généralités . . . 9
I. 2. Les modèles de présence-absence . . . 11
I. 3. Les modèles de présence uniquement . . . 14
I. 4. Les méthodes de présence-background . . . 15
I. 5. Les méthodes de présence-pseudo-absence . . . 16
I. 6. Avantages et limites des différents types de modèles de distribution d'espèces . . . 16
II. Maxent . . . 17
II. 1. Maximum d'entropie . . . 17
II. 2. Implémentation de Maxent . . . 20
II. 3. Avantages et faiblesses de Maxent . . . 24
III. Biais d'échantillonnage . . . 24
IV. Correction de l'effet du biais d'échantillonnage . . . 25
IV. 1. Sélection des données de présence . . . 26
IV. 2. Construction d'un background biaisé . . . 27
IV. 3. Comparaison des méthodes de correction de l'effet du biais d'échantillonnage . . . 28
IV. 4. Avantages et limites des méthodes de correction . . . 29
V. Conclusion . . . 30

2 Méthode de correction de l'effet du biais d'échantillonnage basée sur des critères environnementaux **35**
Introduction . . . 37
I. Description de la méthode de correction de l'effet du biais d'échantillonnage . . . 37
I. 1. Définition de l'espace environnemental . . . 37
I. 2. Définition du voisinage environnemental . . . 38
I. 3. Définition du biais d'échantillonnage . . . 38
I. 4. Sélection des sites de background biaisés . . . 40
II. Discussion . . . 40
III. Conclusion . . . 41

3 Modélisation de la distribution du principal vecteur du paludisme en Guyane française **43**
Introduction . . . 45
I. Le paludisme . . . 45
II. Les *Anopheles* . . . 46
II. 1. Biologie des *Anopheles* . . . 46
II. 2. Comportement trophique des adultes *Anopheles* . . . 47

II. 3. Les vecteurs du paludisme 48
III. La Guyane française 49
III. 1. Géographie et météorologie de la Guyane 49
III. 2. La population guyanaise 51
IV. Le paludisme en Guyane 51
IV. 1. Les acteurs de la surveillance du paludisme et de la lutte antivectorielle et les organismes de recherche 52
IV. 2. Historique du paludisme en Guyane 56
IV. 3. Le paludisme actuellement 56
IV. 4. Les vecteurs du paludisme 58
V. Données de présence des *Anopheles* en Guyane française 61
V. 1. Méthodes de capture d'*Anopheles* 62
V. 2. Données de captures d'*Anopheles* 63
V. 3. Base de données des captures d'*Anopheles* en Guyane 64
VI. Modélisation de la distribution d'*Anopheles darlingi* 67
VI. 1. Données environnementales 67
VI. 2. Caractérisation environnementale 70
VI. 3. Données météorologiques 73
VI. 4. Construction des modèles 75
VII. Résultats 77
VII. 1. Performances de prédiction et contribution des variables environnementales et évaluation 77
VII. 2. Carte de la qualité d'habitat 80
VIII. Discussion 83
VIII. 1. Lien entre les facteurs environnementaux et la qualité d'habitat 83
VIII. 2. Correction de l'effet du biais d'échantillonnage 84
VIII. 3. Le paludisme et la qualité d'habitat d'*An. darlingi* en Guyane française . . 85
VIII. 4. Caractérisation environnementale 86
IX. Conclusion 86

4 Comparaison des méthodes de correction de l'effet du biais d'échantillonnage 89
Introduction 91
I. Méthodologie générale de comparaison 91
II. Simulation des sites de présence 92
II. 1. Simulation des cartes de qualité d'habitat et de présence-absence 92
II. 2. Génération du biais d'échantillonnage 94
II. 3. Sélection des sites de présence 95
III. Correction de l'effet du biais d'échantillonnage 96
III. 1. Sélection des sites de présence basée sur des critères géographiques . . 96
III. 2. Sélection des sites de présence basée sur des critères environnementaux 97
III. 3. Construction d'un background biaisé basé sur des critères géographiques 97
III. 4. Construction d'un background biaisé basé sur des critères environnementaux 97
III. 5. Modélisation de la distribution virtuelle d'*An. darlingi* 98
IV. Évaluation et comparaison des méthodes de correction 98
V. Résultats 101
VI. Discussion 103
VI. 1. Évaluation absolue des méthodes de correction 103
VI. 2. Évaluation relative des méthodes de correction 103
VI. 3. *BGeng vs. BGenv_tg* 104
VI. 4. Paramétrisation de *BGenv* et BGenv_tg 105
VII. Conclusion 105

5 Discussion générale et perspectives **109**
I. Choix du modèle de distribution d'espèce . . . 111
II. Modélisation de la qualité d'habitat d'*An. darlingi* en Guyane . . . 111
II. 1. Précision / incertitude des données de présence . . . 111
II. 2. Résolution des variables environnementales . . . 112
II. 3. Variables météorologiques . . . 112
III. Contributions à la lutte contre le paludisme . . . 114
III. 1. Qualité d'habitat *vs.* risque de transmission . . . 114
III. 2. Qualité d'habitat pour la lutte antivectorielle . . . 115

6 Conclusion générale **117**

Bibliographie **121**

A Requêtes pour l'analyse de publications annuelles **133**

B Données de capture d'*Anopheles* réalisée entre 1902 et 2013 **137**

C Données de capture d'*Anopheles darlingi* précisément géolocalisées entre 2000 et 2013 **143**

D Variables environnementales **147**

E Schéma du principe de fonctionnement de la librairie *virtualspecies* **155**

F Publications et communications **159**
I. Moua Y., Roux E., Seyler F., Girod R., Dusfour I., and Briolant S. Ecological niche modeling for *Anopheles darlingi*, French Guiana. Dans *Amazonian Conference on Emerging and Infectious Diseases*, 26 -28 septembre 2014, Cayenne (France) . . . 161
II. Moua Y., Roux E., Seyler F., and S. Briolant. Sampling bias corrections in Maxent : evaluation of methods. Dans *Mathematical and Computational Epidemiology of Infectious diseases – the interplay between models and public health policies*, 30 août – 5 septembre 2015, Erice (Italie) . . . 163
III. Moua Y., Roux E., Girod R., Dusfour I., de Thoisy B., Seyler F., and Briolant S. Distribution of the habitat suitability of the main malaria vector in French Guiana using Maximum Entropy modeling. *Journal of Medical Entomology*, accepté pour publication. . . . 165

Table des figures

1 **Publications annuelles de document portant sur les modèles de distribution d'espèces répertoriés sur *Web of Science*.** En noir est représenté le nombre de publications annuelles traitant des modèles de distribution d'espèces et en rouge, les publications qui, parmi elles, mentionnent le biais d'échantillonnage. 3

1.1 **Représentation de la calibration des différentes méthodes de présence absences, source : modifié d'après** (Peterson et al., 2011). a) exemple de courbes de réponse des GLM (ligne continue), GAM (ligne en tirets) et MARS (ligne en pointillés) b) exemple d'un arbre de classification où deux variables environnementales, E_1 et E_2, sont partitionnées par des règles t_1, t_2, t_3 et t_4 ; c) exemple d'un réseau neurones ; les neurones sont représentés par les points et les liens pondérés des neurones par les segments. Ce réseau est composé de n neurones d'entrée correspondant aux n variables environnementales E_i ($i \in [1, ..., n]$), de m éléments dérivés sur la couche cachée et d'une sortie. 12

1.2 **Représentation de la méthode SVM, où θ représente la transformation permettant l'existence d'un hyperplan de séparation.** source : Rossi et al. (2015) 14

1.3 **Représentation de la calibration des différentes méthodes de présence-uniquement dans l'espace de deux variables environnementales E_1 et E_2,**(Peterson et al., 2011). Les points représentent les sites de présence, et l'enveloppe environnementale est représentée par : (a) le rectangle pour BIOCLIM, par le polygone convexe pour HABITAT et par l'ellipse pour Mahalanobis ; (b) les lignes pour DOMAIN 15

1.4 **Représentation de l'extrapolation.** La ligne continue noire représente la courbe de réponse générée par le modèle à partir des données d'apprentissage. Au-delà des conditions environnementales représentées par les données d'apprentissage, la ligne en tiret rouge correspond à la situation où l'option *clamp* est choisie, la ligne en tiret vert correspond à la situation où il n'y a pas d'extrapolation, la ligne en tiret bleue correspond à l'extrapolation "fadebyclamping". 22

2.1 **Représentation du voisinage du pixel i dans l'espace environnemental représenté par deux axes factoriels l et k.** Le voisinage environnemental de i est représenté par la fonction d'appartenance de type gaussienne. La droite bleue définit le seuil du degré d'appartenance en deça duquel l'appartenance est jugée non significative. Seul le site j est considéré comme voisin de i dans cet exemple. 39

3.1 **Répartition du paludisme dans le monde, en 2011.** Source : WHO, 2011 . . 45

3.2 **Cycle de transmission et de reproduction de *Plasmodium*.** Source : http://www.dpd.cdc.gov/dpdx. 46

3.3 **Cycle biologique d'*Anopheles*.** Source : Carnevale and Robert (2009) 47

3.4 **Répartition géographique des principaux vecteurs du paludisme dans le monde.** Source : Sinka et al. (2012). 48

3.5 **La Guyane française** 49

3.6 **Carte de la pluviométrie annuelle.** Source : Barret (2001) 50

3.7 **Répartition géographique des différents groupes humains.** Source : Barret (2001) . . . 52
3.8 **Nombre d'habitants en Guyane de 1990 à 2015**, selon les données de l'INSEE 53
3.9 **Système de surveillance du paludisme en Guyane.** . . . 54
3.10 **Nombre de cas de paludisme diagnostiqués en Guyane française entre 2000 à 2014.** Sources : Bulletin de Veille Sanitaire - n°1 / Janvier 2015 – Cire Antilles-Guyane et Chaud et al. (2006) . . . 57
3.11 ***Anopheles darlingi* lors d'un repas de sang.** . . . 59
3.12 **Exemples d'activités humaines** . . . 60
3.13 **Capture sur homme** Crédits : E. Roux . . . 62
3.14 **Piège lumineux** Crédits : E. Roux . . . 63
3.15 **Piège odorant, Mosquito Magnet®.** Crédits : E. Roux . . . 63
3.16 **Exemple de publication issue des Archives de l'Institut Pasteur de la Guyane**, Neveu-Lemaire (1902a) . . . 64
3.17 **Schéma relationnel de la base de données des *Anopheles* en Guyane.** . . . 65
3.18 **Représentation de la base de données de la présence des espèces d'*Anopheles* en Guyane française, de 1901 à 2013.** . . . 67
3.19 **Sites de présence d'*Anopheles* entre 1902 et 1999.** . . . 68
3.20 **Sites de présence d'*Anopheles* entre 2000 et 2013.** . . . 69
3.21 **Répartition des captures d'*Anopheles darlingi* (en rouge) et des captures de Culicidés (sites sur lesquels *An. darlingi* n'a pas été capturé, en vert) ayant des localisations précises, de 2000 à 2013.** . . . 70
3.22 **Répartition de l'ensemble des captures d'*Anopheles darlingi* entre 1980 et 2013.** . . . 75
3.23 **Courbes de réponses des variables catégorielles.** . . . 79
3.24 **Courbes de réponses des variables numériques.** . . . 80
3.25 **Carte de la qualité d'habitat d'*Anopheles darlingi* en Guyane française.** Sept zones caractéristiques du territoire guyanais et présentant des indices de qualité d'habitat élevés (A à F) ou remarquable (G) sont entourées en rouge. . . . 81
3.26 **Zoom sur les grandes zones urbaines.** a), d) et g) : cartes de qualité d'habitat. b), e) et h) : cartes d'occupation du sol. c), f) et i) : pourcentage d'urbanisation dans les pixels voisins. Les rectangles représentent les zones densément urbanisées d'après les critères de la présente étude (la classe de *LS* est *Urban* et *PER_URB_NEIGH* $\geq$ 50 %) . . . 82

4.1 **Schéma général de la simulation pour la comparaison des méthodes de correction de l'effet du biais d'échantillonnage** . . . 93
4.2 **Courbes de réponse des variables continues définies pour la génération des données d'*An. darlingi* virtuelles.** . . . 94
4.3 **Carte de qualité d'habitat simulée.** . . . 95
4.4 **Carte de présence-absence simulée avec une prévalence de 0.2.** . . . 96
4.5 **Schéma des entrées et sorties des méthodes de correction de l'effet du biais d'échantillonnage et des modèles de prédiction dits "corrigés".** . . . 99
4.6 **Calcul de** ΔAUC**.** . . . 99
4.7 **Calcul de** $\Delta Dgeo$**.** . . . 100
4.8 **Calcul de** $\Delta Denv$**, avec** $ni = 500$**.** . . . 101
4.10 **Les boxplots des valeurs des indices d'évaluation des différentes méthodes de correction de l'effet du biais d'échantillonnage.** . . . 102
4.11 **Rang moyen $\pm$ l'erreur-type de chaque méthode de correction de l'effet du biais d'échantillonnage.** La méthode la plus performante est rangée à 1 et la moins performante à 5 . . . 102

5.1 **Spatialisation de la pluviométrie journalière moyenne de 2001 à 2012 à partir des produits satellitaires TMPA V7, TMPA RT, PERSIANN et CMORPH et des données *in situ*.** Source : Ringard et al. (2015) . . . 113
5.2 **Carte de l'irradiation solaire global annuelle pour 2012.** Source : Albarelo et al. (2015) . . . 114

D.1 **Pourcentage d'urbanisation dans les pixels voisins (*PER_URB_NEIGH*)** . . 149
D.2 **Longueur des routes et des pistes dans un pixel de 1km^2 (*ROADS*)** 149
D.3 **Altitude minimale (*ATL_min*)** 150
D.4 **Présence et activités humaine altérant de manière non-permanente l'environnement naturel (*HA_max*)** 150
D.5 **Occupation du sol (*LS*)** 151
D.6 **Paysages géomorphologiques (*GLS*)** 152
D.7 **Unités géomorphologiques (*GLF*)** 153

E.1 **Schéma du principe de fonctionnement de la librairie *virtualspecies*** Source : Leroy et al. (2015) 157

Liste des tableaux

1.1 **Détails des transformations de variables, g étant la variable environnementale initiale et f la variable transformée** . . . 31
1.2 **Valeurs de β en fonction des transformations et du nombre de sites de présence.** . . . 32
1.3 **Méthodes de correction de l'effet du biais d'échantillonnage** . . . 33
1.4 **Les études ayant comparé les méthodes de correction de l'effet du biais d'échantillonnage.** Sont mises en gras les méthodes ayant permis au modèle d'avoir de meilleures performances . . . 34

3.1 **Type de localisation des données de capture entre 1901 et 2013** . . . 66
3.2 **Données environnementales brutes** . . . 71
3.3 **Données environnementales utilisées pour la modélisation.** (/) signifie que l'influence de la variable sur la présence d'*An. darlingi* dépend de la valeur/modalité de la variable (+) signifie que la variable favorise la présence d'*An. darlingi* (-) signifie que la variable limite la présence d'*An. darlingi* . . . 74
3.4 **Contributions moyennes et résultats de Jackknife du modèle construit avec onze variables environnementales** . . . 78
3.5 **Contributions moyennes et résultats de Jackknife du modèle simplifié construit avec sept variables environnementales** . . . 78
3.6 **Caractérisation des zones avec un indice de qualité d'habitat élevé selon les valeurs ou les modalités des variables environnementales** *ns.*, (+) et (-) ne concernent que les variables quantitatives. *ns.* signifie que l'IQH élevé ne dépend pas de variable environnemental, (+) signifie que l'IQH augmente quand la valeur de la variable environnementale augmente, (-) signifie que l'IQH diminue lorsque la valeur de la variable environnementale diminue . 88

4.1 **Tableau récapitulatif des modèles, de leurs données d'entrée et de sortie, ainsi que des métriques d'évaluation absolue et relative.** . . . 107
4.2 **Pourcentages de valeurs de ΔAUC, $\Delta Dgeo$ et de $\Delta Denv$ strictement positives.** Une valeur positive signifie que la méthode de correction permet d'améliorer le modèle biaisé. Les maxima des pourcentages pour chaque valeur de k sont représentés en gras. * Dans ces cas, le pourcentage de valeurs nulles, signifiant que le modèle corrigé est strictement équivalent au modèle biaisé, est mis entre parenthèse. . . . 108

B.1 **Publications mentionnant la présence et la location d'espèce *Anopheles* en Guyane entre 1902 et 1999** . . . 141
B.2 **Publications mentionnant la présence et la location d'espèce *Anopheles* en Guyane entre 2000 et 2013** . . . 142

C.1 **Coordonnées géographiques des sites de présence d'*Anopheles darlingi* précisément géolocalisés** (Système de coordonnées : RGFG95/UTM22N) . . . 145

Listes des sigles et des abréviations

$Denv$	Indice de Schoener calculé dans l'espace environnemental
$Dgeo$	Indice de Schoener calculé dans l'espace géographique
ACM	Analyse des Correspondances Multiples
ACP	Analyse en Composantes Principales
AFDM	Analyse Factorielle de Données Mixtes
AFGM	Analyse Factorielle de Groupe Mixte
AFM	Analyse Factorielle Multiple
ANN	Artificials neural networks
ARS	Agence Régionale de la Santé
AUC, AUCROC	Area Under the Received Operating Characteristic Curve
CDPS	Centres Délocalisés de Prévention et de Soins
CIC-EC	Centre d'Investigation Clinique - Epidémiologie Clinique
CIRAD	Centre de Coopération Internationale en Recherche Agronomique pour le Développement
Cire	Cellule d'intervention en région
CNR	Centres Nationaux de Référence
DAAF	Direction de l'Alimentation, l'Agriculture et de la Forêt
DDAS	Direction de la Démoustication et des Actions sanitaires
EPAT	Équipe Écosystèmes Amazoniens et Pathologie Tropicale
FAG	Forces armées en Guyane
GA	Genetic algorithm
GAM	Generalized additive models
GLF	Unités géomorphologiques, *Geomorphological landform*
GLM	Generalized linear model
GLS	Paysages géomorphologiques, *Geomorphological landscapes*
HA	Activités qui altèrent de manière non-permanente et localement l'environnement naturel, *Human activities*
HFP	Empreinte humaine, *Human Footprint*
IGN	Institut National de l'Information Géographique et Forestière
INSEE	Institut National de la Statistique et des Études Économiques
INSERM	Institut Nationale de la Santé et de la Recherche Médicale
IPG	Insitut Pasteur de la Guyane

IQH	Indice de qualité d'habitat
IRBA	Institut de recherche biomédicale des armées
LHUPM	Laboratoire Hospitalo-Universitaire de Parasitologie et Mycologie
LS	Occupation du sol, *Landscape type*
MARS	Multivariate Adaptative Regression Splines
MDE	Modélisation des distributions d'espèces
NASA	National Aeronautics and Space Administration
ODD	Objectifs de Développement Durable
OMD	Objectifs du Millénaires pour le Développement
OMS	Organisation Mondiale de la Santé
ONF	Office National des Forêts
ONU	Organisation des Nations Unies
PER_URB	Pourcentage d'urbanisation
PER_URB_NEIGH	Pourcentage d'urbanisation dans les pixels voisins, *Percentage of urbanization within the neighbor cells*
SIG	Système d'information géographique
SRTM	Shuttle Radar Topography Mission
SSA	Service de santé des armées
SVM	Support vector machines

Introduction générale

En modélisation écologique, les modèles de distribution d'espèces permettent de spatialiser la connaissance sur la qualité d'habitat, voire sur la présence probable, d'une ou plusieurs espèces d'intérêt (espèces animales ou végétales, bactéries, etc). Pour ce faire, ils mettent en correspondance des données d'observation de la ou des espèces d'intérêt (présence, absence, densité) avec des données environnementales décrivant les milieux. Ces modèles sont utilisés pour des objectifs variés : l'amélioration de la connaissance des niches environnementales des espèces ; la prédiction de la distribution et des aires de prolifération d'espèces invasives ; l'évaluation de l'impact du changement climatique, de l'utilisation et de l'occupation du sol sur la distribution d'espèces ; la prédiction de la distribution des espèces rares, en voie de disparition, en soutien aux plans de conservation et de réintroduction d'espèces en danger. Depuis les années 90, l'utilisation des modèles de distribution d'espèces (toutes espèces confondues) a largement augmenté. En effet, la nombre d'articles scientifiques utilisant ces modèles est passée de 10 à 1093 articles, entre 1990 et 2015 (figure 1, méthodologie en annexe A), montrant l'intérêt de la communauté scientifique pour ce type d'approche.

Le choix du modèle dépend des données d'observations de l'espèce disponibles. Les modèles ne nécessitant que des données de présence ont un avantage important vis-à-vis des modèles requérant à la fois des données de présence et d'absence, compte tenu de la difficulté d'acquérir des informations d'absence fiables. Cependant, les modèles de présence sont très sensibles au biais d'échantillonnage, résultat d'un effort d'échantillonnage inégal d'une zone géographique à une autre. Or, la problématique du biais d'échantillonnage n'est que rarement considérée par les auteurs de la littérature scientifique. L'analyse des publications annuelles d'articles portant sur les modèles de distribution d'espèce le confirment. En effet, la part des

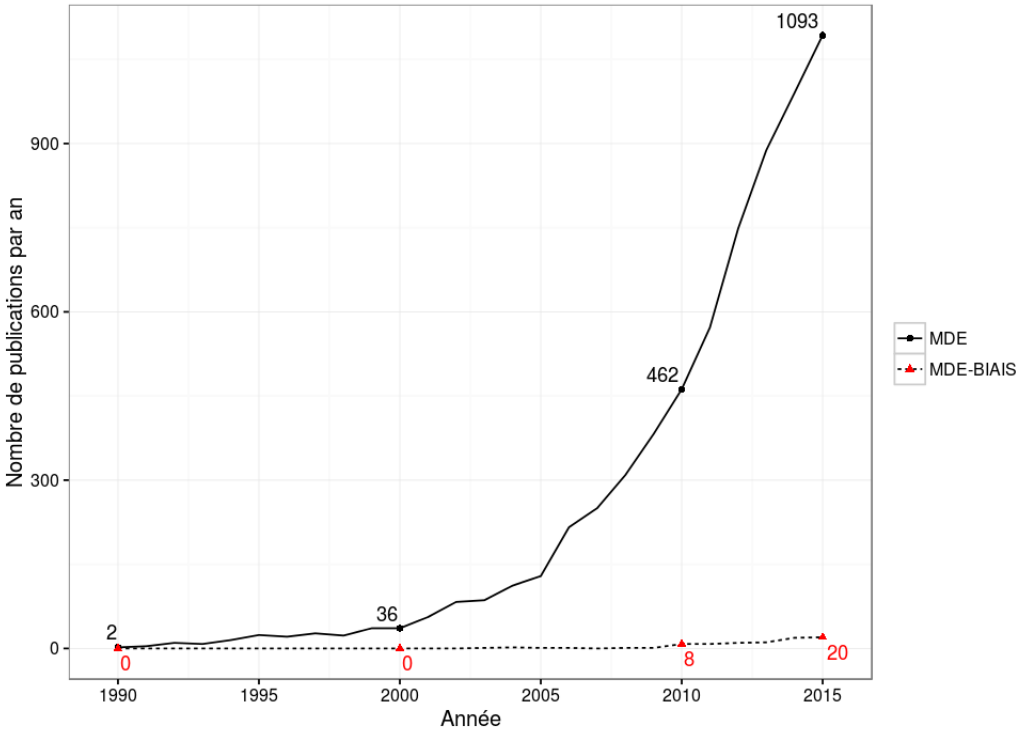


FIGURE 1 – **Publications annuelles de document portant sur les modèles de distribution d'espèces répertoriés sur *Web of Science*.** En noir est représenté le nombre de publications annuelles traitant des modèles de distribution d'espèces et en rouge, les publications qui, parmi elles, mentionnent le biais d'échantillonnage.

articles mentionnant le biais d'échantillonnage ne représentent qu'une faible partie des publications annuelles (figure 1). Si, des méthodes ont été développées pour corriger l'effet du biais d'échantillonnage, elles ne sont pas applicables à toutes les échelles d'étude, et nécessitent souvent un nombre important de site de présence, qui n'est pas toujours à la portée des utilisateurs.

Par la spatialisation de la qualité d'habitat, de la présence voire de la densité d'espèces vectrices de maladies, les modèles de distribution d'espèces peuvent significativement contribuer à l'estimation et à la cartographie du risque de transmission de maladies vectorielles, à la lutte antivectorielle et à la prévention des futures épidémies, en particulier dans le cas du paludisme (Alimi et al., 2015).

Le paludisme est une maladie parasitaire (parasite du genre *Plasmodium*) vectorielle, transmise à l'homme par les moustiques du genre *Anopheles*. Il constitue un problème sanitaire majeur au niveau mondial. Près de la moitié de la population dans le monde est exposée à cette maladie.

En 2000, les pays membres de l'Organisation des Nations Unies (ONU) se sont réunis pour élaborer un projet qui avait pour but de combattre la pauvreté dans le monde. Il s'est traduit par huit Objectifs du Millénaires pour le Développement (OMD) à atteindre en 2015. Une cible principale de l'objectif 6 était de maîtriser le paludisme et d'autres grandes maladies et à inverser la tendance de 2000. En 2000, la transmission du paludisme était active dans 106 pays, cette maladie avait touché près de 262 millions de personnes et avait fait près de 839 000 décès. L'Organisation mondiale de la santé (OMS) a constaté une baisse du nombre de cas de paludisme en 2015, avec 214 millions de cas dont 438 000 de décès (World Health Organization et al., 2015a) et l'OMD 6 a été considéré comme atteint. Le rapport de 2015 des OMD (Nations Unies, 2015) souligne que cette forte diminution du nombre de cas s'explique par un financement international important contre le paludisme depuis 2000. Ceci a permis une amélioration de l'accès à la prévention et au traitement contre le paludisme, en particulier par la distribution de moustiquaires imprégnées d'insecticide dans les pays de l'Afrique subsaharienne (dont le nombre de malades représente près de 80% des cas de paludisme mondiaux) ; la lutte antivectorielle ; un accès aux tests de diagnostic rapide ; l'amélioration des traitements antipaludiques, notamment chez les femmes enceintes. Bien que les résultats soient très significatifs, le paludisme reste toujours un problème sanitaire majeur et un fardeau financier et économique pour les pays touchés. Il nécessite des investissements stratégiques dans les systèmes de santé, la surveillance de la maladie, et impose de développer de nouveaux outils. Pour cela, en 2016, le *Programme de développement durable à l'horizon 2030* a été mis en place par l'ONU. Il a été défini autour de 17 Objectifs de Développement Durable (ODD) à atteindre d'ici 2030. Une des cibles de l'objectif trois qui est de "Permettre à tous de vivre en bonne santé et promouvoir le bien-être de tous à tout âge", est d'éliminer les épidémies de paludisme et d'autres maladies d'ici à 2030 (cible 3.3). Afin d'atteindre cet objectif, l'OMS coordonne, pour ces prochaines années, le projet *Global technical strategy for malaria 2016-2030* qui vise à l'élimination du paludisme (World Health Organization et al., 2015b). Pour cela, ce programme s'appuie sur trois piliers :

1. assurer l'accès à la prévention, au diagnostique et au traitement du paludisme à toutes les populations à risques ;
2. accélérer les efforts pour l'élimination du paludisme et pour parvenir au statut de "malaria-free" ;

3. transformer la surveillance du paludisme en une intervention nationale principale.

Un aspect important, souligné par l'OMS dans le pilier 1, est l'amélioration de lutte antivectorielle. En effet, elle constitue un élément important dans l'élimination du paludisme, car l'absence du vecteur empêche toute transmission du parasite (excepté la transmission due aux transfusions sanguines). Une cartographie de la distribution de la présence des vecteurs basée sur les données d'observation est alors essentielle.

Cette thèse aborde donc la problématique de la distribution de la qualité d'habitat des vecteurs du paludisme avec une approche qui repose sur la modélisation écologique.
L'objectif de cette thèse est double : premièrement, proposer une méthode de correction de l'effet du biais d'échantillonnage originale et générique pouvant être appliquée à un nombre de sites de présence faible et à toute échelle d'étude ; deuxièmement, modéliser la distribution du principal vecteur du paludisme en Guyane française, *Anopheles darlingi*.

Dans le chapitre 1, l'état de l'art des modèles de distribution d'espèces est réalisé en mettant en avant les avantages et les faiblesses de chacun d'eux afin de justifier le choix du modèle Maxent (Phillips et al., 2006) dans cette thèse. Ensuite, la revue des méthodes de correction de l'effet du biais d'échantillonnage, ainsi que de leurs avantages et leurs limites, est effectuée.
Le chapitre 2 porte sur la méthode originale développée durant la thèse afin de corriger l'effet du biais d'échantillonnage et permettant de combler les faiblesses des méthodes existantes, tant sur les aspects théoriques que d'applicabilité.
Le chapitre 3 porte sur la modélisation de la distribution du principal vecteur du paludisme en Guyane française : *Anopheles darlingi*. Située en Amérique du Sud, la Guyane est un département français où le paludisme est endémique. Une première partie de ce chapitre sera consacrée à la justification du choix de la Guyane française comme site d'étude. Le travail décrit dans ce chapitre a été réalisé en partenariat avec l'Institut Pasteur de la Guyane et l'Institut de Recherche Biomédicale des Armées. Ensuite, une description du paludisme et de l'écologie des vecteurs dans cette zone est réalisée afin d'orienter le choix des variables environnementales pour la construction du modèle. Ensuite la modélisation de la distribution d'*An. darlingi* est présentée. Elle intègre la méthode de correction du biais d'échantillonnage présentée au chapitre 2. Les résultats sont ensuite exposés et discutés.
Le chapitre 4 porte sur l'évaluation et la comparaison de la méthode de correction développée dans le chapitre 2 avec les méthodes existantes décrites dans la littérature, en se basant sur des données de présence simulées.
Ce manuscrit se poursuit avec une discussion générale et par les principales perspectives de ce travail. Enfin, il se clôture par une conclusion générale.

Chapitre 1

État de l'art

I. Modèles de distribution d'espèces . . . 9
I. 1. Généralités . . . 9
I. 1. a. Habitat et niche écologique . . . 9
I. 1. b. Terminologie et postulat d'équilibre . . . 9
I. 1. c. Utilisation des modèles de distribution d'espèces . . . 9
I. 2. Les modèles de présence-absence . . . 11
I. 2. a. Méthodes statistiques . . . 11
I. 2. b. Méthodes d'apprentissage . . . 12
I. 3. Les modèles de présence uniquement . . . 14
I. 4. Les méthodes de présence-background . . . 15
I. 5. Les méthodes de présence-pseudo-absence . . . 16
I. 6. Avantages et limites des différents types de modèles de distribution d'espèces . . . 16
II. Maxent . . . 17
II. 1. Maximum d'entropie . . . 17
II. 1. a. L'entropie et son application à la modélisation écologique . . . 17
II. 1. b. Transformations et contraintes . . . 18
II. 1. c. Résolution . . . 19
II. 1. d. Régularisation . . . 19
II. 2. Implémentation de Maxent . . . 20
II. 2. a. Paramétrage de l'algorithme . . . 21
II. 2. b. Sorties de Maxent . . . 21
II. 2. c. Évaluation . . . 22
II. 3. Avantages et faiblesses de Maxent . . . 24
III. Biais d'échantillonnage . . . 24
IV. Correction de l'effet du biais d'échantillonnage . . . 25
IV. 1. Sélection des données de présence . . . 26
IV. 1. a. Sélection des sites de présence basée sur des critères géographiques . . . 26
IV. 1. b. Sélection des sites de présence basée sur des critères environnementaux 26
IV. 2. Construction d'un background biaisé . . . 27
IV. 2. a. Construction d'un background biaisé basée sur des critères géographiques 27
IV. 2. b. Construction d'un background biaisé basée sur des critères environnementaux . . . 27
IV. 2. c. Construction d'un background biaisé basée sur les groupes cibles . . . 28
IV. 3. Comparaison des méthodes de correction de l'effet du biais d'échantillonnage . . . 28
IV. 4. Avantages et limites des méthodes de correction . . . 29
V. Conclusion . . . 30

I. Modèles de distribution d'espèces

I. 1. Généralités

I. 1. a. Habitat et niche écologique

Hall et al. (1997) définit l'habitat comme étant un milieu qui fournit à une espèce les ressources et les conditions nécessaires à sa survie (alimentation, reproduction, abri). Une espèce cherche des habitats propices tant pour la qualité et la quantité alimentaire que pour la présence de structures pour se reposer, se cacher et se reproduire. La *niche écologique* est un concept écologique qui varie selon les auteurs.

— Grinnell (1917) a été le premier a définir cette notion. Selon lui, le terme *niche* décrit la relation entre l'espèce et son environnement. Il a conceptualisé cette notion en schématisant la niche comme étant la somme des habitats requis par l'espèce ;

— Elton (1927) définit la niche écologique comme étant l'ensemble des relations qu'une espèce entretient avec sa nourriture et ses ennemis ;

— Selon Hutchinson (1957), la niche écologique est un hypervolume de n dimensions, défini par n variables environnementales, dans lequel l'espèce peut survivre et se reproduire. Cette notion ne désigne pas uniquement les facteurs abiotiques (facteurs environnementaux) mais également les facteurs biotiques comme la compétition entre espèces, la prédation, et les obstacles géographiques à la dispersion. Un habitat peut être constitué de plusieurs niches écologiques. Une niche écologique est alors une position occupée par l'espèce dans l'écosystème en tenant compte des paramètres physico-chimiques de l'environnement et des paramètres biologiques qui correspondent à l'ensemble des conditions dans laquelle vit et se perpétue la population.

I. 1. b. Terminologie et postulat d'équilibre

Les modèles de distribution d'espèces ont de nombreuses appellations : modèles de distribution d'espèces (*species distribution models*), modèles de niche écologique (*ecological niche models*), ou modèles d'enveloppe climatique (*climate envelop modeling*). Selon Guisan et al. (2013), ces termes sont équivalents car ils utilisent les mêmes données et les mêmes algorithmes, mais la terminologie choisie dépend de l'objectif de l'étude. Le terme "modèle de niche écologique" est utilisée lorsque l'utilisateur cherche à quantifier la niche écologique (environnementale) tandis que "modèle de distribution d'espèces" concerne plutôt les prédictions spatiales des habitats favorables à l'espèce. Dans ce travail, le terme "modèle de distribution d'espèce" sera utilisé.
Les modèles de distribution d'espèces font l'hypothèse que les espèces sont en équilibre avec leur environnement, en d'autres termes, elles sont présentes dans l'ensemble des habitats qui leur est favorable, et absentes dans les habitats qui leur sont défavorables.

I. 1. c. Utilisation des modèles de distribution d'espèces

Les modèles de distribution d'espèce sont des modèles empiriques qui relient les observations des espèces considérées aux données environnementales décrivant le milieu dans lequel elles vivent (Guisan and Thuiller, 2005). Par conséquent, ces modèles nécessitent en entrée des données d'observation de l'espèce et des variables environnementales. En sortie, une carte de qualité d'habitat est produite.

Depuis les années 1990, le nombre d'utilisations de ces modèles n'a pas cessé d'augmenter (figure 1 de l'introduction générale). Ils sont utilisés pour différents objectifs :

— l'amélioration de la connaissance des niches environnementales des espèces (Austin et al., 1990; Piedallu et al., 2016) ;
— la prédiction de la distribution et des aires de prolifération d'espèces invasives (Peterson, 2003; Fraser et al., 2015; Mainali et al., 2015) ;
— l'évaluation de l'impact du changement climatique, de l'utilisation et de l'occupation du sol sur la distribution d'espèces (Thuiller, 2004; Maguire et al., 2015; Prieto-Torres et al., 2016) ;
— la prédiction de la distribution des espèces rares, en voie de disparition (Engler et al., 2004; Stirling et al., 2016), en soutien aux plans de conservation et de réintroduction d'espèces en danger (Pearce and Lindenmayer, 1998; Rovzar et al., 2016).

Les variables environnementales utilisées en entrée sont des données qui informent sur les conditions environnementales de la zone d'intérêt. En modélisation des distributions d'espèces (MDE), elles sont appelées : variables explicatives, prédicteurs, covariables ou entrées. Elles sont souvent issues d'images satellites, de photos aériennes, ou de données extrapolées à partir de données recueillies par des capteurs au sol. Les conditions environnementales qu'elles caractérisent peuvent exercer un effet direct ou indirect sur la répartition des espèces (Guisan and Thuiller, 2005). Elles correspondent à :

— des facteurs de limite (régulateurs) qui contrôlent la dispersion ou la distribution de l'espèce (la température, la pluviométrie, la géologie, ...) ;
— des facteurs de troubles, regroupant toutes les perturbations de l'environnement (actions anthropiques ou naturelles) ;
— des facteurs ressources, qui regroupent tous les éléments qui peuvent être assimilés par l'espèce (alimentation, énergie, ...).

Les données d'observations des espèces sont généralement des informations de la localisation géographique, idéalement les coordonnées géographiques précises des sites où l'espèce a été observée. Ces données sont issues d'un échantillonnage aléatoire et/ou stratifié ou encore opportuniste. Elles représentent soient des sites de présence uniquement soient des sites de présence et d'absence.
Quatre grandes approches existent pour modéliser la distribution des espèces, elles dépendent des données d'observations :

— les méthodes basées sur des sites de présence et d'absence ;
— les méthodes basées sur des sites de présence uniquement ;
— les méthodes basées sur des sites de présence et des sites de *pseudo-absence*, utilisées lorsque l'information d'absence n'est pas disponible. Les sites de pseudo-absence sont censés jouer le rôle des sites d'absence ;
— les méthodes basées sur des sites de présence et des sites de *background*, généralement sélectionnés de manière aléatoire dans la zone d'étude pour refléter toutes les conditions environnementales de cette zone. Ces méthodes sont également utilisées en l'absence de données d'absence fiables.

I. 2. Les modèles de présence-absence

Lorsque la construction du modèle nécessite à la fois des données de présence et des données d'absence, les méthodes sont dites méthodes de présence-absence.

I. 2. a. Méthodes statistiques

De nombreuses méthodes sont basées sur l'analyse de régression. Les méthodes de régression linéaire consistent à trouver une relation linéaire entre la variable expliquée Y et les n variables explicatives $X = (X_1, X_2, ..., X_n)$.

$$Y = \beta_0 + \sum_{j=1}^{n} X_j.\beta_j + \varepsilon \tag{1.1}$$

où β est un vecteur de n coefficients, β_0 une constante et ε le terme d'erreur normalement distribuée avec une moyenne nulle et une variance constante.
En MDE, la variable expliquée Y représente la présence de l'espèce et les variables explicatives X, les variables environnementales. Une courbe de réponse est une fonction décrivant la relation entre Y et les valeurs de X.

Les modèles linéaires généralisés (GLM pour *Generalized linear model* en anglais, Guisan et al. (2002)) sont des extensions des méthodes de régression linéaire permettant une distribution non-normale de la variable expliquée (Venables et Ripley, 1994) (figure 1.1 a). Pour cela, les GLM transforment les variables explicatives via une fonction lien. L'équation (1.1) devient

$$g(E(Y)) = \beta_0 + \sum_{j=1}^{n} X_j.\beta_j + \varepsilon \tag{1.2}$$

où $E(Y)$ est l'espérance mathématique de Y et g la fonction lien. La fonction lien définit la forme de la distribution de probabilité de Y : gaussienne, binomiale, Poisson, ou Gamma.

Les modèles additifs généralisés (GAM pour *generalized additive models*, Guisan et al. (2002)) sont des extensions des GLM. De la même manière que les GLM, les GAM utilisent des fonctions lien mais au lieu d'utiliser le vecteur de n coefficients β, les GAM utilisent une fonction de lissage f (figure 1.1 a). L'équation (1.2) devient

$$g(E(Y)) = \beta_0 + \sum_{j=1}^{n} X_j.f_j + \varepsilon \tag{1.3}$$

Cette fonction lissage permet aux GAM de représenter des relations plus complexes et non linéaires (Yee and Mitchell, 1991).

Une autre méthode, appelée Régression multivariée par splines adaptatives (MARS pour *Multivariate Adaptative Regression Splines*, Elith and Leathwick (2007)) est comparable aux méthodes GLM et GAM, à la différence que la fonction lien est une série de segments de ligne droite connectés entre eux (figure 1.1 a).

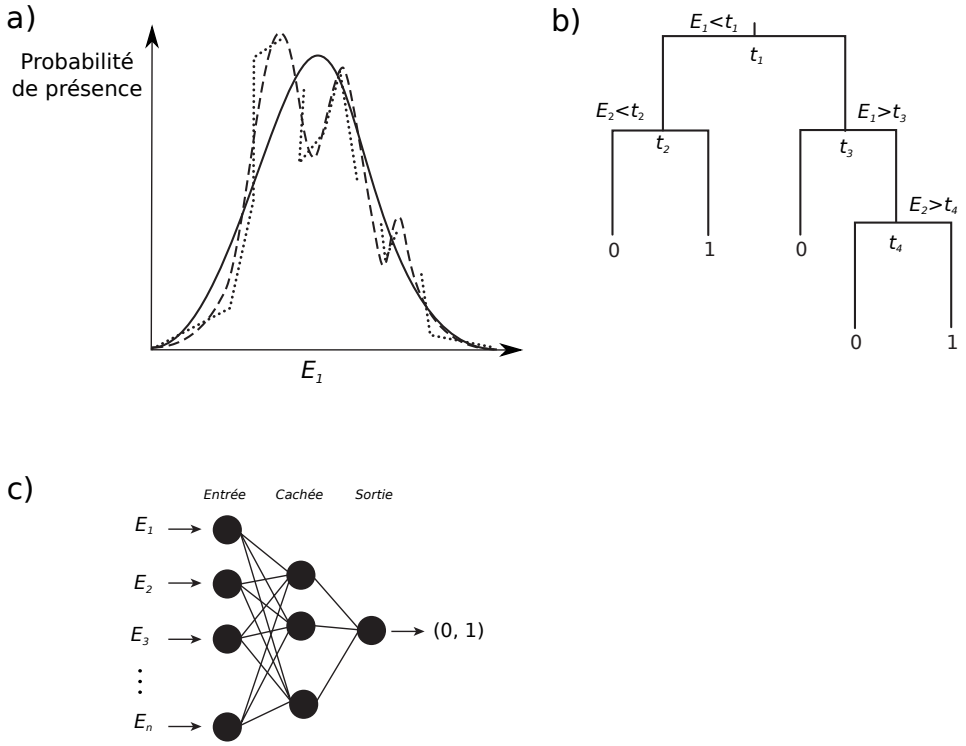


FIGURE 1.1 – **Représentation de la calibration des différentes méthodes de présence absences, source : modifié d'après** (Peterson et al., 2011).
a) exemple de courbes de réponse des GLM (ligne continue), GAM (ligne en tirets) et MARS (ligne en pointillés)
b) exemple d'un arbre de classification où deux variables environnementales, E_1 et E_2, sont partitionnées par des règles t_1, t_2, t_3 et t_4 ;
c) exemple d'un réseau neurones ; les neurones sont représentés par les points et les liens pondérés des neurones par les segments. Ce réseau est composé de n neurones d'entrée correspondant aux n variables environnementales E_i ($i \in [1, ..., n]$), de m éléments dérivés sur la couche cachée et d'une sortie.

I. 2. b. Méthodes d'apprentissage

D'autres méthodes de présence-absence sont basées sur les méthodes d'apprentissage automatique (*machine learning* en anglais) notamment supervisée. L'apprentissage automatique supervisée utilise un ensemble d'exemples, c'est-à-dire d'objets, au sens large, étiquetés (ou labelisés) préalablement et à partir desquels l'algorithme d'apprentissage va induire le modèle. En MDE, les exemples correspondent aux sites échantillonnés et associés à l'information de présence ou d'absence de l'espèce. Différentes méthodes d'apprentissage ont été appliquées à la MDE, telles que :

- les arbres de classification ou de régression ;
- les réseaux de neurones artificiels (*Artificials neural networks*, ANN) ;
- les algorithmes génétiques (*Genetic algorithm*,GA) ;
- l'apprentissage statistique, avec par exemple les machines à support de vecteurs (*support vector machines*, SVM) .

Les méthodes d'apprentissage basées sur les arbres se réfèrent aux arbres de classification et aux arbres de régression. Ces arbres cherchent la relation entre les variables explicatives (variables environnementales) et la variable catégorielle expliquée Y (la présence ou absence d'une espèce). Ils sont construits par partitionnement itératif des données en sous-ensembles qui seront à leur tour partitionnés (figure 1.1 b). Ce partitionnement se fait au regard d'un seuillage d'une variable environnementale continue ou de l'appartenance des exemples aux différentes modalités d'une variable environnementale catégorielle. Le processus de partitionnement est interrompu lorsque les sous-ensembles sont considérés suffisamment homogènes selon un critère prédéfini. Ces sous-ensembles sont alors appelés "feuilles" de l'arbre. Ainsi, le partitionnement permet la construction d'un arbre, où les feuilles terminales correspondent aux différentes classes de Y ((figure 1.1 b), c'est-à-dire la présence et l'absence de l'espèce dans le contexte de la MDE.

Les réseaux neurones artificiels s'inspirent de la structure et des opérations réalisées par le système biologique neuronal. Le réseau neuronal est composé de plusieurs couches : la première couche (entrée) est constituée de p neurones correspondant aux p variables environnementales ; la couche de sortie constituée d'un neurone correspondant à la probabilité de présence de l'espèce (à laquelle un seuil sera appliqué pour obtenir la présence et l'absence de l'espèce) ; et de couches intermédiaires, ou cachées (*hidden layers*), composées de m éléments dérivés de la couche d'entrée (figure 1.1 c, Pearson et al. (2002)). Les neurones sont reliés entre eux ; chaque neurone de la $i^{\text{ième}}$ couche est relié à l'ensemble des neurones de la $(i-1)^{\text{ième}}$ couche. Lors de la phase d'apprentissage, à chaque itération, ces connexions sont ajustées à l'aide d'un poids variable afin d'optimiser le biais et la variance. Le réseau de neurones artificiels estime une réponse (sortie) par la combinaison des éléments issus des couches intermédiaires.

Les algorithmes génétiques sont inspirés de phénomène génétique, tel que la théorie de l'évolution. Cette théorie repose sur le principe que, de générations en générations, seuls sont conservés les gènes les plus adaptés aux besoins de l'espèce par rapport à son environnement. Ce processus d'évolution se traduit par des règles de classification sur les données environnementales (par exemple : l'espèce A est présente si la température mensuelle est supérieur à 20°C et la pluviométrie de Janvier inférieure à 80 mm). Ces règles sont appelées gènes, elles subissent des combinaisons aléatoires avec d'autres, et subissent le principe de sélection naturelle jusqu'à ce qu'une solution optimale soit trouvée, c'est-à-dire jusqu'à trouver les probabilités de présence prédisant la mieux les données observées.

Les méthodes d'apprentissage basées sur les méthodes statistiques sont les machines à support de vecteurs, également appelés séparateurs à vaste marge (*Support Vector Machines*, SVM). À l'origine, la méthode SVM a été développée par Vapnik (1995) et a été appliquée à la reconnaissance de l'écriture manuscrite et à la classification automatique de textes. Les SVM sont utilisés pour résoudre des problèmes de discrimination à deux-classes (faits dits positifs et faits dits négatifs), où l'algorithme recherche tout d'abord un espace de représentation des données dans lequel un hyperplan de séparation existe, puis trouve l'hyperplan de séparation de plus grande marge, c'est-à-dire le plus éloigné possible des deux classes (figure 1.2). Appliquée en écologie et notamment dans la MDE (Guo et al., 2005), les faits

positifs correspondent aux sites de présence de l'espèce étudiée et les faits négatifs aux sites d'absence. Ces méthodes peuvent être utilisées lorsque les sites d'absences ne sont pas disponibles. Dans ce cas, l'algorithme cherche à séparer les conditions environnementales dans lesquelles l'espèce a été trouvée de l'ensemble de l'espace environnemental.

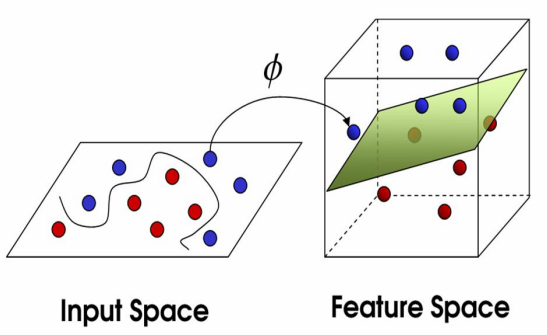


FIGURE 1.2 – **Représentation de la méthode SVM, où θ représente la transformation permettant l'existence d'un hyperplan de séparation.**
source : Rossi et al. (2015)

I. 3. Les modèles de présence uniquement

Les méthodes qui ne nécessitent que l'information sur la présence d'une espèce sont dites méthodes de présence-uniquement. La méthode la plus simple pour modéliser la distribution des espèces est la construction d'une enveloppe environnementale (Peterson et al., 2011).
La méthode BIOCLIM (Busby, 1991) construit dans l'espace environnemental de n dimensions une enveloppe hyper-rectangulaire autour des sites de présence, où n représente le nombre de variables environnementales (figure 1.3 a). Les limites de cette enveloppe sont définies par les valeurs maximale et minimale des variables environnementales pour lesquelles l'espèce a été observée. Cette méthode ne permet pas de modéliser les interactions entre variables environnementales et les limites strictes des enveloppes correspondent à des discontinuités irréalistes.
La méthode HABITAT (Walker and Cocks, 1991) est une autre méthode d'enveloppe environnementale. Au lieu de se baser sur une enveloppe aux limites strictes comme BIOCLIM, cette méthode utilise une enveloppe convexe pour mieux ajuster les limites de l'enveloppe environnementale (figure 1.3 a).
Une autre méthode basée sur la distance de Mahalanobis (Rotenberry et al., 2006) (figure 1.3 a) consiste à calculer la distance de Mahalanobis entre chaque site dans l'espace environnemental de n dimensions. Cette distance est définie par la différence entre les valeurs des n variables environnementales en chacun des sites de la zone d'étude et les valeurs moyennes des n variables environnementales aux sites de présence. La distance d'un site i est définie comme suit :

$$D_i = \sqrt{(y_i - \mu)' C^{-1} (y_i - \mu)} \tag{1.4}$$

où μ est un vecteur des valeurs moyennes des n variables environnementales aux sites de présence, de dimension $n \times 1$, y_i le vecteur des valeurs des variables environnementales au site i, de dimension $n \times 1$, et C la matrice de covariance. Plus la valeur de cette distance est

faible pour un site i, plus les conditions environnementales de ce site sont similaires à celles des sites de présence. De plus, cette méthode prend en compte la colinéarité des variables environnementales via le calcul de la matrice de covariance.
DOMAIN (Carpenter et al., 1993) est une alternative à la méthode basée sur la distance de Mahalanobis, la métrique utilisée étant la distance de Gower (figure 1.3 b). La distance d entre deux sites i et j dans un espace euclidien de n dimensions est définie comme suit :

$$d_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{|y_{i,k} - y_{j,k}|}{range_k}\right) \tag{1.5}$$

où $range$ est le vecteur des amplitudes des domaines de valeurs des n variables environnementales.
La similarité R_{ij} entre i et j est définie par $R_{ij} = 1 - d_{ij}$.
Le maximum de la similarité entre le point i et un ensemble de m sites de présence P est définie par :

$$S_i = max_{k=1}^{m} R_{iP_k} \tag{1.6}$$

S_i varie entre 1 et 0. La valeur 1 signifie que les conditions environnementales de i sont identiques à celles d'au moins un site de présence. Cette valeur est calculée pour l'ensemble des sites de la zone d'étude. Un seuil est ensuite défini sur le maximum de similarité afin de délimiter l'enveloppe.

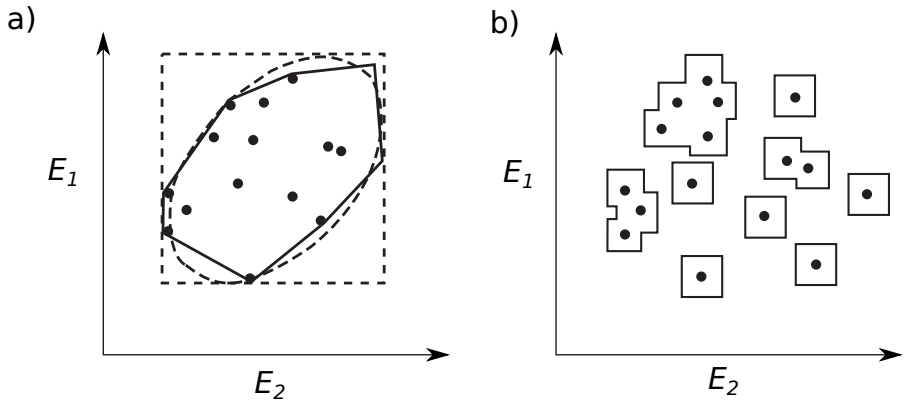


FIGURE 1.3 – **Représentation de la calibration des différentes méthodes de présence-uniquement dans l'espace de deux variables environnementales E_1 et E_2,**(Peterson et al., 2011).
Les points représentent les sites de présence, et l'enveloppe environnementale est représentée par :
(a) le rectangle pour BIOCLIM, par le polygone convexe pour HABITAT et par l'ellipse pour Mahalanobis ;
(b) les lignes pour DOMAIN

.

I. 4. Les méthodes de présence-background

Les méthodes qui utilisent des données de background ne requièrent que des sites de présence, mais sont différentes des méthodes de présence-uniquement du fait qu'elles incorporent une information concernant la variation de l'environnement de la zone d'étude (information appelée *background*) dans le développement du modèle. Les méthodes de présence-background ont souvent une capacité de discrimination plus élevée que les méthodes de présence-uniquement (Peterson et al., 2011).

Une méthode largement utilisée est la méthode *Ecological Niche Factor Analysis* (ENFA, Hirzel et al. (2002)). Cette méthode est un type particulier d'analyse en composantes principales. Elle compare la distribution des sites de présence dans l'espace des variables environnementales à celle de l'ensemble des sites de la zone d'étude via deux paramètres : la marginalité et la spécialisation (Hirzel et al., 2002). La marginalité est la différence entre les conditions environnementales moyennes de l'espèce (les conditions environnementales dans lesquelles l'espèce est présente) et les conditions environnementales moyennes globale (les conditions environnementales de l'ensemble de la zone d'étude). Plus la marginalité est élevée et plus les conditions moyennes de l'espèce s'écartent des conditions moyennes globales. La spécialisation est le ratio entre la variance des conditions environnementales des sites de présence et celles de l'ensemble de la zone d'étude. Elle permet d'évaluer l'étroitesse de l'espace écologique de l'espèce par rapport à l'espace écologique global. La spécialisation correspond ainsi à la tolérance d'une espèce par rapport aux conditions environnementales. L'espace vectoriel de l' ENFA est construit avec la marginalité comme premier axe et la spécialisation comme second axe. L'axe marginalité passe par le barycentre des conditions environnementales globales et par celui des conditions environnementales de l'espèce. L'axe de spécialisation est orthogonal à l'axe de marginalité. Á chaque élément de la zone d'étude est assignée une valeur de qualité selon sa distance aux sites de présence dans l'espace vectoriel.
Une autre méthode très utilisée est celle basée sur le principe de la maximisation de l'entropie, elle sera détaillée dans la partie II.

I. 5. Les méthodes de présence-pseudo-absence

Ces méthodes sont utilisées lorsque l'absence de l'espèce n'a pas été effectivement constatée. Normalement, les données dites de pseudo-absences jouent le rôle des données d'absence, et sont généralement sélectionnées aléatoirement dans la zone d'étude. Les méthodes de presence-absence peuvent être utilisées mais une des méthodes couramment utilisées est l'algorithme génétique pour la production d'ensemble de règles (*Genetic Algorithms for Ruleset Production*, GARP, Stockwell (1999)). Cette méthode produit un ensemble de règles établi par un algorithme génétique. GARP a été développé initialement pour la classification binaire qui génère un ensemble de règles classifiant une espèce comme présente ou absente. Il élabore une série de règles pour résumer l'information associée aux conditions environnementales des sites de présence. Chaque règle est considérée comme un gène, des combinaisons aléatoires des gènes génèrent plusieurs modèles décrivant les conditions environnementales permettant la présence potentielle de l'espèce. Les combinaisons prédisant le mieux la présence sont sélectionnées et se traduisent par différentes cartes de prédiction en sortie du modèle. Une carte globale peut-être obtenue en moyennant les meilleures prédictions.

I. 6. Avantages et limites des différents types de modèles de distribution d'espèces

Les modèles de présence-absence sont plus précis dans la prédiction de la distribution des espèces que les modèles de présence-uniquement (Brotons et al., 2004; Hirzel et al., 2001). Cependant les informations d'absence sont souvent indisponibles, soit parce qu'elles n'existent pas, soit parce qu'elles sont erronées. Selon Peterson et al. (2011) et Hirzel et al. (2002) une donnée d'absence peut être due :

— à la non-détection de l'espèce dans un habitat favorable alors que l'espèce est effectivement présente ;
— à l'absence réelle de l'espèce dans un habitat favorable, pour des raisons historiques (par exemple, le fait que l'espèce n'a pas non encore colonisés le milieu) ou à cause de barrières géographiques (par exemple une chaîne de montagne entre deux habitats favorables) ;
— à la vraie absence de l'espèce due à des conditions écologiques non-favorables à l'espèce.

Face à ces difficultés, une solution est de recourir aux méthodes de présence-uniquement. Les modèle de présence-background sont décrits comme plus discriminants que les modèles de présence-uniquement (Peterson et al., 2011). Pour qu'un modèle de présence-pseudo-absence soit réaliste, il est essentiel d'avoir une connaissance des zones non-favorables à l'espèce afin d'y sélectionner les sites de pseudo-absence.
Ainsi, choisir un modèle de présence-background présente les avantages de ne requérir que des données de présence et ne nécessite pas de connaissance *a priori* sur les zones d'absence de l'espèce. Ce type de modèle sera alors utilisé dans ce travail de thèse.

Certains auteurs ont mené une étude comparative des modèles de présence-background. Elith et al. (2006) ont montré que la méthode basée sur le principe de maximisation de l'entropie, Maxent, est capable de générer des fonctions de réponse complexes. C'est également le modèle le moins sensible au nombre de sites de présence (Pearson et al., 2007; Tognelli et al., 2009; Hernandez et al., 2006). Maxent, un modèle génératif (Phillips et al., 2006) auquel s'ajoute une fonction de régularisation paramétrée en fonction du nombre de site de présence, lui permet de meilleures performances de prédiction par rapport aux méthodes lui étant comparables lorsque le nombre de sites de présence est faible (Wisz et al., 2008; Ng and Jordan, 2002). De plus, Maxent est un modèle pouvant prendre en compte à la fois des variables environnementales continues et catégorielles.
Compte tenu de ces avantages, Maxent sera utilisé dans ce travail et sera décrit en détail dans la suite de ce manuscrit.

II. Maxent

Maxent est une méthode d'apprentissage automatique appliquée à l'écologie. Ce modèle de présence-background ne requiert pas de données d'absence.
L'objectif de cette méthode est d'estimer une distribution de probabilité inconnue en se basant sur le principe de maximisation de l'entropie. L'implémentation de cette méthode est appelée Maxent.

II. 1. Maximum d'entropie

II. 1. a. L'entropie et son application à la modélisation écologique

En théorie de l'information, Shannon (1948) définie l'entropie comme "une mesure de la quantité de choix impliquée dans la sélection d'un événement". Soit $X = x_1, x_2, ..., x_n$ une variable aléatoire discrète associée à la distribution de probabilité p, l'entropie de X est définie

par :

$$H(X) = -\sum_{i=1}^{n} p(x_i).ln(p(x_i)) \tag{1.7}$$

Elle est interprétée comme étant la quantité d'information contenue dans un message reçu par un récepteur. Plus l'entropie est grande et plus le message contient de l'information. À l'inverse, plus l'information d'un message est redondante et plus l'entropie est proche de zéro. L'entropie est également utilisée pour quantifier le degré d'incertitude de l'information. Le principe de maximisation de l'entropie est d'approcher une distribution de probabilité inconnue par la distribution de probabilité qui maximise l'entropie et qui satisfait un ensemble de contraintes relatives aux propriétés, connues partiellement, de la distribution à estimer.
Appliqué à l'écologie, le but est d'estimer la distribution de probabilité d'une espèce quelconque. Nous considérons que seules les données de présence sont disponibles. Soit $y = 1$ lorsque l'espèce est présente, et la probabilité de présence de l'espèce dans un environnement donné est notée $P(y = 1|x)$ où x est un élément de l'ensemble $X = \{x_i\}_{i\in[1,...,l]}$ sur lequel est définie la distribution de probabilité. D'après le théorème de Bayes,

$$P(y = 1|x) = \frac{P(x|y = 1)P(y = 1)}{P(x)} \tag{1.8}$$

où $P(x|y = 1)$, noté ci-après $\pi(x)$, est la probabilité d'occurence de x sachant que l'espèce est présente. $P(y = 1)$ est la prévalence de l'espèce sur l'ensemble X. Dans le cas où les données d'absence n'existent pas, elle ne peut pas être calculée, mais correspond théoriquement à une constante K (Phillips and Dudík, 2008). $P(x) = 1/|l|$ pour tout x (Elith et al., 2011). L'équation (1.8) s'écrit donc :

$$P(y = 1|x) = \pi(x).C \tag{1.9}$$

où $C = K.|l|$.
Estimer $P(y = 1|x)$ revient donc à estimer la probabilité $\pi(x)$ à une constante près. Comme cela va être montré ci-après, $\pi(x)$ peut être estimée par le principe de maximum d'entropie et est interprétée comme un indice de qualité d'habitat.

Soit $\hat{\pi}$ l'estimation de π, l'entropie de $\hat{\pi}$ est définie comme suit :

$$H(\hat{\pi}) = -\sum_{x\in X} \hat{\pi}(x) ln(\hat{\pi}(x)) \tag{1.10}$$

où ln est le logarithme népérien.

II. 1. b. Transformations et contraintes

Considérons un ensemble de n variables environnementales $g_1, ..., g_n$. Souvent, les courbes de réponses de ces variables peuvent être complexes et ne sont pas forcément des fonctions linéaires. Afin d'être en mesure de modéliser ces relations et ainsi de mieux représenter les processus écologiques, des transformations sont appliquées aux variables environnementales (Elith et al., 2011). Elles sont appelées fonctions caractéristiques ou *features functions* en anglais. Les transformations proposées dans la littérature sont présentées dans le tableau 1.1. L'ensemble $f_1, f_2, ..., f_n$ des variables transformées est alors appelé ensemble des caractéristiques ou *features*. Dans la suite, les termes "variable environnementale" et "variable" seront aussi utilisées, même s'il s'agit des transformations des variables environnementales initiales.

À chaque site x de X est donc associé un vecteur $f(x)$ dont les composantes $f_j(x)$ ($j \in [1,...,n]$) correspondent aux valeurs des variables. Pour une variable f_j, nous considérons son espérance mathématique en lien avec $\hat{\pi}$, notée $\hat{\pi}[f_j]$, définie par $\sum_{x \in X} \hat{\pi}(x) f_j(x)$, ainsi que sa moyenne empirique sur les m sites de présence $x_1,...,x_m$ connus sur X, notée $\tilde{\pi}[f_j]$ et définie par $\frac{1}{m}\sum_{i=1}^{m} f_j(x_i)$.

Les contraintes imposées à $\hat{\pi}$ lorsque la transformation linéaire est utilisé sont ainsi définies par l'égalité suivante :

$$\hat{\pi}[f_j] = \tilde{\pi}[f_j] \qquad \text{pour chaque } \textit{feature}\ f_j \tag{1.11}$$

Les contraintes imposées pour les autres fonctions caractéristiques sont répertoriées dans le tableau 1.1.

II. 1. c. Résolution

Considérons la distribution de Gibbs $q_\lambda(x) = \frac{exp(\lambda.f(x))}{Z_\lambda}$ où λ est un vecteur de coefficients réels qui pondèrent les variables transformées et Z_λ une constante de normalisation permettant à la somme de $\sum_{x \in X} q_\lambda(x)$ d'être égale à 1.
Comme l'expliquent Phillips et al. (2006), la distribution de probabilité $\hat{\pi}$ maximisant l'entropie et satisfaisant les contraintes précédentes est égale à la distribution de Gibbs qui maximise la vraisemblance des m sites de présence $x_1,...,x_m$, notée :

$$L = \prod_{i=1}^{m} \frac{exp(\lambda.f(x_i))}{Z_\lambda} \tag{1.12}$$

Maximiser L revient à maximiser l'expression suivante :

$$ln(L) = ln(\prod_{i=1}^{m} \frac{exp(\lambda.f(x_i))}{Z_\lambda}) \tag{1.13}$$

qui peut s'écrire

$$ln(L) = -m.ln(Z_\lambda) + \sum_{i=1}^{m} \lambda.f(x_i) \tag{1.14}$$

Maximiser la vraisemblance est équivalent à minimiser l'expression suivante :

$$-ln(L) = ln(Z_\lambda) - \frac{1}{m}\sum_{i=1}^{m} \lambda f(x_i) \tag{1.15}$$

Cette expression est appelée la fonction du logarithme de la perte (*log loss* pour *logarithm of the loss function* en anglais), qui est l'opposé du logarithme de la vraisemblance.

II. 1. d. Régularisation

En tentant de satisfaire strictement l'égalité (1.11), Maxent a tendance à trop "coller" aux données d'apprentissage (phénomène de surapprentissage), ce qui a pour conséquence de réduire la capacité de généralisation du modèle. Une solution à ce problème est de relâcher les contraintes (équation (1.11)) en considérant, pour la transformation linéaire, l'inégalité suivante :

$$|\hat{\pi}[f_j] - \tilde{\pi}[f_j]| \leq \beta_j \qquad (\text{ pour chaque } \textit{feature} \text{ linéaire } f_j\ ) \tag{1.16}$$

où β_j est une constante.
La distribution qui maximise l'entropie et qui répond aux contraintes imposées (la distribution

Maxent) est alors définie par la distribution de Gibbs qui minimise la fonction du logarithme de la perte régularisée (ou pénalisée) suivante :

$$\underbrace{ln(Z_\lambda) - \frac{1}{m}\sum_{i=1}^{m}\lambda f(x_i)}_{\text{log loss}} + \underbrace{\sum_j \beta_j|\lambda_j|}_{\text{régularisation}} \tag{1.17}$$

Le premier terme est le *log loss* et est la fonction que Maxent optimise (Dudik et al., 2004). Le minimiser revient à maximiser la vraisemblance. Une distribution uniforme sur les m sites correspond à un *log loss* égal à $ln(m)$. Le deuxième terme est appelé terme de régularisation. Minimiser la différence entre ces deux termes peut être vu comme chercher à obtenir la distribution de Gibbs réalisant un compromis entre ajustement aux données et généralisation.
Le vecteur $\beta = \beta_1, ..., \beta_n$ des paramètres de régularisation permet de contrôler ce compromis. Des études se sont intéressées à la détermination des valeurs optimales de β, théoriquement et empiriquement. Phillips and Dudík (2008) définissent $\beta_j = \beta\sqrt{\frac{s^2[f_j]}{m}}$, où β est un paramètre de régularisation indépendant des données et qui ne dépend que de la transformation choisie et $s^2[f_j]$ est l'estimation de la variance de la variable transformée f_j. Théoriquement, un β permettant une bonne performance doit être proportionnel à $\sqrt{log(n)}$, n étant le nombre de variables transformées. Bien qu'en théorie cette valeur permet d'obtenir de très bonnes performances, elle ne le permet pas pour des cas réels. Pour cela, Phillips and Dudík (2008) ont mené une étude empirique pour déterminer les paramètres de régularisation en fonction de la transformation et du nombre de sites de présence. Leurs études reposent sur 39 jeux de données d'espèces et sur 11 à 13 variables environnementales dont une à trois sont catégorielles. Neuf de ces jeux de données contiennent un nombre de sites de présence de 30 à 60. Les valeurs optimales de β obtenues en fonction des transformations et du nombre de sites de présence est utilisée par défaut dans l'implémentation de la méthode Maxent. Elles sont répertoriées dans le tableau 1.2. Toutefois, elles peuvent être modifiées par l'utilisateur.

II. 2. Implémentation de Maxent

L'algorithme implémentant la méthode de maximisation de l'entropie est appelé Maxent. Son implémentation en écologie a été faite en langage Java. Cette application est utilisée dans l'environnement du logiciel *R* avec les librairies *dismo* et *BIOMOD*, dans le système d'information géographique (SIG) ArcGIS®via SDMtoolbox, ou encore dans le logiciel libre et gratuit GRASS GIS. Cette implémentation nécessite en entrée des données environnementales spatialisées (sous forme de rasters et/ou de vecteurs). Les sites de présence sont représentés par leurs coordonnées géographiques. L'ensemble de la zone d'étude représente l'ensemble des pixels non vides pour l'ensemble des couches environnementales. Cependant, en pratique, un sous-ensemble X de ces pixels est utilisé pour la construction du modèle. Ce sous-ensemble est appelé background. Ce dernier est généralement choisi de manière aléatoire dans la zone d'étude et est censé refléter l'ensemble des conditions environnementales de la zone. Le nombre de sites de background est par défaut fixé à 10 000. Cette valeur est issue d'une étude réalisée par Phillips and Dudík (2008), qui ont effectués des tests avec 226 espèces ayant un nombre de site de présence moyen de 57. Leur étude a montré que 10 000 sites de background fournit un bon compromis entre performance du modèle et temps de calcul, et qu'utiliser un nombre supérieur à cette valeur n'améliorait pas les performances du modèle.

L'estimation de la distribution de probabilité $\hat{\pi}$ se fait de manière itérative :

— l'algorithme commence avec une distribution de probabilité uniforme, pour laquelle $\lambda = (0, ..., 0)$;
— à chaque itération, les coefficients du vecteur λ sont ajustés un à un afin de diminuer la valeur de l'expression (1.17) ;
— l'algorithme s'arrête soit lorsque le nombre d'itérations défini par l'utilisateur est atteint, soit lorsque la différence entre les deux valeurs consécutives de la fonction de perte régularisée, d'une itération à une autre, est inférieure au seuil de convergence (fixé par défaut à 10^{-5}, mais pouvant être défini par l'utilisateur).

II. 2. a. Paramétrage de l'algorithme

Afin de permettre à l'utilisateur d'adapter le processus de construction du modèle aux spécificités de son étude, l'implémentation dispose de plusieurs paramètres.

Ainsi, l'utilisateur choisi les transformations à appliquer aux variables environnementales pour permettre de modéliser les relations entre celles-ci et la qualité d'habitat.
Le nombre de sites de background est fixé par défaut à 10 000 (valeur définie à partir de l'étude de Phillips and Dudík (2008)).
Il est possible de permettre ou non au modèle d'extrapoler, c'est-à-dire de prédire la qualité d'habitat au-delà des conditions environnementales représentés par les données d'apprentissage (*cf.* fig. 1.4). Ces dernières correspondent à l'ensemble des données de présence et des données du background utilisées pour construire le modèle. Selon Phillips et al. (2006), Maxent est basé sur une équation exponentielle et peut prédire des valeurs très élevées dans des zones d'extrapolation. Par conséquent, la prédiction dans ces zones doit être considérée avec précaution, à moins de fixer la valeur de qualité d'habitat prédite. Pour cela, il existe deux types d'extrapolation :

— l'utilisateur choisi de fixer la qualité d'habitat des zones d'extrapolation à la valeur de la qualité d'habitat correspondant aux conditions environnementales extrêmes des données d'apprentissage. Ce type d'extrapolation est appelé *clamp* ;
— l'utilisateur peut choisir d'assigner aux zones d'extrapolation la différence entre les valeurs de qualité d'habitat obtenues en extrapolant avec la méthode *clamp* et celles obtenues avec *dontclamp* (correspondant à une extrapolation non soumise à la méthode *clamp*). Ce type d'extrapolation s'appelle *fadebyclamp*.

Lorsque les sites de présence ont les mêmes coordonnées géographiques ou lorsqu'ils se retrouvent dans un même pixel, l'utilisateur peut permettre à Maxent de retirer ces sites de présence redondants. Ainsi, Un seul site de présence sera conservé par pixel.

Comme dit précédent, dans l'implémentation de Maxent, les valeurs des paramètres de régularisation β sont définies par défaut en fonction du nombre de sites de présence et des transformations choisies (*cf.* tableau 1.2), mais l'utilisateur peut également les modifier.

II. 2. b. Sorties de Maxent

En sortie, Maxent produit trois types de résultats :

— la distribution de probabilité $\hat{\pi}$ interprétée comme l'indice de qualité d'habitat. Cette sortie est appelée sortie *brute* ou *raw* en anglais ;

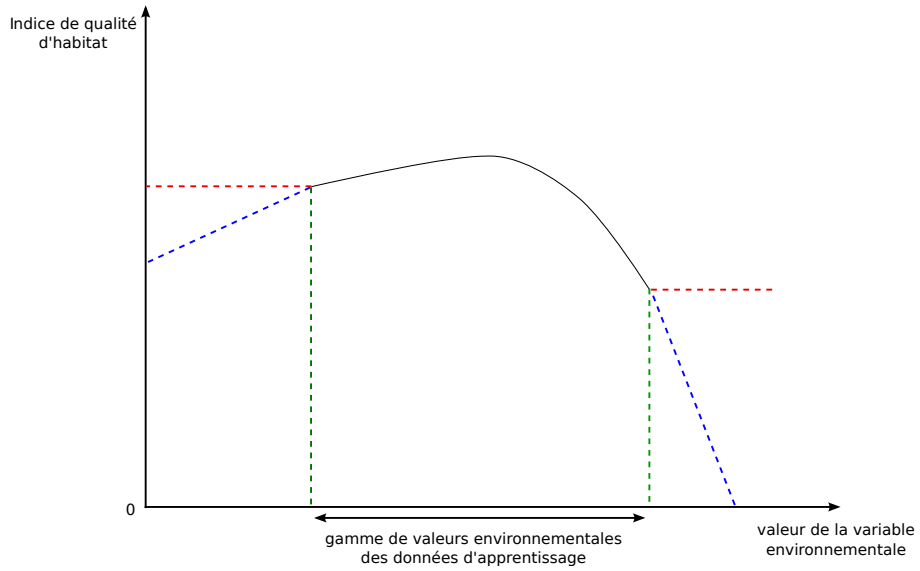


FIGURE 1.4 – **Représentation de l'extrapolation.**
La ligne continue noire représente la courbe de réponse générée par le modèle à partir des données d'apprentissage. Au-delà des conditions environnementales représentées par les données d'apprentissage, la ligne en tiret rouge correspond à la situation où l'option *clamp* est choisie, la ligne en tiret vert correspond à la situation où il n'y a pas d'extrapolation, la ligne en tiret bleue correspond à l'extrapolation "fadebyclamping".

— la somme cumulée des valeurs de $\hat{\pi}$ rangées dans l'ordre croissant, c'est-à-dire la fonction de répartition de $\hat{\pi}$. Cette sortie est dite *cumulative*, elle indique pour un pixel p donné, la probabilité d'avoir une valeur de qualité d'habitat inférieure à celle observée en p ;

— les probabilités de présence de l'espèce, en tout pixel de la zone d'étude. Cette sortie est dite "logistique", correspondant à la transformation de la sortie *brute* par l'équation (1.18) (Phillips and Dudík, 2008) :

$$P(y=1|x) = \frac{\tau e^{\lambda f(x)-r}}{1-\tau+\tau e^{\lambda f(x)-r}} \tag{1.18}$$

où r est l'entropie relative entre $P(x|y=1)$ et $P(x)$ (aussi appelée divergence). Le paramètre τ peut être interprété comme la probabilité d'observer l'espèce dans les conditions environnementales favorables à sa présence. Par défaut, sa valeur est fixée à $\tau = 0.5$ (Elith et al., 2011). Elith et al. (2011) soulignent que la valeur de τ dépend de la rareté de l'espèce et de la difficulté d'observation. Ils proposent de calculer cette valeur à partir de la prévalence au lieu d'utiliser la valeur par défaut. Cependant, cette prévalence, dans le cas des données de présence uniquement, est rarement connue. Merow et al. (2013) déconseillent l'utilisation de cette sortie à cause de la sensibilité à la valeur de τ souvent choisie arbitrairement, et conseillent l'utilisation de la sortie *brute* interprétée comme indice de qualité d'habitat.

II. 2. c. Évaluation

Pour évaluer le modèle, par défaut deux métriques sont utilisées par l'algorithme : le gain et l'aire sous la courbe ROC (AUC ou AUCROC pour *Area Under the Received Operating*

Characteristic Curve).
Le gain d'apprentissage régularisé (*regularized training gain*), aussi appelé gain, correspond à :

$$\text{gain} = ln(l) - \textit{ log loss} \tag{1.19}$$

où l est le nombre de pixel de X. Le gain mesure la vraisemblance aux sites de présence et indique combien de fois la distribution de Maxent ajuste mieux les données de présence par rapport à une distribution uniforme correspondant à un gain nul.
L'AUC est basée sur la courbe ROC. Cette dernière permet de caractériser la performance d'un classifieur binaire (présence *vs.* absence, positif *vs.* négatif, ...). Elle est construite en représentant la *sensibilité* en fonction de *1-spécificité* et en faisant varier la valeur du seuil séparant les deux classes. La *sensibilité* est la proportion de vrais positifs parmi les éléments effectivement positifs. Elle est également appelée taux de vrais positifs (*true positive rate*) et est égale à *1 - l'erreur d'omission* (c'est-à-dire la proportion d'éléments positifs n'ayant pas été détectés). La quantité *1- spécificité* est appelée taux de faux négatif (*false positive rate*. La *spécificité* (ou *erreur de commission*) correspond à la proportion de vrai négatifs parmi les éléments effectivement négatifs. L'aire sous la courbe ROC, l'AUC, est interprétée comme la probabilité globale (indépendemment du seuil) de correctement séparer les éléments positifs et négatifs. Une valeur d'AUC de 1 correspond à 100 % de chance de discriminer les éléments positifs des éléments négatifs. Une valeur de 0.5 correspond à une chance sur deux de discriminer les éléments positifs des éléments négatifs. Autrement dit, la valeur 0,5 correspond à un classifieur aléatoire.
Appliqué à Maxent, étant donné que les données d'absence n'existent pas, les faux positifs ne peuvent pas être identifiés. L'abscisse de la courbe ROC correspond alors à la proportion de la surface de la zone d'étude associée à la prédiction de la présence de l'espèce. L'AUC est alors interprétée comme la probabilité de mettre en évidence les sites de présence parmi l'ensemble des sites définissant la zone d'étude.

Pour évaluer la contribution des différentes variables à la construction du modèle, l'algorithme Maxent dispose de deux méthodes :

- la méthode heuristique, qui consiste à calculer la contribution (en pourcentage) de chaque variable à la construction du modèle final. Durant le processus d'apprentissage, l'augmentation du gain (correspondant à la diminution du *log loss*) à chaque itération est due à l'ajustement un à un des poids des variables transformées. Cette augmentation est assignée à la variable environnementale associée au poids considéré lors de l'itération. La somme des augmentations du gain pour l'ensemble des itérations indique le pourcentage de contribution de chacune des variables environnementales à la construction du modèle ;
- la méthode de jackniffe consiste à retirer une à une chacune des variables environnementales et d'évaluer l'effet de ce retrait sur le modèle. Pour cela, pour chacune des variables, deux modèles sont construits, un modèle avec uniquement la variable concernée et un autre sans celle-ci. La méthode compare ensuite les gains obtenus dans les deux cas.

D'autres indices d'évaluation qui ne sont pas proposées dans les implémentations de Maxent citées précédemment, peuvent être utilisées, tels que le ratio des AUC partiels ou

de l'indice de Boyce. Selon la littérature, ces métriques semblent les plus adaptées aux modèles de présence.
Le ratio des AUC partiels (Peterson et al., 2008) est le ratio entre l'AUC observée et l'AUC correspondant à l'AUC d'un classifieur aléatoire, tout deux calculées sur une partie de la courbe ROC, correspondant à une valeur de taux d'erreur d'ommission choisie par l'utilisateur.
L'indice continu de Boyce (Hirzel et al., 2006) est un indice adapté aux modèles de présence uniquement. Cette méthode consiste à partitionner le domaine de valeurs de qualité d'habitat en b classes $\{c_i\}_{i\in[1,..,b]}$. Pour chaque classe, deux fréquences sont calculées, la fréquence prédite et la fréquence attendu. La fréquence prédite est basée sur les sites de présence conservés pour l'évaluation du modèle. Pour une classe de qualité d'habitat c_i, cette fréquence est le ratio entre le nombre de sites d'évaluation prédits par le modèle comme appartenant à c_i et le nombre total de sites d'évaluation. La fréquence attendue est le ratio entre le nombre de pixel appartenant à la classe de qualité d'habitat c_i et le nombre de pixel de la zone d'étude. Le ratio des fréquences prédite et attendue est calculé pour chaque classe. Hirzel et al. (2006) notent que pour un bon modèle, le ratio des fréquences doit augmenter lorsque la qualité d'habitat (définie par la classe c_i) augmente. Le coefficient de corrélation de Spearman est ensuite calculé entre le ratio des fréquences et la qualité d'habitat (définie par c_i). L'indice obtenu est appelé indice continu de Boyce et varie entre -1 et 1. Une valeur positive signifie que les prédictions sont en concordance avec la distribution des sites de présence, une valeur proche de 0 signifie que le modèle n'est pas plus discriminant qu'un modèle aléatoire et une valeur négative signifie un modèle "incorrect".

II. 3. Avantages et faiblesses de Maxent

Comme indiqué précédemment, Maxent est une méthode de présence-background qui ne requiert que des sites de présence. Ces méthodes sont plus discriminantes que les méthodes de présence-uniquement compte tenu de l'utilisation des données de background qui permettent d'informer sur les conditions environnementales disponibles dans la zone d'étude. Cette méthode peut prendre en compte aussi bien des variables catégorielles que des variables continues, contrairement à ENFA qui ne considère que des variables continues. Les sorties (brutes, logistiques ou cumulées) de ce modèle sont de type numériques et continues, permettant une comparaison quantitative des différents lieux dans une même zone d'étude. L'utilisation du paramètre de régularisation dans la méthode permet d'éviter le surapprentissage. Les travaux d'Elith et al. (2006) et Hernandez et al. (2006) ont également montré que la méthode Maxent est moins sensible au nombre de sites de présence et que, lorsque ce nombre est faible, les performances de prédiction peuvent être supérieures à celles des méthodes comparables.
De plus, les modèles construits avec cette méthode sont faciles à mettre en oeuvre grâce à l'existence de programmes gratuits, libres et relativement aisés à appliquer. Cependant, Phillips et al. (2009) et Elith et al. (2011) soulignent que cette méthode, comme toutes les méthodes qui n'utilisent que les données de présence, est très sensible aux biais d'échantillonnage. La problématique du biais d'échantillonnage fait l'objet des paragraphes suivants.

III. Biais d'échantillonnage

Généralement, les données de présence sont issues de sources différentes (centres d'archives, musées, organismes publics ou privées). Les collectes sont généralement effectuées

avec des protocoles différents en terme de fréquence de capture, de nombre de pièges et/ou de captureurs déployés sur la zone d'intérêt et de technique de capture. Ces données peuvent aussi bien concerner des espèces envahissantes pour lesquelles la détection de l'espèce est souvent aisée, que des espèces rares ou difficiles à capturer. Enfin, certaines localités sont souvent plus échantillonnées que d'autres du fait de facilités de capture ou d'accès, conduisant à un biais d'échantillonnage.

Il existe deux types de biais :

— le biais d'échantillonnage géographique, il est dû à un échantillonnage effectué généralement proche des habitations et des axes de transports. Les données recensées dans la littérature sont très souvent entachées de ce biais ;

— le biais d'échantillonnage environnemental est rencontré lorsque toutes les conditions environnementales de la zone d'étude n'ont pas été échantillonnées. Cette situation est tout à fait possible lorsqu'un jeu de données ne présente pas de biais d'échantillonnage géographique et, réciproquement, un jeu de données biaisé géographiquement peut ne pas l'être dans l'espace environnemental. En effet, cela dépend de la distribution des conditions environnementales dans l'espace géographique de la zone d'étude.

De nombreuses méthodes utilisées pour la modélisation de la distribution d'espèces font l'hypothèse que l'échantillonnage est effectué de manière aléatoire uniforme sur l'ensemble de la zone d'étude (Phillips et al., 2006). La présence d'un biais d'échantillonnage environnemental ou géographique fait que les caractéristiques des données violent cette hypothèse. Cependant Yackulic et al. (2013) notent que parmi les publications parues entre 2008 et 2012, 87 % des publications portant sur Maxent n'utilisent pas des données issues d'un échantillonnage aléatoire et que, parmi ces publications, seules trois utilisent une méthode de correction de l'effet du biais d'échantillonnage. Or, Phillips et al. (2009) soulignent que si le biais d'échantillonnage n'est pas pris en compte, le modèle construit correspond à un modèle d'effort d'échantillonnage et non à un véritable modèle de distribution de l'espèce étudiée. Les différentes méthodes existantes de correction de l'effet du biais d'échantillonnage existants sont décrites dans la partie suivante.

IV. Correction de l'effet du biais d'échantillonnage

Il existe différents types de correction de l'effet du biais d'échantillonnage sur la MDE :

— la sélection des sites de présence (Anderson and Raza, 2010; Carroll, 2010). En effet, le biais d'échantillonnage se traduit souvent par une distribution spatiale des échantillons sous forme de grappes (ou clusters). Ce phénomène, aussi appelé *clumping* en anglais, se traduit par une auto-corrélation spatiale des données de présence provoquant un surapprentissage. La sélection des sites de présence consiste à rééchantillonner les sites de présence afin de tendre vers une distribution spatiales uniforme des sites ;

— la sélection des sites de pseudo-absence, Zaniewski et al. (2002) soulignent que la manière de sélectionner les sites de pseudo-absence influence fortement la qualité du modèle. En effet, des sites de pseudo-absence choisis aléatoirement peuvent se trouver dans une zone favorable à l'espèce, correspondant ainsi à des sites de "fausse absence". Pour cela, certains auteurs proposent de pondérer la probabilité de sélection des sites de pseudo-absence par la qualité d'habitat issue d'un premier modèle (Engler

et al., 2004; Hengl et al., 2009) ;

— la sélection des sites de background selon le même effort d'échantillonnage que celui des sites de présence, afin d'avoir le même biais d'échantillonnage environnemental dans les deux jeux de données. Cependant, l'effort d'échantillonnage est rarement connu.

Deux principales approches de correction de l'effet du biais d'échantillonnage lors de l'utilisation de Maxent ont été recensées dans la littérature :

— la sélection des données de présence ;

— la construction d'un background biaisé.

Ces approches tentent toutes deux d'obtenir le même biais d'échantillonnage dans les données de présence et dans le background, en se basant soit sur des critères géographiques, soit sur des critères environnementaux. Ces approches sont détaillées ci-après.

IV. 1. Sélection des données de présence

Cette approche a été développée pour uniformiser la distribution des sites de présence sur la zone d'étude. Elle consiste à effectuer un filtrage des sites de présence basé soit sur des critères géographiques soit sur des critères environnementaux. Les sites de background sont ensuite sélectionnés de manière aléatoire uniforme sur la zone d'étude.

IV. 1. a. Sélection des sites de présence basée sur des critères géographiques

La méthode de sélection des sites de présence basée sur des critères géographiques consiste à retirer les sites de présence situés à une distance géographique inférieure à une distance d de leur plus proche voisin (Boria et al., 2014; Kramer-Schadt et al., 2013). Cette distance d est déterminée par l'utilisateur et est basée sur les caractéristiques de la zone d'étude ou de l'espèce étudiée. Dans le premier cas (Boria et al., 2014), cette distance dépend de l'hétérogénéité spatiale de la zone d'étude. L'hypothèse de cette méthode est que deux sites de présence situés à au moins une distance d l'un de l'autre auront des conditions environnementales différentes. Dans le deuxième cas (Kramer-Schadt et al., 2013), la distance d est déterminée en fonction du domaine vital de l'espèce, c'est-à-dire de la zone géographique où vit l'espèce et qui répond à ses besoins primaires.
Boria et al. (2014) montrent que les données de présence filtrées permettent d'obtenir un modèle moins spécifique et de meilleures performances de prédiction que les données non-filtrées.

IV. 1. b. Sélection des sites de présence basée sur des critères environnementaux

La méthode de sélection des sites de présence basée sur des critères environnementaux consiste à retirer les sites de captures associés à des conditions environnementales jugées similaires (Varela et al., 2014; Fourcade et al., 2014). Pour cela, Varela et al. (2014) sélectionnent deux variables environnementales (la pluviométrie et la température) pour définir, dans le plan définie par ces deux variables, une grille climatique où la taille de la cellule correspond à un dixième de l'étendue des valeurs dans chaque dimension. Un nombre maximal de sites de présence appartenant à une même cellule est ensuite fixée. Fourcade et al. (2014) proposent quant à eux, de réaliser une analyse en composantes principales (ACP) sur des données environnementales des sites de présence afin de définir l'espace environnemental

avec des axes factoriels indépendants. Ensuite, une classification ascendante hiérarchique des sites de présence est réalisée dans cet espace environnemental, en utilisant la distance Euclidienne. Le nombre de classes est arbitrairement fixée à la moitié du nombre total de sites de présence. Les classes sont alors censées correspondre à des environnements significativement différents et représentatifs de la zone d'étude. Enfin, un seul site de présence est conservé par classe.

IV. 2. Construction d'un background biaisé

Dans ce cas, le principe n'est plus de faire tendre les distributions des sites de présence et du background vers des distributions uniformes. En effet, Phillips et al. (2009) proposent de corriger l'effet du biais d'échantillonnage en sélectionnant les sites de background avec le même biais d'échantillonnage environnemental que celui associé aux sites de présence. Cependant, il est difficile de connaître ce biais d'échantillonnage car l'effort d'échantillonnage réel est rarement connu. Les méthodes suivantes ont été développées pour approcher le biais d'échantillonnage des données de présence.

IV. 2. a. Construction d'un background biaisé basée sur des critères géographiques

Elith et al. (2010) proposent d'estimer l'effort d'échantillonnage à partir des seules données de présence. En chacun des pixels de la zone d'étude, l'effort est obtenu par le ratio du nombre de sites de présence dans son voisinage géographique sur le nombre total de pixels dans ce même voisinage. Le voisinage géographique d'un pixel est défini par une fonction de forme gaussienne centrée sur le pixel et avec un écart-type correspondant à la capacité de déplacement de l'espèce étudiée (dans Elith et al. (2010), l'écart-type est fixé à 200 km). Plus l'effort d'échantillonnage d'un pixel est élevé, plus il aura de chance d'être sélectionné comme site de background.
Kramer-Schadt et al. (2013) proposent de créer une grille de biais (dans leur étude, la grille est de 1km^2), où la valeur de l'effort d'échantillonnage d'une cellule quelconque est égale à la somme du nombre de sites de présence situés dans la cellule elle-même et dans ses huit plus proches voisines.
Une autre méthode consiste à sélectionner les sites de background dans un rayon donné autour des sites de présence (Fourcade et al., 2014).

IV. 2. b. Construction d'un background biaisé basée sur des critères environnementaux

Cette méthode consiste à sélectionner les sites de background avec des conditions bioclimatiques identiques à celles rencontrées sur les sites échantillonnés.
Hill and Terblanche (2014) et Webber et al. (2011) proposent de caractériser l'environnement de la zone d'étude en se basant sur des cartes bioclimatiques. Les cartes utilisées sont :

- la carte de Köppen-Geiger, issue d'une classification des grandes zones bioclimatiques dans le monde, où le nombre de classes est de 30 ;
- une carte bioclimatique, issue d'une stratification des données climatiques appelée stratification environnementale globale (Metzger et al., 2013).

Les sites de background sont sélectionnés dans les classes bioclimatiques où la présence de l'espèce a été observée.

IV. 2. c. Construction d'un background biaisé basée sur les groupes cibles

Phillips et al. (2009) proposent de construire un background biaisé à partir des données relatives aux *groupes cibles* (appelé *target groups* en anglais) (Ponder et al., 2001) pour approcher le biais d'échantillonnage de l'espèce cible. L'espèce d'intérêt, ou espèce cible, est l'espèce dont nous cherchons à estimer l'effort d'échantillonnage. Un *groupe cible* est un groupe d'espèces, incluant l'espèce cible, collectées avec les mêmes méthodes de capture que l'espèce d'intérêt. Un groupe est censé représenter l'activité de collecte de l'espèce d'intérêt. Le *groupe cible* peut rassembler des espèces appartenant à des groupes biologiques différents à condition que leur capture obéisse au même protocole, c'est-à-dire aux mêmes critères de choix des sites et des périodes d'échantillonnage et aux mêmes méthodes de capture (techniques, équipements, ...), que celui utilisé pour l'espèce d'intérêt. Les sites de présence des espèces du groupe cible sont considérés comme étant représentatifs de l'effort d'échantillonnage de l'espèce cible et sont directement utilisés comme sites de background. L'ensemble de ces sites est alors appelé *target-group background* .

Les différentes méthodes de corrections sont répertoriées dans le tableau 1.3, en fonction de l'approche et des critères utilisés.
Une des principales difficultés est de choisir quelle méthode utiliser en fonction du contexte applicatif et des données disponibles. Plusieurs études comparatives ont été menées afin de faciliter ce choix.

IV. 3. Comparaison des méthodes de correction de l'effet du biais d'échantillonnage

À notre connaissance, seuls les travaux publiés par Fourcade et al. (2014), Kramer-Schadt et al. (2013) et Varela et al. (2014) ont comparé entre elles certaines des méthodes décrites précédemment. Les travaux de Varela et al. (2014) ont comparé la méthode de sélection de sites de présence basée sur des critères géographiques (appelée *geographic filter*) à celle basée sur des critères environnementaux (appelé *climatic filter*). Kramer-Schadt et al. (2013) ont comparé la méthode de sélection de sites de présence basée sur des critères géographiques (*spatial filter*) et la méthode de construction d'un background biaisé basée sur les critères géographiques (appelée *background manipulation with bias file*). Fourcade et al. (2014) ont quant à eux comparé les cinq méthodes suivantes :

- la sélection des sites de présence basée sur des critères géographiques (appelée *systematic sampling*) ;
- la sélection des sites de présence basée sur les critères environnementaux (appelée *cluster*) ;
- une méthode de construction d'un background biaisé à partir de critères géographiques basé sur la construction d'une grille de biais, appelée *bias file* ;
- une méthode de construction d'un background biaisé à partir de critères géographiques à partir d'un rayon défini autour des sites de présence, appelée *restricted background*) ;
- une méthode qui consiste à diviser les données de présence en deux : un jeu regroupant uniquement les sites situés au Nord et un deuxième regroupant les données situées au Sud. Deux modèles sont construits sur l'ensemble de la zone d'étude, l'un avec les données du Nord et l'autre avec les données du Sud. Le modèle final est la combinaison des sorties de ces deux modèles. Cette méthode est appelé *split* et ne correspond à aucune des approches de correction identifiées précédemment.

Le tableau 1.4 répertorie les comparaisons effectuées.

Parmi les trois études réalisées, seules celles de Fourcade et al. (2014) et de Kramer-Schadt et al. (2013) comparent au moins une méthode une méthode consistant à sélectionner les sites de présence à au moins une méthode basée la construction d'un background biaisé. Elles montrent que les méthodes consistant à sélectionner les sites de présence permettent aux modèles d'atteindre de meilleures performances de prédiction. Varela et al. (2014) montrent qu'en utilisant une méthode de sélection des sites de présence basée sur des critères climatiques, même lorsque le nombre de sites de présence filtrées est très faible (au nombre de cinq), l'AUC moyen obtenu reste meilleur que celui du modèle dont les sites de présence ont été sélectionnés à partir de critères géographiques. Cependant, cette méthode de sélection de sites de présence n'est basée que sur deux variables environnementales. Dans les cas réels, il est rare que la caractérisation environnementale se réduise à deux variables. La méthode appelée *cluster* (Fourcade et al., 2014), qui consiste a sélectionner les sites de présence selon une classification ascendante hiérarchique, semble plus pertinente car elle se base sur 14 variables environnementales. Fourcade et al. (2014) soulignent que les méthodes de sélection des sites de présence elles tendent à compromettre la précision du modèle, lorsque celui-ci est initialement faible.
Dans ces études, seule la construction d'un background biaisé basée sur des critères géographiques a été comparée aux méthodes de sélection des sites de présence. Elle semble permettre de moins bonnes performances de prédiction, mais, contrairement aux autres méthodes, le nombre initial de sites de présence ne semble pas être un facteur limitant à son utilisation.

IV. 4. Avantages et limites des méthodes de correction

Comme précisé dans la partie IV. 3. , les méthodes de sélection des sites de présence permettent d'obtenir des performances plus élevées que celles basées sur la construction d'un background biaisé. Les méthodes basées sur la sélection des sites de présence sont généralement faciles à mettre en œuvre. La sélection des sites de présence à partir de critères géographiques permet d'uniformiser les données de présence pour approcher un échantillonnage aléatoire uniforme d'un point de vue géographique mais elle ne permet pas d'avoir une distribution uniforme d'un point de vue environnemental. La sélection des sites de présence basée sur les critères environnementaux permet cette uniformisation. Cependant, les méthodes existantes permettent de travailler uniquement avec des variables environnementales continues et quantitatives. La principale limite de ces méthodes de sélection des sites de présence est la nécessité de bénéficier d'un grand nombre de sites de présence, ce qui n'est pas toujours possible.
Les méthodes de construction d'un background biaisé permettent de considérer l'ensemble des données de présence, ce qui est particulièrement intéressant lorsque celles-ci sont en faible quantité. Les méthodes basées sur les critères géographiques sont faciles à mettre en œuvre car un rayon, représentant le domaine vital de l'espèce ou l'hétérogénéité de la zone d'étude, est défini autour des sites de présence et définit les zones de sélection des sites de background. Cependant, dans un rayon donné autour d'un site de présence, plusieurs conditions environnementales significativement différentes peuvent coexister. Les sites de background ainsi sélectionnés présentent un biais d'échantillonnage géographique comparable à

celui des sites de présence, mais peuvent avoir un biais d'échantillonnage environnemental significativement différent. Cette méthode ne permet donc pas d'assurer l'équivalence des biais d'un point de vue environnemental. Les méthodes existantes de construction d'un background biaisé basées sur les critères environnementaux existantes permettent d'assurer cet équivalence. Cependant, le biais d'échantillonnage environnemental est rarement connu. Certaines méthodes tentent d'estimer le biais d'échantillonnage en utilisant des cartes bioclimatiques pré-établie. La méthode de construction d'un background biaisé consistant à sélectionner le background dans les mêmes classes bioclimatiques que celles des sites de présence est basée sur des cartes bioclimatiques pré-établie au niveau mondial. La qualité de ces cartes définie celle du background. De plus, ces cartes bioclimatiques ne tiennent pas forcément compte des corrélations des variables environnementales entres elles. La méthode de construction du background basée sur les groupes cibles semble être une bonne alternative pour estimer le biais d'échantillonnage. Cependant, lorsque le nombre des sites de captures du groupe cible est faible, cela réduit fortement le nombre de sites de background et, par conséquent les performances du modèle.

Les méthodes existants présentent ainsi des limites qui les rendent trop spécifiques ou inadaptées au contexte applicatif. Il apparait donc pertinent de proposer une méthode originale et générique pouvant :

— être appliquée à un nombre de sites de présence faible ;
— assurer l'équivalence des biais d'un point de vue environnemental ;
— être appliquer lorsque les variables environnementales sont qualitatives et quantitatives.

Les avantages et les limites des méthodes existants, présentés dans le présent paragraphe, sont résumés dans le tableau 1.3.

V. Conclusion

Dans ce chapitre, les différents modèles de distribution des espèces ont été présentés. Parmi ceux-ci, nous avons choisi la méthode Maxent basée sur le principe de maximisation de l'entropie, étant donnée qu'elle ne nécessite que des données de présence, qu'elle est peu sensible au nombre de sites de présence et que ses performances de prédiction restent satisfaisantes quand ce nombre est faible. Cependant, comme tous les modèles n'utilisant que les données de présence, Maxent est très sensible au biais d'échantillonnage. Pour corriger l'effet de ce biais, différentes approches décrites dans la littérature ont été présentées. Les études comparatives destinées à évaluer ces méthodes ont été décrites, avant de conclure sur les avantages et les limites propres à chacune des méthodes répertoriées. Enfin, le besoin d'explorer d'autres approches a été identifié.
Dans le chapitre suivant, nous proposons de développer une méthode de correction de l'effet du biais d'échantillonnage générique et originale adaptée à un faible nombre de sites de présence.

Intitulé de la transformation	**Description de la transformation**	**Contraintes imposées sur la distribution de probabilité estimée $\hat{\pi}$**
Linéaire (L)	$f_j(x) = g_j(x)$	$\hat{\pi}[f_j] = \tilde{\pi}[f_j]$
Quadratique (Q)	$f_j(x) = g_j(x)^2$	$\hat{\pi}[f_j^2] - \hat{\pi}[f_j]^2 = \tilde{\pi}[f_j^2] - \tilde{\pi}[f_j]^2$
Produit (P)	$f_{ij}(x) = g_j(x) \times g_i(x)$	$\hat{\pi}[f_j f_i] - \hat{\pi}[f_j]\hat{\pi}[f_i] = \tilde{\pi}[f_j f_i] - \tilde{\pi}[f_j]\tilde{\pi}[f_i]$
Seuil (*Threshold*, T)	Soit h un seuil, $f_j(x) = \begin{cases} 0 \text{ si } g_j(x) < h \\ 1 \text{ sinon.} \end{cases}$ ou l'inverse	$p(\hat{\pi}(g_j(x) > h)_{x \in X}) \approx p(\tilde{\pi}(g_j(x) > h)_{x \in [1,\ldots,m]})$ $p()$ étant la proportion
Linéaire par morceau (*Hinge*, H)	Soit h un seuil, $f_j(x) = \begin{cases} 0 \text{ si } g_j(x) < h \\ \frac{g_j(x)-h}{max(g_j(x))-h} \text{ sinon.} \end{cases}$ ou l'inverse	$\hat{\pi}[f_j(x) > h] = \tilde{\pi}[f_j(x) > h]$
Catégorie (C)	La variable de k classes sera transformée en k variables binaires	$p(\hat{\pi}(g_j(x)_t)_{x \in X}) \approx p(\tilde{\pi}(g_j(x)_t)_{x \in [1,\ldots,m]})$ $p()$ étant la proportion et t étant l'indice de classe parmi les k différentes classes de la nomenclature

TABLEAU 1.1 – **Détails des transformations de variables, g étant la variable environnementale initiale et f la variable transformée**

(a) Linéaire

Nombre de sites de présence	0	10	30	+ 100
β_L	1.0	1.0	0.2	0.05

(b) Linéaire + quadratique

Nombre de sites de présence	0	10	17	30	+100
β_Q	1.3	0.8	0.5	0.25	0.05

(c) Linéaire + Quadratique + Produit

Nombre de sites de présence	0	10	17	30	+100
β_P	2.6	1.6	0.9	0.55	0.05

(d) Seuil (*Threshold*)

Nombre de sites de présence	0	+100
β_T	2.0	1.0

(e) Linéaire par morceau (*Hinge*)

Nombre de sites de présence	+0
β_H	0.5

(f) Catégorique

Nombre de sites de présence	0	10	+17
β_C	0.65	0.5	0.25

TABLEAU 1.2 – **Valeurs de β en fonction des transformations et du nombre de sites de présence.**

		Critères		Avantages	Limites
		Géographiques	**Environnementaux**		
Approche	**Sélection des sites de présence**	Filtrage des sites de présence dans un rayon défini autour des sites de présence (Boria et al., 2014; Kramer-Schadt et al., 2013)	Filtrage des sites de présence ayant des valeurs environnementales similaires — espace environnemental construit à partir de deux variables environnementales (Varela et al., 2014) — espace environnemental construit à partir d'une ACP suivie d'une classification ascendante hiérarchique (Fourcade et al., 2014)	Facilité de mise en œuvre	Nécessité de beaucoup de sites de présence
	Construction d'un background biaisé	— Création d'une grille de biais d'échantillonnage (Elith et al., 2010; Kramer-Schadt et al., 2013) — Sélection des sites de background dans un rayon autour des sites de présence (Fourcade et al., 2014)	— Sélection des sites de background avec les conditions bio-climatiques similaires à celles des sites de présence (Hill and Terblanche, 2014) — *Target-group background* (Phillips et al., 2009)	Permet de considérer l'ensemble des données de présence quand celles-ci sont en faible quantité	Le biais d'échantillonnage est difficile à estimer
Avantages		Facilité de mise en œuvre	Assure l'équilibre des biais d'un point de vue environnmentale		
Limites		Ne permet pas d'assurer l'équivalence des biais d'un point de vue environnemental	Difficulté de mise en œuvre au regard de : — de l'existence de variables qualitatives et quantitatives — des corrélations entre variables — du nombre de variables		

TABLEAU 1.3 – **Méthodes de correction de l'effet du biais d'échantillonnage**

	Sélection de sites de présence basée sur des critères géographiques	Sélection de sites de présence basée sur des critères environnementaux	Construction d'un background biaisé basée sur des critères géographiques	Construction d'un background biaisé basée sur des critères environnementaux
Fourcade et al. (2014)	**Systematic sampling**	Cluster	— Bias file — Restricted background	–
Kramer-Schadt et al. (2013)	**Spatial filter**	–	Background manipulation with bias file	–
Varela et al. (2014)	Geographic filter	**Climatic filter**	–	–

TABLEAU 1.4 – **Les études ayant comparé les méthodes de correction de l'effet du biais d'échantillonnage.**
Sont mises en gras les méthodes ayant permis au modèle d'avoir de meilleures performances

Chapitre 2

Méthode de correction de l'effet du biais d'échantillonnage basée sur des critères environnementaux

Introduction 37
I. Description de la méthode de correction de l'effet du biais d'échantillonnage 37
I. 1. Définition de l'espace environnemental 37
I. 2. Définition du voisinage environnemental 38
I. 3. Définition du biais d'échantillonnage 38
I. 4. Sélection des sites de background biaisés 40
II. Discussion 40
III. Conclusion 41

Introduction

Dans le chapitre précédent, les avantages et les limites des méthodes de correction de l'effet du biais d'échantillonnage ont été présentés. Les études ayant comparé ces méthodes entre elles ont montré que les méthodes de correction qui consistent à sélectionner un sous-ensemble des sites de présence selon des critères géographiques permettaient aux modèles d'obtenir de meilleures performances. Cependant, dans certains cas d'étude, le très faible nombre de sites de présence limite l'utilisation de ces méthodes et, dans ce cas, la construction d'un background biaisé semble être l'approche la plus pertinente. Cependant, ces méthodes présentent certaines limites : soit les méthodes ne tiennent compte que du biais géographique et non du biais environnemental ; soit la construction du background dépend d'une carte bioclimatique ne tenant pas compte des corrélations des variables environnementales entre elles ; soit il est nécessaire d'avoir un grand nombre de sites de présence du groupe cible.
Dans ce chapitre, nous proposons une méthode originale et générique pouvant être appliquée à des données de présence de l'espèce ou du groupe cible en faible quantité, tenant compte du biais environnemental, et en considérant l'ensemble des variables environnementales, quelle que soient leur nature (qualitatif ou quantitative).

I. Description de la méthode de correction de l'effet du biais d'échantillonnage

I. 1. Définition de l'espace environnemental

Afin de baser l'approche sur les caractéristiques environnementales et non géographiques, il est nécessaire de définir une distance environnementale entre les sites (pixels). Or, il existe deux principales difficultés à la définition d'une telle distance : les variables environnementales sont potentiellement corrélées entre elles ; les variables environnementales peuvent être numériques (quantitatives) ou catégorielles (qualitatives). Ainsi, un nouvel espace de représentation des données défini à partir de l'ensemble des variables environnementales, est construit grâce à une Analyse Factorielle de Données Mixtes (AFDM). Une telle analyse est une méthode factorielle qui permet de traiter à la fois des variables catégorielles et des variables numériques (Pagès, 2015). Elle équivaut à réaliser conjointement une ACP sur les variables continues et une Analyse des Correspondances Multiples (ACM) sur les variables catégorielles.
À chaque pixel x de la zone d'étude X sont associées n valeurs environnementales $g_i(x)_{i\in[1,\dots,n]}$.
L'ensemble des l pixels de X est représenté dans un tableau de dimension $l \times n$, où les l pixels représentent ainsi les *individus statistiques*, et les n variables environnementales, les *variables d'analyse*.
L'AFDM est appliquée à ce tableau, et l'ensemble des axes factoriels obtenus, non corrélés entre eux, obtenu va définir l'espace environnemental dans lequel la similitude des individus d'analyse (les pixels) peut être appréhendée au travers de la distance euclidienne. Dans un tel espace, plus deux pixels sont proches l'un de l'autre, et plus leurs conditions environnementales sont comparables.

I. 2. Définition du voisinage environnemental

À chaque pixel x de X est associé un voisinage environnemental défini dans l'espace environnemental décrit précédemment. Le voisinage environnemental d'un pixel i regroupe l'ensemble des pixels ayant des conditions environnementales similaires ou proches des siennes. Le degré d'appartenance d'un pixel quelconque j au voisinage de i ,w_{ij}, est défini par une fonction d'appartenance de type gaussienne. Ce degré d'appartenance s'écrit :

$$w_{ij} = 0.5^{(d_{ij}/D_{min})^2} \tag{2.1}$$

où d_{ij} est la distance euclidienne entre les points i et j dans l'espace environnemental et D_{min} une distance seuil au delà de laquelle le degré d'appartenance du pixel j au voisinage de i passe en dessous de la valeur $0,5$. Au delà de D_{min}, c'est-à-dire en deçà d'un degré d'appartenance de $0,5$, l'appartenance de j au voisinage de i est considéré non significative. La fonction d'appartenance est représentée en figure 2.1. Les propriétés de w_{ij} sont les suivantes :

- $w_{ij} \in]0,1]$;
- $w_{ij} = 1$ lorsque $d_{ij} = 0$;
- $w_{ij} < 0,5$ lorsque $d_{ij} > D_{min}$.

D_{min} ne dépend pas des connaissances liées à l'étendue géographique du domaine vital de l'espèce étudié, comme dans la définition du background biaisé basée sur des critères géographiques, mais elle doit s'appuyer sur des connaissances liées à la bio-écologie de l'espèce. Sa valeur définit la dispersion de la fonction gaussienne et ainsi le voisinage environnemental. Plus sa valeur est faible, plus le voisinage environnemental est réserré autour de i. Le chapitre qui suit donne un exemple de définition de cette distance $Dmin$ en fonction des connaissances sur la bio-écologie de l'espèce d'intérêt.

I. 3. Définition du biais d'échantillonnage

Phillips et al. (2009) proposent de définir un groupe cible pour représenter l'effort d'échantillonnage et constitue le background. Dans ce travail, l'utilisation de groupe cible est également étudiée. Cependant, deux difficultés se présentent : le nombre de sites associés au groupe cible considéré est très faible (bien inférieur au nombre de sites de background par défaut, 10 000) limitant la capacité du groupe cible à représenter les conditions environnementales de la zone d'étude et à générer un modèle suffisamment général ; il n'est pas possible de définir un groupe cible.
Lorsqu'il est possible de définir un groupe cible mais que le nombre de sites associés est en nombre réduit, au lieu d'utiliser directement ces sites comme sites de background, ils sont utilisés pour refléter l'effort d'échantillonnage de l'espèce cible et ainsi estimer le biais d'échantillonnage devant guider la sélection des sites de background.
Lorsqu'il n'est pas possible de définir un groupe cible, l'estimation du biais d'échantillonnage se base uniquement sur les sites de présence, de la même manière que dans la méthode utilisée par Elith et al. (2010) (mais dans l'espace géographique).

Dans la suite, les *sites de capture* correspondent aux sites de présence de l'espèce cible auxquels s'ajoutent, s'ils existent, les sites de présence des autres espèces du groupe cible.
Dans un premier temps, l'espace environnemental est défini avec l'ensemble des pixels de

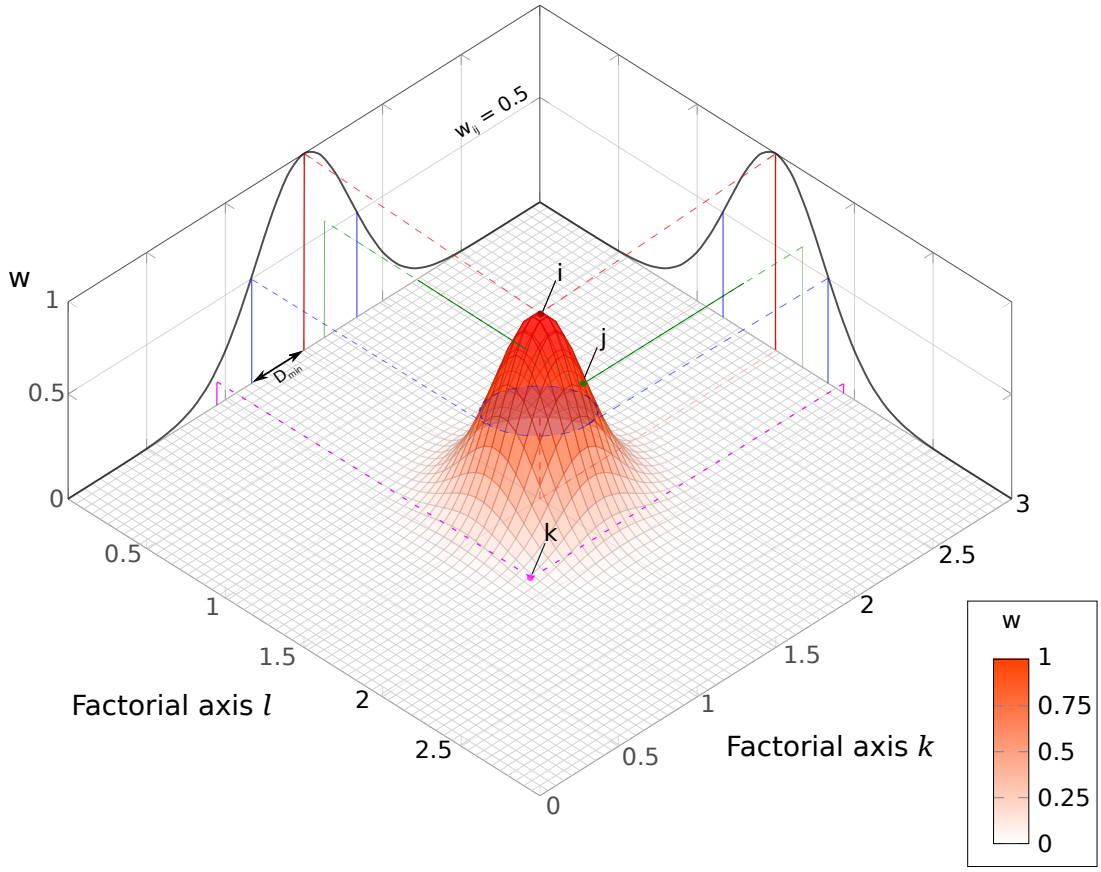


FIGURE 2.1 – **Représentation du voisinage du pixel i dans l'espace environnemental représenté par deux axes factoriels l et k.**
Le voisinage environnemental de i est représenté par la fonction d'appartenance de type gaussienne. La droite bleue définit le seuil du degré d'appartenance en deça duquel l'appartenance est jugée non significative. Seul le site j est considéré comme voisin de i dans cet exemple.

la zone d'étude selon la méthode décrite au paragraphe I. 1. Pour chacun des pixels de X est défini son voisinage environnemental selon l'approche définie au paragraphe I. 2. Le biais d'échantillonnage environnemental d'un pixel i est défini comme étant l'effort d'échantillonnage relatif dans son voisinage environnemental. Il est noté z_i et est défini par le ratio entre le nombre de sites de capture dans le voisinage environnemental de i et le nombre total de sites (pixels) dans ce même voisinage (représentant la disponibilité du contexte environnemental observé en i sur l'ensemble de la zone d'étude). L'effort d'échantillonnage relatif est donc défini par :

$$z_i = \frac{\sum_{j \in X} w_{ij}.c}{\sum_{j \in X} w_{ij}} \tag{2.2}$$

où $c = \{c_k\}_{k \in X}$, tel que $c_k = 1$ si k est échantillonné et $c_k = 0$ sinon.

L'utilisation de la disponibilité environnementale permet de distinguer deux environnements avec un même nombre de capture mais dont l'un est moins représenté que l'autre dans la zone d'étude. L'effort d'échantillonnage associé à l'environnement le plus rare sera plus élevé que celui associé à l'environnement plus fréquent. Les valeurs de z_i sont calculées pour l'ensemble des pixels de la zone d'étude.

I. 4. Sélection des sites de background biaisés

Le biais d'échantillonnage environnemental a été calculé précédemment pour l'ensemble des pixels de la zone d'étude. La sélection du background se fait de manière aléatoire et pondérée par l'effort d'échantillonnage relatif correspondant au biais d'échantillonnage. Ainsi, plus l'effort d'échantillonnage relatif d'un pixel est élevé, plus la chance de sélectionner un site de background dans un pixel sera élevée. Les sites de background auront des conditions environnementales plus ou moins proches de celles des sites de captures en fonction de la valeur attribuée à D_{min}. Ainsi les jeux de données de background et de présence présentent le même biais d'échantillonnage environnemental.

II. Discussion

Dans la méthode proposée dans ce chapitre, une AFDM est proposée pour définir l'espace environnemental à partir des données environnementale afin de se détacher complètement des critères géographiques comme utilisé dans les méthodes existantes. Cet espace environnemental permet la définition du voisinage environnemental de chaque pixel de la zone d'étude, sa paramétrisation est basée sur une connaissance *a priori* correspondant à $Dmin$.
Le choix de l'utilisation de l'AFDM comporte un avantage important, elle permet de tenir compte à la fois des variables environnementales catégorielles et continues. Dans certains contextes, il serait cependant intéressant de faire appel à une Analyse Factorielle Multiple (AFM, Pages and others (2004)) ou à une Analyse Factorielle de Groupe Mixte (AFGM, Pagès (2004)) afin de structurer les variables en groupes, chaque groupe correspondant à un facteur environnemental donné dont on cherche à équilibrer la contribution vis-à-vis des autres facteurs (Roux et al., 2011, 2013).
Dans la méthode de correction proposée, le voisinage environnemental est paramétré par la distance D_{min}. La valeur de cette distance influence la dispersion de la fonction gaussienne et définit ainsi la distribution des probabilités de sélection des sites de background. Plus D_{min} est faible et plus les sites de background sont sélectionnés à proximité des sites de capture. Elle ne doit donc pas être trop faible pour ne pas générer un modèle trop spécifique, c'est-à-dire qui "colle" trop aux données de présence. Il s'agit d'une distance euclidienne environnementale qui est définie à partir des connaissance de la bio-écologie de l'espèce et non sur des connaissances relatives à l'étendue spatiale du domaine vital de l'espèce, souvent difficile à définir.
Phillips et al. (2009) proposent d'utiliser les sites de présence associés au groupe cible comme site de background. Cependant, deux cas de figures peuvent être rencontrés, celui où le nombre de sites associés au groupe cible est faible et un autre cas où il n'est pas possible de définir un groupe cible. Dans le cas où le nombre de sites est faible, les sites associés aux groupes cibles sont utilisés pour estimer le biais d'échantillonnage et lorsqu'il n'est pas possible d'avoir un tel groupe d'espèces, seuls les sites de présence de l'espèce d'intérêt sont utilisés. L'utilisation du groupe cible comporte un avantage par rapport à l'utilisation des sites de présence, car certains points peuvent être équivalent à des sites de pseudo-absence. Ces sites de pseudo-absence sont des sites de présence des espèces appartenant au groupe cible et différentes de l'espèce d'intérêt. Cependant, de telles données ne sont pas toujours disponibles, ce qui contraint l'utilisateur à se baser uniquement sur les sites de présence.

III. Conclusion

Dans ce chapitre, une méthode originale de correction de l'effet du biais d'échantillonnage a été développée en se basant sur l'approche de la construction d'un background biaisé à partir de critères environnementaux. Cette méthode a été spécifiquement développé pour être adaptée au cas où le nombre de sites de présence est initialement faible.
Dans les prochains chapitres, cette méthode de correction sera appliquée à des données de présence associées au principal vecteur du paludisme en Guyane française, *Anopheles darlingi*, dont les captures sont largement biaisées par la proximité aux axes routiers et fluviaux. Ensuite, dans le chapitre 4, elle sera évaluée et comparée aux méthodes de correction existantes présentées dans le chapitre 1, dans un cas de figure où le nombre de sites de présence est faible. Afin de constituer un jeu de données de références nécessaire à une telle évaluation, des données de présence virtuelles seront simulées.

Chapitre 3

Modélisation de la distribution du principal vecteur du paludisme en Guyane française

Introduction . . . 45
I. Le paludisme . . . 45
II. Les *Anopheles* . . . 46
II. 1. Biologie des *Anopheles* . . . 46
II. 2. Comportement trophique des adultes *Anopheles* . . . 47
II. 3. Les vecteurs du paludisme . . . 48
III. La Guyane française . . . 49
III. 1. Géographie et météorologie de la Guyane . . . 49
III. 2. La population guyanaise . . . 51
IV. Le paludisme en Guyane . . . 51
IV. 1. Les acteurs de la surveillance du paludisme et de la lutte antivectorielle et les organismes de recherche . . . 52
IV. 1. a. La surveillance épidémiologique et la lutte antivectorielle . . . 52
IV. 1. b. La recherche sur le paludisme et ses vecteurs . . . 54
IV. 2. Historique du paludisme en Guyane . . . 56
IV. 3. Le paludisme actuellement . . . 56
IV. 4. Les vecteurs du paludisme . . . 58
IV. 4. a. *Anopheles darlingi*, vecteur principal . . . 59
IV. 4. b. Les vecteurs secondaires . . . 61
V. Données de présence des *Anopheles* en Guyane française . . . 61
V. 1. Méthodes de capture d'*Anopheles* . . . 62
V. 2. Données de captures d'*Anopheles* . . . 63
V. 2. a. Données de 1902 à 1999 . . . 64
V. 2. b. Données à partir de l'année 2000 . . . 64
V. 3. Base de données des captures d'*Anopheles* en Guyane . . . 64
VI. Modélisation de la distribution d'*Anopheles darlingi* . . . 67
VI. 1. Données environnementales . . . 67
VI. 2. Caractérisation environnementale . . . 70
VI. 3. Données météorologiques . . . 73
VI. 4. Construction des modèles . . . 75
VI. 4. a. Correction du biais d'échantillonnage . . . 76
VI. 4. b. Paramétrage du modèle . . . 76
VI. 4. c. Évaluation et validation du modèle . . . 77
VII. Résultats . . . 77
VII. 1. Performances de prédiction et contribution des variables environnementales et évaluation . . . 77
VII. 2. Carte de la qualité d'habitat . . . 80
VIII. Discussion . . . 83
VIII. 1. Lien entre les facteurs environnementaux et la qualité d'habitat . . . 83
VIII. 2. Correction de l'effet du biais d'échantillonnage . . . 84
VIII. 3. Le paludisme et la qualité d'habitat d'*An. darlingi* en Guyane française . . . 85
VIII. 4. Caractérisation environnementale . . . 86
IX. Conclusion . . . 86

Introduction

Dans le chapitre précédent, une méthode de correction de l'effet du biais d'échantillonnage a été développée.
Dans ce chapitre, le paludisme est abordé, notamment au travers de la modélisation de la distribution de son principal vecteur en Guyane française, *Anopheles darlingi*. La méthode Maxent ainsi que la méthode de correction de l'effet du biais d'échantillonnage présentée dans le chapitre précédent seront appliquées à ce vecteur sur l'ensemble de la Guyane française.
Le travail effectué dans ce chapitre a été réalisé en collaboration avec l'Unité d'Entomologie Médicale de l'Institut Pasteur de la Guyane et l'Institut de Recherche Biomédicale des Armées.

I. Le paludisme

Le paludisme est une parasitose à transmission vectorielle qui sévit principalement dans les pays de la zone intertropicale (fig. 3.1). En 2000, cette maladie avait touché près de 262 millions de personnes et avait fait près de 839 000 décès. En 2015, l'OMS (World Health Organization et al., 2015a) a recensé 214 millions de cas dont 438 000 décès, ce qui souligne le recul significatif de la maladie depuis 2000, mais également son impact encore exceptionnellement élevé sur la santé mondiale. Quatre-vingt sept pourcent de ces cas ont été recensés en Afrique et principalement chez les enfants de moins de 5 ans.

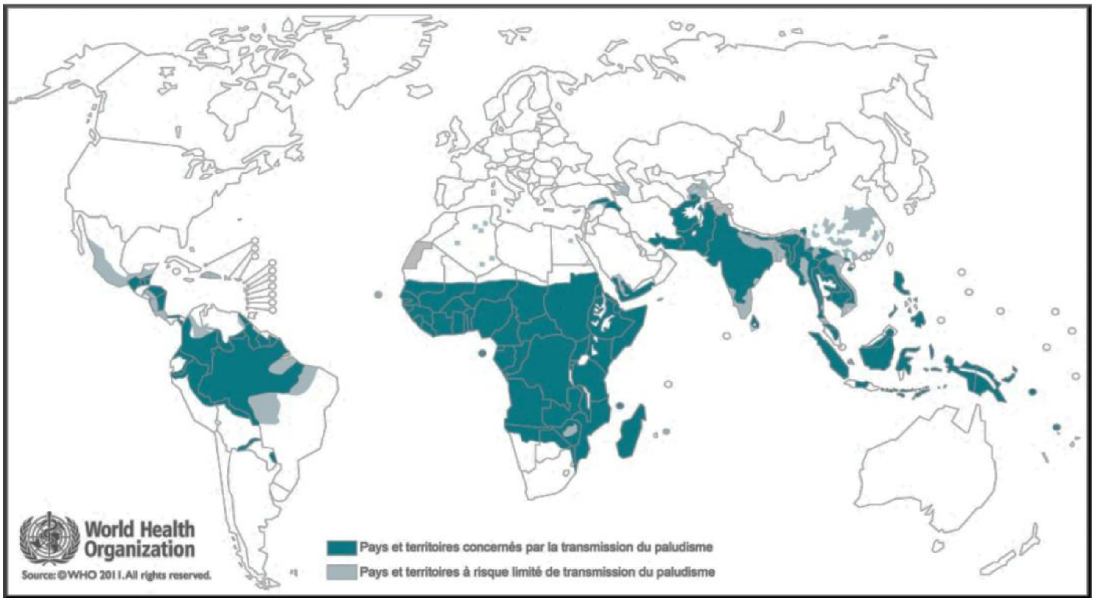


FIGURE 3.1 – **Répartition du paludisme dans le monde, en 2011.**
Source : WHO, 2011

Cette maladie ne peut pas se transmettre d'homme à homme sauf en cas de transfusion sanguine ou de transmission congénitale. Le cycle de transmission de cette maladie exige trois protagonistes : le parasite du genre *Plasmodium* qui évolue à travers un hôte invertébré ; le vecteur du genre *Anopheles* qui transmet le parasite à l'homme par piqûre, lors d'un repas de sang ; l'hôte humain.

Le cycle de développement de *Plasmodium* et de transmission du moustique à l'homme s'effectue selon le schéma en figure 3.2.

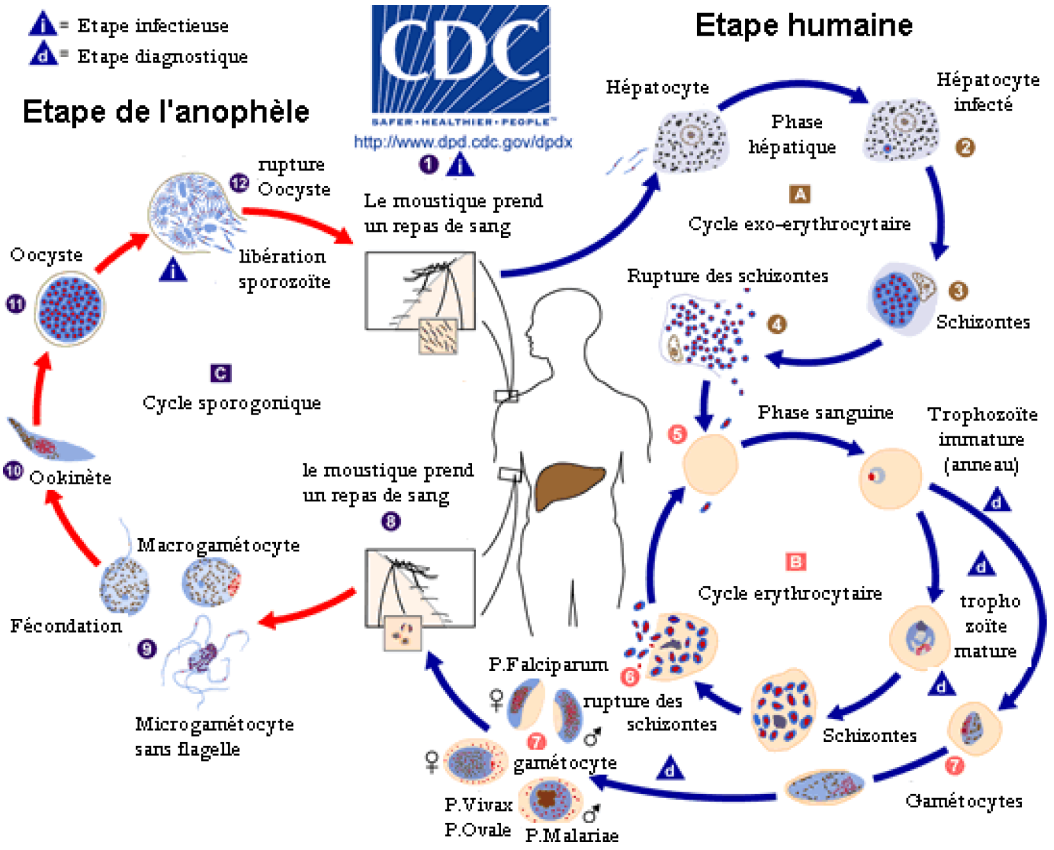


FIGURE 3.2 – **Cycle de transmission et de reproduction de *Plasmodium*.**
Source : http://www.dpd.cdc.gov/dpdx.

Il existe cinq espèces de *Plasmodium* capables d'infecter l'homme : *Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium knowlesi* et *Plasmodium ovale*. *Plasmodium falciparum* est le principal responsable de la mortalité causée par paludisme. *Plamodium vivax* touche plus de monde due à sa répartition géographique plus large, mais les accès palustres sont moins graves que *P. falciparum*. En revanche, contrairement à *P. falciparum*, il peut causer des reviviscences, des semaines ou des mois après la première infection. Les autres espèces sont moins répandues (World Health Organization, 2012).

Actuellement, les moyens de lutte considérés comme les plus efficaces contre cette maladie sont l'utilisation de moustiquaires imprégnées, limitant le contact entre l'homme et le vecteur, et la pulvérisation d'insecticides.

II. Les *Anopheles*

II. 1. Biologie des *Anopheles*

Les moustiques vecteur du paludisme sont du genre *Anopheles*. Harbach (2004) a recencé 484 espèces d'*Anopheles* dans le monde.
Les espèces d'*Anopheles* présentent quatre principaux stades de développement (Carnevale and Robert, 2009). Les trois premiers stades sont dits pré-imaginaux : l'oeuf, la larve et la nymphe, où ils sont exclusivement aquatiques. Le dernier stade, dit adulte ou imago, est aérien (figure 3.3).

La femelle pond entre quarante et cent oeufs à la surface de l'eau. Ils éclosent au bout de 24 à 48 heures. Chaque oeuf donne naissance à une larve. Durant le stade larvaire, la larve se nourrit à la surface de l'eau et mue plusieurs fois jusqu'à devenir une nymphe. Le stade nymphal dure généralement moins de 48 heures. Il correspond à la période où la morphologie évolue pour faire émerger un adulte (imago). Un adulte mâle vit entre une semaine et dix jours, tandis qu'une femelle vit entre deux à quatre semaines. Durant sa durée de vie, la femelle ne s'accouple qu'une seule fois mais peut effectuer plusieurs pontes successives. Lorsque la femelle est fécondée, elle nécessite un apport de protéines pour le développement de ses oeufs, qu'elle obtient au moyen d'un repas de sang, en piquant des vertébrés, dont l'homme. C'est au cours de ce repas sanguin, qu'elle peut ingérer ou transmettre le parasite du paludisme. Les mâles, quant à eux, se nourrissent exclusivement de nectars de fruits.

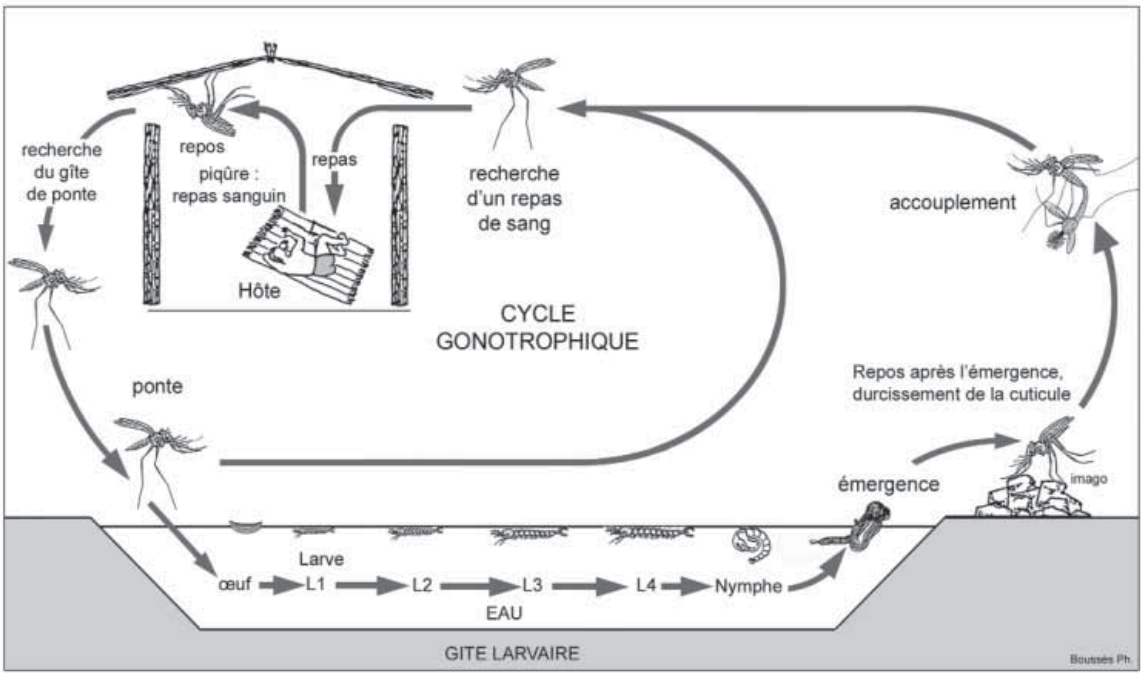


FIGURE 3.3 – **Cycle biologique d'*Anopheles*.**
Source : Carnevale and Robert (2009)

II. 2. Comportement trophique des adultes *Anopheles*

Généralement les espèces anophéliennes ont une activité nocturne, bien qu'une activité diurne de certaines espèces ait été signalée.
Les espèces qui se nourrissent sur l'homme sont dits *anthropophiles* tandis que celles qui se nourrissent sur les animaux sont dits *zoophiles*. Peu d'espèces sont exclusivement anthropophiles, les zoophiles sont les plus fréquentes. Certaines espèces se nourrissent aussi bien sur l'homme que sur le bétail, selon la disponibilité, même si elles ont une préférence pour l'un ou pour l'autre.
En cas d'anthropophilie, lorsque le repas de sang est pris à l'intérieur (respectivement à l'extérieur) des habitations, on parle de comportement *endophage* (respectivement *exophage*). Le cycle entre deux repas de sang est appelé *cycle gonotrophique*. Durant ce cycle, certains anophèles restent dans les habitations, ils sont dits *endophiles*, et ceux qui quittent les habita-

tions pour retrouver un site de repos extérieur sont dits *exophiles*. Le comportement trophique des espèces anophéliennes a un rôle majeur dans la réussite ou l'échec des méthodes de lutte antivectorielle. En effet, l'aspersion des murs des habitations a peu ou pas d'effet sur les espèces exophiles.

II. 3. Les vecteurs du paludisme

Parmi les 484 espèces citées par Harbach (2004), une cinquantaine est capable de transmettre le paludisme. Selon Pages et al. (2007), seules vingt d'entres elles sont des vecteurs avérés et assurent l'essentiel de la transmission du paludisme dans le monde. Les autres espèces ne participent pas à la transmission, soit parce qu'elles sont plutôt zoophiles, soit parce qu'elles sont réfractaires au parasite. Parmi les vecteurs avérés, certains sont considérés comme vecteur primaire et d'autres comme vecteur secondaire.
Un *vecteur primaire* ou *majeur* est une espèce responsable de la majorité des transmissions et présente généralement une grande répartition géographique. Un *vecteur secondaire* est une espèce qui ne transmet que dans certaines localités et sont peu nombreux ou peu infectés. Un *vecteur accidentel* est une espèce dont le rôle dans la transmission est peu connu du fait de la rareté de son infection et du peu de contact qu'elle a avec l'homme (Hamon and Mouchet, 1961).

Les vecteurs sont différents d'un continent à un autre et d'un environnement à un autre. La carte réalisée par Sinka et al. (2012) en figure 3.4, représente la répartition géographique des principaux vecteurs du paludisme dans le monde.

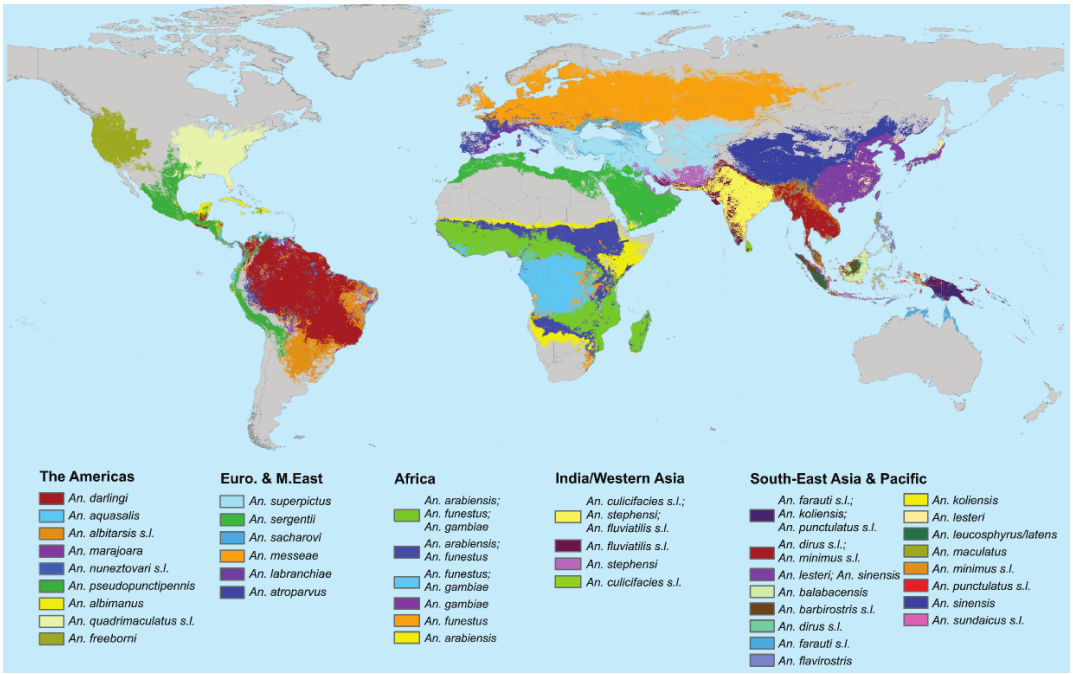


FIGURE 3.4 – **Répartition géographique des principaux vecteurs du paludisme dans le monde.**
Source : Sinka et al. (2012).

III. La Guyane française

III. 1. Géographie et météorologie de la Guyane

La Guyane est une collectivité territoriale située en Amérique du Sud. Elle est séparée du Suriname par le fleuve Maroni et du Brésil par le fleuve Oyapock et les monts Tumuc-Humac. Elle a une superficie d'environ 84 000 km², dont plus de 80 % est recouverte par la forêt équatoriale, et se situe entre les latitudes 2° et 6° Nord et les longitudes 51° et 53 ° Ouest (figure 3.5), dans la zone intertropicale de convergence (ZIC ou ZITC). Cette zone

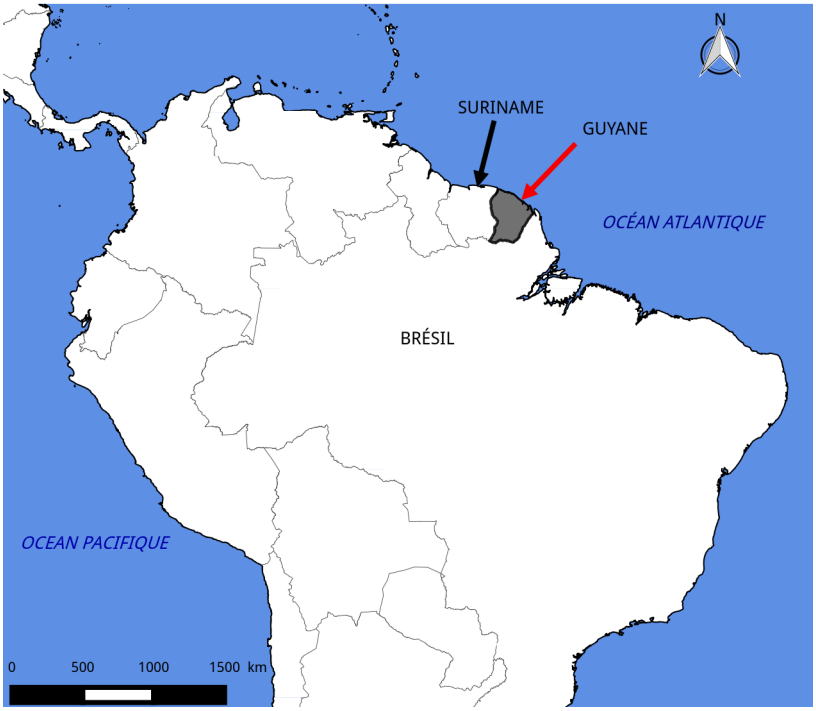


FIGURE 3.5 – **La Guyane française**

météorologique, qui oscille autour de l'équateur, provoque des variations météorologiques qui définissent ainsi les différentes saisons dans cette :

- la saison sèche, entre Juillet et Novembre ;
- la petit saison des pluies, en Décembre et Février-Mars ;
- le petit été de mars, comme son nom l'indique, durant le mois de Mars ;
- la saison des pluies, d'Avril à Juin.

La pluviométrie annuelle moyenne varie de 2 000 mm dans la zone la plus sèche, située au Nord-Est, à 4 000 mm dans la zone la plus humide au Nord-Ouest (figure 3.6). La pluviométrie mensuelle est supérieure à 100 mm tout au long de l'année à l'exception des trois mois les plus secs : septembre, octobre et novembre (Héritier, 2011).

Les températures restent quant à elles assez homogènes sur l'ensemble du territoire tout long de l'année. La température moyenne annuelle est de 26° C et la température journalière varie entre 19° à 37° C. Le taux d'humidité moyen varie entre 80 et 90 %.

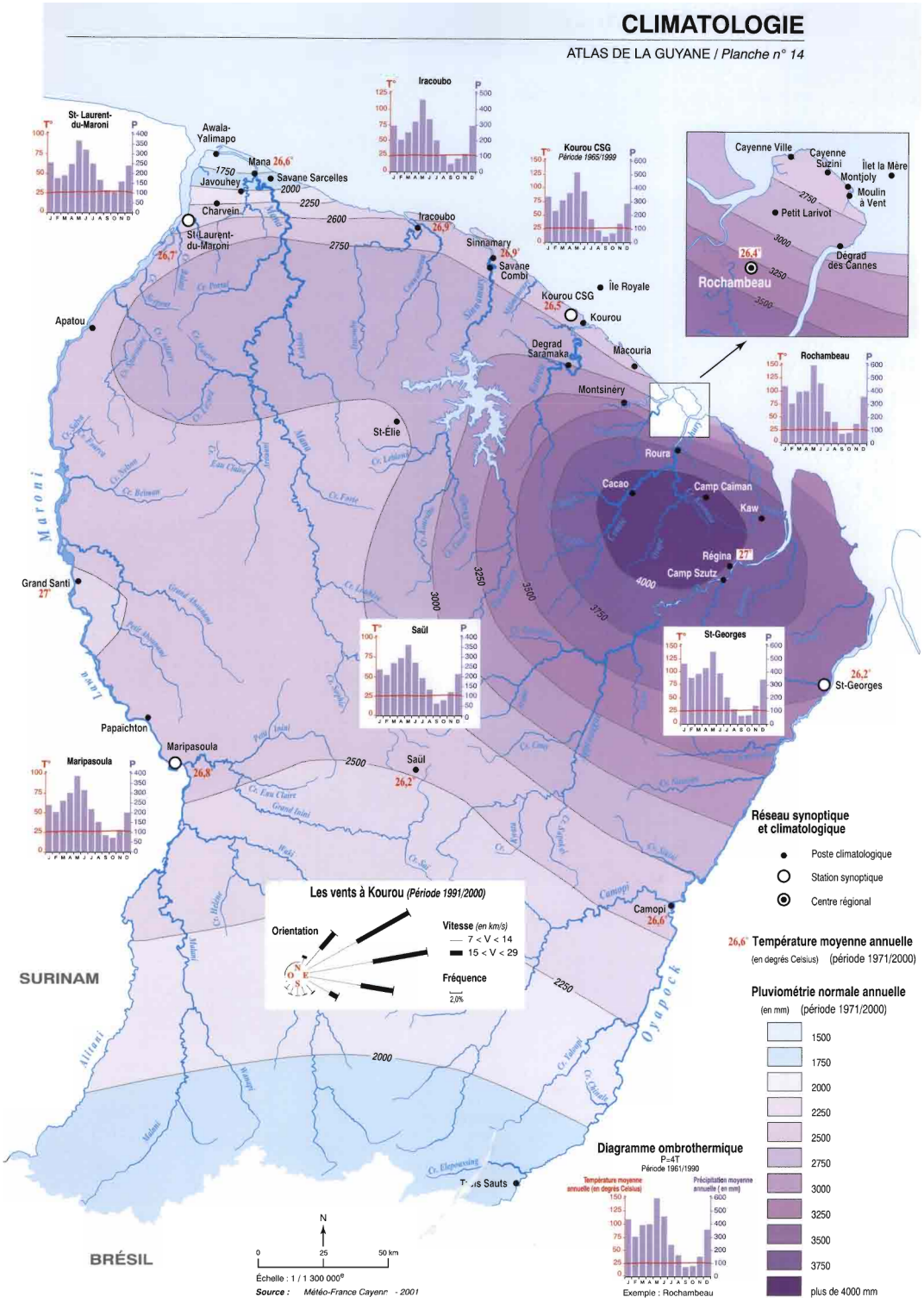


FIGURE 3.6 – **Carte de la pluviométrie annuelle.**
Source : Barret (2001)

Le relief du département est relativement plat, ne dépassant que rarement 200 mètres sur le littoral (Nord-Est) et est traversé par un réseau de hydrographique dense. Quelques monts et inselbergs se distinguent :

— dans le massif des Tumuc-Humac situé au Sud, à la frontière franco-brésilienne, où des

inselbergs peuvent atteindre 600 mètres ;
— dans le massif Inini-Camopi, au centre Ouest, avec un point culminant de 830 mètres ;
— au Nord Ouest, où le relief atteint 500m.

III. 2. La population guyanaise

Depuis les années 80, la population Guyanaise ne cesse d'augmenter (INSEE[1], figure 3.8). En 2015, la population guyanaise a été estimée à environ 260 000 habitants (INSEE). Neuf dixième de cette population vit le long du littoral où sont présentes les principales villes du département (Cayenne, Kourou et Saint-Laurent-du-Maroni). Le reste de la population vit majoritairement le long des deux grands fleuves transfrontaliers, l'Oyapock et le Maroni.
La population guyanaise présente une très forte diversité culturelle, avec plus de vingt groupes ethniques, chacun parlant sa propre langue. La figure 3.7 représente la répartition des différents groupes en Guyane.
Le nombre exact d'habitants vivant en Guyane est mal connu. La population en situation illégale a été estimée à près 40 000 personnes en 2009 (Rapport public annuel de la Cour des Comptes 2011). Les mouvements migratoires des pays voisins vers la Guyane française a commencé dans les années 40 après la fermeture des bagnes. La migration a été marquée en 1965 par l'arrivée de brésiliens fuyant les difficultés économiques de leur pays et profitant du besoin de main d'oeuvre pour la construction du Centre Spatial Guyanais (Institut National de la Statistique et des Etudes Economiques, 2006). Les années 70, et particulièrement le début des années 80, sont marquées par l'arrivée d'une population haïtienne fuyant les troubles politiques de leur pays. La politique de regroupement mise en place par l'OMI[2] au début des années 90 a permis une augmentation de 40 % de l'effectif de natifs haïtiens (Institut National de la Statistique et des Etudes Economiques, 2006). Entre 1986 et 1992, la Guerre civile au Suriname a conduit de nombreux ressortissants surinamais à rejoindre la Guyane française. Quatre camps de réfugiés ont été mis en place dans la commune de Saint-Laurent-du-Maroni pour les accueillir (Bourgarel, 1989). Bien que les crises et troubles politiques aient pris fin, les ressortissants sont restés en Guyane (Institut National de la Statistique et des Etudes Economiques, 2006). Outre ces mobilités dues à des crises majeures, des ressortissants des pays voisins de la Guyane passent les frontières pour la richesse des sous-sol guyanais.

IV. Le paludisme en Guyane

Mayotte et la Guyane française sont les seuls régions françaises où le paludisme est endémique et persiste malgré les actions de prévention et de lutte. En France, le paludisme est une maladie à déclaration obligatoire. En Guyane, le nombre de cas de paludisme a beaucoup évolué au cours du temps.

1. Institut National de la Statistique et des Études Économiques
2. Office des Migrations Internationales

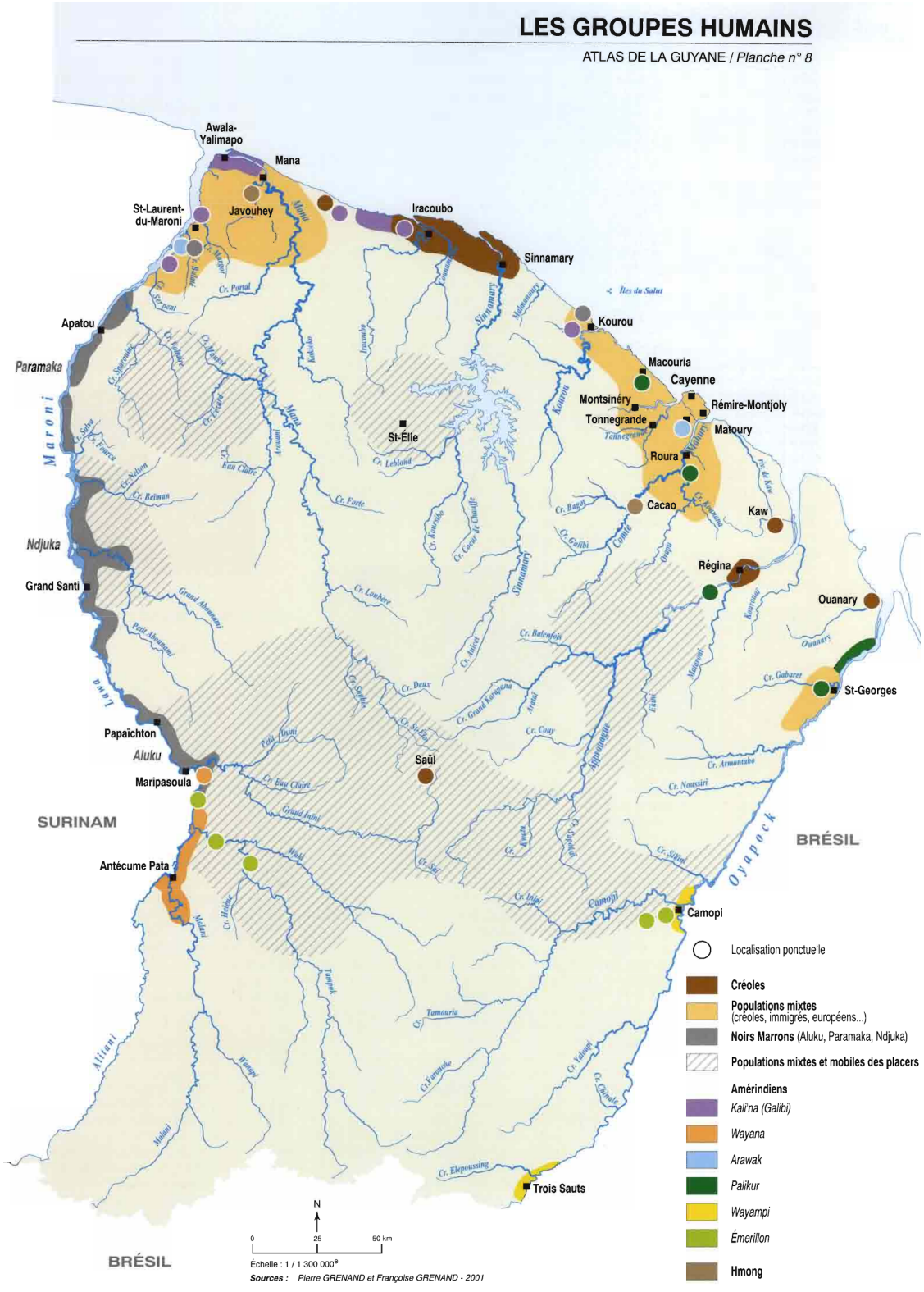


FIGURE 3.7 – **Répartition géographique des différents groupes humains.**
Source : Barret (2001)

IV. 1. Les acteurs de la surveillance du paludisme et de la lutte antivectorielle et les organismes de recherche

IV. 1. a. La surveillance épidémiologique et la lutte antivectorielle

La surveillance épidémiologique du paludisme en Guyane est organisée autour d'un système de surveillance des maladies à déclaration obligatoire mis en place par l'Agence Régio-

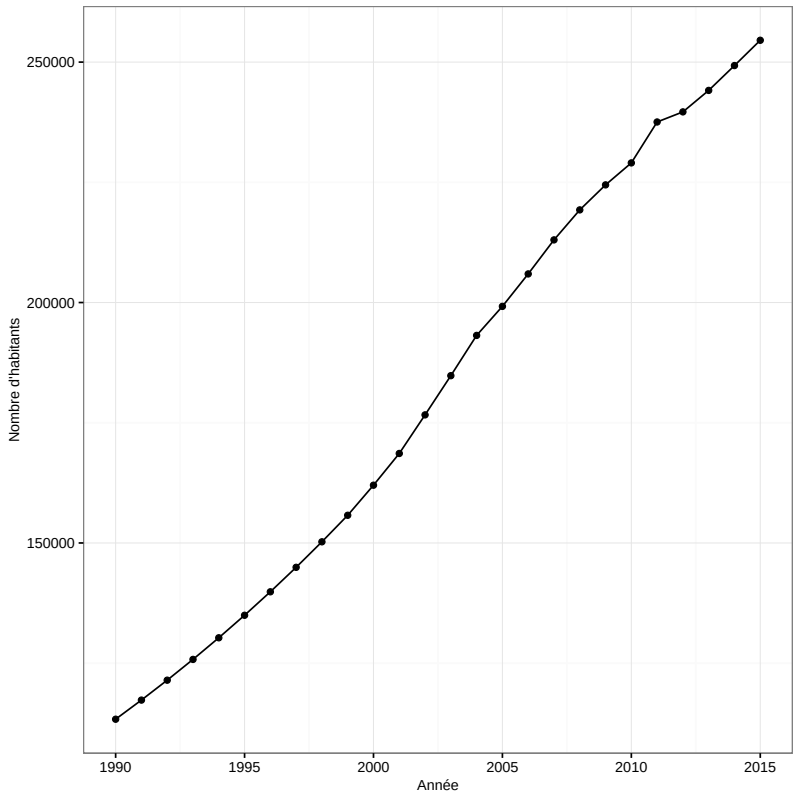


FIGURE 3.8 – **Nombre d'habitants en Guyane de 1990 à 2015**, selon les données de l'INSEE

nale de la Santé (ARS), et assurer par l'Agence Santé Publique France (anciennement l'Institut de Veille Sanitaire (InVS)) et plus spécifiquement par la Cellule d'intervention en région Antilles-Guyane (Cire Antilles-Guyane), récemment séparée en deux entités distincts (Cire-Antilles et Cire-Guyane) .

L'ARS a été crée en avril 2010, sa principale mission est d'assurer le pilotage du système de santé français en organisant la veille et la sécurité sanitaire, en contribuant à la gestion des crises sanitaires et en définissant, en finançant et en évaluant les actions de prévention de santé. L'ARS assure également la régularisation de l'offre de la santé, telle que la répartition de l'offre des soins (les médecins, les établissements médico-sociaux ou hospitaliers) sur le territoire.

L'Agence Santé Publique France repose sur 15 Cire réparties dans les régions françaises, 12 en métropole et trois en Outre-mer. Les missions des Cire sont de surveiller l'état de santé de la population de leur région et d'alerter les pouvoirs publics en cas de menace pour la santé publique. Elles apportent à l'ARS une expertise scientifique et technique pour l'aide à la décision. La CIRE Antilles-Guyane a été crée en 1997 et s'est séparé en deux Cire indépendants en 2016.

En Guyane française, la lutte antivectorielle est assurée par la Direction de la Démoustication et des Actions sanitaires (DDAS) du Conseil Général (jusqu'en 2015) et maintenant de la Collectivité Territoriale. Elle met en oeuvre les opérations de lutte antivectorielle sur l'ensemble de la région grâce, notamment, à ses antennes réparties sur le département. Les interventions

sont la pulvérisation intradomiciliaire d'insecticides et le traitement larvicides des marécages.

Le système de surveillance du paludisme s'organise selon la figure 3.9.

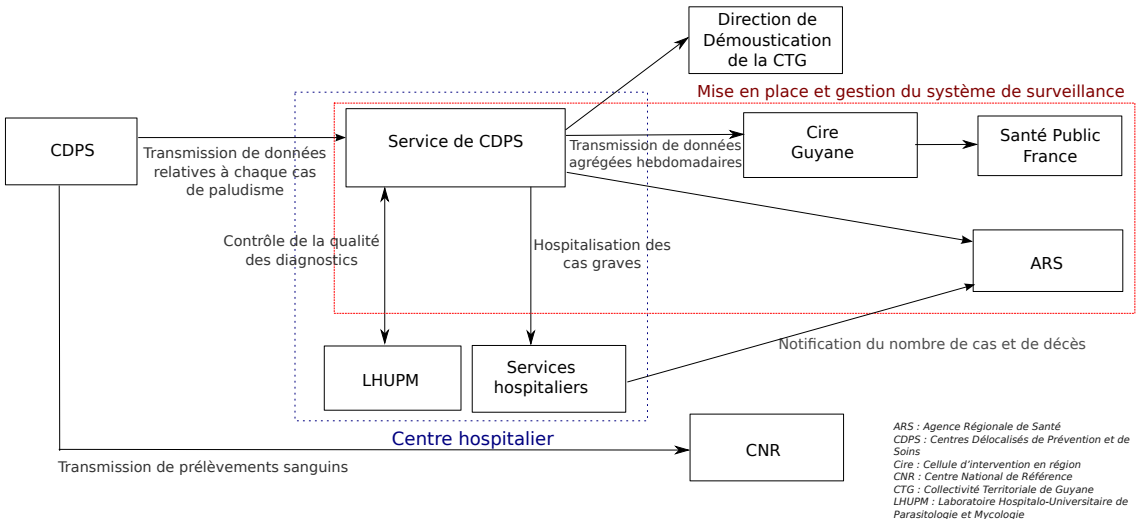


FIGURE 3.9 – **Système de surveillance du paludisme en Guyane.**

IV. 1. b. La recherche sur le paludisme et ses vecteurs

En Guyane, plusieurs organismes de recherche participent à la recherche sur le paludisme via l'étude des Plasmodies, des vecteurs, ou encore de l'interaction homme-vecteur-parasite. Le laboratoire de parasitologie de l'Institut Pasteur de la Guyane (IPG) est un laboratoire associé aux Centres Nationaux de Référence du paludisme. En France, les CNR sont des laboratoires experts dans la microbiologie et des observatoires des maladies transmissibles. Ils centralisent toutes les informations nationales concernant ces maladies et participent à la lutte et au contrôle des maladies. Depuis 2011, 47 CNR ont été crées avec 33 laboratoires associés. Les principaux axes de recherche du laboratoire associé au CNR paludisme sont :

- l'étude des résistances de *P. falciparum* aux anti-paludiques ;
- l'évolution et la diversité des Plasmodies, notamment l'adaptation des Plasmodies aux pressions médicamenteuses et la diversité génétique de *P. vivax* ;
- la recherche opérationnelle, consistant à évaluer les méthodes de diagnostic notamment les tests de diagnostic rapide.

L'Institut de recherche biomédicale des armées (IRBA) est une composante du Service de santé des armées (SSA) qui effectue des recherches biomédicales adaptées à l'armée. La Guyane représente des enjeux importants pour la France et l'Europe dans le domaine spatial, avec la présence du Centre Spatial Guyanais à Kourou, et environnemental, avec la lutte contre l'orpaillage clandestin et la pêche illégale. Les forces armées en Guyane (FAG) garantissent la protection du territoire national et animent la coopération régionale. Les FAG mènent leurs missions dans des milieux où les militaires sont confrontés à des environnements inhospitaliers, où les conditions de vies sont exigeantes et difficiles. L'IRBA étudie la santé des militaires et les interactions avec les environnements extrêmes dans lesquels ceux-ci évoluent. Il participe à l'étude sur la paludisme et à la lutte contre cette maladie. Les travaux de Pommier de Santi et al. (2016b) ont noté que lors d'une opération effectuée par les militaires sur les

sites d'orpaillage en 2011, près de 20% de soldats avaient été infectés par le paludisme. Des travaux menés sur les orpailleurs en situation illégale ont noté une prévalence du paludisme très élevée. Ceci expliquerait le taux d'infection élevé chez les soldats lors des interventions sur ces sites (Pommier de Santi et al., 2016a).

Le Centre d'Investigation Clinique - Epidémiologie Clinique (CIC-EC) Antilles-Guyane est une entité de l'Institut Nationale de la Santé et de la Recherche Médicale (INSERM), créé en janvier 2008. Le besoin de mieux connaître les caractéristiques épidémiologiques propres aux départements français d'Amérique justifie la création de ce centre. En effet, la spécificité de ces départements est lié à la diversité génétique des populations due à l'origine multiple des populations, à la complexité des migrations, à l'environnement tropical et à la persistance d'agents pathogènes qui y sont spécifiques. En Guyane, les thématiques de ce centre sont l'étude des facteurs de risque et de survenue des accès palustres via le suivi d'une cohorte d'enfant de 0 à 7 ans dans le village de Camopi (Hustache et al., 2007; Stefani et al., 2011a) et l'estimation de la prévalence du paludisme dans une population d'orpailleurs illégaux (Douine et al., 2016).

L'équipe associée à l'université de Guyane, Écosystèmes Amazoniens et Pathologie Tropicale (EPAT, anciennement Epidémiologie des Parasitoses Tropicales), créée en 1998, a pour principale mission de mener des activités de recherche opérationnelle sur les différentes pathologies tropicales amazoniennes (telles que le paludisme, la maladie de Chagas, la toxoplasmose, la leishmanioses cutanées, ...) endémiques en Guyane. Son objectif est d'acquérir de meilleures connaissances des contextes épidémiologiques et de leur évolution, des facteurs d'exposition, de risque et de gravité, d'améliorer la surveillance, d'alerte des phénomènes épidémiques, les moyens de diagnostic et les stratégies thérapeutiques et de prévention.

Créé en 2000, le Laboratoire Hospitalo-Universitaire de Parasitologie et Mycologie (LHUPM) est situé au Centre Hospitalier André Rosemon (CHAR) de Cayenne. L'équipe d'accueil EPAT est étroitement liée au LHUPM. Depuis 2008, le LHUPM héberge le CIC-EC. Sa base hospitalière facilite l'accès aux structures de santé à l'intérieur de la Guyane (Centres Délocalisés de Prévention et de Soins, CDPS), qui représentent des postes avancés facilitant la collecte des données et des échantillons pour les études sur le terrain. La relation étroite entre l'EPAT, le LHUPM et le CIC-EC facilite la coordination des activités de recherche.

L'Unité d'Entomologie Médicale de l'Institut Pasteur de Guyane s'intéresse aux insectes vecteurs de maladies. Ses recherches portent notamment sur les vecteurs du paludisme, les moustiques du genre *Anopheles*, et sur celui de la dengue, *Aedes aegypti*. Cette unité, en collaboration avec le SSA, mène des travaux visant à caractériser le rôle des anophèles dans la transmission du paludisme, et à évaluer leur résistance aux insecticides.

L'Unité Mixte de Recherche ESPACE-DEV (UMR 228) développe des recherches sur la spatialisation des dynamiques spatio-temporelles de l'environnement et des sociétés. Ses objectifs de recherche concerne la définition d'indicateurs de ces dynamiques, indicateurs biogéophysiques, indicateurs des évolutions des sociétés, des risques liés aux maladies émergentes en fonction de paramètres environnementaux, indicateurs des changements et de la

vulnérabilité des territoires aux changements globaux. Des projets de cette UMR portent sur l'étude et le suivi du paludisme dans la zone transfrontalière entre la Guyane et le Brésil.

IV. 2. Historique du paludisme en Guyane

De 1940 à 1948 Le nombre de décès causé par le paludisme était estimé à 230 habitants par an, pour une population d'environ 26 500 habitants (Floch, 1954).

De 1949 à 1970 L'introduction de nouvelles méthodes de lutte antipaludique : la chloroquine comme moyen thérapeutique et le dichlorodiphényltrichloroéthane (DDT) comme moyen de lutteantivectorielle, ont permis une importante diminution du nombre de cas (moins de 50 cas annuels). La transmission du paludisme devient très faible sur le littoral mais persiste le long des fleuves frontaliers l'Oyapock et le Maroni (Trape and Cordoliani, 1984).

De 1970 à 1999 Le nombre de cas de paludisme n'a cessé d'augmenter durant toute cette période. Il a atteint 117 cas en 1970, et dépassa 1000 cas en 1982. En 1987, 3 349 cas ont été diagnostiqués, soit une incidence de 38,5 cas pour 1000 habitants.

Cette forte augmentation était due aux facteurs suivants (Lepelletier et al., 1989) :

— un développement des centres de santé périphériques permettant un dépistage parasitologique passif par goutte épaisse ;
— la guerre civile au Suriname entrainant un afflux important d'immigrants vers la Guyane, drainant ainsi de nombreux malades vers les centres de santé français le long du Maroni (en 1987, près de 600 cas étaient originaire du Suriname) ;
— une surveillance active du paludisme, notamment dans les zones transfrontalières (Esterre et al., 1990) qui a consisté en un dépistage actif systématique permettant un prélèvement de lames plus importants et donc une meilleure estimation du nombre de malades

Lepelletier et al. (1989) notaient une répartition inégale des cas de paludisme. Cayenne et Saint-Laurent-du-Maroni ainsi que les communes de l'intérieur étaient considérées comme indemnes, tandis que les fleuves transfrontaliers constituaient des zones endémiques du paludisme. Le nombre de cas de paludisme annuel ne cesse d'augmenter jusqu'en 1999.

IV. 3. Le paludisme actuellement

En 2000, plus de 3000 cas de paludisme ont été diagnostiqués, et en 2005, ce nombre dépassa 4400 cas. En 2006, la zone du bassin du Maroni (incluant les rives surinamienne et française) était considérée comme ayant une incidence des plus élevées en Amérique du Sud (Chaud et al., 2006) et le taux d'incidence annuel chez les enfants de moins de sept ans à Camopi (moyen Oyapock) avait atteint les 100%. À partir de 2010, le nombre de cas en Guyane a fortement diminué (figure 3.10). Plusieurs facteurs pourraient expliquer cette diminution (Ardillon et al., 2012) :

— une large distribution et utilisation des moustiquaires imprégnées ;
— un recours au diagnostic rapide du paludisme grâce à l'utilisation des tests de diagnostic rapide (TDR) aussi bien en Guyane qu'au Suriname, débouchant sur un traitement antimalarique précoce ;
— des actions de lutte antipaludique au Suriname. En 2002 un programme appelé *Medical Mission Malaria Program* (MM-MP), en partenariat avec *Roll Back Malaria*, a été

mis en place afin de réduire la transmission du paludisme dans les communautés à risque vivant à l'intérieur du Suriname (Hiwat et al., 2012) ; en 2008, a débuté le programme *"Looking for gold, finding malaria"*, coordonné par le ministère de la Santé du Suriname, visant à réduire le nombre cas dans les zones d'orpaillage (Breeveld et al., 2012; Heemskerk and Duijves, 2012) en distribuant des moustiquaires imprégnées, en diagnostiquant et en administrant des traitements sur les sites d'orpaillages, en sensibilisant cette population et en formant des résidents des zones d'orpaillage à effectuer des TDR, à collecter du sang pour la recherche, à fournir des médicaments à des cas positifs et à effectuer des rapports. Ces actions ont permis une forte diminution du nombre de cas de paludisme le long du fleuve Maroni et par conséquent en Guyane française.

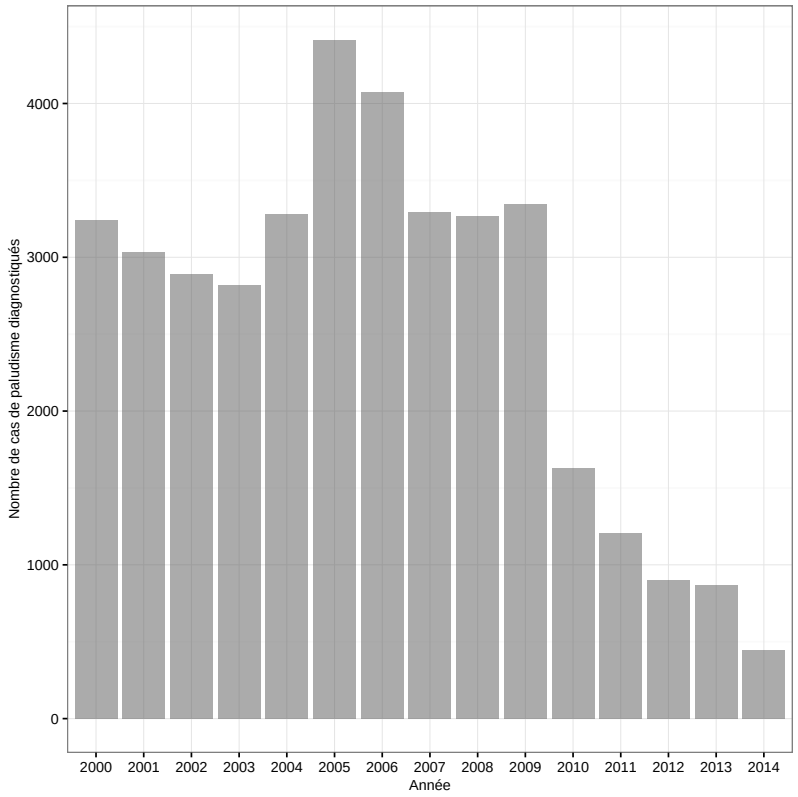


FIGURE 3.10 – **Nombre de cas de paludisme diagnostiqués en Guyane française entre 2000 à 2014.**
Sources : Bulletin de Veille Sanitaire - n°1 / Janvier 2015 – Cire Antilles-Guyane et Chaud et al. (2006)

En 2013, le nombre de cas est inférieur à 1000, avec une incidence globale inférieure à 4 cas pour 1000. Cependant, certaines localités comme Saül, Cacao, Régina, Camopi et Saint-Georges-de-l'Oyapock, ont connu une recrudescence du nombre de cas atteignant une incidence locale de 55,2 cas pour 1 000 (Musset et al., 2014).
En 2015, un plan de lutte contre le Paludisme en Guyane 2015- 2018 a été mis en place (Agence Régionale de Santé Guyane, 2015) dont les objectifs principaux sont de réduire la morbidité du paludisme et limiter les risques d'émergence de résistance en réduisant l'incidence à moins de 1 cas pour 1000 dans chaque localité de Guyane, ce qui correspond à la situation de pré-élimination.
Pour cela, cinq axes ont été définis :

1. Renforcer les surveillances épidémiologique et entomologique ;
2. Renforcer la prévention et la lutte antivectorielle ;

3. Soigner et accompagner sur l'ensemble du territoire ;
4. Développer la recherche et la connaissance ;
5. Renforcer la coopération internationale.

En effet, la répartition du paludisme reste très hétérogène sur le territoire, la bande littorale où vit neuf dixième de la population guyanaise est considérée comme peu touchée, avec une majorité de cas importés, alors que les zones avec une incidence locale élevée se situent plutôt à l'intérieur des terres. Les populations vivants en dehors de la zone littorale sont essentiellement des amérindiens, des noir-marrons et des populations mobiles comme les orpailleurs, les soldats et les exploitants forestiers. Parmi eux, les amérindiens, les orpailleurs et les soldats sont fortement infectés par le paludisme (Berger et al., 2012; Verret et al., 2006; Queyriaux et al., 2011; Hustache et al., 2007; Stefani et al., 2011b; Pommier de Santi et al., 2016a).
Cependant, la connaissance de la répartition géographique des vecteurs est partielle et limitée à l'Île de Cayenne, quelques villes du littoral, quelques villages le long du Maroni et de l'Oyapock. Les populations à risque vivent dans des zones où la distribution des vecteurs est méconnue. Les experts entomologistes estiment que mener un suivi des populations d'*Anopheles* adultes sur l'ensemble du territoire est irréaliste compte tenu du faible réseau routier, des difficultés d'accès à l'ensemble du territoire et des coûts financiers, humains et matériels que cela engendrerait (Agence Régionale de Santé Guyane, 2015). Pour cela, de nouveaux outils et méthodes de surveillance entomologique doivent être développés par la recherche opérationnelle pour la mise en oeuvre d'une surveillance des populations d'anophèles adultes en Guyane.
L'utilisation de données environnementales spatialisées ou issues de satellites a déjà été étudiée pour lutter contre le paludisme (Machault et al., 2011). En effet, l'évolution ou la répartition d'une maladie à transmission vectorielle dépend en grande partie des conditions environnementales déterminant la présence ou l'absence des parasites, des vecteurs et des hôtes. Un aspect important est d'identifier les facteurs susceptibles d'influencer la présence des agents pathogènes et/ou des vecteurs.
Pearson et al. (2007) préconisent l'utilisation des modèles de distribution d'espèces pour extrapoler la connaissance de la distribution des espèces dans le temps et dans l'espace. À partir d'informations recueillies sur quelques sites de captures, il est possible de prédire la distribution de l'espèce concernée sur l'ensemble d'une zone d'étude. En effet, Alimi et al. (2015) soulignent l'importance d'utiliser de tels modèles pour mieux appréhender la distribution des vecteurs du paludisme afin de mieux guider la lutte antivectorielle et d'aider dans la prévention des épidémies. Pour cela, il est nécessaire d'avoir des coordonnées géographiques des sites de captures de l'espèce concernée, ainsi que des données environnementales spatialisées relatives à des facteurs déjà été identifiés comme influençant sur la présence et/ou la densité de l'espèce.

IV. 4. Les vecteurs du paludisme

Le principal vecteur du paludisme en Amériques du Sud et Centrale est *Anopheles darlingi*. Cette espèce, largement distribuée, a été retrouvée du sud du Mexique jusqu'au nord de l'Argentine, et de l'Est de la cordillère des Andes jusqu'à la côte Atlantique. Ayant une aire de répartition très vaste, ce vecteur présente des comportements différents d'une région à une autre.

En Guyane française, 22 espèces d'*Anopheles* ont été recensées, dont *Anopheles darlingi* (*cf.* figure 3.11).

FIGURE 3.11 – ***Anopheles darlingi* lors d'un repas de sang.**

IV. 4. a. *Anopheles darlingi*, vecteur principal

L' efficacité d'*Anopheles darlingi* en tant que vecteur majeur est fortement liée à son comportement anthropophile, exo-endophage et souvent exophile (Mouchet, 2004).
Ce moustique, plutôt diurne, a une activité qui varie selon les lieux et les conditions géographiques. Rozendaal (1987) a observé, le long du Haut-Maroni, une activité élevée d'*An. darlingi* entre 23h30 et 1h30 en zone péridomestique. Hudson (1984) a notifié, dans la même région, trois pics d'activité à l'intérieur des maisons : un principal entre 21h et 23h et deux secondaires entre 18h-19h et 5h-6h. Pajot et al. (1977b) ont noté des pics d'activité différentes en fonction du lieu de capture. Le principal pic d'activité extérieur se situe entre 18h et 19h, suivi d'un pic secondaire entre 7h et 8h et une faible activité entre 9h et 16h. Sous les vérandas des habitations, les pics d'activité se situent entre 1h et 2h, entre 18h et 19 et dans une moindre mesure entre 7h et 8h. Le nombre de femelles capturées sous une véranda était deux fois supérieure à celui obtenu à l'intérieur des habitations. La véranda semblait favoriser l'activité d'*An. darlingi* en étant en adéquation avec son comportement endo-exophage et exophile, qui nécessite un retour à son site de repos en dehors des habitations après chaque repas sanguin.
La présence de ce moustique est directement liée aux conditions environnementales définissant son habitat écologique. Il choisit ses gîtes larvaires en fonction de leur composition chimique et de leur condition physique stable afin d'assurer la survie des larves.
Ce vecteur nécessite un gîte ensoleillé mais suffisamment ombragé pour conserver une température entre 20 et 28 °C (Hiwat and Bretas, 2011). Il a été retrouvé dans des plans d'eau douce larges, propres, avec de la végétation tels que les berges de cours d'eau, les criques, les retenues d'eau formées à proximité des cours d'eau après inondation, les marécages, les savanes et les forêts inondées ou inondables (Rozendaal, 1987; Hiwat et al., 2010). Les gîtes larvaires se situent rarement dans la forêt dense, à cause de l'acidité de l'eau et le manque d'ensoleillement sous le couvert forestier (Richard, 1985). Ils sont généralement situés en basse altitude et uniquement en eau douce (Deane et al., 1948).
Les adultes, fortement anthropophiles, sont capables de se déplacer jusqu'à 7 km (Charlwood, 1980) pour trouver leur hôte humain et effectuer leur repas de sang. Plutôt exophile, *An. darlingi* nécessite un lieu de repos à l'extérieur des habitations après chaque repas de sang. Ces zones de repos sont souvent des aires herbacées avec de la végétation secondaire et/ou arbustive. Vittor et al. (2006) soulignent le lien entre les zones où le taux de piqûres par personne est élevé et ces zones de repos.

(a) Activités forestières (crédits : E. Roux)

(b) Activités minières (crédits : J. Gaudet)

FIGURE 3.12 – **Exemples d'activités humaines**

Girod et al. (2011) ont montré un lien entre la densité d'*An. darlingi* et la variation de la pluviométrie en Guyane. Selon Martens et al. (1995), les températures optimales pour le développement de cette espèce se situent entre 20 et 30 °C, avec une humidité supérieure à 60%. Smith et al. (2013) ont recensé les seuils minimaux de pluie mensuelle pour garantir un site favorable aux larves des *Anopheles*. Les différents seuils se situent entre 10 et 80 mm, la quantité de pluie devant se maintenir au-dessus de ces seuils durant trois à quatre mois d'affilés.
Singer and Castro (2001) notent que les routes non goudronnées et les pistes avec des caniveaux de part et d'autres constituent des gîtes larvaires idéaux pour *An. darlingi*. Cependant, certains auteurs (Singer and Castro, 2001; Tadei et al., 1998) ont noté qu'une amélioration de la qualité des routes, en les goudronnant, en comblant les caniveaux et en mettant des trottoirs de part et d'autres des routes, a permis l'élimination d'*An. darlingi* dans certaines villes en forêt Amazonienne. Ils notent également que la forte déforestation créant une grande distance entre la forêt dense et les habitations et en exposant les gîtes larvaires au soleil, réduisait fortement la densité des vecteurs.

Stefani et al. (2013) se sont basés sur une recherche bibliographique systématique pour formaliser les connaissances sur l'impact du processus de déforestation, au travers d'un modèle conceptuel. Ce modèle distingue deux aspects importants :

— l'ouverture de la forêt dense et le maintien de l'interaction entre les aires déforestées et celles non-déforestées impliquent une diminution de la distance entre les gîtes larvaires, les sites de repos et les sites de "repas", favorisant ainsi la présence et la forte densité de ce vecteur et les rencontres homme-vecteur (exemple d'activités humaines favorisant la présence d'*An. darlingi* en figure 3.12) ;
— l'intensification de la déforestation, pour une urbanisation et/ou pour le développement d'une agriculture de grande surface, tend à augmenter la distance entre les gîtes larvaires et les sites de repos et/ou de repas, limitant les interactions homme-vecteur et les risques de piqûre.

Ces deux aspects constituent des caractéristiques importantes pour différencier une zone favorable d'une zone non-favorable à la rencontre homme-vecteur et donc à la transmission du paludisme.

IV. 4. b. Les vecteurs secondaires

D'autres espèces d'*Anopheles* peuvent intervenir dans la transmission du paludisme en Guyane. Les études de Dusfour et al. (2012a) et Pommier de Santi et al. (2016b) ont mis en évidence l'infection naturelle d'*An. nuneztovari, An. oswaldoi, An. intermedius* et de *An. marajoara*. Ces derniers sont considérés comme des vecteurs secondaires ou occasionnels dans la région mais peuvent être des vecteurs majeures dans d'autres pays d'Amazonie :

— *Anopheles nuneztovari* est une espèce forestière dont les larves ont été retrouvées dans les mêmes gîtes que ceux d'*An. darlingi*. Son comportement est variable en fonction du lieu géographique. Cette espèce peut être soit fortement anthropophile et exophile, et être considérée comme vecteur primaire (au Vénézuela et au nord de la Colombie, Gabaldon (1983); Hamon et al. (1970)) soit fortement zoophile et être considérée comme vecteur secondaire (au Surinam et au Brésil, Tadei and Dutary Thatcher (2000); Panday (1977)). Rozendaal (1987) a trouvé que 14,8% des moustiques capturés avec appât humain en zone péridomestique, le long du Maroni sur la rive surinamienne, correspondaient à *An. nuneztovari*, contre 63,9% sous-couvert forestier. En Guyane, Dusfour et al. (2012a) ont trouvé cette espèce naturellement infectée ;

— *An. oswaldoi* est décrit comme une espèce zoophile et exophile. Cependant il a été trouvé naturellement infecté par *Plasmodium* au Brésil (Branquinho et al., 1996) et en Guyane (Dusfour et al., 2012a) ;

— *An. intermedius* est une espèce souvent trouvée en forêt ou en lisière de forêt et est plutôt exophile et zoophile. Le contact entre cette espèce et l'homme est occasionnelle et peut se produire dans les habitations ou les camps à proximité de la forêt. Dusfour et al. (2012a) ont retrouvé cette espèce naturellement infectée par *Plasmodium* en Guyane ;

— *An. marajoara* est considéré comme un vecteur important du paludisme dans l'état brésilien d'Amapa (frontalier avec la Guyane). Sa présence est associée à la végétation secondaire et aux actions anthropiques (Conn et al., 2002). Certains auteurs notent que les activités humaines ont une influence plus importante sur la présence d'*An. marajoara* que sur *An. darlingi* (Moreno et al., 2007; Conn et al., 2002). En effet, l'installation de l'homme et de ses activités en forêt créé des zones favorables à *An. darlingi*. Cependant, la déforestation importante et la pollution des eaux réduisent le nombre de sites de pontes d'*An. darlingi* mais correspondent à des conditions favorables à *An. marajoara*. Ses larves se développent généralement dans des gîtes d'eau clair et ensoleillées tels que les rizières, les bassins de pisciculture ou les retenues d'eau des sites d'orpaillage. En Guyane, cette espèce a uniquement été capturée sur des sites d'orpaillage et a été retrouvée infectée par *Plasmodium* (Pommier de Santi et al., 2016b; Dusfour et al., 2012b).

V. Données de présence des *Anopheles* en Guyane française

Afin de modéliser la distribution des vecteurs du paludisme en Guyane, il est nécessaire d'avoir des localisations géographiques précises de la présence et de l'absence de ces espèces. Pour cela, une base de données a été mise en place à partir des données de présence des moustiques du genre *Anopheles*. Sa réalisation est décrite dans cette partie.

V. 1. Méthodes de capture d'*Anopheles*

Afin d'estimer l'intensité de la transmission du paludisme, le nombre de piqûres infectantes reçues par une personne durant un période donnée est très importante. Cette mesure est appelée taux d'inoculation entomologique (Entomological inoculation rate (EIR)). Cette valeur est importante dans la quantification du risque potentiel de transmission du paludisme et de la dynamique de transmission (Mathenge et al., 2002).
Une méthode directe pour obtenir le taux de piqûre des moustiques anthropophagiques est la capture sur homme (Davis et al., 1995). Muni d'un aspirateur à bouche ou mécanique, l'homme est à la fois l'appât et le capturer (*cf.* figure 3.13). Cependant, cette méthode est très critiquée

FIGURE 3.13 – **Capture sur homme**
Crédits : E. Roux

compte tenu du fort risque d'exposition du captureur aux piqûres infectées des moustiques (bien que la prophylaxie contre le paludisme est proposée aux captureurs, ces derniers restent exposés à d'autres maladies tels que le Zika, la dengue, ...), des reproductibilité et répétitivité faibles, et d'une efficacité jugée parfois insuffisante. En effet, les captures dépendent fortement de l'attractivité du captureur et du nombre de captureurs (Davis et al., 1995). Les captures sur homme se font le plus souvent de nuit et nécessitent une organisation conséquente, du temps et de la main d'oeuvre. Pour pallier ces inconvénients, des méthodes de captures moins contraignantes ont été mise en place :

- les pièges lumineux tel que le piège lumineux miniature (*cf.* figure 3.14) des Centres de Contrôle et Prévention des Maladies (Centers for Disease Control and Prevention miniature light traps, Odetoyinbo (1969)). Cependant certaines études ont montré que les pièges lumineux sous-estimaient le nombres d'*Anopheles* anthropophagiques (Zaim et al., 1986; Hii et al., 2000) ;
- les pièges à doubles moustiquaires comme le piège Mbita, où un homme est placé sous une première moustiquaire qui est elle-même placée sous une moustiquaire extérieure (Mathenge et al., 2002) ;
- les pièges odorants tels que les Mosquito Magnet® (*cf.* figure 3.15) utilisés avec différents attractants (octénol, Lurex®)

Ces pièges ont été utilisés pour les captures en Guyane. Dusfour et al. (2010) et Vezenegho et al. (2014) ont montrés qu'en Guyane les pièges à attractants (Mosquito Magnet® à octénol) étaient une bonne alternative à la capture sur homme.
Les *Anopheles* capturés sont ensuite morphologiquement identifiés en utilisant les clés taxo-

FIGURE 3.14 – **Piège lumineux**
Crédits : E. Roux

FIGURE 3.15 – **Piège odorant, Mosquito Magnet®.**
Crédits : E. Roux

nomiques spécifiques à la région (Floch and Abonnenc, 1951; Faran and Linthicum, 1981; Forattini, 1962).

V. 2. Données de captures d'*Anopheles*

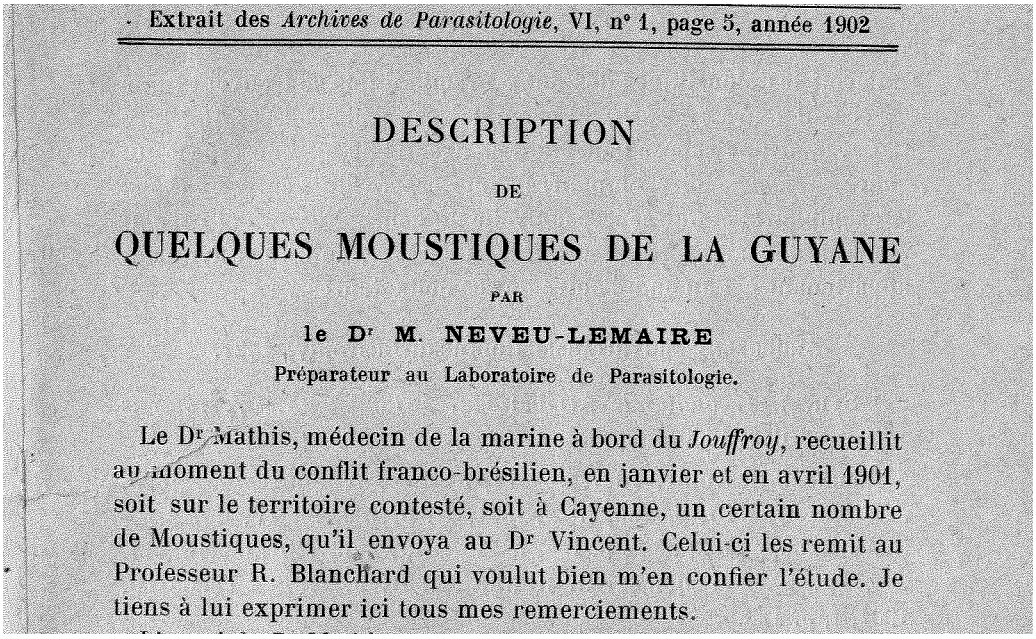
Les données de captures des espèces d'*Anopheles* rassemblées et étudiées dans cette thèse sont issues de différentes sources : l'Institut Pasteur de la Guyane, la DDAS, le SSA et l'ORSTOM[3]. Seules les données de présence étaient disponibles.
Le travail sur les données s'est fait en plusieurs étapes :

1. le recueil d'informations sur la présence des espèces *Anopheles* en Guyane française : nom de l'espèce capturée ; localité de capture et ses coordonnées géographiques si elles sont disponibles ; le stade (adulte ou larvaire) ; type de piège utilisé ;
2. la construction d'une base de données de présence d'*Anopheles* ;
3. la recherche des coordonnées géographiques des localités sans localisation précise ;
4. l'ajout de l'extension géographique à la base de données.

3. Actuellement l'Institut de Recherche pour le Développement (IRD)

V. 2. a. Données de 1902 à 1999

Les premières captures effectuées et notifiées en Guyane datent de 1901 et ont été publiées en 1902 dans les Archives de l'Insitut Pasteur d'Algérie (Neveu-Lemaire, 1902a,b). En 1940, est inauguré l'Institut Pasteur de la Guyane et du territoire de l'Inini à Cayenne, et devient l'Institut Pasteur de la Guyane en 1946. Les données de captures issues de l'Institut Pasteur proviennent essentiellement des archives, disponibles sous format papier (*cf.* 3.16).



Extrait des *Archives de Parasitologie*, VI, n° 1, page 5, année 1902

DESCRIPTION

DE

QUELQUES MOUSTIQUES DE LA GUYANE

PAR

le Dr M. NEVEU-LEMAIRE

Préparateur au Laboratoire de Parasitologie.

Le Dr Mathis, médecin de la marine à bord du *Jouffroy*, recueillit au moment du conflit franco-brésilien, en janvier et en avril 1901, soit sur le territoire contesté, soit à Cayenne, un certain nombre de Moustiques, qu'il envoya au Dr Vincent. Celui-ci les remit au Professeur R. Blanchard qui voulut bien m'en confier l'étude. Je tiens à lui exprimer ici tous mes remerciements.

FIGURE 3.16 – **Exemple de publication issue des Archives de l'Institut Pasteur de la Guyane**, Neveu-Lemaire (1902a)

Les rapports d'activités du DDAS de 1990 à 1999 ont également été utilisés pour recueillir les informations de présence d'*Anopheles*.
Ont également été utilisées certaines publications de l'ORSTOM mentionnant la présence d'*Anopheles*. Les documents mentionnant la présence d'au moins une espèce d'*Anopheles* et du lieu de capture sont répertoriés dans le tableau B.1 situé en Annexe B.

V. 2. b. Données à partir de l'année 2000

Les données à partir de 2000, sont issues des rapports d'activités du DDAS, des rapports d'activités de l'Unité d'entomologie Médicale de l'Institut Pasteur de Guyane, des relevés bruts lors des campagnes de captures, des relevés effectués sur les sites d'orpaillage par le SSA et des données issues des captures effectuées dans le cadre de la thèse d'Antoine Adde (Adde et al., 2016).
Ces informations sont répertoriées dans le tableau B.2 situé en Annexe B.

V. 3. Base de données des captures d'*Anopheles* en Guyane

Une base de données a été construite à partir de l'ensemble des captures effectuées entre 1901 et 2013. Elle a été créée sous PostgreSQL. Cette base de données est composée de 3880 entrées. Le schéma relationnel de cette base de données est représenté à la figure 3.17.

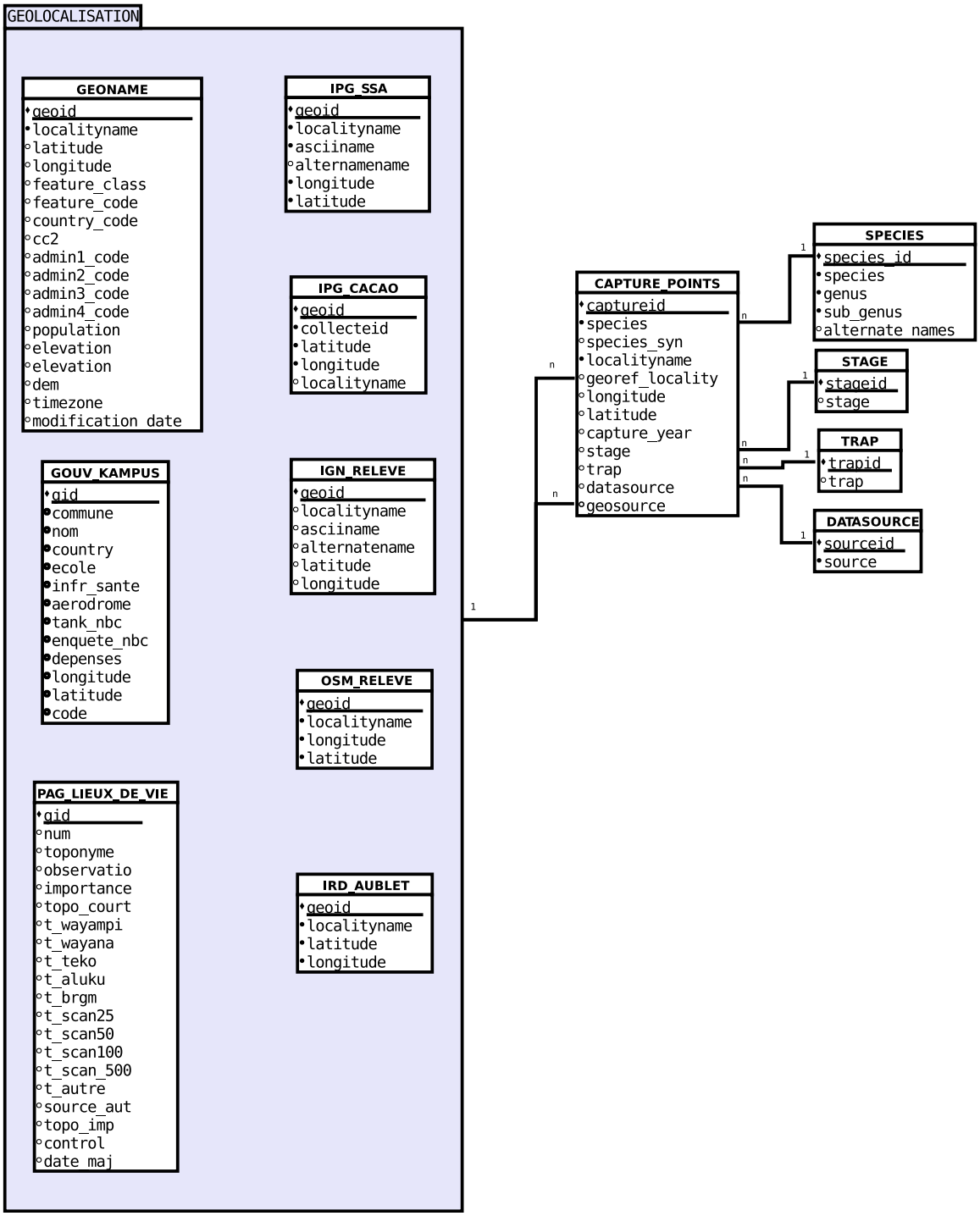


FIGURE 3.17 – **Schéma relationnel de la base de données des *Anopheles* en Guyane.**

Les descriptions utilisées pour la localisation des sites de captures étaient différentes d'un organisme à un autre et/ou d'une période à une autre. Les différents types de localisation sont répertoriés dans le tableau 3.1.

Pour spatialiser l'ensemble des données, il a fallu assigner à chaque site de capture n'ayant pas de localisation précise des coordonnées géographiques. Pour cela, les bases de données existantes suivantes ont été utilisées :

— la base de données des kampus de la Préfecture de la Guyane, qui répertorie tous

Types de localisation	Exemples
Surface en eau	Lac du Rorata, Étang des américains
Cours d'eau	Tour de l'Île, crique Anguille, Haut-Mana
Saut	Saut Maripa, Saut-Tigre
Plusieurs villes	Île de Cayenne
Ville	Cayenne, Tonate
Village	Village de Montjoly (1943), Trois-Sauts
Quartier	Baduel, Châton
Route, Piste, Chemin	Route nationale 1, piste de Paul Isnard
Rue	Rue Nationale
Bâtiment	Hopital Jean Martial, Gendarmerie de Cayenne, Boulangerie de Saül
Pont	Pont des Cascades , Pont vert et jaune de Montsinéry-Tonnegrande
Site d'orpaillage	Mine d'or de Saint-Elie, Placer Paul-Isnard
Ile	Ilet la Mère
Dégrad	Dégrad Kwata
Adresse postale	10 Lotissement Lony Remire-Montjoly
Coordonnées géographiques	Cacao, lon : 336279 / lat : 505360
Autre	Savane de Régina, Montagne de Maringouins

TABLEAU 3.1 – **Type de localisation des données de capture entre 1901 et 2013**

les *kampus* le long des fleuves Lawa et Maroni. Un *kampu* étant un village noir-marron d'ancienneté et de taille diverses, allant de l'exploitation d'un abattis (culture sur brûlis) par une famille, à des villages plus étendus (Léobal, 2013) ;

— la base de données *Lieux de vie* du Parc Amazonien de Guyane, qui répertorie les lieux d'habitation de la population guyanaise ;

— GeoNames [4], une base de données géographiques disponible gratuitement, distribuée sous une licence Creative Commons Attribution, et qui contient plus de 10 millions de noms géographiques dans le monde ;

— AUBLET2, une base de données de l'IRD, qui réunit les données de collecte des spécimens botaniques sur le plateau des Guyanes, surtout en Guyane française (Hoff et al., 2007). Cette base de données thématique a été utilisée lorsque le nom du site de capture n'existe pas dans d'autres bases (par exemple : Saut-Tigre, Saut-Japigny).

Lorsque le site de capture n'a pas de coordonnée géographique mais que sa description est précise et récente, à l'exemple d'une adresse postale ou d'un bâtiment unique dans un village (boulangerie de Saül), la localisation a été faite en relevant les coordonnées à partir :

— des cartes de l'Institut national de l'information géographique et forestière (IGN) (SCAN 025 et 050) ;

— du site Open Street map [5]. Ce site a pour but de créer une carte mondiale libre où les internautes contribuent à l'alimentation de la base de données en utilisant des données libres ou des données GPS.

Une des difficultés est lorsque l'orthographe des toponymes des sites lors de la capture est différente de celle actuelle (exemple : Caux et Kaw) ou lorsque le toponyme ou le village n'existe dans aucune des bases de données actuelles (par exemple Tonate-Sophie ou Marie-Anne). Ces derniers sites ne sont alors pas géolocalisés. Tous les sites de capture décrits par le nom

4. http://www.geonames.org
5. http://openstreetmap.fr/

d'un cours d'eau, par le nom d'une route ou d'une piste ne sont également pas géolocalisés. L'extension PostGIS à été ajoutée à la base de données PostgreSQL afin de prendre en compte la spatialisation des sites de capture. Une vue de la base de données est représentée en figure 3.18.

captureid	species	species_syn	locality	georef_locality	longitude	latitude	capture_year	stage	trap	datasource	geosource
2127	darlingi	darlingi	Pointe Combi	Pointe Combi	283897.0570...	588011.4987...	2007	Adult	Human	Service Départemental de Désinfection	GEONAME
2128	darlingi	darlingi	Sablance	Sablance	344545.6936...	546278.4860...	2007	Adult	Human	Service Départemental de Désinfection	RELEVE_IGN
2129	darlingi	darlingi	La Chaumière	La Chaumière	349516.1675...	540310.6965...	2007	Adult	Human	Service Départemental de Désinfection	GEONAME
2130	darlingi	darlingi	Cacao	Cacao	336895.0863...	505521.8001...	2007	Adult	Human	Service Départemental de Désinfection	PAGLIEUXDE...
2131	aquasalis	aquasalis	Tonate	Bourg de Tonate - ...	336689.2332...	554845.7528...	2007	Adult	Human	Service Départemental de Désinfection	IRDAUBLET
2132	aquasalis	aquasalis	Sablance	Sablance	344545.6936...	546278.4860...	2007	Adult	Human	Service Départemental de Désinfection	RELEVE_IGN
2133	aquasalis	aquasalis	La Chaumière	La Chaumière	349516.1675...	540310.6965...	2007	Adult	Human	Service Départemental de Désinfection	GEONAME
2134	braziliensis	braziliensis	Pointe Combi	Pointe Combi	283897.0570...	588011.4987...	2007	Adult	Human	Service Départemental de Désinfection	GEONAME
2135	braziliensis	braziliensis	Sablance	Sablance	344545.6936...	546278.4860...	2007	Adult	Human	Service Départemental de Désinfection	RELEVE_IGN
2136	triannulatus	triannulatus	Cacao - routes des cha...	Cacao	336895.0863...	505521.8001...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	PAGLIEUXDE...
2137	triannulatus	triannulatus	Saint-Jean	Saint Jean	158347.1783...	598749.0471...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	PAGLIEUXDE...
2138	triannulatus	triannulatus	Route de Saint Jean	Saint Jean	158347.1783...	598749.0471...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	PAGLIEUXDE...
2139	darlingi	darlingi	Soula	Soula	344043.0103...	542972.3460...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	GEONAME
2140	braziliensis	braziliensis	Soula	Soula	344043.0103...	542972.3460...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	GEONAME
2141	aquasalis	aquasalis	Soula	Soula	344043.0103...	542972.3460...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	GEONAME
2142	darlingi	darlingi	La Chaumière	La Chaumière	349516.1675...	540310.6965...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	GEONAME
2143	darlingi	darlingi	Copaya	Copaya	353373.0111...	535581.7734...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	RELEVE_IGN
2144	darlingi	darlingi	Concorde	Concorde	350473.2038...	534717.3053...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	RELEVE_IGN
2145	darlingi	darlingi	Chemin Moges	Chemin Mogès	352032.3295...	529456.7176...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	RELEVE_IGN
2146	triannulatus	triannulatus	Chemin Moges	Chemin Mogès	352032.3295...	529456.7176...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	RELEVE_IGN
2147	darlingi	darlingi	Rémire	Remire	358489.6953...	540291.4574...	2007	Larvae	Larvae capture	Service Départemental de Désinfection	PAGLIEUXDE...
2148	darlingi	darlingi	Langatabiki	Langa Tabiki	118126.5783...	552923.8260...	2008	Adult	Human	Rapport Annuel - Institut Pasteur de la ...	GOUVKAMPUS
2149	oswaldoi	oswaldoi	Alikéné - village forestie...	Alikéné	322929.9782...	360135.5278...	2008	Adult	other trap	Institut Pasteur de la Guyane	IPGSSA
2150	oswaldoi	oswaldoi	Alikéné - village forestie...	Alikéné	322929.9782...	360135.5278...	2008	Adult	Human	Institut Pasteur de la Guyane	IPGSSA
2151	darlingi	darlingi	Alikéné - village forestie...	Alikéné	322929.9782...	360135.5278...	2008	Adult	Human	Institut Pasteur de la Guyane	IPGSSA

FIGURE 3.18 – **Représentation de la base de données de la présence des espèces d'*Anopheles* en Guyane française, de 1901 à 2013.**

La spatialisation des sites de capture, toutes sources confondues entre 1902 et 1999, est représentée en figure 3.19, Ces sites sont au nombre de 1444. La spatialisation des captures (toutes sources confondues) faites à partir de 2000 est représentée en figure 3.20. Le nombre de sites de capture à partir de 2000 est de 2097. Parmi les 2097 sites de captures recensés en Guyane après 2000, 665 sites correspondent à la présence d'*An. darlingi*. Parmi ces 665 sites, seuls 48 sites de capture d'*An. darlingi* ont des coordonnées géographiques précises. Ces derniers sont représentés sur la carte de la figure 3.21

VI. Modélisation de la distribution d'*Anopheles darlingi*

La modélisation de la distribution d'*An. darlingi* requiert des données environnementales reflétant les facteurs environnementaux favorisant ou, au contraire, pénalisant la présence de l'espèce.

VI. 1. Données environnementales

Le choix des données environnementales a été fait en se basant sur les connaissances issues de la littérature et sur l'expertise des entomologistes de l'IPG. Ceci a permis de distinguer trois grands types de milieu :

- les milieux naturels, pour lesquels la présence d'*An. darlingi* dépend de la valeur ou de la classe de la variable environnementale associée (*MIL_NAT*) ;
- les milieux associés aux activités anthropiques qui altèrent de manière non-permanente et de manière localisée l'environnement naturel, influençant positivement la présence d'*An. darlingi* (*ANTHR_NON_PERM*) ;
- les milieux urbains, correspondant à la présence et aux activités humaines qui altèrent de manière permanente l'environnement naturel sur de larges surfaces et qui agissent comme un frein à la présence d'*An. darlingi* (*URBAIN*).

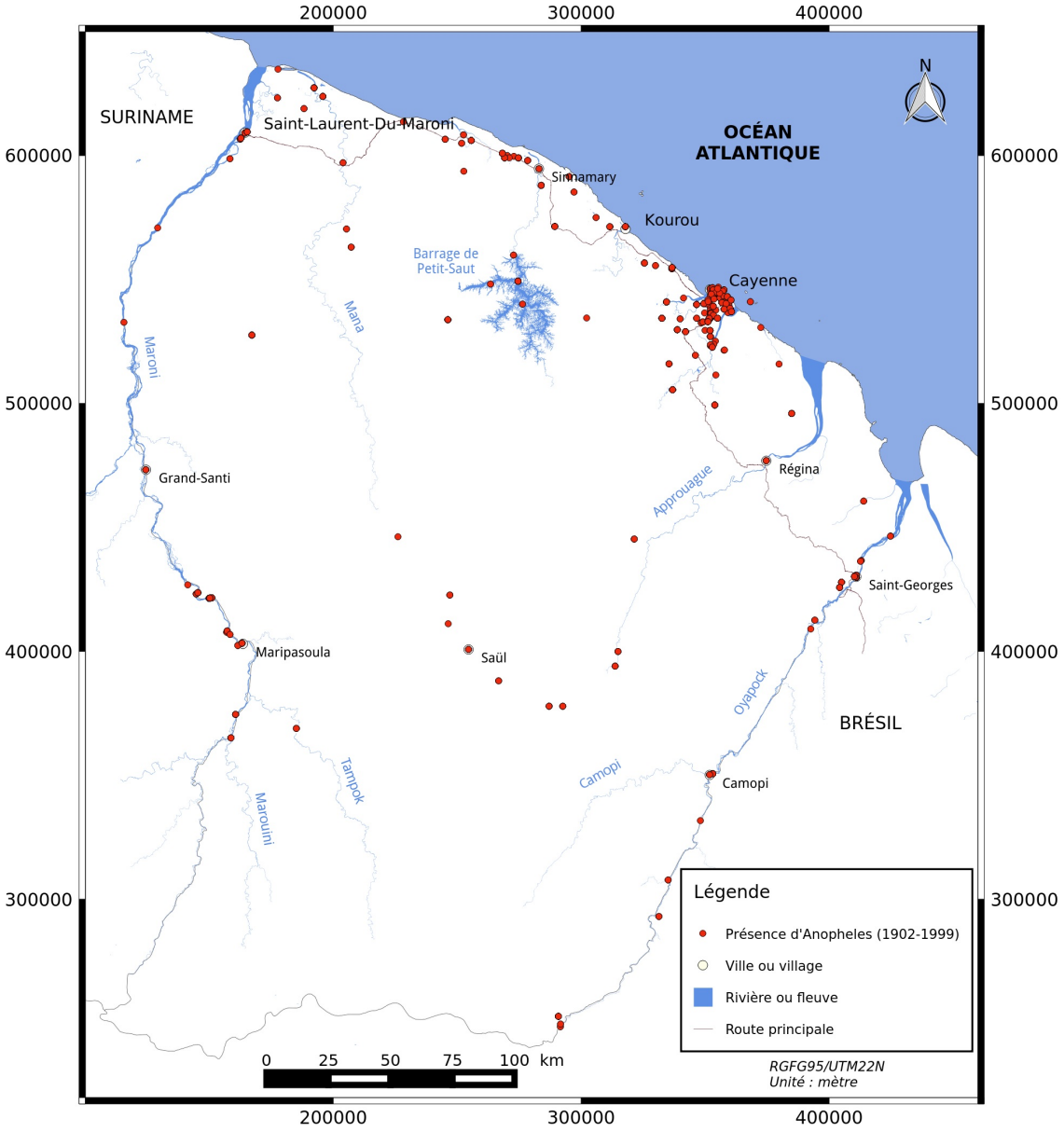


FIGURE 3.19 – **Sites de présence d'*Anopheles* entre 1902 et 1999.**

Pour chacun des types décrits précédemment, des données environnementales ayant un lien avéré ou supposé avec la présence et/ou l'absence de ce vecteur en zone amazonienne ont été choisies :

— l'altitude (*ALT*) issue des données de SRTM1 de la NASA. Le SRTM1, Shuttle Radar Topography Mission, de résolution d'une seconde d'arc (soit 31 m à l'équateur), est un modèle numérique de terrain qui a été établi à partir de données RADAR ;

— les paysages et unités géomorphologiques (*GLS : Geomorphological landscape* ; *GLF : Geomorphological landform*, Guitet et al. (2013)) , caractérisent la forme du relief comme résultat de l'effet du climat sur le substrat géologique. Ces cartes ont été obtenues à partir des données SRTM1. Smith et al. (2013) ont montré qu'il y a un lien entre la géomorphologie et la création de gîtes larvaires ;

— l'occupation du sol (*LS : Landscape type*, Gond et al. (2011)). Il s'agit d'une carte de

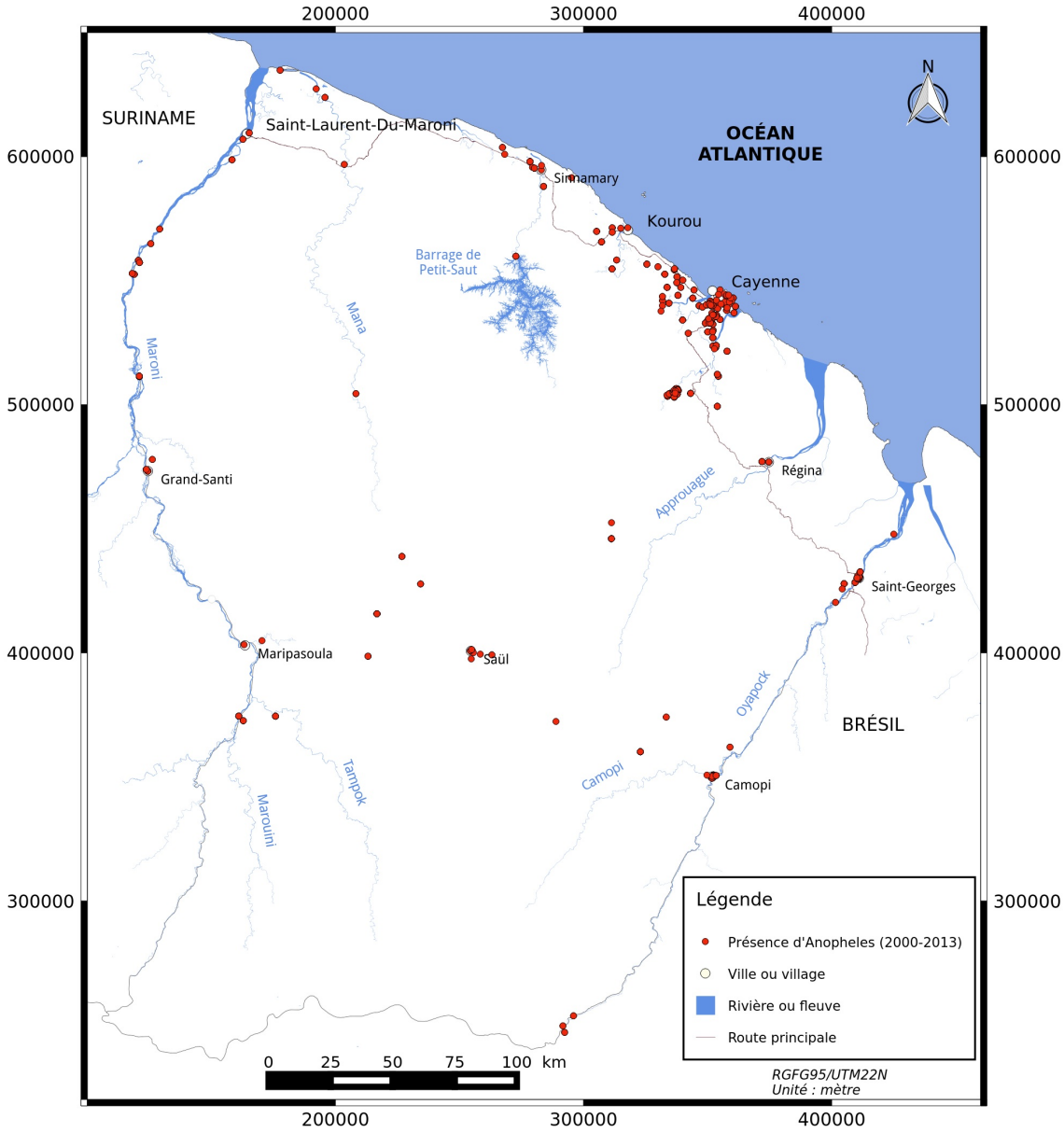


FIGURE 3.20 – **Sites de présence d'*Anopheles* entre 2000 et 2013.**

la distribution des différents types d'occupation et d'utilisation du sol en Guyane en particulier des écosystèmes forestiers. Elle a été produite à partir d'images issues du capteur VEGETATION du satellite SPOT4 ;

— l'empreinte humaine (*HFP : Human Footprint*, de Thoisy et al. (2010)). Cette carte correspond à la spatialisation d'un indice de l'activité humaine. Cet indice est une mesure générale de l'étendue de la menace de l'homme sur la biodiversité, en attribuant une note selon la nature de la perturbation. Cette variable combine plusieurs sous-couches : la densité de la population humaine ; les zones urbaines ; les sites d'activités minières (légaux et illégaux) ; les zones de chasse potentielles correspondant à une zone de deux kilomètres de part et d'autres des routes, des pistes ou des cours d'eau pouvant être empruntés par l'homme ; les exploitations forestières et les camps forestiers ou touristiques. La somme de ces notes donne l'indice de l'activité humaine ;

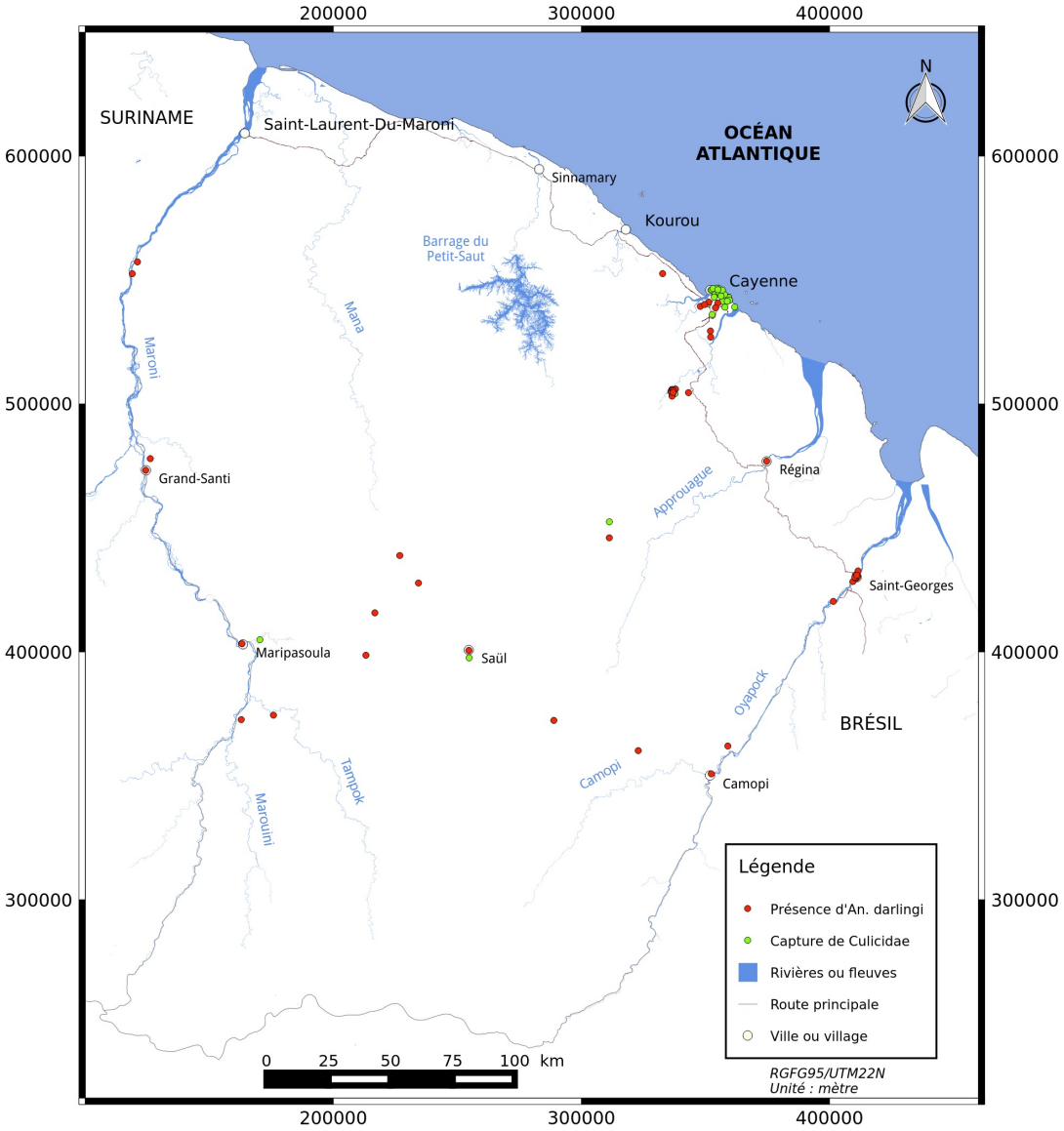


FIGURE 3.21 – **Répartition des captures d'*Anopheles darlingi* (en rouge) et des captures de Culicidés (sites sur lesquels *An. darlingi* n'a pas été capturé, en vert) ayant des localisations précises, de 2000 à 2013.**

— les routes et les pistes (*ROADS*) issue de la BD TOPO®. Il s'agit du réseau routier présent en Guyane (asphaltées ou non).

Les données environnementales exploitées dans cette thèse sont décrites dans le tableau 3.2.

VI. 2. Caractérisation environnementale

Afin de mieux caractériser l'environnement et de refléter les hypothèses écologiques et les connaissances liées à *An. darlingi*, des variables ont été extraites des données environnementales décrites précédemment. La résolution spatiale la plus basse, est égale à 1 x 1 km, correspond à celle de la carte de l'occupation du sol. Cette résolution a été choisie comme résolution de référence, afin d'homogénéiser les résolution spatiales des caractérisations environnementales. Les sous-couches constituant l'empreinte humaine (*HFP*), ayant un élément

Données environnementales	Producteurs	Sources des données	Date	Résolution spatiale ou échelle d'interprétation	Type de variable
Paysages géomorphologiques (*GLS*)	ONF [a] , Guitet et al. (2013)	SRTM30	2000	$\geq$ 5000m	Catégorielle
Unités géomorphologiques (*GLF*)	ONF , Guitet et al. (2013)	SRTM30	2000	$\geq$ 200m	Catégorielle
Occupation du sol (*LS*)	CIRAD [b], Gond et al. (2011)	données du capteur VEGETATION du satellite SPOT-4	2000	1 000 m	Catégorielle
Empreinte humaine (*HFP*)	KWATA, de Thoisy et al. (2010)	INSEE, DAAF [c] , DDE [d], Hammond et al. (2007)	2005	$\geq$ 1000m	Continue
Routes et pistes	IGN [e]	BD TOPO®	2011	$\geq$ 1000m	Continue
Altitude (*ALT*)	NASA [f]	SRTM30	2000	30 m	Continue

TABLEAU 3.2 – **Données environnementales brutes**

a. Office National des Forêts
b. Centre de Coopération Internationale en Recherche Agronomique pour le Développement
c. Direction de l'Alimentation, l'Agriculture et de la Forêt
d. Direction Départementale de l'Équipement
e. Institut National de l'Information Géographique et Forestière, anciennement Institut Géographique National
f. National Aeronautics and Space Administration

de dimension minimale de 10 x 40 m, ont été rasterisées à une résolution de 30 mètres. Les variables suivantes en ont ensuite été extraites :

— Le pourcentage d'urbanisation (*PER_URB*). Cette variable est obtenue en calculant le pourcentage d'urbanisation dans les pixels de 1 km^2 ;

— Le pourcentage d'urbanisation dans les pixels voisins (*PER_URB_NEIGH*). Cette variable permet de distinguer les zones urbanisées de grande taille des petites zones urbanisées. Elle est obtenue en se basant sur la variable *PER_URB*. Un pixel p est considéré comme urbain si sa valeur de *PER_URB* est supérieure ou égale à 50%. Son voisinage correspond aux huit pixels contigus. Pour chaque pixel urbain p, la moyenne du pourcentage d'urbanisation de ses voisins est calculée ;

— La variable caractérisant les activités qui altèrent de manière non-permanente et localement l'environnement naturel (*HA*) a été obtenue en sommant les scores des sous-couches suivantes : les camps touristiques et forestiers, les activités minières et forestières et les zones de chasse potentiel le long des cours d'eau. Pour homogénéiser cette variable à la résolution de référence, le maximum, le minimum et la médiane ont été calculés pour chaque pixel de 1 x 1km.

Afin de refléter la présence humaine et en particulier son potentiel de déplacement en dehors des villes, la longueur des routes dans un pixel de 1 x 1 km a été calculée. Seules les routes et pistes en dehors des zones urbaines ont été prises en compte. En effet, dans les zones urbaines, les routes ont un effet à long terme qui est contenu dans la couche *PER_URB*. La valeur des pixels représente le potentiel de transit humain et la modification du milieu naturel pouvant créer des gîtes larvaires. Il est important de noter que dans de nombreuses localités de Guyane, les routes correspondent à des pistes forestières entourées de forêt.

Quelques corrections ont été apportées à la carte d'occupation du sol (*LS*). Cette carte ne mentionne pas certaines zones urbaines (par exemple le centre ville de Cayenne) et ne fait pas la distinction entre les forêts inondées associées à l'eau douce et la mangrove exclusivement associée à l'eau saumâtre. La correction de l'urbanisation a été réalisée en utilisant la couche *PER_URB*. Les pixels de *LS* ayant une valeur de *PER_URB* supérieure ou égale à 50% sont reclassifiés comme pixels urbains.
La correction des pixels de mangrove a été réalisée à partir de la couche d'occupation du sol issue du projet "Expertise Littoral 2011" (ONF, 2013). Il s'agit d'une carte faisant l'état des lieux de l'occupation du sol de la bande littoralle guyanaise par l'analyse d'images satellite et de photographies aériennes. Elle a permis une meilleure connaissance du développement anthropique sur le littoral où vit 90% de la population. Les pixels de *LS* classifiés comme *forêt inondée* et correspondant à de la mangrove selon l'Expertise Littoral sont ainsi reclassifiés en *Mangrove*.

Les données relatives aux paysages géomorphologiques (*GLS*) et les unités géomorphologiques (*GLF*) correspondent à des variables catégorielles. Elles ont été rasterisées à 30 mètres puis agrégées pour obtenir des pixels de 1km auxquels est attribuée la classe majori-

taire.
Pour obtenir des raster de 1 $\times$ 1 km, le maximum, le minimum et la médiane des données d'altitude (*ALT*) et de *HA* ont été calculées.
La longueur des routes et/ou des pistes a été calculée dans un pixel de 1km^2.
Pour éviter une redondance d'information avec la classe *urbain* de *LS*, la couche *PER_URB* a été retiré du jeu de variables utilisé pour construire le modèle.
L'ensemble des variables environnementales qui ont été extraites ou corrigées pour permettre de construire le modèle est récapitulé dans le tableau 3.3 (leur représentation est situé en annexe D).

VI. 3. Données météorologiques

La température moyenne annuelle en Guyane est de 26,5°C, elle est homogène tout au long de l'année et sur l'ensemble du territoire. Elle varie annuellement entre 24°C et 29°C. La variation thermique journalière est plus importante, elle peut varier de 16°C à 34°C selon la saison (Barret, 2001). Par exemple, pour Maripasoula, situé sur le fleuve à la frontière surinamienne et pour Camopi, situé sur le fleuve à la frontière brésilienne, les amplitudes thermiques annuelles minimales et maximales sont égales à 4,3°C et 9,6°C (moyennées sur la période de 2001 à 2008), tandis que l'amplitude thermique journalière moyenne est de 9,8°C (moyennée sur la même période). Olson et al. (2009) soulignent que dans le bassin amazonnien, les températures sont situées entre 24,6°C et 29,4°C pour 95% des observations et, par conséquent, ils n'ont pas inclus les températures dans leur modèle, constatant qu'elles sont dans la très grande majorité des cas comprises dans l'intervalle de valeurs optimales pour le développement du vecteur, et supposant qu'elles ne permettent pas d'expliquer les différences de niveau de transmission. De la même manière dans notre modèle, les données de température n'ont pas été incluses dans la construction du modèle, en faisant l'hypothèse qu'elles ne sont pas suffisamment discriminantes pour expliquer la présence d'*An. darligni*.

En Guyane française, les travaux de Girod et al. (2008) et Vezenegho et al. (2015) ont montré un lien entre la densité d'*An. darlingi* et la variation de la pluviométrie. Cette variation de densité est fortement influencée par les variations intra-annuelles de la pluviométrie. Celles-ci diffèrent selon les zones avec des amplitudes allant de 2 000 mm dans la zone la plus sèche à 4 000 mm dans la zone la plus pluvieuse. Cependant, ce lien n'a pas été systématiquement montré sur l'ensemble de la Guyane. En effet, Girod et al. (2011) ont montré l'existence de corrélation entre la densité d'*An. darlingi* et la pluviométrie pour Camopi, mais cette corrélation n'existe pas pour Apatou et Régina. Bien que la raison de cette absence de relation soit encore floue, les auteurs supposent que ceci est dû à la différence des types forestiers d'Apatou et de Camopi et à la présence de zones humides permanentes (marécages et vallées humides) maintenant la présence de gîtes larvaires tout au long de l'année à Régina. De plus, la Guyane française est caractérisée par un réseau hydrographique dense (Barret, 2001). Compte tenu de cette densité, il est convenable de faire l'hypothèse que, malgré les différences de pluviométrie annuelle observées d'une zone à l'autre du territoire, *An. darlingi* peut trouver un site favorable à la ponte sur l'ensemble du territoire.
L'ensemble des sites de présence d'*An. darlingi* (ayant une localisation précise ou non) notifié entre 1980 et 2013 a été cartographié et présenté en figure 3.22. Cette carte montre que les sites de présence se trouvent aussi bien dans la zone la plus sèche (Awala-Yalimapo, Saint-

Numéro de variable	Intitulé de la variable	Données environnementales initiales	Type d'information pour chaque pixel de 1 x 1 km	Impact sur *An. darlingi* et références bibliographiques sur cette connaissance *a priori*	Type de milieu	Type de variables
1	Paysages géomorphologiques ***GLS***	*GLS*	Classe majoritaire	(/) Smith et al. (2013)	*MIL_NAT*	Catégorielle
2	Unités géomorphologiques ***GLF***	*GLF*	Classe majoritaire	(/) Smith et al. (2013)	*MIL_NAT*	Catégorielle
3	Occupation du sol ***LS***	LS	Correction des pixels *Urbain* et de *Mangrove*	(/) Stefani et al. (2013); Charlwood (1996); Girod et al. (2011); Rozendaal (1992); Hiwat et al. (2010); Vittor et al. (2006, 2009)	*MIL_NAT*	Catégorielle
4 5 6	Présence et activités humaines qui influent de manière non-permanente et localement l'environnement naturel ***HA_max*** ***HA_med*** ***HA_min***	*HFP*	Maximum, Médiane, Minimum	(+) Vittor et al. (2009)	*ANTHR_NON_PERM*	Continue
7	Pourcentage d'urbanisation dans les pixels voisins ***PER_URB_NEIGH***	*HFP (PER_URB)*	Calcul du pourcentage d'urbanisation dans les huit plus proches voisins	(-) Stefani et al. (2013)	*URBAiN*	Continue
8	Longueur des routes des pistes ***ROADS***	routes et pistes de la BD TOPO®	Calcul de la longueur des routes et des pistes	(+) Singer and Castro (2001)	*ANTHR_NON_PERM*	Continue
9 10 11	Altitude ***ALT_max*** ***ALT_med*** ***ALT_min***	SRTM	Maximum, Médiane, Minimum	(-) Zeilhofer et al. (2007)	*MIL_NAT*	Continue

TABLEAU 3.3 – **Données environnementales utilisées pour la modélisation.**
(/) signifie que l'influence de la variable sur la présence d'*An. darlingi* dépend de la valeur/modalité de la variable
(+) signifie que la variable favorise la présence d'*An. darlingi*
(-) signifie que la variable limite la présence d'*An. darlingi*

Laurent-du-Maroni, Iracoubo) que dans la zone la plus humide (Régina, Cacao et Kaw). Par conséquent, nous faisons l'hypothèse qu'à l'échelle de la Guyane, et à une échelle temporelle pluri-annuelle, la pluviométrie n'influe pas de manière significative sur la présence d'*An. darlingi* (les variations intra-annuelles de la quantité de pluie va influer la densité des vecteurs). Les données de pluviométrie n'ont pas été utilisées dans ce travail.

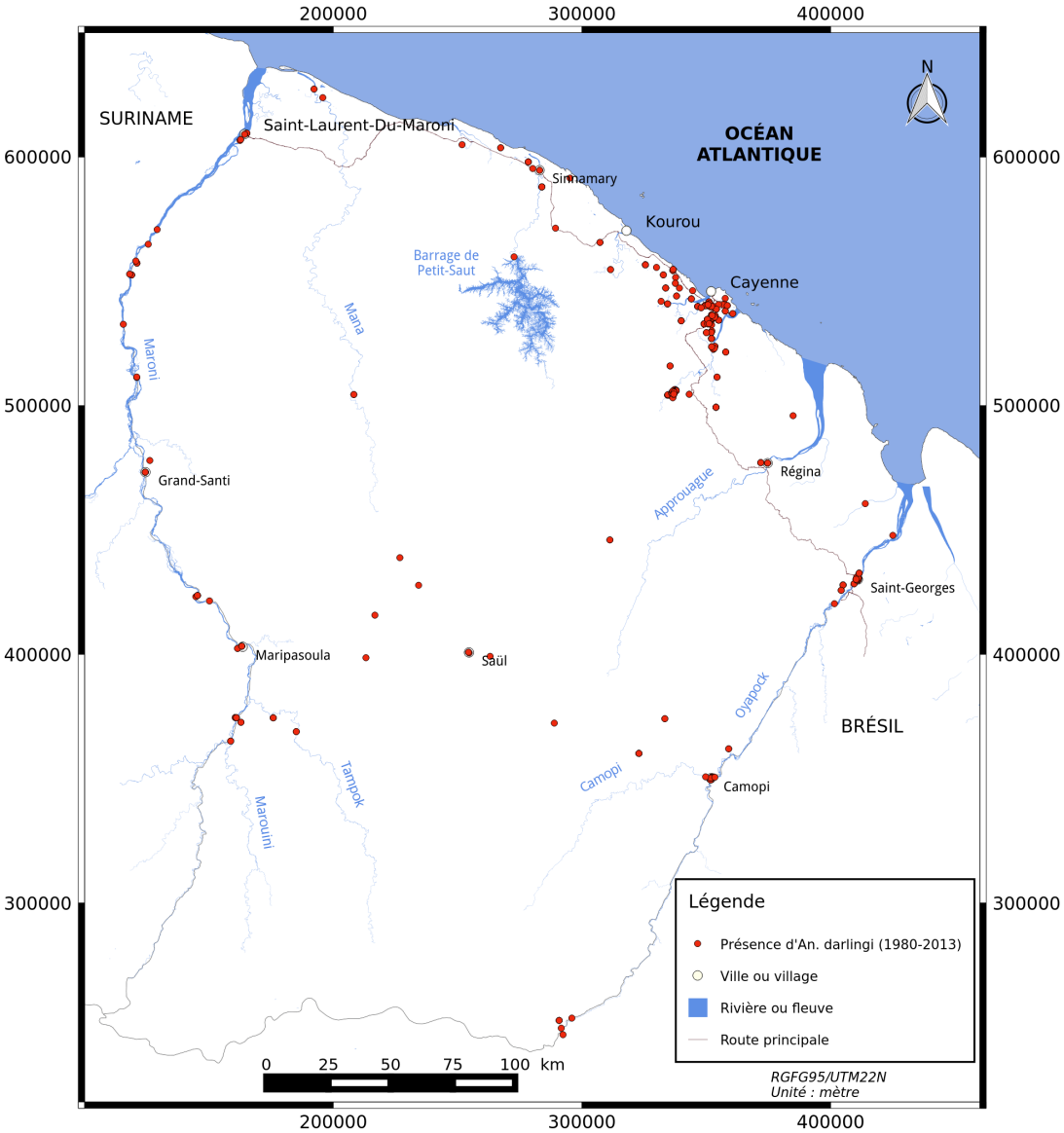


FIGURE 3.22 – **Répartition de l'ensemble des captures d'*Anopheles darlingi* entre 1980 et 2013.**

VI. 4. Construction des modèles

La construction du modèle nécessite des sites de présence et des données environnementales. Dans ce travail, seuls les sites de captures datant de 2000 - 2013 ont été retenus compte-tenu du fait que les données environnementales datent des années 2000. Parmi ces sites, seuls ceux qui ont une géolocalisation précise (coordonnées géographiques, adresse postale) sont conservés pour la construction du modèle. Au total, 48 sites de captures d'*Anopheles*

darlingi sont conservés pour la construction du modèle (les coordonnées géographiques se trouvent en Annexe C.1). Ils sont représentés sur la figure 3.21. En Guyane française, l'accès aux différentes localités se fait par voies fluviales, routières ou aériennes. Ce réseau ne recouvrant pas l'intégralité du territoire, il rend difficile la réalisation de captures de *Anopheles* sur l'ensemble de la Guyane. Ceci implique un échantillonnage d'*Anopheles* biaisé se traduisant par un effort d'échantillonnage important sur le littoral où les routes sont plus nombreuses et un effort d'échantillonnage faible voir quasi-inexistant à l'intérieur des terres. Pour corriger ce biais d'échantillonnage, la méthode de correction présentée dans le chapitre 2 a été appliquée avant la construction du modèle.

VI. 4. a. Correction du biais d'échantillonnage

Phillips et al. (2009) proposent de corriger le biais d'échantillonnage en sélectionnant les points de background avec le même biais d'échantillonnage environnemental que celui ayant entaché l'obtention des points de présence. La méthode proposée au chapitre 2 a pour but de corriger le biais d'échantillonnage en réalisant cette opération.
Pour estimer le biais d'échantillonnage d'*An. darlingi*, un groupe cible constitué de l'ensemble des espèces de moustique (*Culicidæ*) a été utilisé. Les sites de présence de *Culicidæ* proviennent de l'IPG. Seuls les sites précisément géolocalisés et pour lesquels l'ensemble des espèces (quelque soit le genre) a été identifié, ont été considérées. Ces sites sont au nombre de 74. Les sites de capture associés au groupe cible sont appelés par la suite *sites de capture* tandis que les sites de présence d'*An. darlingi* sont appelés *sites de présence*.
La définition de la fonction d'appartenance w nécessite d'assigner une valeur à D_{min}. Le choix de la valeur de $Dmin$ ne dépend pas de critères géographiques mais reflète les connaissances *a priori* sur la bio-écologie du vecteur. Dans cette thèse, sa définition se base sur la connaissance qu'*An. darlingi* n'est pas présent dans les zones densément urbanisées, ce qui permet d'émettre l'hypothèse qu'un site de présence d'*An. darlingi* ne peut être situé dans le voisinage environnemental d'une zone urbaine dense. Par conséquent, la valeur de $Dmin$ est fixée à la valeur de la distance euclidienne minimale observée entre les sites de présence et les pixels de zone urbaine dense (pixels où *PER_URB* est supérieure ou égale à 50%) dans l'espace environnemental. Avec une telle valeur de $Dmin$, d'après l'équation 2.1 du chapitre 2, la valeur d'appartenance de tout pixel de présence d'*An. darlingi* au voisinage de tout pixel de zone urbaine dense est inférieure à 0,5. L'attribution d'une valeur à $Dmin$ permet la définition de la fonction d'appartenance et par conséquent le calcul de l'effort d'échantillonnage relatif, représentant le biais d'échantillonnage, pour chacun des pixels de la zone d 'étude. Plus un pixel est associé à un effort d'échantillonnage élevé, plus il aura de chance d'être sélectionné comme pixel de background.

VI. 4. b. Paramétrage du modèle

Lorsqu'un pixel comprend plusieurs sites de présence, un seul site est sélectionné pour la construction du modèle. Ainsi, parmi les 48 sites de présence, seuls 39 sont conservés après élimination des sites redondants. Ces 39 sites sont utilisés avec les onze variables environnementales présentées dans le tableau 3.3, pour la construction du modèle. Les fonctions caractéristiques "linéaire par morceau" et "catégorielle" ont été choisies. Selon Elith et al. (2011) et Phillips and Dudík (2008), la transformation "linéaire par morceau" est un bon compromis

entre la simplicité et la qualité d'approximation des courbes de réponse des espèces.
Les valeurs des paramètres de régularisation recommandées par Phillips and Dudík (2008) ont été utilisées pour construire le modèle et du nombre de sites de background. En effet, les études effectuées par Phillips and Dudík (2008) sont basées sur des données de présence dont les caractéristiques sont proches de celles relatives à *An. darlingi* dans la présente étude (57 sites de présence pour la détermination du nombre de sites de background ; entre 30 et 60 sites de présence et entre 11 et 13 variables environnementales pour la détermination des valeurs optimales des coefficients de régularisation). Ainsi, les coefficients de régularisation sont fixés à 0,25 et 0,5, pour les transformations "linéaire par morceau" et "catégorielle", respectivement. Dix mille sites de background ont été sélectionnés de manière aléatoire et pondéré par l'effort d'échantillonnage relatif défini dans le paragraphe précédent.
Enfin, afin de ne pas prédire la qualité d'habitat d'*An. darlingi* dans les zones ayant des conditions environnementales trop différentes de celles associées aux données d'apprentissage, l'extrapolation n'est pas permise.

VI. 4. c. Évaluation et validation du modèle

L'évaluation de la performance de prédiction du modèle a été réalisée avec une validation croisée à dix ensembles de test (*10-fold cross-validation*). Les courbes ROC, les aires sous la courbe ROC (AUC) et l'indice continue de Boyce ont été calculés. Les rapports des AUC partiels ont également été calculés pour des taux d'erreur d'ommission égalent à 20 %, 20 % et 5%. Le gain (*regularized training gain*) a également été calculé. La contribution de chacune des variables environnementales à la construction du modèle a été estimée avec deux méthodes : la méthode heuristique proposée dans l'algorithme de Maxent et le test de Jackknife.

VII. Résultats

VII. 1. Performances de prédiction et contribution des variables environnementales et évaluation

L'AUC moyen obtenu est égale à 0,93. Les rapports d'AUC partiels moyens sont égaux à 1,08, 1,03 et 1,01 pour des taux d'erreur de commission maximaux de 20%, 10% et 5%, respectivement. L'indice continue de Boyce moyen est égal à 0,356. Le gain moyen atteint 3,14. Les résultats de contribution de chacune des variables sont répertoriés dans le tableau 3.4.

L'ensemble des variables environnementales peut être divisé en sous-ensemble (ou groupe) en fonction de contributions considérées comme élevées, modérées et faibles.
Le premier groupe est composé des trois variables les plus contributives et dont la somme des contributions dépasse à 80% : *ROADS*, *PER_URB_NEIGH* et *LS*. Le deuxième groupe réunit *HA_max*, *GLS*, *ALT_min* et *GLF*, dont les contributions sont considérées comme modérées. Enfin le troisième groupe rassemble *HA_min*, *HA_med*, *ALT_med* et *ALT_max* dont les contributions au modèle sont considérées comme négligeables. Les résultats du test de Jackknife montrent également une contribution non-significative des variables du dernier groupe.

Au regard des résultats précédents, un deuxième modèle simplifié a été construit avec uniquement les sept variables du groupe un et deux, c'est-à-dire avec les variables : *ROADS*,

	Méthode heuristique		Test de Jackknife	
Variables environnementales	**Contribution (%)**	**Contribution cumulée (%)**	**Gain avec une seule variable**	**Diminution du gain sans la variable (%)**
ROADS	51,45	51,45	2,20	-7,98
PER_URB_NEIGH	17,17	68,62	1,86	-0,41
LS	15,32	83,94	2,23	-4,67
HA	7,43 min : 0,35 med : 0,24 max : 6,84	91,37	 min : 0,02 med : 0,15 max : 0,43	 min : -0,06 med : -0,22 max : -2,10
GLS	5,35	96,72	1,40	-2,59
ALT	2,09 min :1,34 med : 0,69 max : 0,06	98,81	 min : 1,12 med : 1,04 max : 0,766	 min : -1,04 med : -0,39 max : -0,03
GLF	1,19	100	0,80	-0,21

TABLEAU 3.4 – **Contributions moyennes et résultats de Jackknife du modèle construit avec onze variables environnementales**

PER_URB_NEIGH, *LS*, *HA_max*, *GLS*, *ALT_min* et *GLF*.
L'AUC moyen de ce second modèle est de 0,93. Les rapports d'AUC partiels moyens sont de 1,11, 1,05 et 1,03 pour des taux d'erreur de commission maximaux de 20%, 10% et 5%, respectivement. Le gain moyen est de 3,19 et l'indice continue de Boyce moyen est de 0,421. Les résultats de contribution des variables sont présentés dans le tableau 3.5.

	Méthode heuristique		Test de Jackknife	
Variables environnementales	**Contribution (%)**	**Contribution cumulée (%)**	**Gain avec la variable seule**	**Diminution du gain sans la variable (%)**
ROADS	62,61	62,61	2,31	-8,61
LS	14,10	76,71	2,35	-6,23
PER_URB_NEIGH	11,15	87,86	2,05	-0,58
HA_max	5,39	93,25	0,37	-1,74
GLS	3,84	97,09	1,44	-1,90
GLF	2,1	99,19	1,01	-0,32
ALT_min	0,88	100	1,27	-1,29

TABLEAU 3.5 – **Contributions moyennes et résultats de Jackknife du modèle simplifié construit avec sept variables environnementales**

Les courbes de réponses de chacune des variables sont représentées en figures 3.23 et 3.24. La courbe 3.24(c) montre que l'indice de qualité d'habitat (IQH) est maximal lorsque le pourcentage d'urbanisation des pixels voisins (*PER_URB_NEIGH*) a une valeur inférieure à 8%. Au delà de cette valeur, l'IQH diminue progressivement vers une valeur nulle. D'après la courbe 3.24(d), l'IQH augmente avec longueur de routes et de pistes (jusqu'a 7 000 mètres) puis se stabilise avant de décroître au delà de 10 000 mètres de longueur de routes. D'après la courbe 3.24(a), l'IQH est au maximal lorsque l'altitude *ALT* est proche de 0. Il diminue rapidement lorsque l'altitude augmente. La courbe de réponse de la variable relative à l'activité humaine maximale (*HA_max*) (courbe 3.24(b)) présente un profil plus complexe. L'indice de qualite d'habitat augmente pour atteindre son maximum lorsque la valeur de *HA_max* se situe entre 0 et 8. Au delà de 8, l'IQH diminue jusqu'à ce que *HA_max* soit de 24, puis augmente de nouveau. La courbe de réponse 3.23(a) montre que parmi les différentes classes de la couche *LS*, les classes *Woodland savanna/dry forest* et *Open forest* contribuent le plus à une valeur élevé de IQH. Les figures 3.23(b) et (c) montre que les paysages géomorphologiques *Coastal*

flat plain et *Plain with residual relief* et les unités géomorphologiques *Small-size and flat wet land*, *Small-size rounded hill* et *Lowered half-orange* – appelé *demi-orange* en français, correspondant à des collines ayant une forme grossièrement hémisphérique à flancs convexes, et généralement lié à des bas-fonds plats et marécageux drainés par des cours d'eau riches en méandres (George and Verger, 1972) – contribuent également à un IQH élevé.

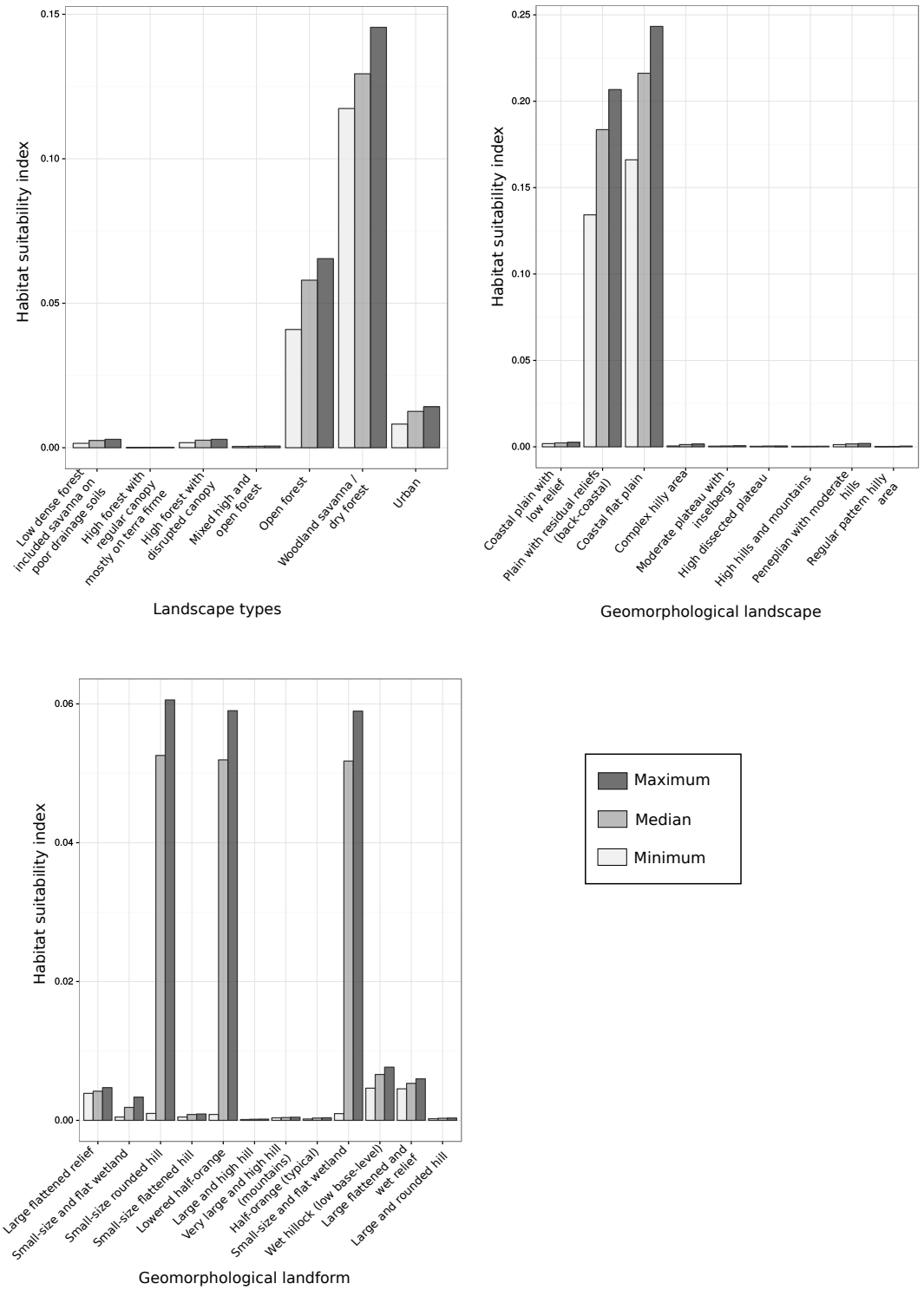


FIGURE 3.23 – **Courbes de réponses des variables catégorielles.**

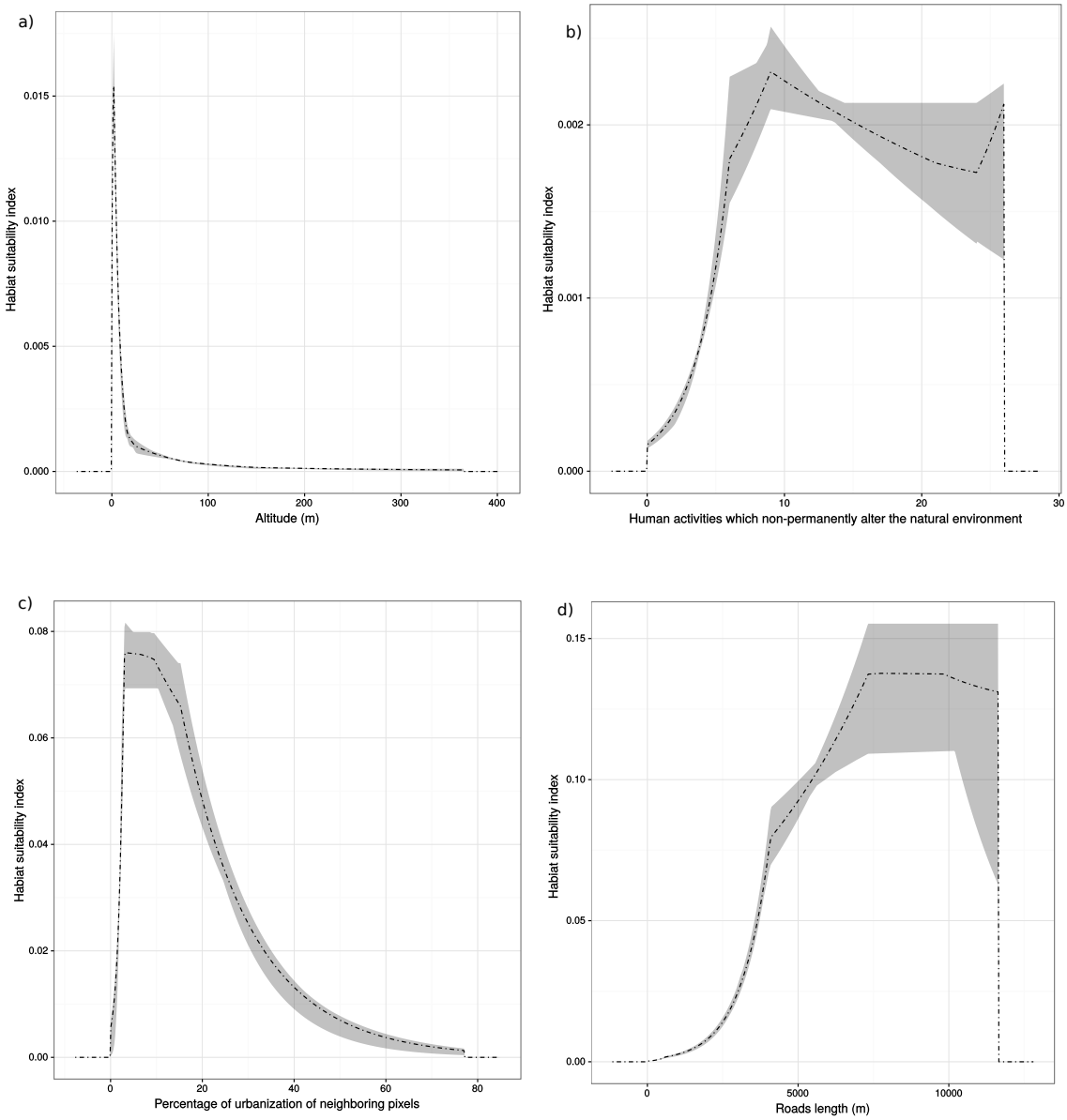


FIGURE 3.24 – **Courbes de réponses des variables numériques.**

VII. 2. Carte de la qualité d'habitat

La carte de qualité d'habitat obtenue en sortie du modèle est présentée en figure 3.25. Sept zones sont mises en évidence, de par leur IQH élevé (zones A, B, C, D, E et F) ou leur intérêt d'un point de vue épidémiologique (Camopi, zone G).

Une analyse qualitative de cette carte a été réalisée afin de déterminer les caractéristiques environnementales des zones à forts IQH (tableau 3.6). La zone A correspond à la bande littorale où vit la majorité de la population guyanaise et où se trouvent les zones urbaines denses, Cayenne, Kourou et Saint-Laurent-du-Maroni. Cette zone est caractérisée par un relief bas et par la présence de zones humides. Les principales routes du département sont situées dans cette bande littorale, et les classes d'occupation du sol, de par et d'autres des routes, sont majoritairement de la savane dans l'Ouest (de Cayenne à Saint-Laurent-du-Maroni) et de la forêt dense dans l'Est (de Cayenne à Saint-Georges).

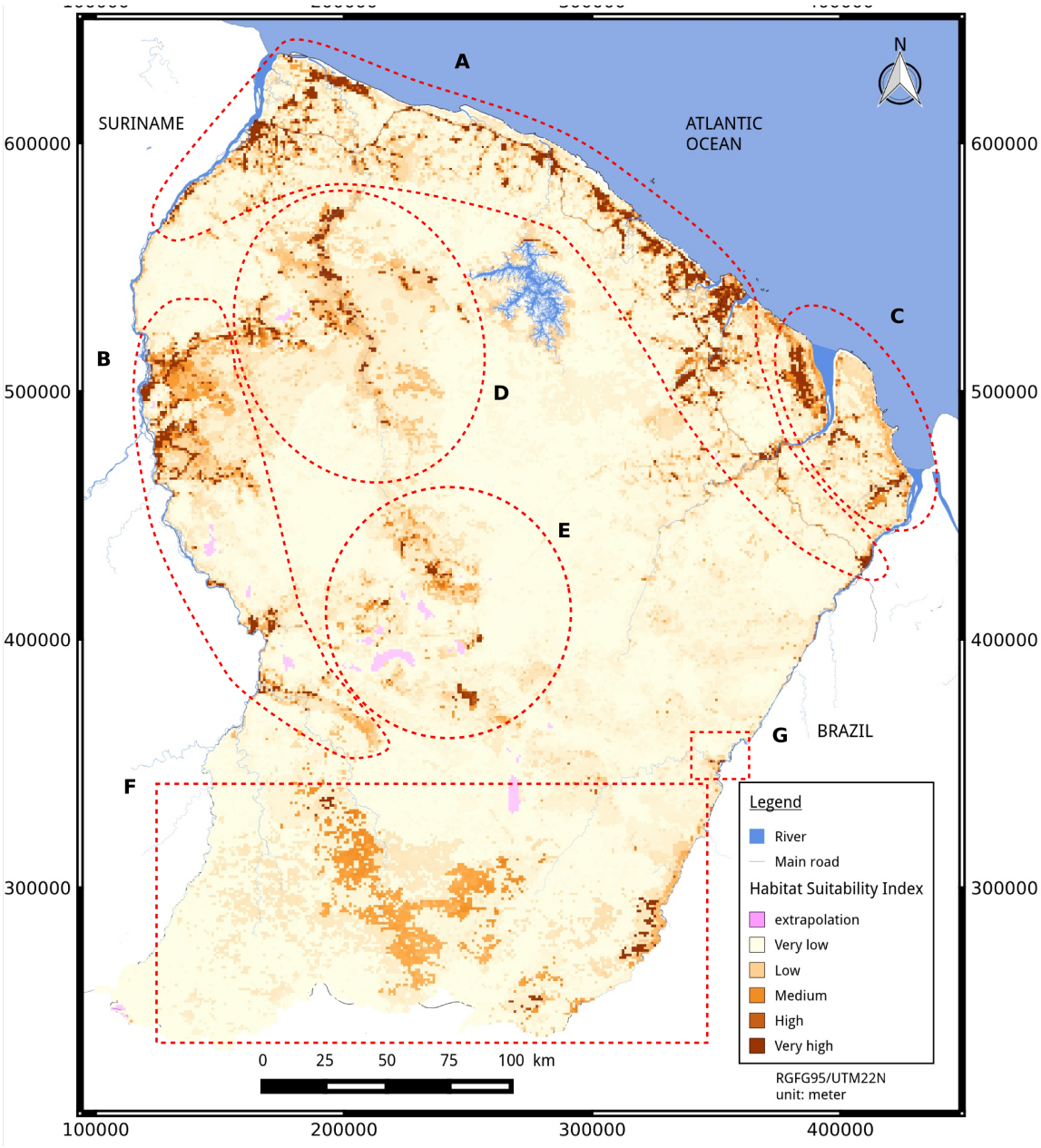


FIGURE 3.25 – **Carte de la qualité d'habitat d'*Anopheles darlingi* en Guyane française.**
Sept zones caractéristiques du territoire guyanais et présentant des indices de qualité d'habitat élevés (A à F) ou remarquable (G) sont entourées en rouge.

La zone B est une zone où vit la majorité du dixième de population guyanaise restante. Le principal axe de communication est le fleuve. Cette zone est caractérisée par la présence d'activités humaines et par des ouvertures de la forêt dense.
La zone C est une zone quasiment inhabitée. Elle est caractérisée par un relief plat et la présence de marais, en particulier le marais de Kaw.
La zone D est une zone avec une forte activité humaine, dont l'orpaillage et l'agriculture. Elle se caractérise aussi par la présence d'un nombre important de pistes forestières.
La zone E comprend le village Saül. En dehors de ce village, l'activité humaine est essentiellement l'activité minière. L'accès à cette zone se fait essentiellement par voie aérienne.
La zone F comprend la majeure partie du Parc Amazonien de Guyane. Seuls quelques vil-

lages se trouvent le long des fleuves. L'intérieur de cette zone n'est, à notre connaissance, pas ou très peu habité.
La zone G est le village de Camopi. Ce village se situe à la conférence du fleuve Oyapock et de la rivière Camopi. Il est caractérisé par une présence humaine et un faible réseau routier.

Pour vérifier que les résultats sont en concordance avec la connaissance *a priori* qu'*An. darlingi* n'est pas présent en zone densément urbaine, des zooms sur les principales zones urbaines sont présentés en figure 3.26. À Cayenne et à Kourou, dans la zone urbaine (en

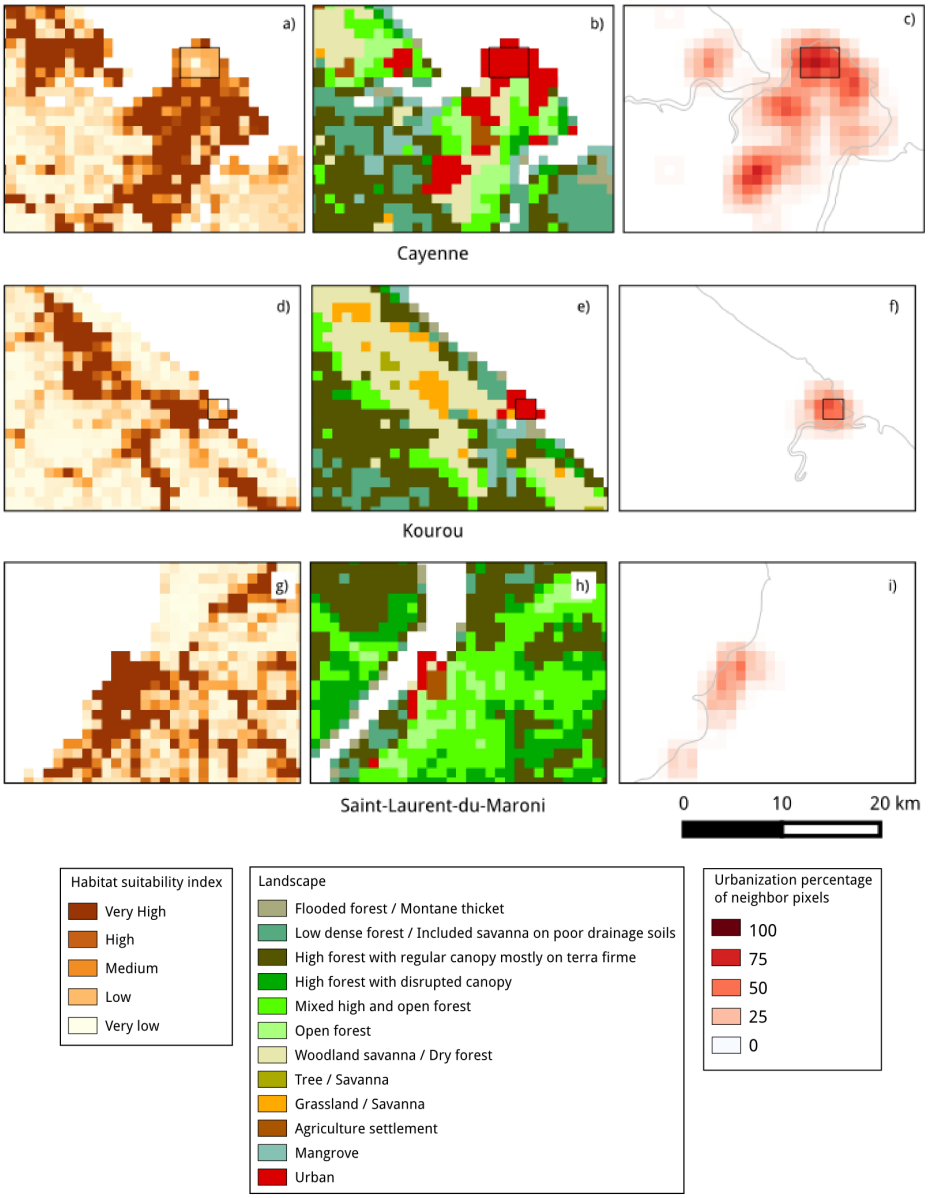


FIGURE 3.26 – **Zoom sur les grandes zones urbaines.**
a), d) et g) : cartes de qualité d'habitat. b), e) et h) : cartes d'occupation du sol. c), f) et i) : pourcentage d'urbanisation dans les pixels voisins. Les rectangles représentent les zones densément urbanisées d'après les critères de la présente étude (la classe de *LS* est *Urban* et *PER_URB_NEIGH* $\geq$ 50 %)

rouge sur les cartes b, e et h de la figure 3.26) les pixels considérés comme densément urbains (pixels dans le rectangle de la figure 3.26) ont une qualité d'habitat inférieure à celles des pixels voisins. Á Saint-Laurent-du-Maroni, la qualité d'habitat est élevée sur toute la zone

urbaine et il n'y a pas de pixel considéré comme densément urbanisé.

VIII. Discussion

L'objectif de ce chapitre était double et consistait à appliquer la méthode de correction du biais d'échantillonnage présentée au chapitre 2 aux données réelles d'*Anopheles darlingi*, et de modéliser la distribution de la qualité des habitats écologiques de ce vecteur à l'échelle de la Guyane française.
La carte de qualité d'habitat obtenue est en cohérence avec la connaissance actuelle des entomologistes de l'Unité d'entomologie médicale de l'Institut Pasteur de la Guyane et ce malgré le faible nombre de points de présence et le fort biais d'échantillonnage. Selon les zones géographiques, les IQH élevés sont caractérisés par des variables environnementales différentes.

VIII. 1. Lien entre les facteurs environnementaux et la qualité d'habitat

Dans la plupart des zones (A, B, D et E), l'IQH élevé est lié à la présence et aux activités humaines, caractérisées par les variables *HA_max* et *ROADS*. Dans les zones D, E et B, les routes sont très souvent non-pavées et correspondent à des pistes.
La corrélation significative entre la longueur de routes et de pistes (*ROADS*) et l'IQH tend à confirmer que la construction des routes et des pistes, accompagnée d'une déforestation, crée une accumulation d'eau le long des voies, favorisant la formation des gîtes larvaires (Singer and Castro, 2001). La courbe de réponse de la variable *ROADS* montre une forte augmentation de l'IQH jusqu'à 7 000 m par km^2 pour atteindre un plateau au dessus de 7 000 m et ensuite diminuer. Cette diminution après 7 000 m peut signifier que la densité du réseau routier provoque une perte de la forêt, éloignant les sites de repos de la population humaine. La courbe de réponse de la variable *PER_URB_NEIGH* montre que l'urbanisation (intense) ne permet pas une qualité habitat favorable à *An. darlingi*. En effet, l'urbanisation intensive implique le bétonnage et le goudronnage des routes, diminuant ou retirant les espaces verts ou arborés, détruisant ainsi les gîtes larvaires favorables et les sites de repos privilégiés d'*An. darlingi* (Stefani et al., 2013). Ce phénomène est observé dans les zones de Cayenne et de Kourou (figure 3.26). En revanche, Saint-Laurent-du-Maroni, la seconde plus grande zone urbaine en terme de taille et de densité d'urbanisation, apparaît avec un IQH très élevé. Bien que corrigée par la méthode de correction du biais d'échantillonnage basée sur la connaissance *a priori* de la bio-écolologie du vecteur, la qualité d'habitat reste élevée dans cette zone. Cependant, contrairement à Cayenne et Kourou, aucun des pixels de Saint-Laurent-du-Maroni n'est considéré comme très fortement urbanisé selon les critères utilisés dans ce chapitre (*i.e* *PER_URB_NEIGH* $\geq$ 50%). La caractérisation de l'occupation du sol devrait être actualisée. Des captures d*Anopheles* pourraient également être planifiées dans cette zone pour y confirmer ou non la présence d'*An. darlingi*. Enfin, la sensibilité du modèle à la définition des zones densément urbanisées pourrait être étudiée.

Les valeurs de la variable quantifiant les activités humaines (*HA_max*) dans les zones D et E sont essentiellement associées à l'activité minière. En Guyane française, cette activité est responsable de plus 2 000 hectares de déforestation annuelle (près 1 200 hectares en 2012, due à l'activité illégale). Entre 2001 et 2013, Alvarez-Berríos and Aide (2015) estiment qu'en forêt humide tropicale et subtropical d'Amérique du Sud, la plus grande proportion de

déforestation pour l'activité aurifère se trouve dans la forêt des quatre Guyanes (le Vénézuela, le Guyana, le Suriname et la Guyane française). Ceci indique que l'activité aurifère, associée à une déforestation, à des retenues d'eau et à la présence d'un nombre important d'hommes au milieu de la forêt, créé une situation particulièrement favorable à *An. darlingi*.
Dans les zones D et E de la carte en figure 3.25, l'IQH élevé est également lié à la classe d'occupation du sol *mixed high and open forest* qui est associée à de la forêt avec de la végétation secondaire et dégradée du fait de l'anthropisation (Gond et al., 2011). Ces résultats confirment l'importance du rôle de l'Homme et de ses activités dans la création des conditions favorables à *An. darlingi*.

Certaines classes d'occupation du sol sont associées à un IQH élevé. C'est la cas de la classe *woodland savanna / dry forest* (figure 3.23). Elle correspond à des zones sèches qui peuvent être régulièrement inondées créant ainsi des gîtes larvaires (Gond et al., 2011; Rosa-Freitas et al., 2007). Ceci est en cohérence avec les captures effectuées dans les savanes le long du littoral (Vezenegho et al., 2015; Dusfour et al., 2013) ayant confirmé la présence d'*An. darlingi* dans ce type de milieu en quantité parfois élevée. Les zones F et C, connues pour être inhabitées, apparaissent avec des IQH élevés et sont liées à la classe d'occupation sol *open forest*. La zone C est associée à des zones humides (ou considérés comme des forêts inondées selon l'expertise littoral (ONF, 2013)) tandis que la zone F, située au Sud, correspond à de la cambrouse (Gond et al., 2011). La cambrouse est un terme guyanais désignant des formations de graminées bambusiformes denses et difficilement pénétrables (De Granville, 1990). Elle ne permet pas à d'autres espèces végétales de germer ni de croître pour reconstituer la forêt. *An. darlingi* a déjà été observé en forêt inondée. Cependant, aucune information de la littérature n'a fait mention de sa présence en cambrouse. Une nomenclature d'occupation du sol différenciant des deux types d'occupation du sol s'avère nécessaire à l'avenir. Par conséquent, la prédiction de l'IQH dans la zone F doit être prise avec précaution.

VIII. 2. Correction de l'effet du biais d'échantillonnage

La méthode de correction de l'effet du biais d'échantillonnage, consistant à sélectionner les points de background avec le même biais environnemental que celui des points de capture, a été appliquée aux données réelles d'*An. darlingi*. Un modèle (dit non-corrigé) a également été généré avec les mêmes données d'entrée que le modèle simplifié, au paragraphe VII. 1. , mais sans l'application de la méthode de correction du biais d'échantillonnage. Les résultats de ce modèle non-corrigé ne sont pas présentés dans le détail, mais confirment l'intérêt de la correction. En effet, cette dernière semble appropriée car sans la correction, le modèle prédit une qualité d'habitat très élevée dans les zones urbaines denses alors que ces zones devraient être défavorables au développement d'*An. darlingi*. Les sites de background biaisés (c'est-à-dire en appliquant la méthode de correction du biais d'échantillonnage) sont sélectionnés dans le voisinage environnemental des sites de capture, signifiant que leurs conditions environnementales sont plus proches de celles des sites de captures que si les sites de background avaient été choisis selon un tirage aléatoire uniforme. La carte de qualité d'habitat issue du modèle corrigé présente ainsi des zones exclues de la prédiction tandis que celle résultant du modèle non corrigé n'en présente pas, malgré le choix de ne pas extrapoler dans les deux cas. Ceci signifie que le modèle construit avec le background uniforme prédit dans des zones

ayant des conditions environnementales très différentes de celles des sites de captures ce qui peut mener à une carte de qualité d'habitat erronée, tout au moins dans ces zones.
Quantitativement, les modèles corrigé et non-corrigé ont un AUC moyen et des rapports d'AUC partiels moyens similaires. Cependant, le gain moyen et l'indice continu de Boyce moyen du modèle avec la correction sont supérieurs à celui sans correction. Ils sont en effet de 3,18 (contre 2,81) et de 0,421 (contre 0,284), respectivement.
Les résultats du modèle corrigé montrent une forte contribution de la variable *ROADS*. Bien que cette contribution ait été précédemment expliquée au regard de la connaissance des caractéristiques des gîtes larvaires d'*An. darlingi*, elle peut cependant résulter d'un effet du biais d'échantillonnage résiduel. En effet, les points de capture sont situés dans des zones accessibles, notamment le long des routes ou à proximité des villes, des villages ou des zones d'orpaillage. Bien que la méthode de correction ait tenté de corriger l'effet de ce biais, il se peut que cela ne suffise pas.

VIII. 3. Le paludisme et la qualité d'habitat d'*An. darlingi* en Guyane française

Alimi et al. (2015) soulignent l'importance de faire appel à des modèles de distribution d'espèces pour permettre une meilleure compréhension de la distribution des vecteurs et aussi faciliter l'élimination du paludisme et la prévention des épidémies. En Guyane française, la bande littoral est généralement considérée comme épargnée par la transmission du paludisme, bien que quelques cas persistent (Ardillon et al., 2015). Toutefois, le travail de ce chapitre, ainsi que les travaux de Vezenegho et al. (2015), montrent que les savanes de Guyane, essentiellement présents sur le littoral, semblent très favorables à *An. darlingi*.
En forêt, Pommier de Santi et al. (2016b) ont trouvé un lien fort entre l'activité d'orpaillage, les cas de paludisme et la présence d'*An. darlingi*. En effet, plus 74 % des cas diagnostiqués chez les soldats sont associés à des opérations de lutte contre l'orpaillage illégal (Pommier de Santi et al., 2016a). Cependant, selon les résultats de ce chapitre, certaines zones associées à une activité minière intense et connues pour être des foyers de transmission importants, ne correspondent pas forcément à des zones avec un IQH d'*An. darlingi* élevé. Dans le village de Camopi, la prévalence annuelle était de 70% chez les enfants de moins de sept ans entre 2000 et 2002 (Carme et al., 2005), et atteignait 100% en 2006 (Hustache et al., 2007). Cependant, seuls quelques pixels situés le long de la rivière Camopi et dy fleuve Oyapock présentent des IQH élevés. Bien que cela semble contredire les situations épidémiologiques décrites précédemment, ceci est en cohérence avec les travaux de Girod et al. (2011), qui montrent que le nombre d'*An. darlingi* capturés dans ce village est très faible au regard du nombre de cas de paludisme et des densités observées en d'autres lieux de Guyane. Zanini et al. (2014) montrent également que les densités d'*An. darlingi* sont relativement très faibles à Vila Brazil, une localité brésilienne faisant face au village.
Ces remarques soulignent deux points importants : premièrement, la carte de qualité d'habitat (figure 3.25) n'est pas une carte de risque de transmission du paludisme. En effet, le risque de transmission dépend de différents facteurs qui n'ont pas été pris en compte dans ce travail, tels que la charge parasitaire de la population locale, la configuration et la composition des paysages (Stefani et al., 2013; Li et al., 2016) et le comportement de la population humaine ; deuxièmement, la transmission du paludisme peut avoir lieu dans une zone où la densité d'*An. darlingi* est relativement faible et où la qualité d'habitat est également faible. Ceci peut aussi

s'expliquer par la présence d'autres espèces d'*Anopheles* telles que *An. nuneztovari, An. oswaldoi, An. intermedius* ou *An. marajoara*, ayant déjà été trouvées naturellement infectées par des espèces *Plasmodium* et/ou décrites comme espèces vectrices secondaires (Dusfour et al., 2012a; Pommier de Santi et al., 2016b).

VIII. 4. Caractérisation environnementale

Une limite importante de ce travail concerne la résolution spatiale des données environnementales d'entrée. Les campagnes de capture sont généralement effectuées à une échelle locale (échelle du village ou d'un camp d'orpaillage, ou d'un quartier) (Dusfour et al., 2013; Vezenegho et al., 2015) alors que la résolution spatiale de référence des données environnementales est de 1 x 1km. Cette résolution n'est pas suffisamment fine pour prendre en compte l'hétérogénéité de l'environnement à l'échelle des points de captures. L'utilisation de données environnementales avec une résolution spatiale plus fine, telles que les données de hauteur de la canopé à 250 m de résolution (Fayad et al., 2016a) ainsi que la carte de biomasse Fayad et al. (2016b) (publiée en août 2016) pourraient être utilisée dans de futurs travaux.

IX. Conclusion

La méconnaissance de la distribution spatiale du principal vecteur du paludisme sur l'ensemble du territoire guyanais constitue un obstacle important à l'estimation du risque de transmission, à la lutte antivectorielle et plus généralement à la lutte contre le paludisme dans cette région. Les données de présence disponibles et exploitables à l'échelle de la Guyane sont relativement rares et présentent un fort biais d'échantillonnage dû à un accès limité à l'ensemble du territoire. L'objectif principal du travail réalisé dans ce chapitre était de modéliser la distribution d'*An. darlingi* à l'échelle de la Guyane compte tenu de ces limites. La méthode décrite dans le chapitre 2 a été utilisée pour corriger l'effet du biais d'échantillonnage des données de présence d'*An. darlingi*. Le modèle de distribution d'*An. darlingi* a ensuite été construit à partir de données environnementales qui, selon la littérature, influencent de manière positive ou négative la présence de ce vecteur. Pour ce faire, le modèle Maxent a été utilisé. La carte de qualité d'habitat résultante est en cohérence avec la connaissance actuelle des entomologistes travaillant sur la Guyane, ayant participé à ces travaux, et fournie des résultats de prédictions très satisfaisants (notamment un AUC de 0,93).
Les résultats de ce chapitre aide à compléter la connaissance actuelle sur la distribution spatiale du principal vecteur du paludisme en Guyane et d'identifier les principaux facteurs environnementaux favorisant sa présence. Ces résultats peuvent être exploités pour guider la lutte antivectorielle en Guyane afin d'atteindre la situation de pré-élimination et peuvent être extrapoler à l'échelle de la région amazonienne. Cette méthodologie pourrait dans le futur être appliquée aux vecteurs secondaires du paludisme en Guyane ou à des vecteurs d'autres maladies.
Bien que la méthode de correction de l'effet du biais d'échantillonnage, appliquée à des données réelles fortement biaisées, ait permis d'obtenir une carte de qualité d'habitat satisfaisante, l'impact réel d'une telle méthode, en absolue et relativement aux autres méthodes décrites dans la littérature (*cf.* chapitre 1), reste à être évaluer. Dans le prochain chapitre, cette méthode sera ainsi comparée aux méthodes existantes dans un cadre où les données de présence sont en faible quantité. La génération de jeux de données simulées s'appuyant sur les

résultats du présent chapitre, permettra de bénéficier de jeux de données de références pour ces évaluations objectives et quantitatives.

Zone	ROADS	Modalités de *LS*	*PER_URB_NEIGH*	*HA_max*	Modalités *GLS*	Modalités de *GLF*	*ALT*
A	(+)	- Woodland savanna / dry forest - Mixed high and open forest	(-)	(+)	- Coastal plain with low relief - Plain with residual reliefs (back coastal)	- Small size and flat wetland - Large flattened and wet relief - Wet hillock (low base-level)	(-)
B	(+)	- Open forest - Mixed high and open forest	*ns.*	(+)	- Peneplain with moderate hills	- Wet hillock (low base-level) - Large flattened relief	(-)
C	*ns.*	- Open forest	*ns.*	*ns.*	- Coastal flat plain	- Large flattened and wet relief	(-)
D	(+)	- Mixed high and open forest	*ns.*	(+)	*ns*	*ns*	(-)
E	(+)	- Mixed high and open forest	*ns.*	(+)	- Peneplain with moderate hills	- Large flattened relief	(-)
F	*ns.*	- Open forest	*ns.*	*ns.*	- Peneplain with moderate hills	-Lowered half-orange	(-)
G	(+)	- Mixed high and open forest	*ns.*	(+)	*ns.*	*ns.*	(-)

TABLEAU 3.6 – **Caractérisation des zones avec un indice de qualité d'habitat élevé selon les valeurs ou les modalités des variables environnementales**
ns., (+) et (-) ne concernent que les variables quantitatives. *ns.* signifie que l'IQH élevé ne dépend pas de variable environnemental, (+) signifie que l'IQH augmente quand la valeur de la variable environnementale augmente, (-) signifie que l'IQH diminue lorsque la valeur de la variable environnementale diminue

Chapitre 4

Comparaison des méthodes de correction de l'effet du biais d'échantillonnage

Introduction . . . 91
I. Méthodologie générale de comparaison . . . 91
II. Simulation des sites de présence . . . 92
II. 1. Simulation des cartes de qualité d'habitat et de présence-absence . . . 92
II. 2. Génération du biais d'échantillonnage . . . 94
II. 3. Sélection des sites de présence . . . 95
III. Correction de l'effet du biais d'échantillonnage . . . 96
III. 1. Sélection des sites de présence basée sur des critères géographiques . . . 96
III. 2. Sélection des sites de présence basée sur des critères environnementaux . . . 97
III. 3. Construction d'un background biaisé basé sur des critères géographiques . . . 97
III. 4. Construction d'un background biaisé basé sur des critères environnementaux . . . 97
III. 5. Modélisation de la distribution virtuelle d'*An. darlingi* . . . 98
IV. Évaluation et comparaison des méthodes de correction . . . 98
V. Résultats . . . 101
VI. Discussion . . . 103
VI. 1. Évaluation absolue des méthodes de correction . . . 103
VI. 2. Évaluation relative des méthodes de correction . . . 103
VI. 3. *BGeng vs. BGenv_tg* . . . 104
VI. 4. Paramétrisation de *BGenv* et BGenv_tg . . . 105
VII. Conclusion . . . 105

Introduction

Dans le chapitre 1, l'état de l'art des méthodes existantes de corrections de l'effet du biais d'échantillonnage a été réalisé.
Dans le chapitre 2, une méthode de correction de l'effet du biais d'échantillonnage originale et générique a été proposée. Elle consiste à construire un background biaisé sur la base de critères environnementaux. Cette méthode a été développée afin de pouvoir être appliquée spécifiquement à des cas d'étude où le nombre de sites de présence de l'espèce (ou du groupe d'espèces) cible est faible. Elle relève de critères environnementaux et est paramétrée d'après des connaissances sur la bio-écologie de l'espèce cible, indépendamment des propriétés des données dans l'espace géographique.
Dans le chapitre 3, Maxent à été appliqué à des données réelles d'*An. darlingi* qui présentaient un biais d'échantillonnage important. La méthode de correction proposée dans le chapitre 2 a été appliquée afin de réduire l'effet de ce biais sur la prédiction de la qualité d'habitat. Les résultats de modélisation obtenus sont très bons et la carte de qualité d'habitat est en concordance avec la connaissance actuelle des entomologistes sur la distribution de cette espèce.
Dans le présent chapitre, la méthode de correction de l'effet du biais d'échantillonnage développée dans le chapitre 2 est comparée aux méthodes existantes. En effet, à notre connaissance, aucune étude n'a comparé l'ensemble des quatre types de méthodes de correction de l'effet du biais d'échantillonnage décrites dans la littérature et présentées dans le tableau 1.3. Ainsi, l'objectif de ce chapitre est d'évaluer, de manières relative et absolue, la capacité à corriger l'effet du biais d'échantillonnage de la méthode de correction proposée dans cette thèse. Cependant, l'absence de données de référence fiables rend cette évaluation problématique. Pour répondre à cet objectif, des données de présence non-biaisées, considérées comme des données de référence, et des données de présences biaisées seront simulées afin de pouvoir appliquer les différents modèles de correction et obtenir une évaluation fiable. La simulation doit satisfaire les conditions suivantes :

— un faible nombre de sites de présence (correspondant aux conditions pour lesquelles la méthode a été développée) ;
— un cadre expérimental réaliste et le plus proche possible de l'application à *An. darlingi* en Guyane française (en termes d'échelle et de résolution d'étude, de caractérisation environnementale et de biais d'échantillonnage).

I. Méthodologie générale de comparaison

La méthodologie générale permettant d'évaluer les méthodes de correction est présentée ci-après. Elle est inspirée de celle de Fourcade et al. (2014). Elle se décompose en quatre grandes parties :

1. n sites de présence sont simulés. Parmi eux sont sélectionnés, d'une part, k ($k \in \{20, 50, 100, 150, 200\}$) sites de présence non-biaisés qui sont considérés comme des données de référence et, d'autre part, k sites de présence biaisés ;
2. Cinq méthodes de correction sont appliquées aux données biaisées ;
3. Maxent est appliqué, séparément, aux données non-biaisées, aux données biaisées et aux données corrigées selon les différentes méthodes de correction. En sortie des modèles sont obtenues des cartes de qualité d'habitat ;

4. Les sorties de chaque modèle corrigé (c'est-à-dire obtenu à partir des données corrigées) sont ensuite comparées à celles du modèle non-biaisé (c'est-à-dire obtenu à partir des données non-biaisées) puis à celles des autres modèles corrigés.

Cette méthodologie est schématisée en figure 4.1. La méthode est répliquée 100 fois pour chacune des cinq valeurs de k. Ainsi, sont générés au total :

— 500 modèles non-biaisés ;
— 500 modèles biaisés ;
— 2500 modèles corrigées (5 méthodes de corrections $\times$ 500 modèles).

Chaque étape de la méthode est détaillée par la suite.

II. Simulation des sites de présence

II. 1. Simulation des cartes de qualité d'habitat et de présence-absence

La simulation les données de présence d'*An. darlingi* sur le territoire de la Guyane française est basée sur les résultats de modélisation du chapitre précédent, portant sur l'application de Maxent aux données réelles d'*An. darlingi* sur ce même territoire. Pour cela, la librairie *virtualspecies* de R, dont le principe de fonctionnement est schématisé en annexe E, a été utilisée. Pour générer ces sites de présence, il est nécessaire d'avoir une carte de présence-absence, construite à partir d'une carte de qualité d'habitat obtenue au moyen de courbes de réponse des variables environnementales. Pour cela, certaines variables environnementales entrant dans la construction du modèle du chapitre 3 ont été retenues pour la simulation :

— le pourcentage d'urbanisation (*PER_URB*) ;
— le pourcentage d'urbanisation des pixels voisins (*PER_URB_NEIGH*) ;
— la longueur des routes et des pistes dans un pixel (*ROADS*) ;
— la présence et les activités humaines qui altèrent de manière non-permanente l'environnement naturel (*HA_max*) ;
— les classes correspondant aux marécages, aux savanes arbustives et aux forêts mixtes (*mixed high and open forest*) issues de la carte d'occupation (*LS*) décrit dans le chapitre précédent.

Ces variables ont été choisies car elles sont apparues comme ayant des contributions fortes ou modérées au modèle de distribution d'*An. darlingi* d'après les résultats du chapitre précédent. Les courbes de réponse associées à ces variables ont été définies d'après les courbes obtenues dans le chapitre précédent. Celles des variables continues sont présentées en figure 4.2. La combinaison de ces courbes de réponse via une fonction additive (implémentant le OU logique) permet d'obtenir une carte de qualité d'habitat qui a été normalisée afin que la somme des valeurs de qualité d'habitat sur l'ensemble de la zone d'étude soit égale à un (figure 4.3). La conversion de cette dernière en une carte de présence-absence se fait via une fonction logistique, qui représente la relation entre la qualité d'habitat et la probabilité de présence. Dans cette étude, une prévalence de 0.2 a été choisie, représentant la proportion de pixel de présence parmi l'ensemble des pixels de la zone d'étude. Les pixels de présence sont alors choisis aléatoirement, en pondérant leur probabilité de choix par la probabilité de présence. Plus un pixel a une qualité d'habitat élevée (et donc une probabilité de présence élevée), plus il a de chance d'être sélectionné comme pixel de présence. Un pixel n'étant pas choisi comme pixel de présence est désigné comme pixel d'absence. La carte de présence-absence obtenue est présentée en figure 4.4.

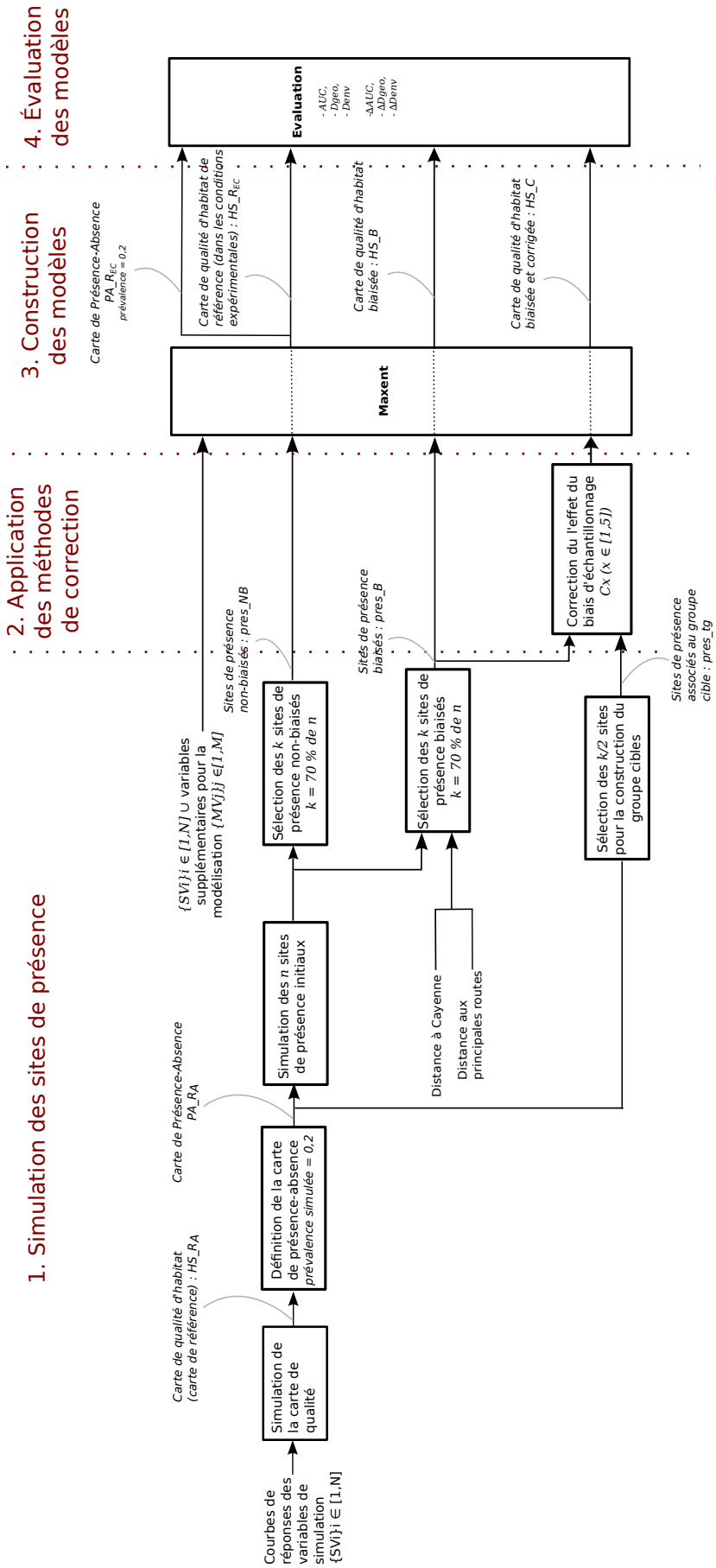


FIGURE 4.1 – **Schéma général de la simulation pour la comparaison des méthodes de correction de l'effet du biais d'échantillonnage**

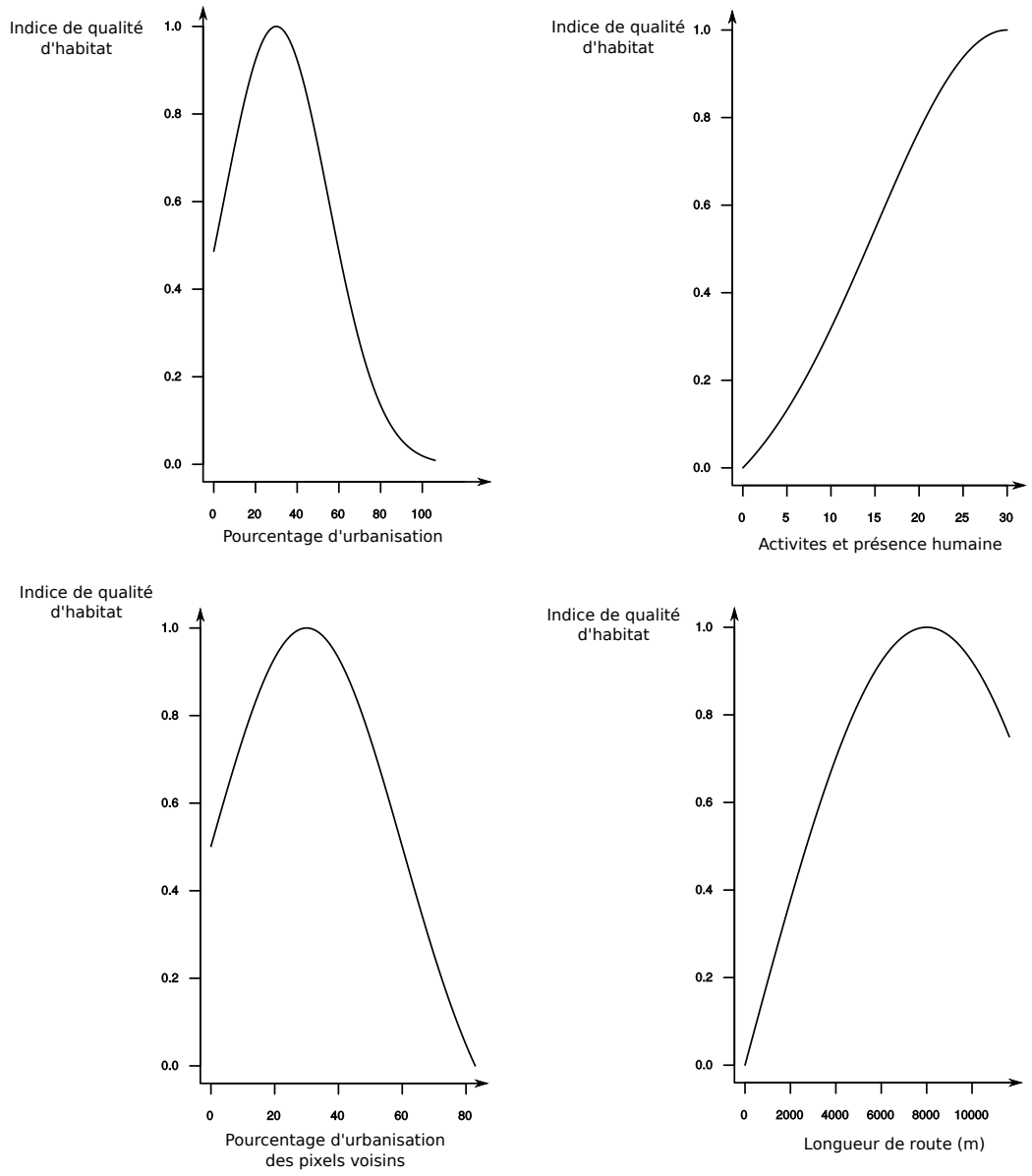


FIGURE 4.2 – **Courbes de réponse des variables continues définies pour la génération des données d'*An. darlingi* virtuelles.**

II. 2. Génération du biais d'échantillonnage

Pour représenter le biais d'échantillonnage lors des captures des espèces, une carte de biais a été réalisée. En Guyane, les routes nationales situées le long du littoral constituent les principaux axes routiers et représentent un accès facile aux captures. La distance à la ville de Cayenne, où se situent les équipes d'entomologie médicale ayant recueilli les données de présence d'*An. darlingi*, représente également une source de biais dans les captures. En effet, la figure 3.21, représentant les captures réalisées entre 2000 et 2013, montre qu'une partie des captures de culicidés est effectuée à proximité immédiate de Cayenne. Ainsi, dans le présent chapitre, le biais d'échantillonnage est déterminé en fonction à la fois de la distance à Cayenne et de la distance aux routes principales : plus un pixel est situé proche de Cayenne *et* des routes principales, plus il a de chance de contenir un site de capture.

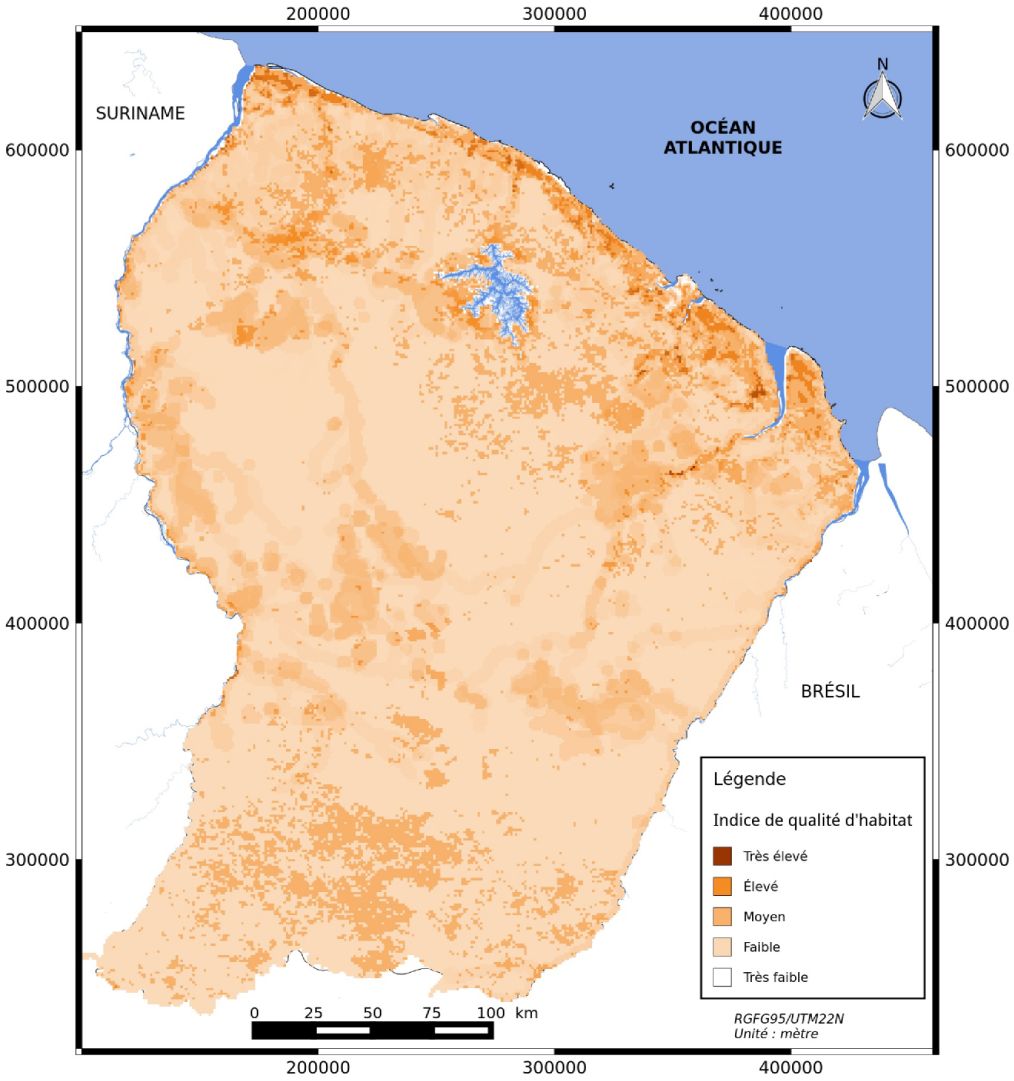


FIGURE 4.3 – **Carte de qualité d'habitat simulée.**

II. 3. Sélection des sites de présence

Afin de simuler un modèle non-biaisé et un modèle biaisé, deux types de sites de présence sont requis :

— des sites de présence non-biaisés (*pres_NB*) et
— des sites de présence biaisés (*pres_B*).

Tout d'abord, n sites de présence dits initiaux, notés *pres_NB_init*, sont sélectionnés de manière aléatoire uniforme parmi les pixels de présence de la carte de présence-absence obtenue précédemment.
Les k sites *pres_NB* et les k sites *pres_B* sont sélectionnés parmi les n sites de *pres_NB_init*. La valeur de n est fixée pour chaque valeur de k afin que les ensembles *pres_NB* et *pres_B* aient une proportion significative de points en commun. Ainsi, k est fixé à 70 % de n. Dans le cas où *pres_NB* (respectivement *pres_B*) contiendrait tous les sites non sélectionnés pour construire *pres_B* (respectivement *pres_NB*), au moins 40 % des n sites de *pres_NB_init* se trouvent à la fois dans *pres_NB* et dans *pres_NB*. Cette proportion, fixée arbitrairement, per-

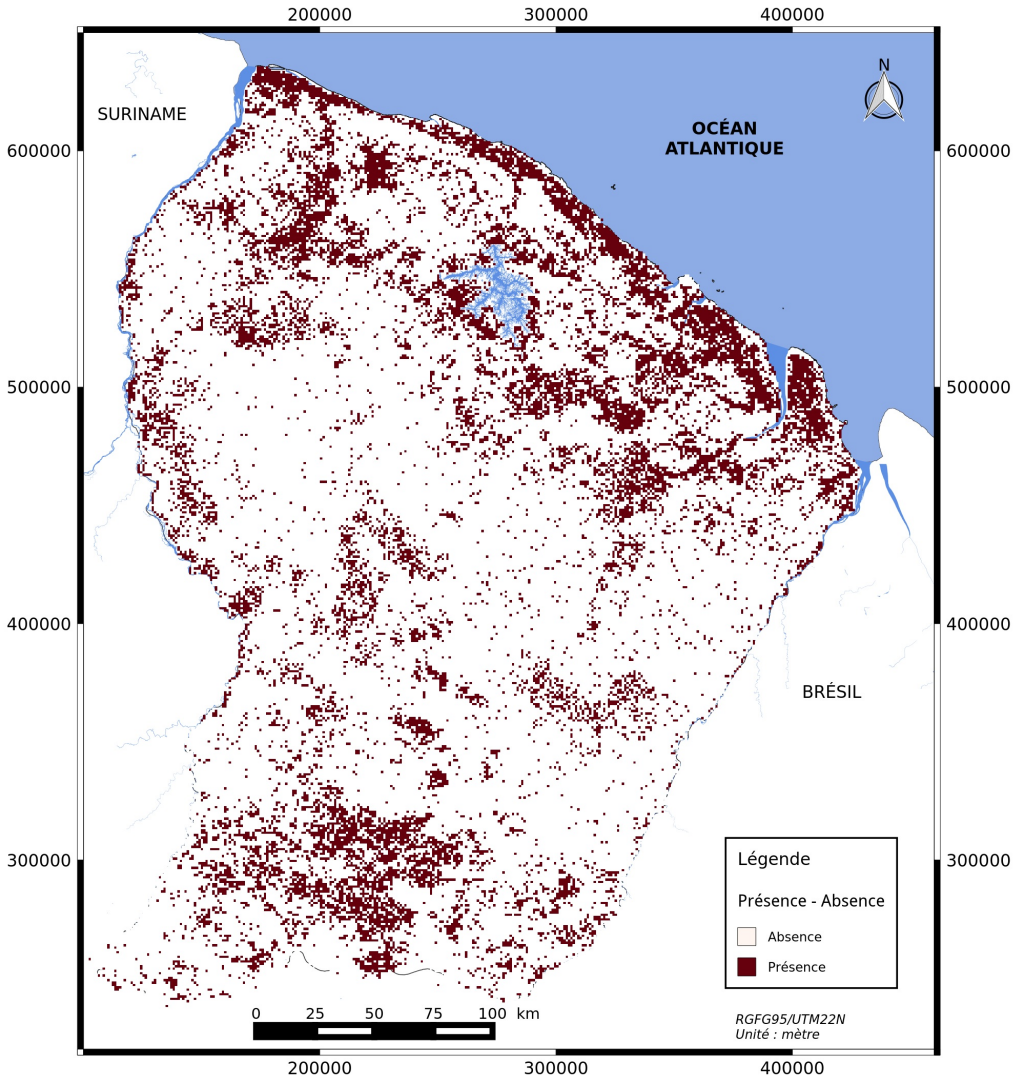


FIGURE 4.4 – **Carte de présence-absence simulée avec une prévalence de 0.2.**

met de limiter l'effet aléatoire dans le choix des sites *pres_NB* et *pres_B* et assure une certaine similarité, au biais d'échantillonnage près, entre ces deux jeux de données. Pour chaque valeur de k, 100 réplicats ont été générés. Les sites *pres_NB* sont sélectionnés de manière aléatoire uniforme parmi l'ensemble *pres_NB_init*. Les sites *pres_B* sont sélectionnés parmi l'ensemble *pres_NB_init* de manière aléatoire et pondérée par la carte de biais générée précédemment (§II. 2.).

III. Correction de l'effet du biais d'échantillonnage

Pour corriger l'effet du biais d'échantillonnage inséré précédemment, quatre types de méthodes de correction ont été appliqués. Les méthodes utilisées ont été décrites dans le chapite 1 et leur application aux données simulées est détaillée ci-après.

III. 1. Sélection des sites de présence basée sur des critères géographiques

Cette méthode consiste à filtrer les sites de présence situés à une distance inférieure ou égale à r d'un autre (*cf.* chap. 1. §IV. 1. a.). Elle est notée *Fgeo*. Dans la simulation, r est égal à

7 000 mètres, correspondant à la distance de déplacement maximale d'*An. darlingi* observée dans l'état de Rondônia en Amazonie brésilienne (Charlwood and Alecrim, 1989). Les sites de présence sélectionnés par cette méthode sont appelés *pres_C*$_{fgeo}$.

III. 2. Sélection des sites de présence basée sur des critères environnementaux

Selon la méthode présentée par Fourcade et al. (2014) (*cf.* chap. 1. §IV. 1. b.), une ACP des sites de présence caractérisés par les variables environnementales est réalisée. Dans cette partie, au lieu de l'ACP, une AFDM est réalisée afin de prendre en compte les variables qualitatives et quantitatives. Ensuite, une classification ascendante hiérarchique est effectuée en se basant sur la distance euclidienne entre les sites de présence, et ce, dans l'espace environnemental créé par les différents axes factoriels. Le nombre de classes est égal à la moitié du nombre de sites de présence. Une sélection aléatoire est effectuée afin de ne retenir qu'un site de présence par classe. Cette méthode est notée *Fenv* et les sites de présence sélectionnés sont appelés *pres_C*$_{fenv}$.

III. 3. Construction d'un background biaisé basé sur des critères géographiques

Cette méthode consiste à définir un voisinage géographique à partir d'une fonction de type gaussienne avec un écart-type égal à une valeur d, correspondant à la distance de déplacement ou au rayon du domaine vital de l'espèce (*cf.* chap. 1. §IV. 2. a.). La densité de présence dans le voisinage géographique de chacun des pixels, correspondant au nombre de pixels de présence sur le nombre total de pixels dans le voisinage, est calculée. Elle reflète le biais d'échantillonnage et sera utilisée pour sélectionner les sites de background. Cette méthode est notée *BGgeo* et les sites de background sélectionnés *bg_C*$_{geo}$. Dans la simulation, d est fixée à 7 000 mètres pour la même raison que pour le choix de r dans le paragraphe III. 1.

III. 4. Construction d'un background biaisé basé sur des critères environnementaux

Cette méthode est décrite dans le chapitre 2. Une AFDM est réalisée sur l'ensemble X des pixels de la zone d'étude, décrits par l'ensemble des variables environnementales. Deux cas de figure sont simulés :

— le cas où un groupe cible ne peut être défini, du fait de l'absence de donnée d'autres espèces échantillonnées de manière similaire à l'espèce cible. L'estimation du biais d'échantillonnage est réalisée à partir des sites de présence d'*An. darlingi* uniquement. Cette méthode est notée *BGenv* ;

— le cas où les sites de présence d'autres espèces permettent de définir un groupe cible mais sont peu nombreux. La simulation de ces sites consiste à sélectionner des sites dans les zones d'absence de la carte de présence-absence d'*An. darlingi*. Leur sélection est faite de manière aléatoire et pondérée par la carte de biais utilisée lors de la sélection des sites de présence d'*An. darlingi*. L'estimation du biais d'échantillonnage est réalisée à partir des sites de présence du groupe cible. Cette méthode est notée *BGenv_tg*.

Tout comme dans le chapitre précédent, le paramètre D_{min} est défini en se basant sur la connaissance *a priori* qu'*An. darlingi* n'est pas présent en zone urbaine dense. Sa valeur est alors égale à la distance euclidienne minimale entre les pixels de présence simulés et les pixels considérés comme densément urbanisés, dans l'espace environnemental. L'ensemble

des sites de background sélectionnés en se basant sur le biais d'échantillonnage estimé uniquement à partir des sites de présence est appelé bg_C_{env}. Celui sélectionné à partir du biais d'échantillonnage estimé avec le groupe cible est appelé $bg\_C_{env\_tg}$.

III. 5. Modélisation de la distribution virtuelle d'*An. darlingi*

La modélisation de la distribution virtuelle d'*An. darlingi* prend en entrée la totalité des variables utilisées dans le chapitre précédent :

- les paysages et les unités géomorphologique (*GLS* et *GLF*) ;
- l'occupation du sol modifiée (*LS*) ;
- l'altitude minimale (*ATL_min*) ;
- la présence et les activités humaines altérant de manière non-permanente l'environnement naturel (*HA_max*) ;
- la longueur de route et pistes dans un pixel de 1 km^2 (*ROADS*) ;
- le pourcentage d'urbanisation dans les pixels voisins (*PER_URB_NEIGH*).

Les différents modèles de distribution construits sont :

- les modèles "non-biaisés", c'est-à-dire construits à partir des 100 réplicats de l'ensemble de sites de présence non-biaisés ;
- les modèles "biaisés", c'est-à-dire construits à partir des 100 réplicats de l'ensemble de sites de présence biaisés ;
- les modèles biaisés puis corrigés, appelés modèles "corrigés", obtenus en considérant les données simulées biaisées et corrigées par les cinq méthodes décrites précédemment.

La construction des modèles "corrigés" est schématisée en figure 4.5. L'ensemble des entrées-sorties des différents modèles et leurs notations sont répertoriées dans le tableau 4.1

IV. Évaluation et comparaison des méthodes de correction

Pour évaluer chacune des méthodes de correction, trois métriques d'évaluation absolues des résultats de prédiction sont calculées pour chacun des modèles corrigés et biaisés :

- l'AUC qui prend en entrée, pour chacun des modèles, la carte de présence-absence de référence *PA_REC* et la carte de qualité d'habitat obtenue en sortie, soit du modèle biaisé *HS_B*, soit du modèle corrigé, HS_C_x (figure.4.6), avec $x \in$ {Fgeo, Fenv, BGgeo, BGenv, BGenv_tg}.

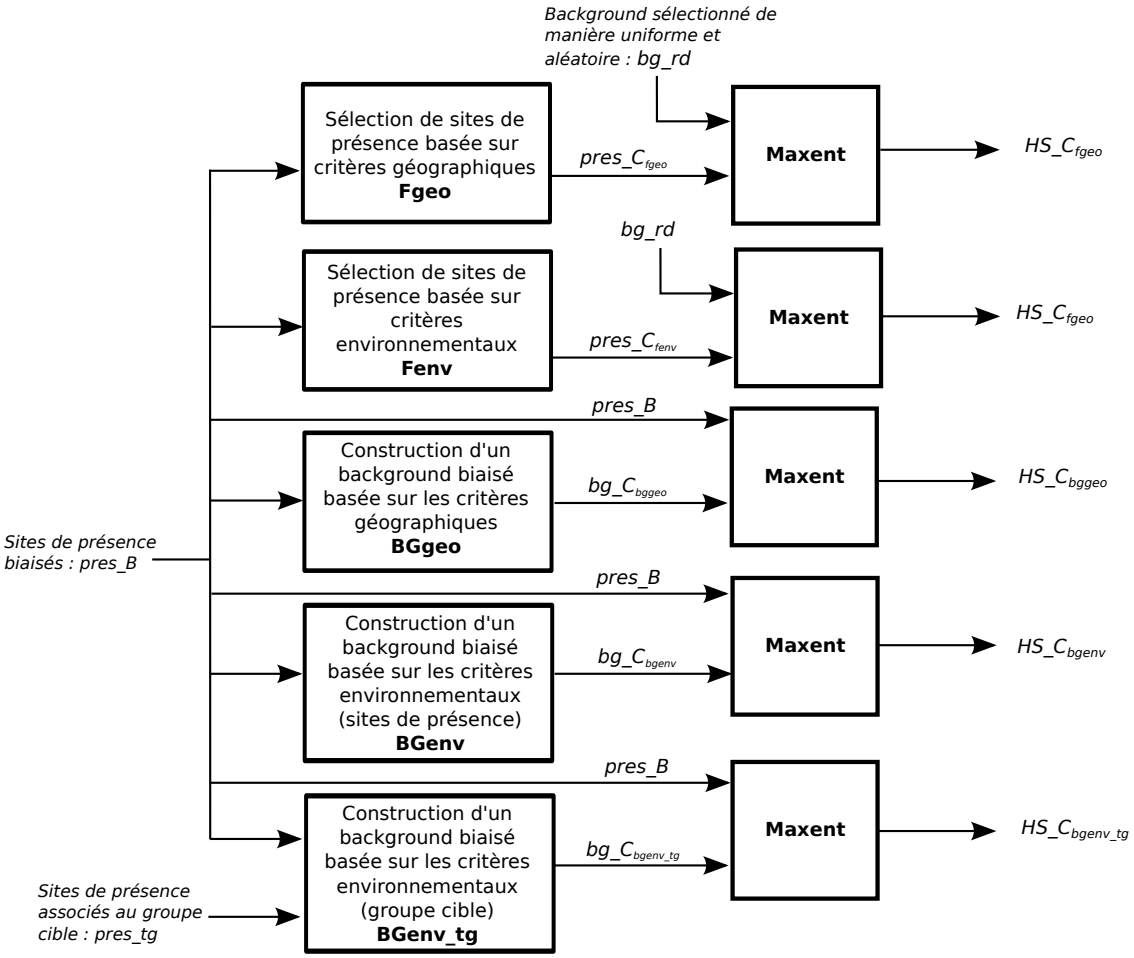


FIGURE 4.5 – **Schéma des entrées et sorties des méthodes de correction de l'effet du biais d'échantillonnage et des modèles de prédiction dits "corrigés".**

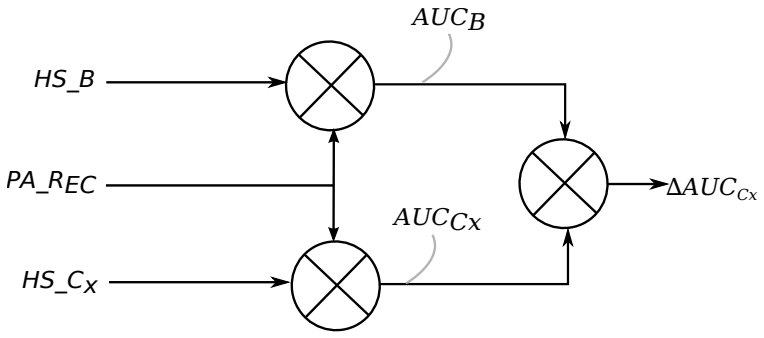


FIGURE 4.6 – **Calcul de ΔAUC.**

— l'indice de Schoener, $Dgeo$ (Schoener, 1968), qui permet de quantifier le chevauchement des habitats des espèces Y et Z dans l'espace géographique.

$$Dgeo(p_Y, p_Z) = 1 - \frac{1}{2}\sum_{i \in X} |p_{Y,i} - p_{Z,i}| \tag{4.1}$$

où p_Y et p_Z représentent, respectivement, les distributions de probabilité de l'espèce Y

et Z, obtenues en sortie des modèles de distribution, et $p_{Y,i}$ et $p_{Z,i}$ les probabilités de présence des mêmes espèces assignées au pixel i. Une valeur de 0 signifie qu'il y n'a aucun chevauchement des habitats écologiques des deux espèces et la valeur 1 signifie que les habitats des deux espèces sont strictement similaires.
Dans ce travail, l'espèce Y correspond aux données de référence, c'est-à-dire aux données non-biaisées, et l'espèce Z correspond soit aux données biaisées soit aux données corrigées (figure 4.7) ;

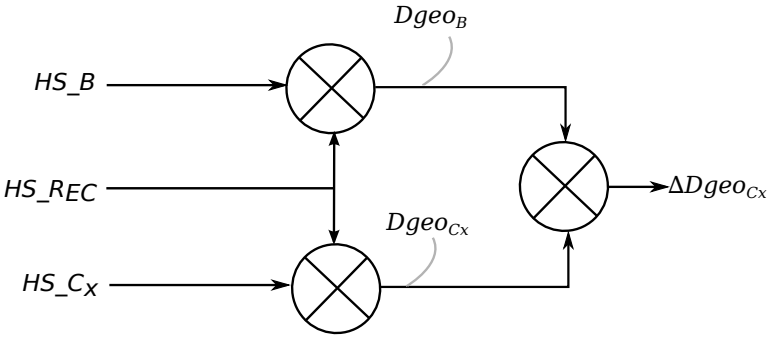


FIGURE 4.7 – **Calcul de** $\Delta Dgeo$**.**

— l'indice de Schoener, $Denv$ (Broennimann et al., 2012), qui permet de quantifier le chevauchement des habitats des espèces Y et Z dans l'espace environnemental. Ce dernier est défini par les deux premiers axes factoriels issus d'une ACP ou, dans cette étude, d'une AFDM. Ensuite, cet espace environnemental est divisé en $m \times m$ cellules, où m est défini par l'utilisateur (ici, $m = 100$). Chacune des cellules correspond a un contexte environnemental unique noté v_{ij} où i et j se réfèrent à la position de la cellule dans l'espace environnemental.
Le chevauchement environnemental des habitats des espèces Y et Z est défini comme suit :

$$Denv(z_Y, z_Z) = 1 - \frac{1}{2} \sum_{ij \in X} |z_{Y,ij} - z_{Z,ij}| \tag{4.2}$$

où $z_{Y,ij}$ et $z_{Z,ij}$ sont les estimations des densités de présence des espèces Y et Z, respectivement, dans l'environnement v_{ij}. Dans cette simulation, comme pour le calcul de $Dgeo$, l'espèce Y correspond aux données non-biaisées et l'espèce Z, aux données biaisées ou corrigées (figure 4.8). Les sites de présence représentant l'espèce Y sont sélectionnés de manière aléatoire et pondérée par la carte de qualité d'habitat non-biaisées et ceux de l'espèce Z par la sélection aléatoire et pondérée par les cartes de qualité d'habitat biaisée ou corrigée (Fourcade et al., 2014). Le nombre de sites sélectionnés, ni, est arbitrairement fixé à 500 sites.

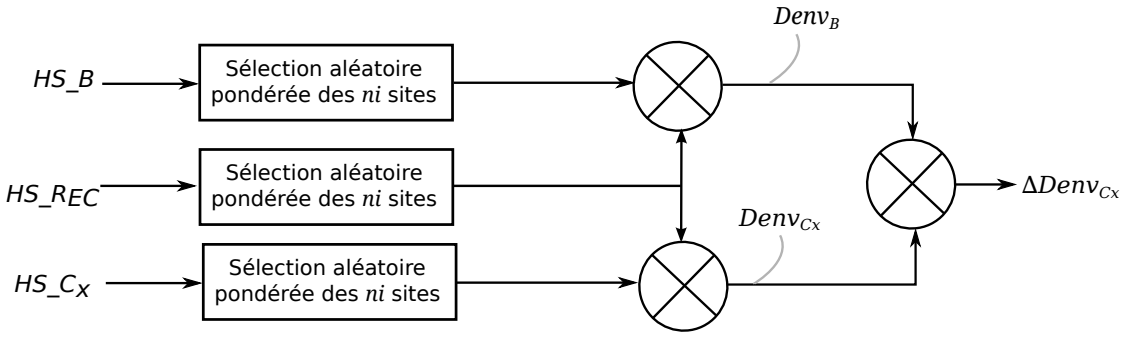


FIGURE 4.8 – **Calcul de $\Delta Denv$, avec $ni = 500$.**

Par la suite, ces métriques seront notées de la manière suivante :

— l'AUC d'un modèle corrigé avec la méthode x sera noté AUC_{Cx}, avec $x \in$ {Fgeo, Fenv, BGgeo, BGenv, BGenv_tg}, et l'AUC d'un modèle biaisé sera noté AUC_B ;

— $Dgeo(p_Y, p_Z)$ sera notée $Dgeo_Z$ (*e.g.* $Dgeo(p_{NB}, p_{Cx})$ *sera noté* $Dgeo_{Cx}$, *avec* $x \in$ *{Fgeo, Fenv, BGgeo, BGenv, BGenv_tg}) ;*

— *de la même manière,* $Denv(z_Y, z_Z)$ *sera noté* $Denv_Z$.

Pour comparer les méthodes de correction entre elles, trois indices d'évaluation relative (Fourcade et al., 2014) sont calculés (figures 4.6, 4.7 et 4.8) :

$$\Delta AUC_{Cx} = \frac{AUC_{Cx} - AUC_B}{1 - AUC_B} \tag{4.3}$$

$$\Delta Dgeo_{Cx} = \frac{Dgeo_{Cx} - Dgeo_B}{1 - Dgeo_B} \tag{4.4}$$

$$\Delta Denv_{Cx} = \frac{Denv_{Cx} - Denv_B}{1 - Denv_B} \tag{4.5}$$

Ces indices varient entre -∞ et 1. Une valeur négative signifie que le modèle corrigé produit des résultats moins bons que ceux du modèle biaisé. Une valeur positive signifie que les résultats du modèle corrigé sont meilleurs que ceux du modèle biaisé. Une valeur de 1 correspond à un cas "parfait", où les résultats du modèle corrigé sont strictement similaire à ceux du modèle de référence.

V. Résultats

Les résultats de l'*AUC*, $Dgeo$ et de $Denv$ sont représentés par des boxplots en figure 4.10.

Les valeurs minimales et maximales d'*AUC* varient entre 0.806 et 0.994 pour 20 sites de présence et de 0.905 à 0.998 pour 200 sites de présence. Celles de $Dgeo$ varient de 0.504 à 0.788 pour 20 sites de présences et de 0.685 à 0.829 pour 200 sites de présence. Enfin, les valeurs minimales et maximales de $Denv$ varient de 0.639 à 0.948 pour 20 sites de présences et de 0.726 à 0.915 pour 200 sites de présences.

Les résultats de l'*AUC* et de $Dgeo$ montrent que les performances des modèles augmentent avec le nombre k de sites de présence, jusqu'à une valeur de k égale à 100. Au-delà de 100 sites, les méthodes offrent des résultats comparables. En revanche, les résultats de $Denv$ sont assez variables et ne permettent pas d'identifier une tendance. Du point de vue de l'*AUC*, la

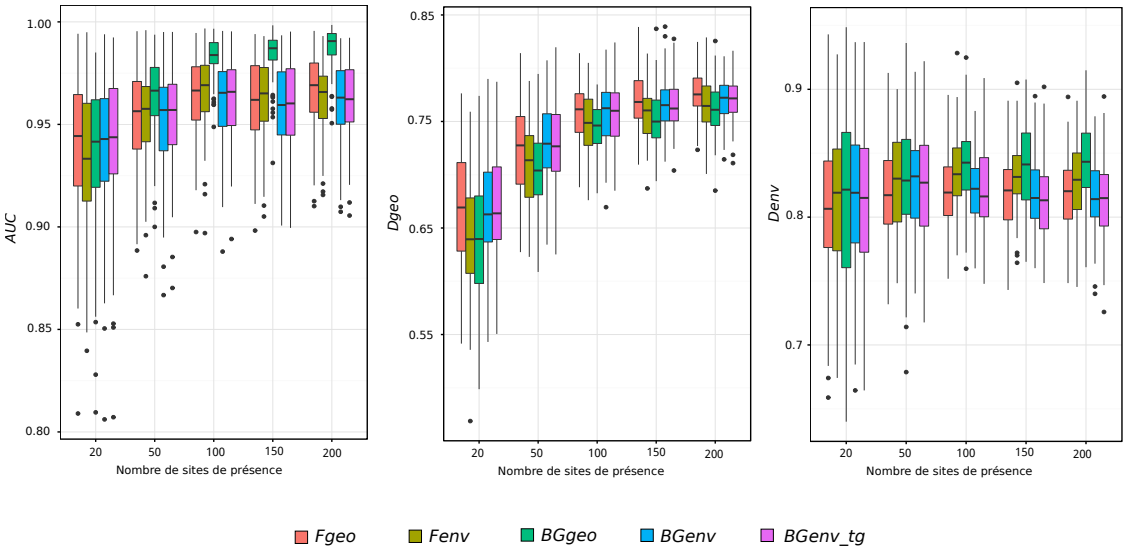


FIGURE 4.10 – **Les boxplots des valeurs des indices d'évaluation des différentes méthodes de correction de l'effet du biais d'échantillonnage.**

méthode *BGgeo* fournit des résultats nettement supérieurs aux autres méthodes à partir de 100 sites de présence.

Le tableau 4.2 répertorie les pourcentages des valeurs strictement positifs des indices d'évaluation relative, selon la méthode de correction et le nombre de sites de présence. Il s'agit, pour chaque condition expérimentale, du nombre de fois (sur 100 réplicats) que le modèle corrigé a permis de meilleurs résultats que le modèle biaisé. Ces résultats montrent que lorsque le nombre de sites de présence est de 20 sites, *BGenv* ou *BGenv_tg* procure le pourcentage le plus élevé tout en étant relativement faible, proche de 50 %.

Le classement des modèles corrigés selon leurs performances et en fonction des différentes méthodes de correction est présenté en figure 4.11. Le classement varie en fonction

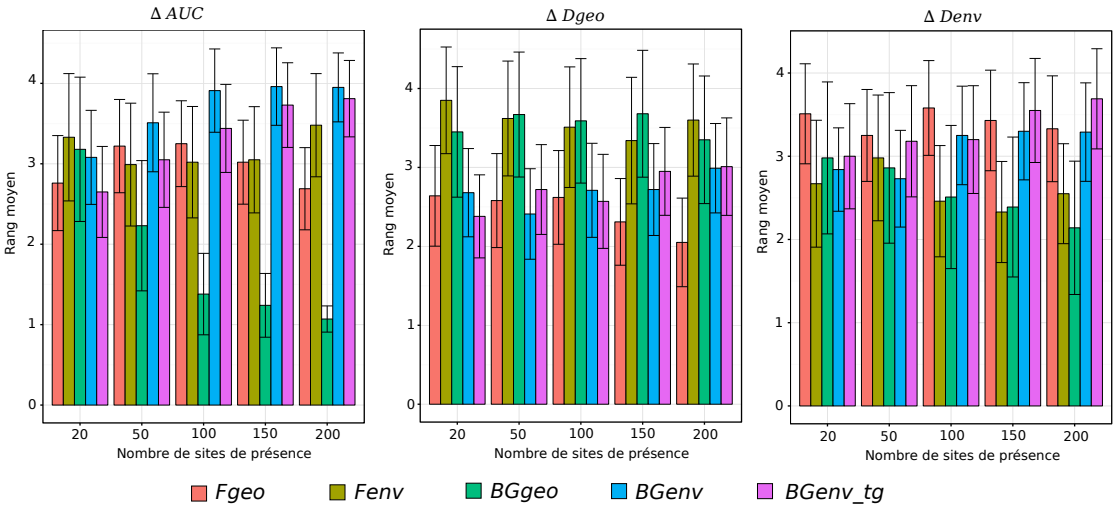


FIGURE 4.11 – **Rang moyen ± l'erreur-type de chaque méthode de correction de l'effet du biais d'échantillonnage.** La méthode la plus performante est rangée à 1 et la moins performante à 5

du type de correction et du nombre de sites de présence. Lorsque le nombre de sites de présence est de 20 sites, *BGenv_tg* est en moyenne classée en meilleure position selon ΔAUC et

$\Delta Dgeo$. Selon $\Delta Denv$, *Fenv* est la méthode la mieux classée, devant *BGenv*, pour 20 sites de présence. Pour $k = 50$, *BGenv* semble offrir les meilleurs résultats pour les indices $\Delta Dgeo$ et $\Delta Denv$.
Lorsque le nombre de sites augmente :

— la méthode de correction *BGgeo* est la mieux classée selon ΔAUC ;
— les méthodes *BGgeo* et *Fenv* sont les mieux classées selon $\Delta Denv$;
— la méthode *Fgeo* est la mieux classée selon $\Delta Dgeo$.

VI. Discussion

L'objectif de ce chapitre était de comparer les différentes de méthodes de correction entres elles dans le cas où le nombre de de présence est faible. Globalement, les résultats apparaissent très variables d'une métrique à une autre. Cependant, lorsqu'on se focalise sur le classement des méthodes lorsque le nombre de sites de présence est faible, les résultats des modèles corrigés avec la méthode développée dans cette thèse présentent les meilleurs rangs.

VI. 1. Évaluation absolue des méthodes de correction

Les modèles corrigés ont été évalués par rapport aux modèles non-biaisés à partir de l'*AUC* et des indices de Schoener calculés dans les espaces géographique ($Dgeo$) et environnemental ($Denv$). La figure 4.10 montrent que pour l'*AUC* et $Dgeo$, plus le nombre de sites de présence augmente, plus les performances des modèles corrigés augmentent et plus la variabilité des résultats diminue. Bien que Wisz et al. (2008) aient montré que Maxent était peu sensible au nombre de sites de présence, il apparaît que les performances sont fortement altérées en deçà de 100 sites de présence. Quelque soit la métrique d'évaluation, la méthode de correction proposée dans cette thèse (indifféremment *BGenv_tg* et *BGenv*) donne des résultats satisfaisants et comparables aux autres méthodes. Lorsque k est égale ou supérieur à 100, *BGgeo* offre des résultats significativement meilleurs, selon les résultats d'AUC. Il est, cependant, difficile de conclure définitivement sur la ou les méthodes à privilégier, lorsque le nombre de site de présence est inférieur à 100, en se basant uniquement sur une évaluation absolue.

VI. 2. Évaluation relative des méthodes de correction

Le tableau 4.2 indique que pour 20 et 50 sites de présence, 71 % et 11% des modèles corrigés avec *Fgeo* sont similaires aux modèles biaisés, selon ΔAUC et $\Delta Dgeo$. Ceci s'explique par le fait que lorsque les sites de présence biaisés se situent à une distance supérieure à r les uns des autres, aucun filtrage de données n'est réalisé et les sites de présence "corrigés" sont exactement les mêmes que les sites de présence biaisés. Lorsque le nombre de sites est faible ($k = 20$), les pourcentages de valeurs positifs sont tous inférieurs à 50% pour ΔAUC et $\Delta Dgeo$, à l'exception de la méthode *BGenv_tg* qui permet 53% de valeurs positives pour $\Delta Dgeo$. Ces valeurs négatives signifient que les modèles corrigés ne permettent pas de meilleurs résultats que les modèles biaisés. De la même manière, les travaux de Fourcade et al. (2014) ont conduit à des pourcentages de valeurs positives faibles, dans un cas de figure où le nombre de sites de présence simulées était de 2 000 sites et le biais d'échantillonnage très significatif. Dans la simulation effectuée dans ce chapitre, les valeurs négatives peuvent être liées au pro-

cessus de simulation des ensembles de sites de présence biaisés (*pres_B*) et non-biaisées (*pres_NB*). En effet, ces ensembles ont en commun un nombre significatif de sites issues de leur tirage parmi *pres_NB_init* (de cardinal n). La proportion de *pres_NB_init* sélectionnée pour construire *pres_NB* et *pres_B* étant à 70%, au moins 40% des sites de *pres_NB_init* sont communs aux deux jeux de données sélectionnées, soit près de 60% de ces derniers (pour $k = 20$, au moins 11 sites sont communs à *pres_NB* et à *pres_B*). Ce pourcentage de sites en commun est possiblement trop élevé, impliquant une distribution des sites de présence considérée comme biaisée trop proche de celle considérée comme non-biaisée et minimisant les effets des méthodes de correction. La réduction du nombre de sites de présence en commun est envisageable (à 40 % par exemple), afin d'augmenter le biais d'échantillonnage et permettre aux méthodes de correction de fournir des résultats plus significatifs.

Les résultats de classement des modèles corrigés avec différentes méthodes montrent une variation des indices d'évaluation relatives selon le nombre de sites. Les méthodes de construction de background biaisé *BGenv* et *BGenv_tg* permettent aux modèles de distribution d'être bien classés lorsque le nombre de sites de présence est de 20 et de 50, et ce pour l'ensemble des indicateurs. Lorsque le nombre de sites de présence augmente, les différents indices d'évaluation relative ne montrent pas une tendance identique. ΔAUC et $\Delta Denv$ montrent que la méthode *BGgeo* permet d'obtenir des modèles mieux classés, alors que $\Delta Dgeo$ montre que les modèles corrigés avec *Fgeo* et *BGenv* permettent de meilleurs résultats.
Cette différence entre les indices d'évaluation relative est possiblement liée à une composante aléatoire trop importante dans le processus de simulation, résultant notamment du faible de nombre de sites de présence. En effet, pour le calcul de l'*AUC*, la carte de présence-absence utilisée comme carte de référence, PA_R_{ec}, est issue de d'une conversion logistique de la carte de qualité d'habitat issue du modèle non-biaisé, avec une prévalence de 0,2. Le calcul de $Denv$ est réalisé en sélectionnant 500 sites de manière aléatoire et pondérée par la carte de qualité d'habitat. Malgré la pondération par la carte de qualité d'habitat, l'effet aléatoire de la sélection est important. En revanche, $Dgeo$ est calculé à partir d'une distribution de probabilité sur l'ensemble de la zone d'étude. La différence de sélection des données pour l'évaluation peut être à l'origine de cette différence. Les travaux de Fourcade et al. (2014) montrent également cette différence de résultats. Ceci rend difficile la conclusion sur la meilleure méthode de correction quelque soit le nombre de site de présence. Cependant, les résultats montrent que la méthode de correction développée (*BGenv* et *BGenv_tg*) permet de meilleurs classements lorsque le nombre de sites est égale ou inférieur à 50.

VI. 3. *BGeng vs. BGenv_tg*

Dans ce chapitre, deux cas de figures ont été simulés pour la méthode de correction basée sur la construction d'un background biaisé à partir de critères environnementaux : d'une part, le cas où l'absence de donnée de présence d'autres espèces ne permet pas de définir un groupe cible (*BGenv*) – dans un tel cas, le biais d'échantillonnage est estimé uniquement à partir des sites de présence virtuels d'*An. darlingi* ; et d'autre part, le cas où il est possible de définir un groupe cible (*BGenv_tg*).
Bien que les modèles corrigés avec *BGenv_tg* ne présentent pas toujours le meilleur classement par rapport à *BGenv*, l'utilisation de données de présence d'autres espèces pour former un groupe cible (dans *BGenv_tg*) présente un avantage. Les données de présence d'autres

espèces utilisées pour le groupe cible sont équivalentes à des données de pseudo-absence de l'espèce cible. Phillips et al. (2009) notent également que les données associées aux groupes cibles, qui ne correspondent pas à des sites de présence de l'espèce d'intérêt, pourraient être utilisées comme données d'absence. Cependant, Hirzel et al. (2002) soulignent qu'une donnée d'absence peut correspondre à une fausse absence, c'est-à-dire que l'absence est notifiée alors que l'espèce est réellement présente, dans un habitat qui lui est favorable, mais que pour diverses raisons elle n'a pas été capturée et notent que ces données peuvent considérablement biaiser la modélisation. Ainsi, nous nous positionnons sur le fait que les données de présence d'autres espèces constituent un apport important à la construction du background biaisé mais ne peuvent pas être considéré comme des données d'absence.

VI. 4. Paramétrisation de *BGenv* et BGenv_tg

Dans ce travail, la paramétrisation de *BGenv* et *BGenv_tg* est basée sur la définition de $Dmin$ à partir de la connaissance *a priori* qu'*An. darlingi* n'est pas présent en zone densément urbanisée, tandis que *Fgeo* et *BGgeo* sont basée sur r et d qui sont définies à partir de connaissance des distances de déplacement d'*An. darlingi*. L'évaluation absolue note un détachement de *BGgeo* par rapport aux autres. Ceci est probablement lié à la définition de d qui a été fixé à 7km correspondant à la distance maximale de déplacement d'*An. darlingi* relevée dans la littérature. En effet, Barve et al. (2011) ont montré que plus la surface dans laquelle la sélection de background est effectuée augmente, plus les modèles permettaient des AUC significativement meilleurs qu'un modèle aléatoire. Cependant, la définition de d ne semble pas aisé car en région amazonienne, le rayon de déplacement d'*An. darlingi* est très variable (entre 0m et 7km) selon les sites d'études. Ainsi, la définition de $Dmin$ dans l'espace environnemental semble plus robuste que celle des paramètres dépendant de distance géographique lié au déplacement de l'espèce, d'où l'intérêt dans ce cas d'étude, de la définir comme étant la distance euclidienne, dans l'espace environnemental, entre les pixels de présence et les pixels considérés comme fortement urbanisées pour lesquelles *An. darlingi* n'est pas présente. Dans cette thèse, une étude détaillée du comportement de la correction par rapport à ce paramètre $Dmin$ a été initiée. Elle devra être poursuivie dans les travaux futurs.

VII. Conclusion

L'objectif de ce chapitre était de comparer la méthode de correction de l'effet du biais d'échantillonnage proposée dans cette thèse aux méthodes existantes dans un cas de figure où le nombre de site de présence est faible. La méthode proposée, a été développée pour être adaptée à un faible nombre de données de présence, a été appliquée avec succès à des données réelles d'*An. darlingi* peu nombreuses et fortement biaisées (chap. 3).
Dans ce chapitre, des jeux de données de présence d'*An. darlingi* ont été simulés afin de comparer différentes méthodes de correction de l'effet du biais d'échantillonnage. La simulation a consisté à définir des sites de présence non-biaisés qui ont été utilisés pour la construction de modèles de distribution non-biaisés, dont les cartes de qualité d'habitat ont été considérées comme cartes de référence. Des données biaisées ont également été simulées et ont été utilisées pour construire les modèles biaisées. Ces données biaisées ont été corrigées avec différentes méthodes afin de construire les modèles corrigées.

Les sorties de ces modèles ont été évaluées de manière absolue en calculant l'*AUC*, les indices de Schoener $Dgeo$ et $Denv$, puis de manière relative en calculant les indices d'évaluation relative ΔAUC, $\Delta Dgeo$ et $\Delta Denv$. Bien que les résultats de cette étude ne permettent pas de conclure sur la meilleure méthode de correction de l'effet du biais d'échantillonnage quelque soit le nombre de sites de présence, ils permettent de montrer que la méthode de correction développée dans cette thèse (décrite dans le chapitre 2) est effectivement adaptée à des cas d'études où le nombre de sites de présence est faible – avec un classement de *BGeng_tg* à 1, 1 et 4 selon ΔAUC, $\Delta Dgeo$ et $\Delta Denv$, respectivement, et un classement de *BGenv* à 3, 3 et 2 et une proportion de valeurs positives la plus élevée lorsque le nombre de sites de présence est de 20.
Cette méthode de correction a été appliqué au modèle de présence-background Maxent, mais elle peut être utiliser pour corriger l'effet du biais d'échantillonnage dans d'autres modèles de présence-background tel l'ENFA.
Enfin, ce chapitre propose des pistes de recherche afin :

— de développer des méthodes alternatives de corrections du biais d'échantillonnage ;
— d'obtenir des résultats plus significatifs et ainsi faciliter le choix de la ou des méthodes de correction en fonction du contexte d'étude.

		Entrées des modèles		Sorties des modèles		Métriques d'évaluation	
Modèles	**Correction appliquée**	**Données de présence en entrée du modèle**	**Données de background en entrée du modèle**	**Carte de qualité d'habitat résultante**	**Carte de présence-absence résultante**	**Métriques d'évaluation absolue des résultats de prédiction**	**Métriques pour l'évaluation relative**
Modèle de référence	–	*pres_NB*	*bg_rd*	*HS_R*$_{EC}$	*PA_R*$_{EC}$	–	–
Modèle biaisé	–	*pres_B*	*bg_rd*	*HS_B*	–	AUC_B ; $Dgeo_B$; $Dgeo_B$	–
Modèle biaisé et corrigé	*C*$_x$*	*pres_C*$_x$*	*bg_rd* ou *bgenv* ou *bgenv_tg*	*HS_C*$_x$*	–	AUC_{Cx}* ; $Dgeo_{Cx}$* ; $Denv_{Cx}$*	ΔAUC_{Cx}* ; $\Delta Dgeo_{Cx}$* ; $\Delta Denv_{Cx}$*

* *Cx* représente les différentes méthodes de correction : *Fgeo*, *Fenv*, *BGgeo*, *BGenv* ou *BGenv_tg*

TABLEAU 4.1 – **Tableau récapitulatif des modèles, de leurs données d'entrée et de sortie, ainsi que des métriques d'évaluation absolue et relative.**

	Nombre de sites de présence k	*Fgeo*	*Fenv*	*BGgeo*	*BGenv*	*BGenv_tg*
ΔAUC	20	10$^{\star(71)}$	38	40	35	**43**
	50	45$^{\star(11)}$	52	**71**	43	54
	100	70	65	**91**	38	52
	150	77	68	**94**	46	51
	200	86	52	**98**	38	45
$\Delta Dgeo$	20	13$^{\star(71)}$	23	35	45	**53**
	50	31$^{\star(11)}$	27	28	**46**	41
	100	**59**	29	31	47	53
	150	**77**	35	30	57	51
	200	**78**	37	41	53	51
$\Delta Denv$	20	53	61	52	**68**	65
	50	58	66	58	**69**	60
	100	67	**82**	71	67	67
	150	69	**84**	70	67	59
	200	66	**79**	72	65	54

TABLEAU 4.2 – **Pourcentages de valeurs de ΔAUC, $\Delta Dgeo$ et de $\Delta Denv$ strictement positives.** Une valeur positive signifie que la méthode de correction permet d'améliorer le modèle biaisé. Les maxima des pourcentages pour chaque valeur de k sont représentés en gras.
* Dans ces cas, le pourcentage de valeurs nulles, signifiant que le modèle corrigé est strictement équivalent au modèle biaisé, est mis entre parenthèse.

Chapitre 5

Discussion générale et perspectives

I. Choix du modèle de distribution d'espèce 111
II. Modélisation de la qualité d'habitat d'*An. darlingi* en Guyane 111
II. 1. Précision / incertitude des données de présence 111
II. 2. Résolution des variables environnementales 112
II. 3. Variables météorologiques 112
III. Contributions à la lutte contre le paludisme 114
III. 1. Qualité d'habitat *vs.* risque de transmission 114
III. 2. Qualité d'habitat pour la lutte antivectorielle 115

Introduction

Ce chapitre présente une discussion relative aux principaux choix méthodologiques effectués dans cette thèse ; à leurs modalités d'application à la modélisation des habitats du principal vecteur du paludisme en Guyane et en Amazonie, ainsi qu'à l'impact potentiel des résultats obtenus en termes de lutte antivectorielle et plus généralement de santé publique. Il ne revient pas sur les discussions menées dans les chapitres précédents, il les complète et permet de dégager les perspectives de recherche les plus importantes.

I. Choix du modèle de distribution d'espèce

Dans ce travail, le modèle Maxent a été choisi. Plusieurs auteurs ont montré sa faible sensibilité au nombre de sites de présence et sa meilleure performance de prédiction lorsque ce nombre est faible, par rapport aux modèles existants lui étant comparables. Cependant, il n'existe pas de modèle idéal adapté à toutes situations, et d'autres modèles auraient pu être étudiés et appliqués dans cette thèse à la fois pour modéliser la distribution d'*An. darlingi* et pour étudier l'impact du biais d'échantillonnage et de sa correction.
Pour déterminer le modèles le plus adapté, une approche consiste à générer plusieurs modèles et à ne conserver que ceux qui permettent les meilleurs résultats. Cette approche a, par exemple, été utilisée par Le Roux et al. (2016) pour étudier la répartition d'espèces, à partir de la plateforme *Biomod 2*. Une telle approche pourrait être utilisée pour la spatialisation de la qualité d'habitat d'*An. darlingi*, afin de définir le modèle le plus adapté de manière exploratoire sans faire un choix basé sur les connaissances des modèles.
Dans cette thèse, la méthode de correction a été développée pour une application à Maxent et a été également évaluée en utilisant ce même modèle. Cette méthode pourrait s'appliquer indifféremment à d'autres modèles de présence-background tel que l'ENFA. Dans l'avenir, il serait intéressant et utile d'étudier l'impact de la correction sur une telle méthode qui est également fortement impactée par le biais d'échantillonnage.

II. Modélisation de la qualité d'habitat d'*An. darlingi* en Guyane

II. 1. Précision / incertitude des données de présence

En Guyane, des données de captures d'*Anopheles* existent depuis 1902. Dans ce travail de thèse, seules les données ayant une géolocalisation précise ont été utilisées. Ces données ne sont disponibles qu'à partir de 2000. Le choix de ne considérer que les localités précises s'est avéré très restrictif. En effet, les données de présence d'*An. darlingi* n'ayant pas de géolocalisation précise entre 2000 et 2013 représentent près de 90% des données recueillies auprès de l'Institut Pasteur de la Guyane, du Service de Santé des Armées et de la Direction de la Démoustication de la Collectivité Territoriale. Ces données sont décrites par des noms de localité, de route ou de rue, de zone ou de cours d'eau dont les localisations sont imprécises ou incertaines et ne peuvent pas être utilisées telles quelles pour la modélisation. Des travaux réalisés par Graham et al. (2008) ont évalué différents modèles en se basant sur des données précises (données originales) auxquelles des erreurs liées aux positions géographiques des sites de capture ont été ajoutées. Ces erreurs étaient tirées aléatoirement selon une distribution de probabilité normale de moyenne nulle et d'écart-type de 5 km (correspondant à l'erreur spatiale supposée des données issues des musées). Avec des données environnementales

de 100 mètres de résolution, les auteurs estiment que le modèle Maxent était peu impacté par ces erreurs spatiales. Cependant, l'association d'une forte hétérogénéité spatiale de l'environnement dans la zone de capture avec des informations de localisation des sites de présence imprécises ou incertaines, peut fortement dégrader la prédiction. Les travaux de Naimi et al. (2011) ont montré que les performances des modèles étaient stables (proches des modèles construits avec les données ne présentant pas d'erreur de localisation) lorsque la distance d'autocorrélation des variables environnementales était supérieure à trois fois l'écart-type de l'erreur de localisation.
Pour pouvoir utiliser les données de capture d'*An. darlingi* imprécises en Guyane, il serait donc nécessaire, dans un premier temps, d'estimer l'erreur de localisation. Ainsi, une capture réalisée dans un village présentera une erreur de localisation moindre qu'une capture réalisée dans une ville de plus grande taille ou le long des routes nationales de Guyane. Dans un deuxième temps, l'étude de l'hétérogénéité spatiale de l'environnement de la zone de capture, au travers, notamment, de l'analyse des variogrammes des variables environnementales, permettrait d'intégrer à la modélisation les données imprécises satisfaisant les conditions définies par Naimi et al. (2011).

II. 2. Résolution des variables environnementales

Les captures d'*Anopheles* réalisées en Guyane sont souvent faites à une échelle très locale (à l'échelle d'un village, d'un quartier). En revanche, dans cette thèse, la résolution spatiale utilisées est de 1 000 m. À cette résolution, l'hétérogénéité spatiale de l'environnement à l'échelle de la zone d'échantillonnage ne peut être caractérisés de manière totalement satisfaisante. Une étude devrait donc être menée sur l'apport de la haute résolution spatiale (30 mètres voire 10 mètres) sur la modélisation d'*An. darlingi* en Guyane. Ces données permettraient-elles de cartographier des micro-habitats des espèces cibles et quelle serait la conséquence sur les cartes de prédiction ? Phillips et al. (2009) soulignent que le choix de la résolution spatiale des variables environnementales affecte le degré de généralisation des modèles. Les travaux de Guisan et al. (2007) ont comparé différents modèles de distribution d'espèces en utilisant des données environnementales de résolutions spatiales différentes. Ils ont montré que les arbres de régression (*boosted regression tree, BRT*) et Maxent permettaient d'obtenir les meilleurs AUC parmi l'ensemble des modèles évalués, lorsqu'une résolution spatiale grossière était utilisée. Cependant, seul le modèle BRT restait insensible à l'amélioration de la résolution spatiale, alors que les performances de Maxent se sont dégradées. Ainsi, comme suggéré dans la partie I. , une modélisation de la distribution d'*An. darlingi* avec d'autres modèles peut être envisagée, notamment avec la méthode de BRT lorsqu'un passage à une résolution spatiale plus fine sera envisageable.

II. 3. Variables météorologiques

La modélisation de la qualité d'habitat d'*An. darlingi* a été réalisée à partir des variables suivantes :

- les paysages et les unités géomorphologiques (*GLS* et *GLF*) ;
- l'occupation du sol modifiée (*LS*) ;
- l'altitude minimale (*ATL_min*) ;

— la présence et les activités humaines altérant de manière non-permanente l'environnement naturel (*HA_max*) ;
— la longueur de route dans un pixel de 1 km^2 (*ROADS*) ;
— le pourcentage d'urbanisation dans les pixels voisins (*PER_URB_NEIGH*).

Bien que le choix de ne pas avoir utiliser les variables météorologiques ait été discuté dans le chapitre 3, une réflexion sur l'exploitation des séries temporelles de données météorologiques peut être menée. Entre 2000 et 2010, seuls 35 de postes de mesure de la pluviométrie au sol étaient disponibles [1]. De plus, la majeure partie des postes étaient situés le long du littoral. Ringard et al. (2015) ont évalué la qualité de quatre produits satellitaires en les comparant aux données *in situ* sur le plateau des Guyanes, afin de pallier au manque de données *in situ* dans certaines zones. Une représentation de la pluviométrie journalière moyenne issue de produits satellites et d'une interpolation de données *in situ* est en figure 5.1. Bien que des erreurs d'estimation de la pluviométrie en fonction du régime hydro-climatique aient été notées lors de cette comparaison, de telles données satellites seraient intéressantes pour cartographier les différents régimes de pluviométrie en Guyane. En particulier, l'estimation mensuelle du nombre de jours consécutifs pour lesquels la quantité de précipitation dépasse certaine valeur peut constituer une information importante, car une pluviométrie trop conséquente dans une zone peut lessiver les gîtes larvaires et constituer un frein à la présence d'*An. darlingi*.

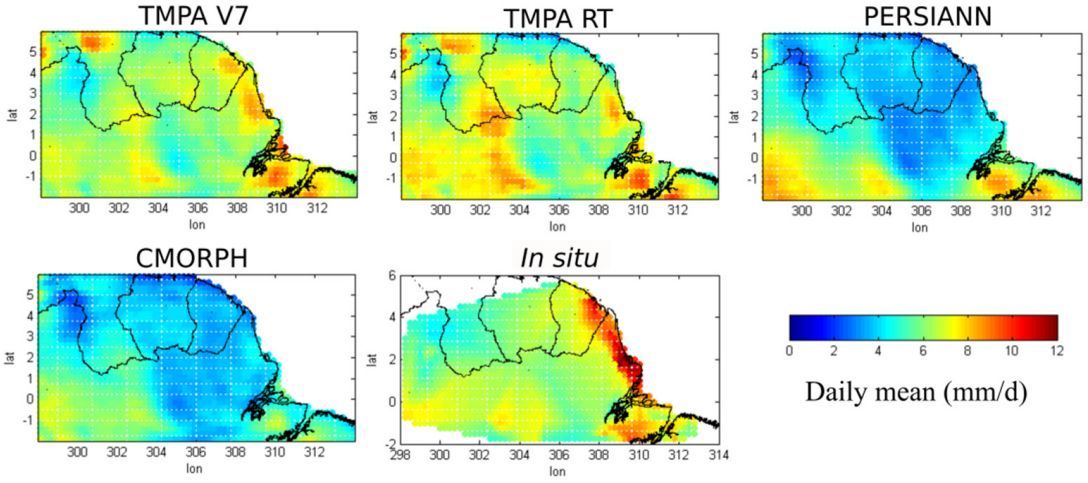


FIGURE 5.1 – **Spatialisation de la pluviométrie journalière moyenne de 2001 à 2012 à partir des produits satellitaires TMPA V7, TMPA RT, PERSIANN et CMORPH et des données *in situ*.**
Source : Ringard et al. (2015)

An. darlingi nécessite des gîtes larvaires ensoleillés mais présentant suffisamment d'ombre pour maintenir la température de l'eau dans un intervalle de valeurs favorables. Les travaux de Albarelo et al. (2015) ont permis d'estimer l'irradiation solaire en Guyane à partir d'images satellite. La carte de la distribution spatiale de l'irradiation solaire annuelle en Guyane (figure 5.2) permet d'identifier différentes zones plus ou moins ensoleillées et pouvant éventuellement correspondre à des qualités d'habitat différents. Une telle carte pourrait donc être utilisée pour la modélisation de la qualité d'habitat d'*An. darlingi*.

1. http://pluiesextremes.meteo.fr/guyane/IMG/sipex_pdf/carte_reseau_dep973.pdf

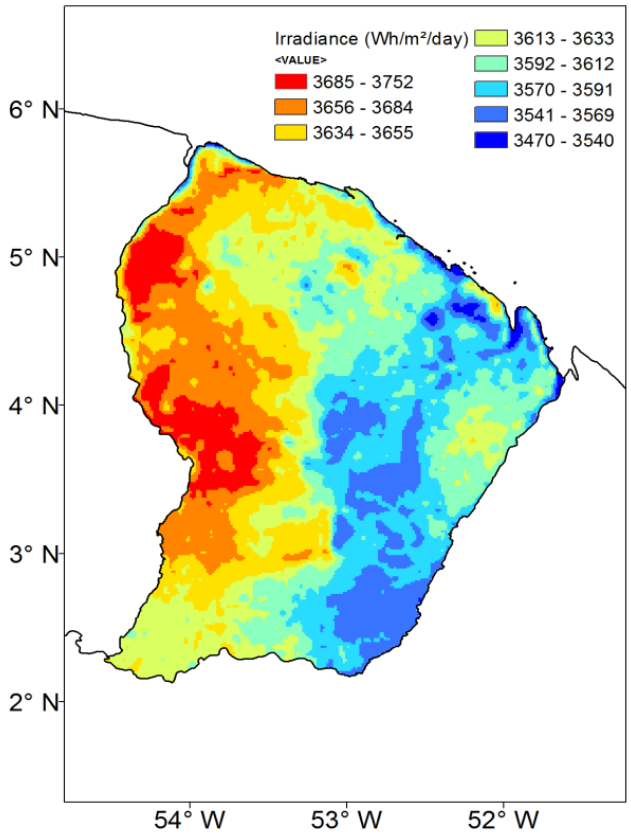


FIGURE 5.2 – **Carte de l'irradiation solaire global annuelle pour 2012.**
Source : Albarelo et al. (2015)

III. Contributions à la lutte contre le paludisme

III. 1. Qualité d'habitat *vs.* risque de transmission

Dans le chapitre 3, une carte de qualité d'habitat du principal vecteur du paludisme en Guyane, *An. darlingi*, a été réalisée. Les travaux de Girod et al. (2011), Pommier de Santi et al. (2016b) et Dusfour et al. (2012b) ont montré la participation potentiellement significative d'autres espèces d'*Anopheles* à la transmission de la maladie, dans certaines zones de la Guyane. La modélisation de la distribution spatiale de ces vecteurs secondaires pourra être réalisée afin d'avoir une cartographie complète des vecteurs potentiels de cette maladie. Cependant, une réflexion, similaire à celle menée dans le cas d'*An. darlingi*, devra être menée afin de choisir les variables environnementales les plus appropriées à la caractérisation et à la modélisation des habitats de ces vecteurs.
Bien que la transmission du paludisme soit directement liée à la présence de ces vecteurs, les seules distributions spatiales de ces derniers ne suffisent pas à déterminer le risque de transmission du paludisme. Alimi et al. (2015) notent que la notion du risque des maladies à transmission vectorielle est complexe et diffère de celle considérée en santé public et définie par "la probabilité que la maladie se développe chez un individu dans un intervalle de temps spécifique". Ils notent que l'estimation du risque ne dépend pas uniquement de la présence ou de la densité des vecteurs et des parasites, mais également des investissements publics consacrés au contrôle de la maladie, de la mise en œuvre de la lutte antivectorielle et bien sûr

des caractéristiques des populations humaines exposées. Des cartes de risque, produites par l'Organisation panaméricaine de la santé (OPS, *Pan American Health Organization (PAHO)*), sont disponibles à l'échelle de l'Amérique latine. Cependant, da Silva-Nunes et al. (2012) estiment que ces cartes sont trop généralisées et aggrégées à des échelles trop grossières pour être appliquées à des cas concrets.
Bien que la cartographie d'un risque de transmission du paludisme ne soit pas envisageable à court terme à l'échelle de la Guyane, l'estimation du risque d'exposition vectorielle peut, dans un premier temps, être envisagée. En effet, les travaux de Li et al. (2016) ont permis de définir un indicateur reflétant le "danger" que peuvent représenter certains paysages, de par leur capacité à faciliter les rencontres hommes-vecteurs, en fonction du degré d'interaction entre milieux forestiers et non-forestiers. La combinaison de cet indicateur et de la carte de qualité d'habitat des vecteurs du paludisme en Guyane pourrait fournir une carte de risque individuel d'exposition vectorielle, correspondant au risque qu'un humain soit exposé à une piqûre d'*Anopheles*, en tout point de l'espace. Dans un deuxième temps, la combinaison de ce risque individuel avec une carte d'empreinte humaine, identifiant les lieux d'habitation et d'activité humaines (de Thoisy et al., 2010), permettrait de définir une carte de risque d'exposition vectorielle au niveau populationnel. Cette carte permettrait d'identifier les lieux où la population guyanaise est la plus exposée aux piqûre d'*Anopheles*. Toutefois, la méthode de combinaison de ces couches d'information, ainsi que l'évaluation des résultats, restent des questions de recherche ouvertes.

III. 2. Qualité d'habitat pour la lutte antivectorielle

Ce travail de thèse apporte des éléments de connaissance complémentaires en ce qui concerne la distribution des vecteurs du paludisme, pouvant aider à guider la lutte antivectorielle. Mais pour qu'une lutte antivectorielle soit optimale, il est nécessaire de connaître le lieu et le moment de la lutte. Cependant, la carte de qualité d'habitat ne renseigne pas sur les moments propice à lutte. Les travaux réalisés par Adde et al. (2016) ont permis de construire un modèle spatio-temporel définissant les lieux et les moments où la population anophélienne est la plus dense. Pour cela, ils se basent sur des données de densité d'espèces, des variables météorologiques et des indices paysagers, à une échelle très locale de la ville de Saint-Georges située à la frontière Guyane-Brésil. Compte tenu des données utilisées, il n'est pas envisageable actuellement de construire un tel modèle à l'échelle de la Guyane. Bien que la carte de qualité d'habitat d'*An. darlingi* obtenue dans cette thèse ne permette pas de planifier la lutte anti-vectorielle, elle permet d'avoir une vision globale de la distribution du principal vecteur en Guyane et pourrait être utilisé comme point de départ pour définir les prochaines zones de modélisation spatio-temporelle locale. Ces prochaines zones d'études à fine échelle pourrait correspondre à des zones où l'indice de qualité d'habitat est apparu très élevé.
Il est important de noter que le comportement des populations locales influencent fortement la réussite de la lutte antivectorielle. En effet, certaines populations mènent une vie traditionnelle, vivant dans des carbets, des constructions typiques amérindiennes, correspondant à des abris en bois dépourvus de murs. Ces populations vivent de l'agriculture, de la pêche et de la chasse, les amenant à rester à l'extérieur et à proximité de la forêt. Or, une des méthodes de lutte contre les vecteurs du paludisme principalement utilisées en Guyane est la pulvérisation intra-domicilaire (Fontenille et al., 2009), qui ne semble pas pertinente dans ces cas de figure. Ainsi, dans ces zones, bien qu'une cartographie fine puisse être réalisée, le

comportement et le style de vie de la population sont des facteurs importants à prendre en compte dans la planification de la lutte antivectorielle.

Chapitre 6

Conclusion générale

Cette thèse s'est intéressée à la modélisation de la distribution d'espèces, dans le cas particulier, mais fréquent, où le nombre de données de présence de l'espèce considérée est très faible et où les collectes sont entachées d'un biais d'échantillonnage significatif. Bien que des méthodes de correction de l'effet du biais d'échantillonnage aient été développées ces dernières années, soit ces méthodes sont qu' adaptées à un grand nombre de sites de présence ; soit elles ne sont pas assez génériques pour être appliquées quels que soient le nombre et le type (continue ou catégoriel) des variables environnementales.
Les objectifs de cette thèse étaient premièrement, de proposer une méthode de correction de l'effet du biais d'échantillonnage originale et générique pouvant être appliquée à un nombre de sites de présence faible et quelles que soient les variables environnementales d'entrée et, deuxièmement, de modéliser la distribution spatiales du principal vecteur du paludisme en Guyane, *Anopheles darlingi*, dont les données de présence sont en faible quantité et fortement biaisées.

Premièrement, un inventaire des modèles de distribution d'espèces a été réalisé. Maxent a été choisi parmi les modèles existants car il ne requiert que des données de présence, pour ses performances par rapport aux autres approches de modélisation lorsque le nombre de sites de présence est faible, et pour sa facilité d'utilisation. Cependant, comme pour tous les modèles ne nécessitant que les données de présence, il est très sensible au biais d'échantillonnage. L'inventaire des méthodes existantes de correction de l'effet du biais d'échantillonnage pour Maxent a été réalisé. Ceci a permis de relever leurs lacunes méthodologiques et d'applicabilité, notamment par rapport au nombre de sites de données de présence requis.

Dans un deuxième temps, une méthode originale de correction du biais d'échantillonnage a été développée pour combler les lacunes relevées lors de l'inventaire. La méthode de correction proposée est générique. Elle est, théoriquement, capable de corriger l'effet du biais d'échantillonnage environnemental et d'être utilisée lorsque le nombre de site de présence est faible. Elle est basée sur une Analyse Factorielle de Données Mixtes permettant de définir l'espace environnemental et la notion de voisinage environnemental requis pour l'estimation de l'effort d'échantillonnage et la sélection des sites de background selon le même biais que celui entachant les données de présence. L'AFDM rend la méthode générique et applicable à n'importe quel type de variables environnementales. La sélection du background biaisé permet de corriger le biais y compris lorsque le nombre de sites de présence est faible.

Dans un troisième temps, cette méthode a été appliquée à des données réelles d'*An. darlingi*, issues de captures fortement biaisées et en nombre faible (39 sites de présence). Tout d'abord, la zone d'étude (Guyane française) ainsi que les différents aspects liés au paludisme dans cette région (situation épidémiologique, vecteurs et leur bioécologie, actions de contrôle, etc.) ont été présentés en détails. Ensuite, la construction du modèle de distribution d'*An. darlingi* s'est faite en plusieurs étapes. Une première étape a consisté en une étude bibliographique afin de répertorier les variables environnementales ayant une influence sur la présence d'*An. darlingi* et étant disponibles sur l'ensemble de la Guyane. Ces variables ont une résolution allant de 30 m à 1 000 mètres. Pour éviter les incohérences spatiales, la résolution spatiale de l'ensemble de ces couches environnementales a été ramenée à 1 000 m. Dans une deuxième étape, la méthode de correction proposé au chapitre 2 a été appliquée et le mo-

dèle de distribution d'*An. darlingi* a été construit. Les résultats se sont avérés très satisfaisants et en cohérence avec la connaissance actuelle des entomologistes. Ainsi, le travail effectué dans cette partie a permis de modéliser la distribution d'*An. darlingi* à l'échelle de la Guyane française, malgré le fort biais d'échantillonnage et le faible nombre des sites de présence. Ce travail complète donc la connaissance sur la distribution de ce vecteur en Guyane française connaissance qui, jusqu'alors, restait partielle et n'avait pas donné lieu à une cartographie à de telles résolution et échelle.

Bien que l'application de la méthode de correction de l'effet du biais d'échantillonnage à des données réelles ait permis d'obtenir des résultats très satisfaisants, une étude a ensuite été réalisée afin de comparer cette méthode à d'autres méthodes de correction existantes, dans un contexte où le nombre de sites de présence est en faible quantité. Pour cela, une simulation de données de présence d'*An .darlingi*, s'appuyant sur les résultats obtenus au chapitre 3, a été réalisée. Les résultats de ce travail montrent que la méthode proposée permet aux modèles d'obtenir de meilleures performances lorsque le nombre de sites de présence est de 20 et 50 sites. Cependant, lorsque ce nombre augmente, d'autres méthodes fournissent de meilleures performances.

Enfin, une discussion générale a permis de prendre du recul par rapport aux choix méthodologiques, à leur modalités d'application à la spatialisation de la qualité d'habit d'*An. darlingi*, ainsi qu'aux impacts potentiels de ces travaux de thèse dans le cadre de la lutte contre le paludisme en Guyane et plus généralement en Amazonie. Des perspectives de recherche permettant de dépasser les limites de ces travaux ont également été proposées.

Ainsi, ce travail de thèse interdisciplinaire, et réalisé en partenariat avec les experts en entomologie médicale en Guyane française, a permis non-seulement de combler certaines lacunes méthodologiques et d'applicabilité des méthodes de correction de l'effet du biais d'échantillonnage mais également de spatialiser la qualité d'habitat du principal vecteur du paludisme en Amérique à l'échelle de Guyane française. Cette thèse a également fait émerger de nouvelles problématiques méthodologiques et des solutions possibles à exploiter.

Bibliographie

Adde, A., Roux, E., Mangeas, M., Dessay, N., Nacher, M., Dusfour, I., Girod, R., and Briolant, S. (2016). Dynamical Mapping of Anopheles darlingi Densities in a Residual Malaria Transmission Area of French Guiana by Using Remote Sensing and Meteorological Data. *PLOS ONE*, 11(10) :e0164685. *pp. 64 and 115*

Agence Régionale de Santé Guyane (2015). Plan de lutte contre le paludisme en Guyane 2015-2018. Technical report, ARS Guyane, Cayenne, French Guiana. *pp. 57 and 58*

Albarelo, T., Marie-Joseph, I., Primerose, A., and Linguet, L. (2015). Application of Kernel Density Estimation for Mapping of Solar Potential in French Guiana. In *Conference Proceedings Paper – Remote Sensing*, Online. Remote Sensing. *pp. xiv, 113, and 114*

Alimi, T. O., Fuller, D. O., Quinones, M. L., Xue, R.-D., Herrera, S. V., Arevalo-Herrera, M., Ulrich, J. N., Qualls, W. A., and Beier, J. C. (2015). Prospects and recommendations for risk mapping to improve strategies for effective malaria vector control interventions in Latin America. *Malaria Journal*, 14(1) :1–15. *pp. 4, 58, 85, and 114*

Alvarez-Berríos, N. L. and Aide, T. M. (2015). Corrigendum : Global demand for gold is another threat for tropical forests (2014 *Environ. Res. Lett.* **10** 014006). *Environmental Research Letters*, 10(2) :029501. *p. 83*

Anderson, R. P. and Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution : preliminary tests with montane rodents (genus Nephelomys) in Venezuela : Effect of study region on models of distributions. *Journal of Biogeography*, 37(7) :1378–1393. *p. 25*

Ardillon, V., Carvalho, L., Prince, C., Abboud, P., and Djossou, F. (2015). Bilans 2013 et 2014 de la situation épidémiologique du paludisme en Guyane. *Le bulletin de veille sanitaire*, 1. *p. 85*

Ardillon, V., Eltges, F., Chocho, A., Chantilly, S., Carvalho, L., Flamand, C., and Carme, B. (2012). Evolution de la situation épidémiologique du paludisme en Guyane de 2005 à 2011. *Le bulletin de veille sanitaire*, 1(1-2) :5–11. *p. 56*

Austin, M. P., Nicholls, A. O., and Margules, C. R. (1990). Measurement of the Realized Qualitative Niche : Environmental Niches of Five Eucalyptus Species. *Ecological Monographs*, 60(2) :161–177. *p. 10*

Barret, J. (2001). *Atlas illustré de la Guyane*. Laboratoire de cartographie de la Guyane : Institut d'Enseignement Supérieur de Guyane, French Guiana, ird ed edition. *pp. xiii, xiv, 50, 52, and 73*

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., and Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11) :1810–1819. *p. 105*

Berger, F., Flamand, C., Musset, L., Djossou, F., Rosine, J., Sanquer, M.-A., Dusfour, I., Legrand, E., Ardillon, V., Rabarison, P., Grenier, C., and Girod, R. (2012). Investigation of a Sudden Malaria Outbreak in the Isolated Amazonian Village of Saul, French Guiana, January-April 2009. *American Journal of Tropical Medicine and Hygiene*, 86(4) :591–597. *p. 58*

Boria, R. A., Olson, L. E., Goodman, S. M., and Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275 :73–77. *pp. 26 and 33*

Bourgarel, S. (1989). Migration sur le Maroni : les réfugiés surinamais en Guyane. *Revue européenne de migrations internationales*, 5(2) :145–153. *p. 51*

Branquinho, M. S., Araújo, M. S., Natal, D., Marrelli, M. T., Rocha, R. M., Taveira, F. A., and Kloetzel, J. K. (1996). Anopheles oswaldoi a potential malaria vector in Acre, Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 90(3) :233. *p. 61*

Breeveld, F. J., Vreden, S. G., and Grobusch, M. P. (2012). History of malaria research and its contribution to the malaria control success in Suriname : a review. *Malaria Journal*, 11(1) :95. *p. 57*

Broennimann, O., Fitzpatrick, M. C., Pearman, P. B., Petitpierre, B., Pellissier, L., Yoccoz, N. G., Thuiller, W., Fortin, M.-J., Randin, C., Zimmermann, N. E., Graham, C. H., and Guisan, A. (2012). Measuring ecological niche overlap from occurrence and spatial environmental data : Measuring niche overlap. *Global Ecology and Biogeography*, 21(4) :481–497. *p. 100*

Brotons, L., Thuiller, W., Araújo, M. B., and Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, 27(4) :437–448. *p. 16*

Carme, B., Lecat, J., and Lefebvre, P. (2005). Le paludisme dans le foyer de l'Oyapock (Guyane) : incidence des acces palustres chez les Amerindiens de Camopi. *Médecine tropicale*, 65(2) :149–154. *p. 85*

Carnevale, P. and Robert, V. (2009). *Les anophèles biologie, transmission du plasmodium et lutte antivectorielle*. IRD, Marseille. OCLC : 690760559. *pp. xiii, 46, and 47*

Carpenter, G., Gillison, A. N., and Winter, J. (1993). DOMAIN : a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity & Conservation*, 2(6) :667–680. *p. 15*

Carroll, C. (2010). Role of climatic niche models in focal-species-based conservation planning : Assessing potential effects of climate change on Northern Spotted Owl in the Pacific Northwest, USA. *Biological Conservation*, 143(6) :1432–1437. *p. 25*

Charlwood, J. D. (1980). Observations on The bionomics of Anopheles darlingi Root (Diptera : Culicidae) from Brazil. *Bulletin of Entomological Research*, 70(04) :685. *p. 59*

Charlwood, J. D. (1996). Biological variation in Anopheles darlingi Root. *Memórias do Instituto Oswaldo Cruz*, 91(4) :391–398. *p. 74*

Charlwood, J. D. and Alecrim, W. A. (1989). Capture-recapture studies with the South American malaria vector Anopheles darlingi, Root. *Annals of tropical medicine and parasitology*, 83(6) :569–576. *p. 97*

Chaud, P., Paquet, C., Huguet, P., and Cottrelle, B. (2006). Surveillance épidémiologique du paludisme en Guyane. *pp. xiv, 56, and 57*

Conn, J. E., Wilkerson, R. C., Segura, M. N. O., de Souza, R. T., Schlichting, C. D., Wirtz, R. A., and Póvoa, M. M. (2002). Emergence of a new neotropical malaria vector facilitated by human migration and changes in land use. *The American Journal of Tropical Medicine and Hygiene*, 66(1) :18–22. *p. 61*

da Silva-Nunes, M., Moreno, M., Conn, J. E., Gamboa, D., Abeles, S., Vinetz, J. M., and Ferreira, M. U. (2012). Amazonian malaria : Asymptomatic human reservoirs, diagnostic challenges, environmentally driven changes in mosquito vector populations, and the mandate for sustainable control strategies. *Acta Tropica*, 121(3) :281–291. *p. 115*

Davis, J. R., Hall, T., Chee, E. M., Majala, A., Minjas, J., and Shiff, C. J. (1995). Comparison of sampling anopheline mosquitoes by light-trap and human-bait collections indoors at Bagamoyo, Tanzania. *Medical and veterinary entomology*, 9(3) :249–255. *p. 62*

De Granville, J. (1990). *Les formations végétales primaires de la zone intérieure de Guyane*. ORSTOM. *p. 84*

de Thoisy, B., Richard-Hansen, C., Goguillon, B., Joubert, P., Obstancias, J., Winterton, P., and Brosse, S. (2010). Rapid evaluation of threats to biodiversity : human footprint score and large vertebrate species responses in French Guiana. *Biodiversity and Conservation*, 19(6) :1567–1584. *pp. 69, 71, and 115*

Deane, L. M., Causey, O. R., and Deane, M. P. (1948). Notas sobre a distribui\textbackslashccão ea biologia dos anofelinos das regiões nordestina e amazônica do Brasil. *Revista do Serviço Especial de Saúde Pública*, 1 :827–965. *p. 59*

Degallier, N., Le Pont, F., and Claustre, J. (1983). Description d'un piège à moustiques avec appat animal, utilisé en Guyane française. *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*, 11(2) :103–109. *p. 140*

Degallier, N., Pajot, F., Kramer, R., Claustre, J., Bellony, S., and Le Pont, F. (1978). Rythmes d'activité des Culicidés de la Guyane française *(Diptera*, Culicidae). *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*, 16(1) :73–84. *p. 140*

Douine, M., Musset, L., Corlin, F., Pelleau, S., Pasquier, J., Mutricy, L., Adenis, A., Djossou, F., Brousse, P., Perotti, F., Hiwat, H., Vreden, S., Demar, M., and Nacher, M. (2016). Prevalence of Plasmodium spp. in illegal gold miners in French Guiana in 2015 : a hidden but critical malaria reservoir. *Malaria Journal*, 15(1). *p. 55*

Dudik, M., Phillips, S. J., and Schapire, R. E. (2004). Performance guarantees for regularized maximum entropy density estimation. In *International Conference on Computational Learning Theory*, pages 472–486. Springer. *p. 20*

Dusfour, I., Carinci, R., Gaborit, P., Issaly, J., and Girod, R. (2010). Evaluation of Four Methods for Collecting Malaria Vectors in French Guiana. *Journal of Economic Entomology*, 103(3) :973–976. *p. 62*

Dusfour, I., Carinci, R., Issaly, J., Gaborit, P., and Girod, R. (2013). A survey of adult anophelines in French Guiana : enhanced descriptions of species distribution and biting responses. *Journal of Vector Ecology*, 38(2) :203–209. *pp. 84 and 86*

Dusfour, I., Issaly, J., Carinci, R., Gaborit, P., and Girod, R. (2012a). Incrimination of Anopheles (Anopheles) intermedius Peryassú, An.(Nyssorhynchus) nuneztovari Gabaldón, An.(Nys.) oswaldoi Peryassú as natural vectors of Plasmodium falciparum in French Guiana. *Memórias do Instituto Oswaldo Cruz*, 107(3) :429–432. *pp. 61 and 86*

Dusfour, I., Jarjaval, F., Gaborit, P., Mura, M., Girod, R., and Pagès, F. (2012b). Confirmation of the Occurrence of Anopheles (Nyssorhynchus) Marajoara in French Guiana. *Journal of the American Mosquito Control Association*, 28(4) :309–311. *pp. 61 and 114*

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K. S., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S., and Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29 :129–151. *pp. 17 and 24*

Elith, J., Kearney, M., and Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4) :330–342. *pp. 27, 33, and 38*

Elith, J. and Leathwick, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, 13(3) :265–275. *p. 11*

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists : Statistical explanation of MaxEnt. *Diversity and Distributions*, 17(1) :43–57. *pp. 18, 22, 24, and 76*

Elton, C. S. (1927). *Animal ecology*. University of Chicago Press, Chicago. *p. 9*

Engler, R., Guisan, A., and Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41(2) :263–274. *pp. 10 and 25*

Esterre, P., Cordoliani, G., Germanetto, P., and Robin, Y. (1990). Epid{\'e}miologie du paludisme en Guyane fran{\c{c}}aise. *Bulletin de la Société de pathologie exotique*, 83(2) :193–205. *p. 56*

Faran, M. E. and Linthicum, K. J. (1981). A handbook of the Amazonian species of Anopheles (Nyssorhynchus) (Diptera : Culicidae). *Mosquito Systematics*, 13 :1–81. *p. 63*

Fayad, I., Baghdadi, N., Bailly, J.-S., Barbier, N., Gond, V., Hérault, B., El Hajj, M., Fabre, F., and Perrin, J. (2016a). Regional Scale Rain-Forest Height Mapping Using Regression-Kriging of Spaceborne and Airborne LiDAR Data : Application on French Guiana. *Remote Sensing*, 8(3) :240. *p. 86*

Fayad, I., Baghdadi, N., Guitet, S., Bailly, J.-S., Hérault, B., Gond, V., El Hajj, M., and Minh, D. H. T. (2016b). Aboveground biomass mapping in French Guiana by combining remote sensing, forest inventories and environmental data. *International Journal of Applied Earth Observation and Geoinformation*, 52 :502–514. *p. 86*

Floch, H. (1954). La lutte antipaludique en Guyane française. *Bulletin of the World Health Organization*, 11(4-5) :579–633. *p. 56*

Floch, H. and Abonnec, E. (1942). Espèces de Moustiques signalées pour la première fois en Guyane française. *Institut Pasteur de la Guyane et du Territoire de l'Inini*, 41 :1–6. *p. 139*

Floch, H. and Abonnenc, E. (1951). Anophèles de la Guyane Française. *Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini*, 236 :1–92. *p. 63*

Fontenille, D., Lagneau, C., and Lecollinet, S. (2009). *La lutte antivectorielle en France*. IRD, Paris. *p. 115*

Forattini, O. (1962). *Entomologia médica : Vol. 1. Parte geral, Diptera, Anophelini*. Faculdade de Higiene e Saúde Pública, Departamento de Parasitologia, Universidade de São Paulo, São Paulo. *p. 63*

Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (2014). Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data : A Performance Assessment of Methods for Correcting Sampling Bias. *PLoS ONE*, 9(5) :e97122. *pp. 26, 27, 28, 29, 33, 34, 91, 97, 100, 101, 103, and 104*

Fraser, E. J., Lambin, X., Travis, J. M. J., Harrington, L. A., Palmer, S. C. F., Bocedi, G., and Macdonald, D. W. (2015). Range expansion of an invasive species through a heterogeneous landscape - the case of American mink in Scotland. *Diversity and Distributions*, 21(8) :888–900. *p. 10*

Gabaldon, A. (1983). Malaria eradication in Venezuela : doctrine, practice, and achievements after twenty years. *Am J Trop Med Hyg*, 32. *p. 61*

George, P. and Verger, F. (1972). *Dictionnaire de la géographie*. PUF, Paris. OCLC : 830087761. *p. 79*

Girod, R., Gaborit, P., Carinci, R., Issaly, J., and Fouque, F. (2008). Anopheles darlingi bionomics and transmission of Plasmodium falciparum, Plasmodium vivax and Plasmodium malariae in Amerindian villages of the Upper-Maroni Amazonian forest, French Guiana. *Memórias do Instituto Oswaldo Cruz*, 103(7) :702–710. *p. 73*

Girod, R., Roux, E., Berger, F., Stefani, A., Gaborit, P., and Carinci, R. (2011). Unravelling the relationships between Anopheles darlingi (Diptera : Culicidae) densities, environmental factors and malaria incidence : understanding the variable patterns of malarial transmission in French Guiana (South America). *Ann Trop Med Parasitol*, 105. *pp. 60, 73, 74, 85, and 114*

Gond, V., Freycon, V., Molino, J.-F., Brunaux, O., Ingrassia, F., Joubert, P., Pekel, J.-F., Prévost, M.-F., Thierron, V., Trombe, P.-J., and Sabatier, D. (2011). Broad-scale spatial pattern of forest landscape types in the Guiana Shield. *International Journal of Applied Earth Observation and Geoinformation*, 13(3) :357–367. *pp. 68, 71, and 84*

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A., and The Nceas Predicting Species Distributions Working Group (2008). The influence of spatial errors in species occurrence data used in distribution models : Spatial error in occurrence data for predictive modelling. *Journal of Applied Ecology*, 45(1) :239–247. *p. 111*

Grinnell, J. (1917). The niche-relationships of the california thrasher. *The Auk*, 34(4) :427–433. *p. 9*

Guisan, A., Edwards, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions : setting the scene. *Ecological modelling*, 157(2) :89–100. *p. 11*

Guisan, A., Graham, C. H., Elith, J., Huettmann, F., and the NCEAS Species Distribution Modelling Group (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, 13(3) :332–340. *p. 112*

Guisan, A. and Thuiller, W. (2005). Predicting species distribution : offering more than simple habitat models. *Ecology Letters*, 8(9) :993–1009. *pp. 9 and 10*

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M. R., Possingham, H. P., and Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12) :1424–1435. *pp. 9 and 135*

Guitet, S., Cornu, J.-F., Brunaux, O., Betbeder, J., Carozza, J.-M., and Richard-Hansen, C. (2013). Landform and landscape mapping, French Guiana (South America). *Journal of Maps*, 9(3) :325–335. *pp. 68 and 71*

Guo, Q., Kelly, M., and Graham, C. H. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling*, 182(1) :75–90. *p. 13*

Hall, L. S., Krausman, P. R., and Morrison, M. L. (1997). The habitat concept and a plea for standard terminology. *Wildlife Society Bulletin*, pages 173–182. *p. 9*

Hammond, D. S., Gond, V., de Thoisy, B., Forget, P.-M., and DeDijn, B. P. E. (2007). Causes and Consequences of a Tropical Forest Gold Rush in the Guiana Shield, South America. *AMBIO*, 36(8) :661–670. *p. 71*

Hamon, J. and Mouchet, J. (1961). Les vecteurs secondaires du paludisme humain en Afrique. *Medecine tropicale*, 21(spécial) :643–660. *p. 48*

Hamon, J., Mouchet, J., Brengues, J., and Chauvet, G. (1970). Problems facing anopheline vector control : vector ecology and behavior before, during, and after application of control of measures. *Entomol Soc Amer Misc Publ*, 7 :28–41. *p. 61*

Harbach, R. (2004). The classification of genus Anopheles (Diptera : Culicidae) : a working hypothesis of phylogenetic relationships. *Bulletin of Entomological Research*, 94(06). *pp. 46 and 48*

Heemskerk, M. and Duijves, C. (2012). Looking for gold finding malaria. Technical report, Social Solution, Suriname. *p. 57*

Hengl, T., Sierdsema, H., Radović, A., and Dilo, A. (2009). Spatial prediction of species' distributions from occurrence-only records : combining point pattern analysis, ENFA and regression-kriging. *Ecological Modelling*, 220(24) :3499–3511. *p. 26*

Héritier, P. (2011). *Le climat guyanais ; petit atlas climatique de la Guyane française*. Météo France, [Paris]. *p. 49*

Hernandez, P. A., Graham, C. H., Master, L. L., and Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5) :773–785. *pp. 17 and 24*

Hii, J., Smith, T., Mai, A., Ibam, E., and Alpers, M. (2000). Comparison between anopheline mosquitoes (Diptera : Culicidae) caught using different methods in a malaria endemic area of Papua New Guinea. *Bulletin of Entomological Research*, 90(03). *p. 62*

Hill, M. P. and Terblanche, J. S. (2014). Niche Overlap of Congeneric Invaders Supports a Single-Species Hypothesis and Provides Insight into Future Invasion Risk : Implications for Global Management of the Bactrocera dorsalis Complex. *PLoS ONE*, 9(2) :e90121. *pp. 27 and 33*

Hirzel, A. H., Hausser, J., Chessel, D., and Perrin, N. (2002). Ecological-niche factor analysis : how to compute habitat-suitability maps without absence data ? *Ecology*, 83(7) :2027–2036. *pp. 16 and 105*

Hirzel, A. H., Helfer, V., and Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological modelling*, 145(2) :111–121. *p. 16*

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., and Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199(2) :142–152. *p. 24*

Hiwat, H. and Bretas, G. (2011). Ecology of Anopheles darlingi Root with respect to vector importance : a review. *Parasit Vectors*, 4 :177. *p. 59*

Hiwat, H., Hardjopawiro, L. S., Takken, W., and Villegas, L. (2012). Novel strategies lead to pre-elimination of malaria in previously high-risk areas in Suriname, South America. *Malar J*, 11(10). *p. 57*

Hiwat, H., Issaly, J., Gaborit, P., Somai, A., Samjhawan, A., Sardjoe, P., Soekhoe, T., and Girod, R. (2010). Behavioral heterogeneity of Anopheles darlingi (Diptera : Culicidae) and malaria transmission dynamics along the Maroni River, Suriname, French Guiana. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 104(3) :207–213. *pp. 59 and 74*

Hoff, M., Cremers, G., Chevillotte, H., De Granville, J., Guérin, V., and Molino, J. (2007). Base de données botaniques Aublet2 de l'Herbier de Guyane française (CAY). *p. 66*

Hudson, J. (1984). Anopheles darlingi Root (Diptera : Culicidae) in the Suriname rain forest. *Bulletin of Entomological Research*, 74(1). *p. 59*

Hustache, S., Nacher, M., Djossou, F., and Carme, B. (2007). Malaria risk factors in amerindian children in French Guiana. *The American Society of Tropical Medicine and Hygiene*, 76(4) :619–625. *pp. 55, 58, and 85*

Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22(0) :415–427. *p. 9*

Institut National de la Statistique et des Etudes Economiques, editor (2006). *Atlas des populations immigrées en Guyane*. l'ACSÉ, Paris. OCLC : 255123870. *p. 51*

Juminer, B., Robin, Y., Pajot, F., and Eutrope, R. (1981). Physionomie du paludisme en Guyane Française. *Bull. Soc. Path. Ex*, 74(2) :176–192. *p. 140*

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J. W., Breitenmoser-Wuersten, C., Belant, J. L., Hofer, H., and Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 19(11) :1366–1379. *pp. 26, 27, 28, 29, 33, and 34*

Le Roux, M., Redon, M., Vincent, S., Tillon, L., Bouix, T., Archaux, F., and Luque, S. (2016). La modélisation spatiale des habitats et des corridors : un outil pour la conservation et la gestion des chauves-souris. *Symbiose*, (34) :28–34. *p. 111*

Léobal, C. (2013). *Saint-Laurent-du-Maroni : une porte sur le fleuve*. Espace outre-mer. Ibis Rouge Éditions, Matoury (Guyane). *p. 66*

Lepelletier, L., Gay, F., Nadire-Galliot, M., Poman, J. P., Bellony, S., Claustre, J., Traore, B. M., and Mouchet, J. (1989). Le paludisme en Guyane I. Situation général de l'endémie. *Bull. Soc. Path. Ex*, 82 :385–392. *p. 56*

Leroy, B., Meynard, C. N., Bellard, C., and Courchamp, F. (2015). virtualspecies, an R package to generate virtual species distributions. *Ecography*, pages n/a–n/a. *pp. xv and 157*

Li, Z., Roux, E., Dessay, N., Girod, R., Stefani, A., Nacher, M., Moiret, A., and Seyler, F. (2016). Mapping a Knowledge-Based Malaria Hazard Index Related to Landscape Using Remote Sensing : Application to the Cross-Border Area between French Guiana and Brazil. *Remote Sensing*, 8(4) :319. *pp. 85 and 115*

Machault, V., Vignolles, C., Borchi, F., Vounatsou, P., Briolant, S., Lacaux, J.-P., and Rogier, C. (2011). The use of remotely sensed environmental data in the study of malaria. *Geospatial Health*, 5(2) :151–168. *p. 58*

Maguire, K. C., Nieto-Lugilde, D., Fitzpatrick, M. C., Williams, J. W., and Blois, J. L. (2015). Modeling Species and Community Responses to Past, Present, and Future Episodes of Climatic and Ecological Change. *Annual Review of Ecology, Evolution, and Systematics*, 46(1) :343–368. *p. 10*

Mainali, K. P., Warren, D. L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., Karki, D., Shrestha, B. B., and Parmesan, C. (2015). Projecting future expansion of invasive species : comparing and improving methodologies for species distribution modeling. *Global Change Biology*, 21(12) :4464–4480. *p. 10*

Martens, W., Niessen, L., Rotmans, J., Jetten, T., and McMichael, A. (1995). Potential impact of global climate change on malaria risk. *Environmental Health Perspectives*, 103(5) :458–464. *p. 60*

Mathenge, E. M., Killeen, G. F., Oulo, D. O., Irungu, L. W., Ndegwa, P. N., and Knols, B. G. J. (2002). Development of an exposure-free bednet trap for sampling Afrotropical malaria vectors. *Medical and veterinary entomology*, 16(1) :67–74. *p. 62*

Merow, C., Smith, M. J., and Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions : what it does, and why inputs and settings matter. *Ecography*, 36(10) :1058–1069. *p. 22*

Metzger, M. J., Bunce, R. G. H., Jongman, R. H. G., Sayre, R., Trabucco, A., and Zomer, R. (2013). A high-resolution bioclimate map of the world : a unifying framework for global biodiversity research and monitoring : High-resolution bioclimate map of the world. *Global Ecology and Biogeography*, 22(5) :630–638. *p. 27*

Molez, J.-F. (1999). Les mythes représentant la transmission palustre chez les Indiens d'Amazonie et leurs rapports avec deux modes de transmissions rencontrés en forêt. *Cahier Santé*, 9 :157–162. *p. 141*

Moreno, J. E., Rubio-Palis, Y., Páez, E., Pérez, E., and Sánchez, V. (2007). Abundance, biting behaviour and parous rate of anopheline mosquito species in relation to malaria incidence in gold-mining areas of southern Venezuela. *Medical and veterinary entomology*, 21(4) :339–349. *p. 61*

Mouchet, J., editor (2004). *Biodiversité du paludisme dans le monde*. Libbey [u.a.], Montrouge. OCLC : 836234059. *p. 59*

Mouchet, J., Nadire-Galliot, M., GAY, P., and Poman, J. P. (1989). Le paludisme en Guyane II. Les caractéristiques des différents foyers et la lutte antipaludique. *Bull. Soc. Path. Ex*, 82 :393–405. *p. 140*

Musset, L., Pelleau, S., Girod, R., Ardillon, V., Carvalho, L., Dusfour, I., Gomes, M. S., Djossou, F., and Legrand, E. (2014). Malaria on the Guiana Shield : a review of the situation in French Guiana. *Memórias do Instituto Oswaldo Cruz*, 109(5) :525–533. *p. 57*

Naimi, B., Skidmore, A. K., Groen, T. A., and Hamm, N. A. S. (2011). Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling : Spatial autocorrelation and positional uncertainty. *Journal of Biogeography*, 38(8) :1497–1509. *p. 112*

Nations Unies (2015). Objectifs du Millénaire pour le développement Rapport 2015. *ONU,< http ://www. un. org/fr/millenniumgoals/>, consulté le*, 25. *p. 4*

Neveu-Lemaire, M. (1902a). Description de quelques moustiques de la Guyane. *Archives de Parasitologie*, 6(1) :5–9. *pp. xiv, 64, and 139*

Neveu-Lemaire, M. (1902b). Note additionnelle sur quelques moustiques de la Guyane. *Archives de Parasitologie*, 6(4) :613–615. *pp. 64 and 139*

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 2 :841–848. *p. 17*

Odetoyinbo, J. A. (1969). Preliminary investigation on the use of a light-trap for sampling malaria vectors in the Gambia. *Bulletin of the World Health Organization*, 40(4) :547. *p. 62*

Olson, S., Gangnon, R., Elguero, E., Durieux, L., Guégan, J., Foley, J., and Patz, J. (2009). Links between climate, malaria, and wetlands in the Amazon Basin. *Emerging infectious diseases*, 15(4) :659. *p. 73*

ONF, O. N. d. F. D. R. d. G. (2013). Projet « EXPERTISE LITTORAL 2011 » Occupation du sol et sa dynamique sur la bande côtière de la Guyane de 2005 à 2011. Technical report, ONF et le ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt. *pp. 72 and 84*

Pages, F., Orlandipradines, E., and Corbel, V. (2007). Vecteurs du paludisme : biologie, diversité, contrôle et protection individuelle. *Médecine et Maladies Infectieuses*, 37(3) :153–161. *p. 48*

Pagès, J. (2004). Analyse factorielle de donnees mixtes : principe et exemple d'application. *Montpellier SupAgro, http ://www. agro-montpellier. fr/sfds/CD/textes/pages1. pdf*. *p. 40*

Pagès, J. (2015). *Multiple factor analysis by example using R*. Chapman & Hall/CRC the R series. CRC Press, Taylor & Francis Group, Boca Raton. OCLC : ocn903630995. *p. 37*

Pages, J. and others (2004). Multiple factor analysis : Main features and application to sensory data. *Revista Colombiana de Estadística*, 27(1) :1–26. *p. 40*

Pajot, F., Le Pont, F., and Molez, J.-F. (1975). Données sur l'alimentation non sanguine chez *anopheles (Nyssorhynchus*) darlingi Root, 1926 (*(Diptera* Culicidae) en Guyane française. *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*, 13(3) :131–134. *p. 140*

Pajot, F., Le Pont, F., and Molez, J.-F. (1977a). Utilisation des pièges lumineux "C.D.C. miniature light trap" comme moyen d'échantillonnage des populations anophéliennes dans un village du littoral de la Guyane-française. *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*, 15(3) :233–240. *p. 140*

Pajot, F. X., Molez, J. F., and Pont, F. (1978). Anophèles et paludisme sur le haut Oyapock (Guyane française). *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*, 16. *p. 140*

Pajot, F. X., Pont, F., Molez, J. F., and Degallier, N. (1977b). Agressivité d'Anopheles (Nyssorhynchus) darlingi Root, 1926 (Diptera, Culicidae) en Guyane française. *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*, 15. *pp. 59 and 140*

Panday, R. (1977). Anopheles nuneztovari and malaria transmission in Surinam. *Mosq. News*, 37 :728–737. *p. 61*

Pearce, J. and Lindenmayer, D. (1998). Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (Lichenostomus melanops cassidix) in southeastern Australia. *Restoration Ecology*, 6(3) :238–243. *p. 10*

Pearson, R. G., Dawson, T. P., Berry, P. M., and Harrison, P. A. (2002). SPECIES : a spatial evaluation of climate impact on the envelope of species. *Ecological Modelling*, 154(3) :289–300. *p. 13*

Pearson, R. G., Raxworthy, C. J., Nakamura, M., and Townsend Peterson, A. (2007). Predicting species distributions from small numbers of occurrence records : a test case using cryptic geckos in Madagascar : Predicting species distributions with low sample sizes. *Journal of Biogeography*, 34(1) :102–117. *pp. 17 and 58*

Peterson, A. T. (2003). Predicting the geography of species' invasions via ecological niche modeling. *The quarterly review of biology*, 78(4) :419–433. *p. 10*

Peterson, A. T., Papeş, M., and Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, 213(1) :63–72. *p. 24*

Peterson, A. T., Soberón, J., Pearson, R. G., Martinez-Meyer, E., Nakamura, M., and Araújo, M. B. (2011). *Ecological niches and geographic distributions*. Princeton University Press, Princeton, NJ (USA). *pp. xiii, 12, 14, 15, 16, and 17*

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4) :231–259. *pp. 5, 17, 19, 21, and 25*

Phillips, S. J. and Dudík, M. (2008). Modeling of species distributions with Maxent : new extensions and a comprehensive evaluation. *Ecography*, 31(2) :161–175. *pp. 18, 20, 21, 22, 76, and 77*

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only distribution models : implications for background and pseudo-absence data. *Ecological Applications*, 19(1) :181–197. *pp. 24, 25, 27, 28, 33, 38, 40, 76, 105, and 112*

Piedallu, C., Gégout, J.-C., Lebourgeois, F., and Seynave, I. (2016). Soil aeration, water deficit, nitrogen availability, acidity and temperature all contribute to shaping tree species distribution in temperate forests. *Journal of Vegetation Science*, 27(2) :387–399. *p. 10*

Pommier de Santi, V., Dia, A., Adde, A., Hyvert, G., Galant, J., Mazevet, M., Nguyen, C., Vezenegho, S. B., Dusfour, I., Girod, R., and Briolant, S. (2016a). Malaria in French Guiana Linked to Illegal Gold Mining. *Emerging Infectious Disease journal*, 22(2) :344–346. *pp. 55, 58, and 85*

Pommier de Santi, V., Girod, R., Mura, M., Dia, A., Briolant, S., Djossou, F., Dusfour, I., Mendibil, A., Simon, F., Deparis, X., and Pagès, F. (2016b). Epidemiological and entomological studies of a malaria outbreak among French armed forces deployed at illegal gold mining sites reveal new aspects of the disease's transmission in French Guiana. *Malaria Journal*, 15(1). *pp. 54, 61, 85, 86, and 114*

Ponder, W. F., Carter, G. A., Flemons, P., and Chapman, R. R. (2001). Evaluation of museum collection data for use in biodiversity assessment. *Conservation biology*, 15(3) :648–657. *p. 28*

Prieto-Torres, D. A., Navarro-Sigüenza, A. G., Santiago-Alarcon, D., and Rojas-Soto, O. R. (2016). Response of the endangered tropical dry forests to climate change and the role of Mexican Protected Areas for their conservation. *Global Change Biology*, 22(1) :364–379. *p. 10*

Queyriaux, B., Texier, G., Ollivier, L., Galoisy-Guibal, L., Michel, R., and Meynard, J. B. (2011). Plasmodium vivax malaria among military personnel, French Guiana, 1998–2008. *Emerg Infect Dis*, 17 :1280–1282. *p. 58*

Richard, A. (1985). Le paludisme en forêt. In *Connaissance du milieu amazonien*, Colloques et Séminaires, pages 249–250, Paris. ORSTOM. *p. 59*

Ringard, J., Becker, M., Seyler, F., and Linguet, L. (2015). Temporal and Spatial Assessment of Four Satellite Rainfall Estimates over French Guiana and North Brazil. *Remote Sensing*, 7(12) :16441–16459. *pp. xiv and 113*

Rosa-Freitas, M. G., Tsouris, P., Peterson, A. T., Honório, N. A., de Barros, F. S. M., de Aguiar, D. B., Gurgel, H. C., de Arruda, M., Vasconcelos, S. D., and Luitgards-Moura, J. F. (2007). An ecoregional classification for the state of Roraima, Brazil : the importance of landscape in malaria biology. *Memórias do Instituto Oswaldo Cruz*, 102(3) :349–358. *p. 84*

Rossi, M., Benatti, S., Farella, E., and Benini, L. (2015). Hybrid EMG classifier based on HMM and SVM for hand gesture recognition in prosthetics. pages 1700–1705. IEEE. *pp. xiii and 14*

Rotenberry, J. T., Preston, K. L., and Knick, S. T. (2006). GIS-BASED NICHE MODELING FOR MAPPING SPECIES'HABITAT. *Ecology*, 87(6) :1458–1464. *p. 14*

Roux, E., Gaborit, P., Romaña, C. A., Girod, R., Dessay, N., and Dusfour, I. (2013). Objective sampling design in a highly heterogeneous landscape-characterizing environmental determinants of malaria vector distribution in French Guiana, in the Amazonian region. *BMC ecology*, 13(1) :45. *p. 40*

Roux, E., Venâncio, A., Girres, J., and Romaña, C. (2011). Spatial patterns and eco-epidemiological systems–part II : characterising spatial patterns of the occurrence of the insect vectors of Chagas disease based on remote sensing and field data. *Geospatial health*, 6(1) :53–64. *p. 40*

Rovzar, C., Gillespie, T. W., and Kawelo, K. (2016). Landscape to site variations in species distribution models for endangered plants. *Forest Ecology and Management*, 369 :20–28. *p. 10*

Rozendaal, J. A. (1987). Observations on the biology and behaviour of Anophelines in the Suriname rainforest with special reference to Anopheles darlingi Root. *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*. *pp. 59 and 61*

Rozendaal, J. A. (1992). Relations between Anopheles darlingi breeding habitats, rainfall, river level and malaria transmission rates in the rain forest of Suriname. *Medical and veterinary entomology*, 6(1) :16–22. *p. 74*

Schoener, T. W. (1968). The Anolis Lizards of Bimini : Resource Partitioning in a Complex Fauna. *Ecology*, 49(4) :704–726. *p. 99*

Sevenet, G. and Abonnec, E. (1940). Les moustiques de la Guyane Française VI. Les anophelinés 2. Le sous-genre stethomyia. *Archives de l'Institut Pasteur d'Algérie*, 16(4) :486–512. *p. 139*

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27 :379–423. *p. 17*

Silvain, J.-F. and Pajot, F.-X. (1981). Ecologie d'Anopheles (Nyssorhynchus) aquasalis Curry, 1932 en Guyane Française. *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*, 19(1) :11–21. *p. 140*

Singer, B. H. and Castro, M. C. (2001). Agricultural colonization and malaria on the Amazon frontier. *Annals of the New York Academy of Sciences*, 954(1) :184–222. *pp. 60, 74, and 83*

Sinka, M. E., Bangs, M. J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., Mbogo, C. M., Hemingway, J., Patil, A. P., Temperley, W. H., and others (2012). A global map of dominant malaria vectors. *Parasit Vectors*, 5(1) :69. *pp. xiii and 48*

Smith, M., Macklin, M., and Thomas, C. (2013). Hydrological and geomorphological controls of malaria transmission. *Earth-Science Reviews*, 116 :109–127. *pp. 60, 68, and 74*

Stefani, A., Dusfour, I., Corrêa, A. P. S. A., Cruz, M. C., Dessay, N., Galardo, A. K. R., Galardo, C. D., Girod, R., Gomes, M. S. M., and Gurgel, H. (2013). Land cover, land use and malaria in the Amazon : a systematic literature review of studies using remotely sensed data. *Malaria journal*, 12(1) :1–8. *pp. 60, 74, 83, and 85*

Stefani, A., Hanf, M., Nacher, M., Girod, R., and Carme, B. (2011a). Environmental, entomological, socioeconomic and behavioural risk factors for malaria attacks in Amerindian children of Camopi, French Guiana. *Malar J*, 10 :246. *p. 55*

Stefani, A., Roux, E., Fotsing, J., and Carme, B. (2011b). Studying relationships between environment and malaria incidence in Camopi (French Guiana) through the objective selection of buffer-based landscape characterisations. *International journal of health geographics*, 10(1) :65. *p. 58*

Stirling, D. A., Boulcott, P., Scott, B. E., and Wright, P. J. (2016). Using verified species distribution models to inform the conservation of a rare marine species. *Diversity and Distributions*, 22(7) :808–822. *p. 10*

Stockwell, D. (1999). The GARP modelling system : problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, 13(2) :143–158. *p. 16*

Tadei, W. P. and Dutary Thatcher, B. (2000). Malaria vectors in the Brazilian Amazon : Anopheles of the subgenus Nyssorhynchus. *Revista do Instituto de Medicina Tropical de São Paulo*, 42(2) :87–94. *p. 61*

Tadei, W. P., Thatcher, B. D., Santos, J. M., Scarpassa, V. M., Rodrigues, I. B., and Rafael, M. S. (1998). Ecologic observations on anopheline vectors of malaria in the Brazilian Amazon. *The American journal of tropical medicine and hygiene*, 59(2) :325–335. *p. 60*

Thuiller, W. (2004). Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, 10(12) :2020–2027. *p. 10*

Tognelli, M. F., Roig-Juñent, S. A., Marvaldi, A. E., Flores, G. E., and Lobo, J. M. (2009). An evaluation of methods for modelling distribution of Patagonian insects. *Revista chilena de historia natural*, 82(347-360). *p. 17*

Trape, J.-F. and Cordoliani, G. (1984). Lutte antipaludique en Guyane française : 1. Problèmes actuels. *Cahiers ORSTOM. Série Entomologie Médicale et Parasitologie*. *pp. 56 and 140*

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer, New York. *p. 13*

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, pages no–no. *pp. 26, 28, 29, 33, and 34*

Verret, C., Cabianca, B., Haus-Cheymol, R., Lafille, J.-J., Loran-Haranqui, G., and Spiegel, A. (2006). Malaria outbreak in troops returning from French Guiana. *Emerging Infectious Diseases*, 12(11) :1794–5. *p. 58*

Vezenegho, S. B., Adde, A., Gaborit, P., Carinci, R., Issaly, J., Pommier de Santi, V., Dusfour, I., Briolant, S., and Girod, R. (2014). Mosquito magnet® liberty plus trap baited with octenol confirmed best candidate for Anopheles surveillance and proved promising in predicting risk of malaria transmission in French Guiana. *Malaria journal*, 13(1) :384. *p. 62*

Vezenegho, S. B., Carinci, R., Gaborit, P., Issaly, J., Dusfour, I., Briolant, S., and Girod, R. (2015). Anopheles darlingi (Diptera : Culicidae) Dynamics in Relation to Meteorological Data in a Cattle Farm Located in the Coastal Region of French Guiana : Advantage of Mosquito Magnet Trap. *Environmental Entomology*, 44(3) :454–462. *pp. 73, 84, 85, and 86*

Vittor, A. Y., Gilman, R. H., Tielsch, J., Glass, G., Shields, T. I. M., Lozano, W., Pinedo-Cancino, V., and Patz, J. A. (2006). The effect of deforestation on the human-biting rate of Anopheles darlingi, the primary vector of falciparum malaria in the Peruvian Amazon. *American Journal of Tropical Medicine and Hygiene*, 74(1) :3–11. *pp. 59 and 74*

Vittor, A. Y., Pan, W., Gilman, R. H., Tielsch, J., Glass, G., Shields, T., Sánchez-Lozano, W., Pinedo, V. V., Salas-Cobos, E., and Flores, S. (2009). Linking deforestation to malaria in the Amazon : characterization of the breeding habitat of the principal malaria vector, Anopheles darlingi. *American Journal of Tropical Medicine and Hygiene*, 81(1) :5–12. *p. 74*

Walker, P. A. and Cocks, K. D. (1991). HABITAT : A Procedure for Modelling a Disjoint Environmental Envelope for a Plant or Animal Species. *Global Ecology and Biogeography Letters*, 1(4) :108. *p. 14*

Webber, B. L., Yates, C. J., Le Maitre, D. C., Scott, J. K., Kriticos, D. J., Ota, N., McNeill, A., Le Roux, J. J., and Midgley, G. F. (2011). Modelling horses for novel climate courses : insights from projecting potential distributions of native and alien Australian acacias with correlative and mechanistic models : Modelling Australian acacias. *Diversity and Distributions*, 17(5) :978–1000. *p. 27*

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., and NCEAS Predicting Species Distributions Working Group (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14(5) :763–773. *pp. 17 and 103*

World Health Organization (2012). *World malaria report 2012.* World Health Organization, Geneva. *p. 46*

World Health Organization (2016). World malaria report 2016. Technical report, WHO, Switzerland. *pas de citations*

World Health Organization, Global Malaria Programme, and World Health Organization (2015a). World Malaria Report 2015. Technical report, WHO. OCLC : 948838960. *pp. 4 and 45*

World Health Organization, World Health Organization, and Global Malaria Programme (2015b). *Global technical strategy for malaria, 2016-2030*. OCLC : 921272942. *p. 4*

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., and Veran, S. (2013). Presence-only modelling using MAXENT : when can we trust the inferences ? *Methods in Ecology and Evolution*, 4(3) :236–243. *p. 25*

Yee, T. W. and Mitchell, N. D. (1991). Generalized additive models in plant ecology. *Journal of Vegetation Science*, 2(5) :587–602. *p. 11*

Zaim, M., Ershadi, M. R., Manouchehri, A. V., and Hamdi, M. R. (1986). The use of CDC light traps and other procedures for sampling malaria vectors in southern Iran. *J Am Mosq Control Assoc*, 2(4) :511–15. *p. 62*

Zaniewski, A. E., Lehmann, A., and Overton, J. M. (2002). Predicting species spatial distributions using presence-only data : a case study of native New Zealand ferns. *Ecological modelling*, 157(2) :261–280. *p. 25*

Zanini, V. M., Maragarete do Socorro Mendonça Gomes, Allan Kardec Ribeiro Galardo, Ana Cristina da Silva Ferreira Lima, Ana Paula Sales de Andrade Correa, Aurélia Stefani, Emmanuel Roux, Raimundo Tadeu Lessa de Souza, Póvoa, M. M., and Raimundo Nonato Picanço Souto (2014). Potencial de transmissão de malária no município de Oiapoque - Amapá - Brasil. At Acre, Brazil. *p. 85*

Zeilhofer, P., Santos, E., Ribeiro, A. L., Miyazaki, R. D., and Santos, M. (2007). Habitat suitability mapping of Anopheles darlingi in the surroundings of the Manso hydropower plant reservoir, Mato Grosso, Central Brazil. *International Journal of Health Geographics*, 6(1) :7. *p. 74*

Annexes

Annexe A

Requêtes pour l'analyse de publications annuelles

L'analyse des publications a été réalisé sur *Web of Sciences*. Les requêtes suivantes ont été réalisées sur la période de 1990 à 2015 (selon Guisan et al. (2013)) :

— pour les publications portant les modèles de distribution :
`"Species distribution model*" OR "habitat model*" OR "niche model*" OR "habitat distribution model*" OR "habitat suitability model*" OR "ecological niche model*" OR "niche-based model*" OR "bioclimatic envelope model*" OR "resource selection function*"`

— pour les publication portant sur les modèles de distribution d'espèces mentionnant le biais d'échantillonnage :
`("Species distribution model*" OR "habitat model*" OR "niche model*" OR "habitat distribution model*" OR "habitat suitability model*" OR "ecological niche model*" OR "niche-based model*" OR "bioclimatic envelope model*" OR "resource selection function*" ) AND "sampl* bias*"`

Annexe B

Données de capture d'*Anopheles* réalisée entre 1902 et 2013

Année de publication	Publication (volume, numéro)	Source et références
1902	4	Institut Pasteur d'Algérie (Neveu-Lemaire, 1902a,b)
1940	16,4	Institut Pasteur d'Algérie (Sevenet and Abonnec, 1940)
1942	41	Institut Pasteur de la Guyane et du Territoire de l'Inini (Floch and Abonnec, 1942)
1942	43	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1942	47	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1943	71	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1943	72	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1944	77	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1943	75	Rapport sur le fonctionnement technique de l"Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1942
1945	103	Rapport sur le fonctionnnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1944
1945	116	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1946	124	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1946	126	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1946	139	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1946	125	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1945
1947	144	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1947	163	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1947	151	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1946
1948	16	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1948	173	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1947
1949	188	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1948
1950	213	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1950	207	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1949
1951	236	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1951	229	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1950
1952	257	Archives de l'Institut Pasteur de la Guyane et du territoire de l'Inini
1952	262	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1951
1953	288	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane et du territoire de l'Inini pendant l'année 1952
1954	326	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane
1954	345	Archives de l'Institut Pasteur de la Guyane
1954	346	Archives de l'Institut Pasteur de la Guyane
1954	348	Archives de l'Institut Pasteur de la Guyane
1955	352 b	Archives de l'Institut Pasteur de la Guyane
1955	354	Archives de l'Institut Pasteur de la Guyane

Année de publication	Publication (volume, numéro)	Source et références
1955		Archives de l'Institut Pasteur de la Guyane
1955	366	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1954
1956	392	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1955
1957	428	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1956
1958	443	Archives de l'Institut Pasteur de la Guyane
1958	453	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1957
1961	468	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1958
1961	469	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1959
1962	470	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1960
1963	480	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1961
1964	481	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1962
1964	486	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1963
1965	494	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1964
1966	510	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1965
1967	516	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1966
1968	520	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1967
1969	525	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1968
1969		Rapport Annuel 1969
1969	527	Rapport sur le fonctionnement technique de l'Institut Pasteur de la Guyane pendant l'année 1969
1970		Rapport annuel
1971		Rapport annuel
1971		Rapport annuel 1971 – activités du groupe recherche U79
1972		Rapport annuel
1973		Rapport annuel
1974		Rapport annuel
1975	13,3	Cahiers ORSTOM (Pajot et al., 1975)
1975		Rapport annuel
1976	Rapport annuel	
1977	15,1	Cahiers ORSTOM (Pajot et al., 1977b)
1977	15,3	Cahiers ORSTOM (Pajot et al., 1977a)
1977		Rapport annuel
1978	16,2	Cahiers ORSTOM (Pajot et al., 1978)
1978	16,1	Cahiers ORSTOM (Degallier et al., 1978)
1979		Rapport annuel
1981	74,2	Juminer et al. (1981)
1981	19,1	Cahiers ORSTOM (Silvain and Pajot, 1981)
1982		Rapport annuel
1983	11,2	Cahiers ORSTOM (Degallier et al., 1983)
1983		Rapport annuel
1984		Cahiers ORSTOM(Trape and Cordoliani, 1984)
1987		Connaissance du milieu amazonien : actes du séminaire Le paludisme en forêt
1989	82	Mouchet et al. (1989)
1990		Rapport d'Activité, Service Départemental de Désinfection
1991		Rapport d'Activité, Service Départemental de Désinfection
1993		Rapport annuel
1994		Rapport d'Activité, Service Départemental de Désinfection
1995		Rapport annuel
1995		Rapport d'Activité, Service Départemental de Désinfection
1996		Laboratoire d'Entomologie Médicale Institut Pasteur, Etude de la Transmission du paludisme en Guyane française
1996		Rapport Annuel d'Activité Année 1996, Institut Pasteur de la Guyane
1996		Rapport d'Activité, Service Départemental de Désinfection

Année de publication	Publication (volume, numéro)	Source et références
1997		Rapport Annuel d'Activité du Laboratoire d'Entomologie Médicale Année 1997, Institut Pasteur de la Guyane
1997	57,4	Med. TropPaludisme, anopheles, lutte anti-paludiqye en Guyane-française
1997		Rapport d'Activité, Service Départemental de Désinfection
1998		Rapport d'activités du Laboratoire d'Entomologie Médicale Année 1998, Institut Pasteur de la Guyane
1999		Rapport d'activités du Laboratoire d'Entomologie Médicale Année 1999, Institut Pasteur de la Guyane
1999	9	Molez (1999)
1999		Rapport d'Activité, Service Départemental de Désinfection

TABLEAU B.1 – **Publications mentionnant la présence et la location d'espèce *Anopheles* en Guyane entre 1902 et 1999**

Année de publication	Publication (volume, numéro)	Source et références
2000		Rapport d'Activité, Service Départemental de Désinfection
2000		Rapport d'activités du Laboratoire d'Entomologie Médicale Année 2000, Institut Pasteur de la Guyane
2000		Institut Pasteur de la Guyane, Identification de(s) l'espèce(s) de moustique provoquant de fortes nuisances dans les zones de Kourou et du Centre Spatial Guyanais
2001		Rapport d'activités du Laboratoire d'Entomologie Médicale Année 2001, Institut Pasteur de la Guyane
2002		Rapport d'activités du Laboratoire d'Entomologie Médicale Année 2002, Institut Pasteur de la Guyane
2003		Rapport d'activités du Laboratoire d'Entomologie Médicale Année 2003, Institut Pasteur de la Guyane
2004		Rapport d'Activité, Service Départemental de Désinfection
2004		Rapport d'activités du Laboratoire d'Entomologie Médicale Année 2004, Institut Pasteur de la Guyane
2005		Rapport d'Activité, Service Départemental de Désinfection
2005		Rapport d'activités du Laboratoire d'Entomologie Médicale Année 2005, Institut Pasteur de la Guyane
2007		Rapport d'Activité, Service Départemental de Désinfection
2007		Rapport Annuel, Institut Pasteur de la Guyane
2007		Institut Pasteur de la Guyane
2008		Rapport d'Activité, Service Départemental de Désinfection
2008		Rapport Annuel, Institut Pasteur de la Guyane
2008		Institut Pasteur de la Guyane
2009		Rapport d'Activité, Service Départemental de Désinfection
2009		Institut Pasteur de la Guyane
2010		Rapport d'Activité, Service Départemental de Désinfection
2010		Institut Pasteur de la Guyane
2011		Rapport d'Activité, Service Départemental de Désinfection
2011		Institut Pasteur de la Guyane
2012		Institut Pasteur de la Guyane et le Service de Santé des Armées
2012		Institut Pasteur de la Guyane
2013		Institut Pasteur de la Guyane et le Service de Santé des Armées

TABLEAU B.2 – **Publications mentionnant la présence et la location d'espèce *Anopheles* en Guyane entre 2000 et 2013**

Annexe C

Données de capture d'*Anopheles darlingi* précisément géolocalisées entre 2000 et 2013

Numéro de site	Localité	Longitude	Latitude
1	Cayodé	175866.785122426	374453.1653833
2	Taluène	162884.684665956	372648.888139721
3	Cacao	336895.086300153	505521.800163598
4	Midenangalanti	121041.015986301	557342.72581239
5	Grand Santi	124374.105534813	473261.338110987
6	Bois Martin	118927.138507594	552587.496776593
7	Flavien Campou	126212.611562883	477960.065899793
8	Régina	374698.048396995	476928.730177523
9	Camopi	352395.046055802	350764.216480451
10	Alikéné	322929.978284572	360135.527835386
11	Mine Boulanger	343189.582357019	504590.723307373
12	Carbet Légion crique Sikini	359046.553479	361996.097871
13	Cogneau - 23. Lot. Aquavilla	354091.053713437	538837.801956188
14	41. rue des Ixoras - Lot. Cogneau Larivot	351322.960529913	541116.10235073
15	Attila Cabassou	354987.967015231	540678.87835953
16	La Chaumière	347961.540051952	539402.864452834
17	1228. Ch. de La Chaumière	349626.610158034	540136.658859129
18	Saint-Georges	411052.118458134	429833.606330522
19	Quartier Espérance - Saint-Georges	411113.960068634	430709.665437749
20	Village Martin - Saint-Georges	411519.804866871	432628.481492214
21	Boulangerie – Saül	254644.262333224	400558.641871421
22	Chemin Mogès	352032.329517247	529456.717653458
23	Dorlin	216742.863044	415732.862195
24	Maripasoula	163199.637191727	403271.422431979
25	Repentir	234327.380239	427769.677349
26	Stoupan	352127.057631364	527017.037636427
27	Camp Pararé – Nouragues	311267.546406115	446010.858172732
28	Village Blondin - Saint-Georges	409492.529993713	428392.815435875
29	Quartier Adimo - Saint-Georges	410525.149504269	430989.662548913
30	Camp Bernet/ Légion étrangère	410437.343147834	429860.701524782
31	La ferme de Lait-Quateur	332757.212761399	552628.404024885
32	Grand Usine	288898.531939378	372330.073251345
33	Dagobert	226797.35968	438864.200286
34	Cacao	336515.199831583	505801.969900718
35	Saut-Maripa – Camp militaire	401579.145107468	420354.782186144
36	Eau-Claire	213145.824694	398626.838699
37	Cacao	336956.199791048	504651.969979654
38	Cacao	337909.199733996	506008.969874801
39	Cacao	336576.199808207	503413.970071365
40	Cacao	336555.199807874	503184.970087929
41	Impasse de la rafinerie – Cogneau	354091.053713437	538837.801956188
42	Quartier Bambou	411277.041477492	430178.275876772
43	Quartier Maripa - Saint-Georges	410252.974346512	430188.463058511
44	Quartier Savane - Saint-Georges	411024.737954334	430940.054473617
45	Cacao	336826.199809151	505764.969900894
46	Cacao	337091.199787723	505441.969921949
47	Cacao	336186.199848992	505049.969957266
48	Quartier Onozo- Saint-Georges	411339.040785181	430641.896115415

TABLEAU C.1 – **Coordonnées géographiques des sites de présence d'*Anopheles darlingi* précisément géolocalisés**
(Système de coordonnées : RGFG95/UTM22N)

Annexe D

Variables environnementales

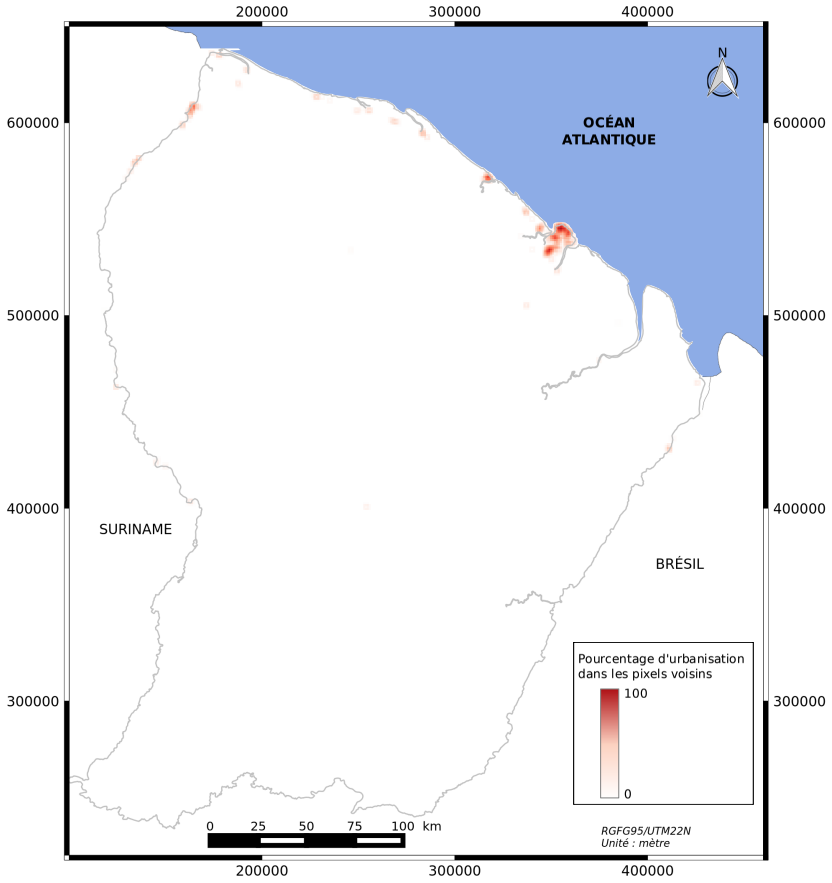


FIGURE D.1 – **Pourcentage d'urbanisation dans les pixels voisins (*PER_URB_NEIGH*)**

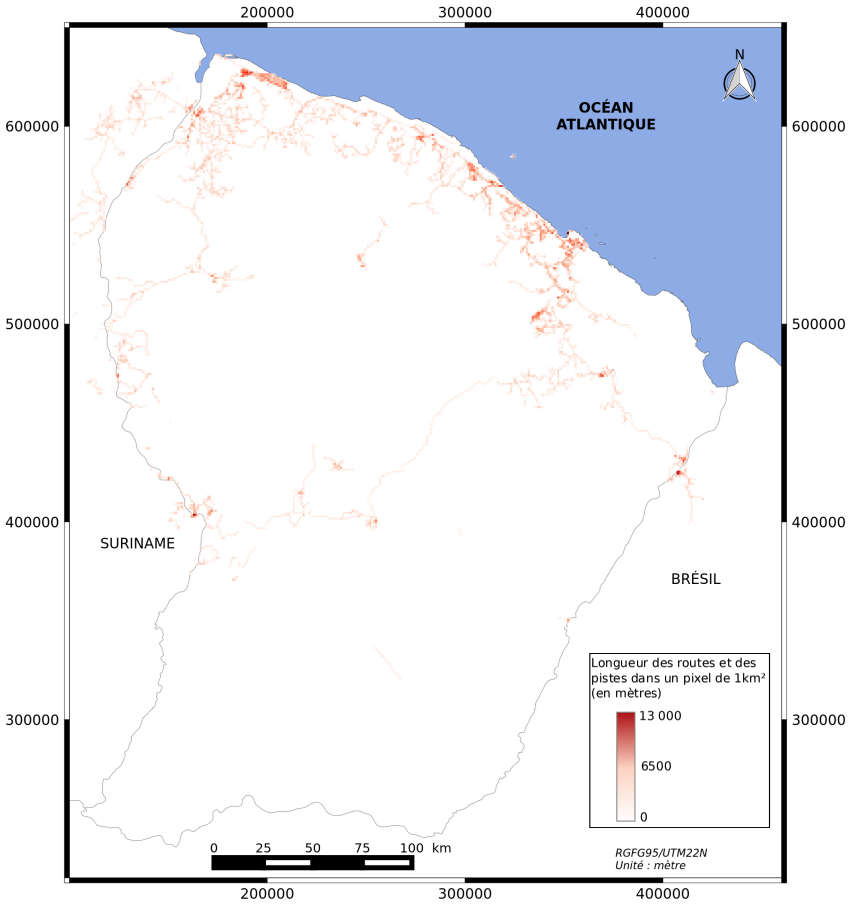


FIGURE D.2 – **Longueur des routes et des pistes dans un pixel de 1km² (*ROADS*)**

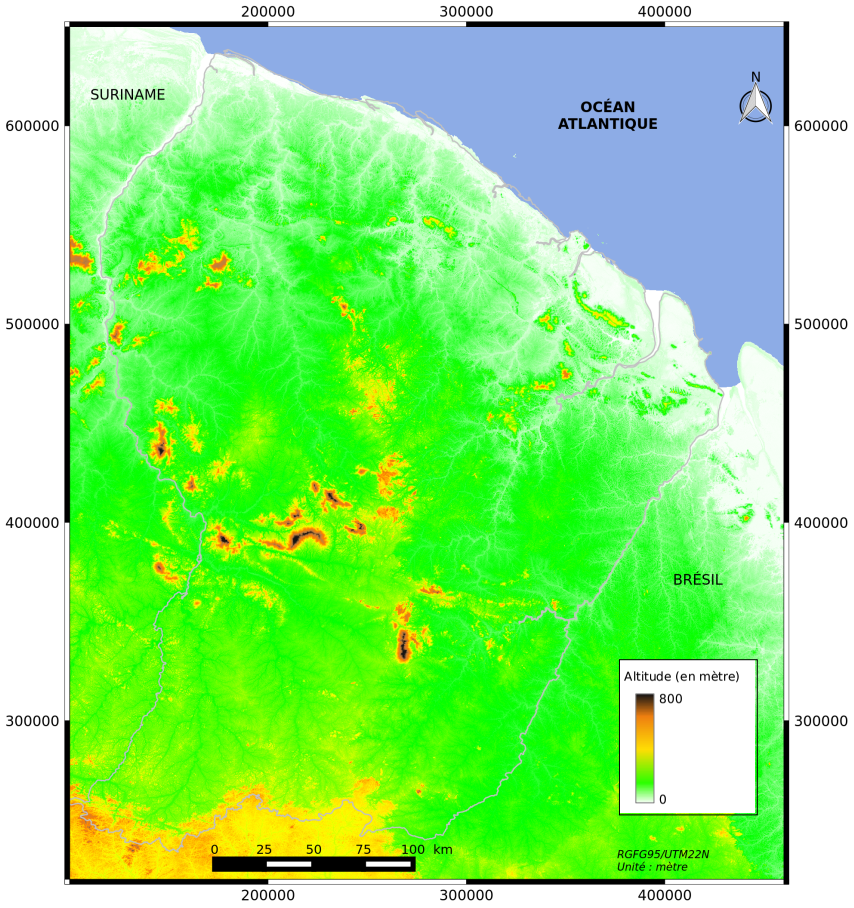


FIGURE D.3 – **Altitude minimale (*ATL_min*)**

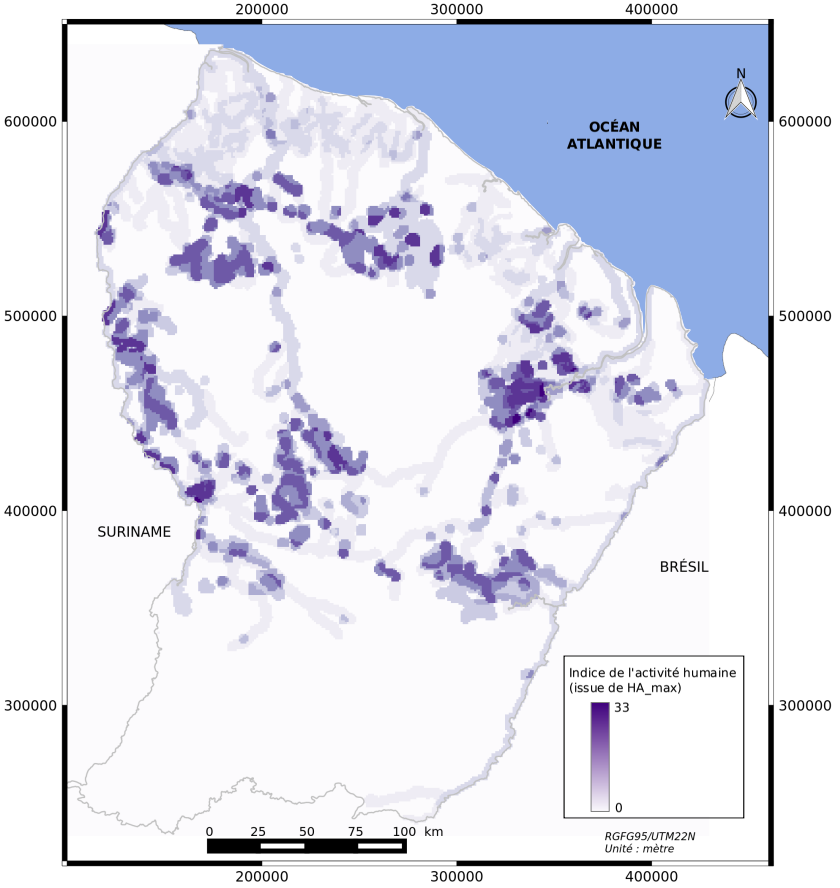


FIGURE D.4 – **Présence et activités humaine altérant de manière non-permanente l'environnement naturel (*HA_max*)**

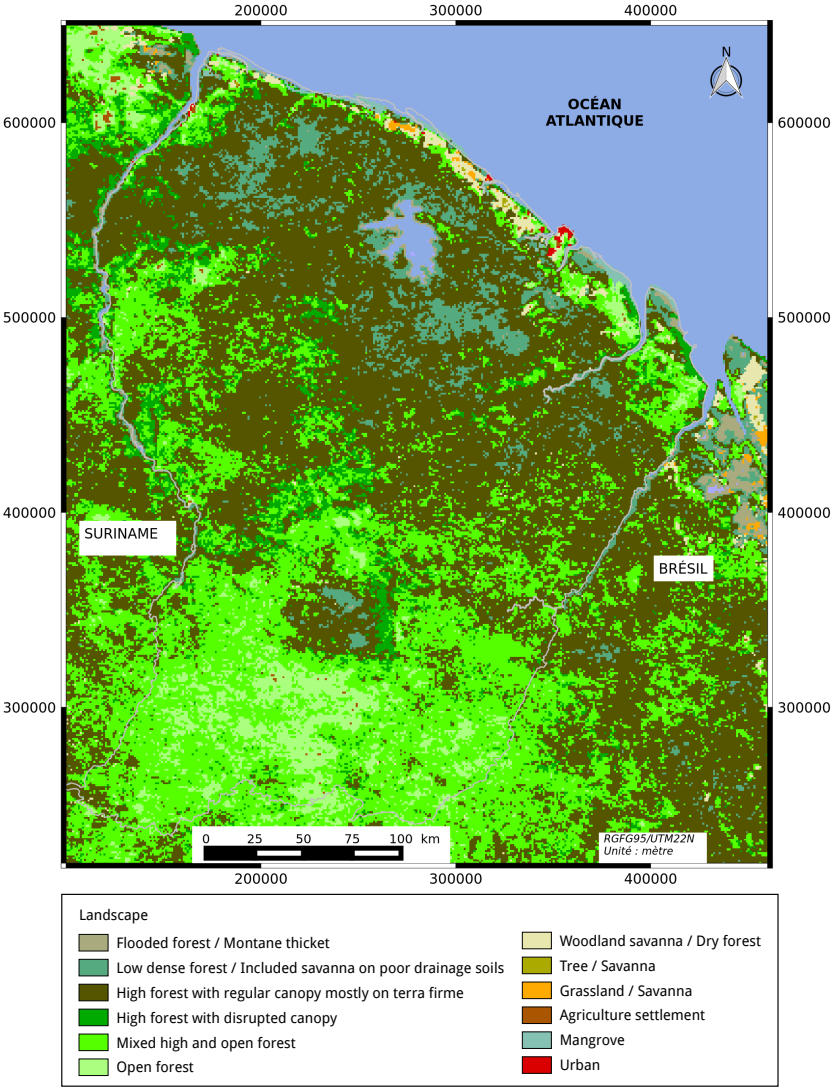
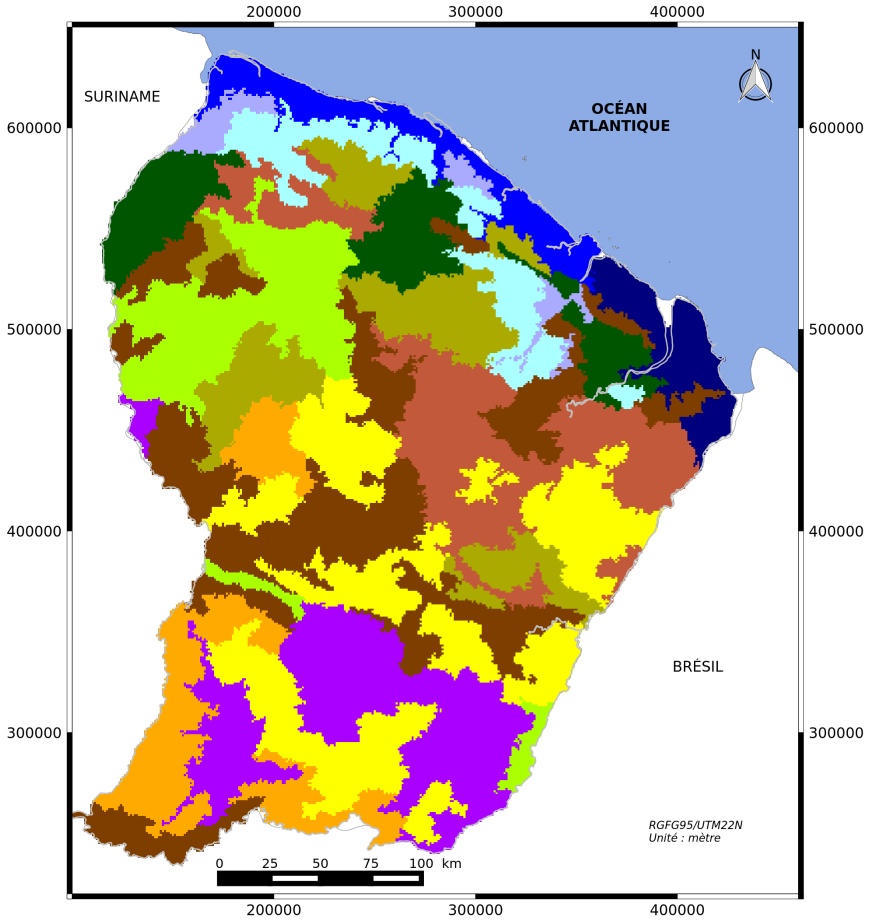


FIGURE D.5 – **Occupation du sol (*LS*)**



Geomorphological landscape

- Coastal plain with low relief
- Plain with residual reliefs (back-coastal)
- Coastal fat plain
- Complex hilly area
- Low hilly area and large valleys
- Inland plain
- Moderate plateau with inselbergs
- Hilly plateau with inselbergs
- High dissected plateau
- High hills and mountains
- Peneplain with moderate hills
- Regular pattern hilly area

FIGURE D.6 – **Paysages géomorphologiques (*GLS*)**

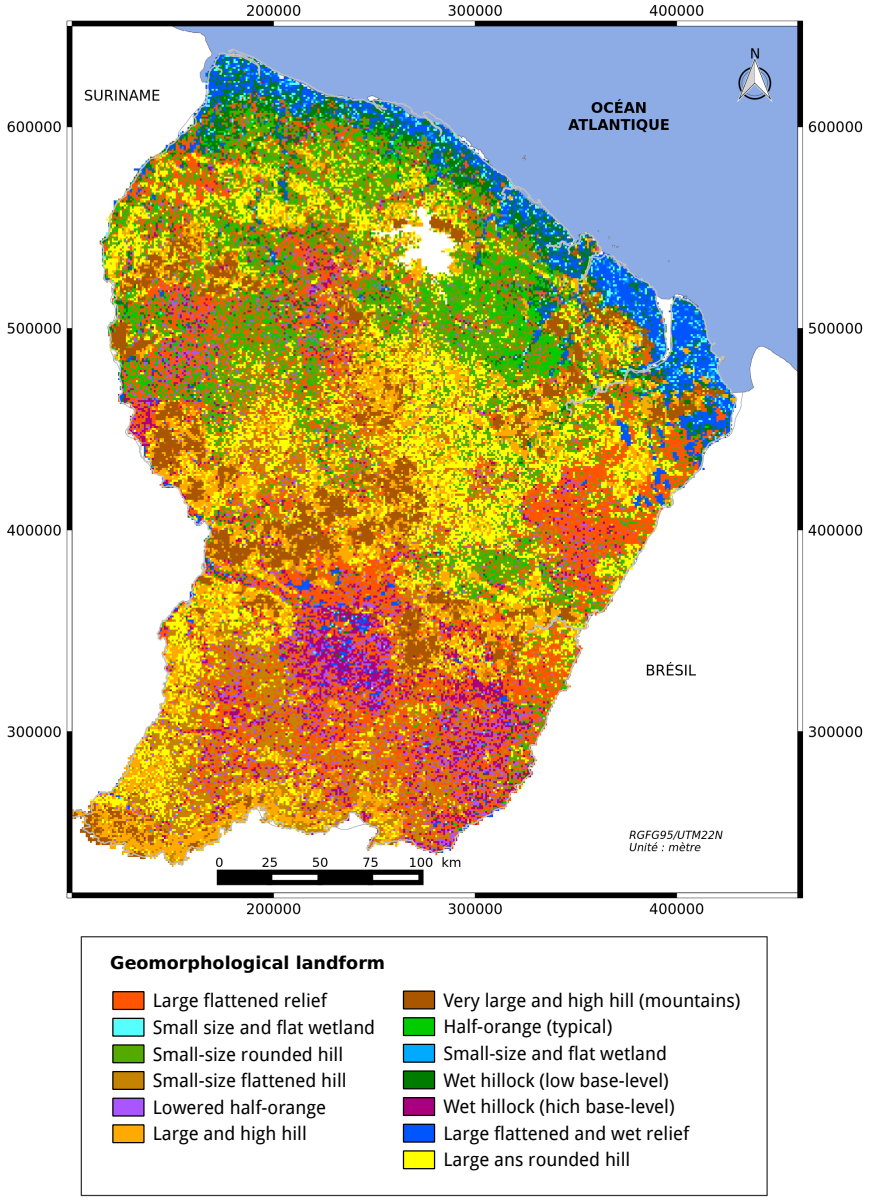


FIGURE D.7 – **Unités géomorphologiques (*GLF*)**

Annexe E

Schéma du principe de fonctionnement de la librairie *virtualspecies*

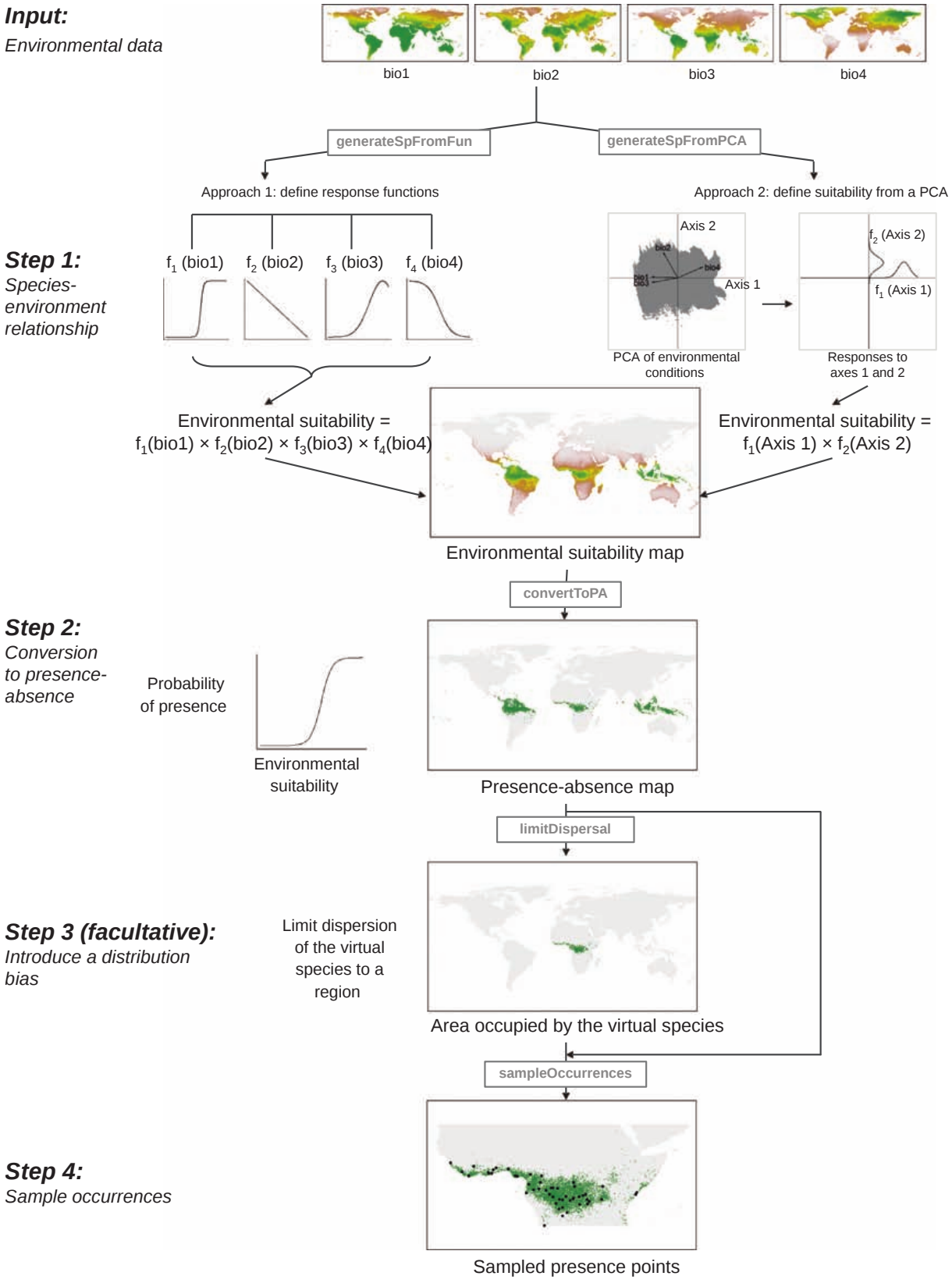


FIGURE E.1 – **Schéma du principe de fonctionnement de la librairie *virtualspecies***
Source : Leroy et al. (2015)

Annexe F

Publications et communications

I. Moua Y., Roux E., Seyler F., Girod R., Dusfour I., and Briolant S. Ecological niche modeling for *Anopheles darlingi*, French Guiana. Dans *Amazonian Conference on Emerging and Infectious Diseases*, 26 -28 septembre 2014, Cayenne (France)

“AMAZONIAN CONFERENCE ON EMERGING AND INFECTIOUS DISEASES”

CAYENNE , FRENCH GUIANA

SEPTEMBER 26TH TO 28TH , 2014

Ecological niche modeling for *Anopheles darlingi*, French Guiana

Yi Moua[a], Emmanuel Roux[b], Frédérique Seyler[b], Romain Girod[c], Isabelle Dusfour[c], Sébastien Briolant[d,e]

[a] *ESPACE-DEV, UMR 228, Université des Antilles et de la Guyane, Cayenne, French Guiana (yi.moua@ird.fr)*

[b] *ESPACE-DEV, UMR 228, Institut de Recherche pour le Développement, Montpellier, France*

[c] *Unité d’entomologie médicale, Institut Pasteur de la Guyane, Cayenne, French Guiana*

[d] *Direction Interarmées du Service de Santé en Guyane et Institut de Recherche Biomédicale des Armées, France.*

[e] *Laboratoire de parasitologie, Institut Pasteur de la Guyane, Cayenne, French Guiana*

Malaria remains an important health issue in French Guiana. The main vector of the disease in this amazonian region is *Anopheles darlingi*. Due to the extent of the territory and the difficulties in accessing remote areas, the knowledge on the spatial distribution of this species is very partial. However, previous works showed significant spatial variations in *An. darlingi* presence and density at the French Guiana scale [1, 2]. Consequently, the design of a malaria transmission risk map at the French Guiana scale requires to spatialize, at the same scale, the habitat suitability of the *An. darlingi* species.

We propose to reach such an objective by applying ecological niche modeling, more precisely the Maxent model, a presence-only modeling approach based on the maximum entropy principle [3].

The model was built with geo-localized *An. darlingi* presence data obtained from 2000 to 2013 and environmental variables (vegetation indices, landform, landscape, forest ecosystem and rainfall). The prediction performance of the model is evaluated by means of a 10-fold cross validation procedure, by taking into account the area under the receiver operating characteristic curve (AUC) values.

The resulting map is an estimation of environmental conditions suitable for *An. darlingi* presence and survival. This work shows how the habitat suitability for malaria vectors can be assessed using environmental data and modeling tool at the French Guiana scale.

[1] Hiwat, H., Issaly, J., Gaborit, P., Somai, A., Samjhawan, A., Sardjoe, P., Soekhoe, T., Girod, R., 2010. Behavioral heterogeneity of Anopheles darlingi (Diptera: Culicidae) and malaria transmission dynamics along the Maroni River, Suriname, French Guiana. *Trans. R. Soc. Trop. Med. Hyg* **104**, 207–213.

[2] Girod, R., Roux, E., Berger, F., Stefani, A., Gaborit, P., Carinci, R., Issaly, J., Carme, B., Dusfour, I., 2011. Unravelling the relationships between Anopheles darlingi (Diptera: Culicidae) densities, environmental factors and malaria incidence: understanding the variable patterns of malarial transmission in French Guiana (South America). *Annals of Tropical Medicine and Parasitology* **105**, 107–122.

[3] Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**, 231–259.

II. Moua Y., Roux E., Seyler F., and S. Briolant. Sampling bias corrections in Maxent : evaluation of methods. Dans *Mathematical and Computational Epidemiology of Infectious diseases – the interplay between models and public health policies*, 30 août – 5 septembre 2015, Erice (Italie)

Sampling bias correction in Maxent: evaluation of methods

Y. Moua[1], E . Roux[2], F. Seyler[2], S. Briolant[5,6]

Contact: yi.moua@ird.fr

Context

- 17 % of current diseases are vector-borne diseases (World Health Organisation)
- Malaria is the most deadly vector-borne disease (over 627 000 deaths in 2012)
- Its repartition is directly linked to the distribution of its vectors which depends on environmental conditions
- *Anopheles darlingi* is the main malaria vector in South America
- Ecological niche modeling (ENM) can be used to **predict the distribution of species** by using environmental data
- Among existing ENM approaches, **Maxent** (Phillips et al. 2006) is widely used because (1) requires presence-only data (2) outperforms other presence-only methods (Elith et al, 2006) (3) is adapted to small number of presence points
- Maxent estimates, on the entire study area, an unknown probability distribution interpreted as **habitat suitability**. Such a probability distribution is the one with the **maximum entropy** (the most uniform) that **satisfies constraints** defined by the environmental variables features observed at presence sites and in the entire study area.
  Only a random selection of the pixels of the study area (denoted as "background") is considered for the model building. Several selection strategies can be adopted.

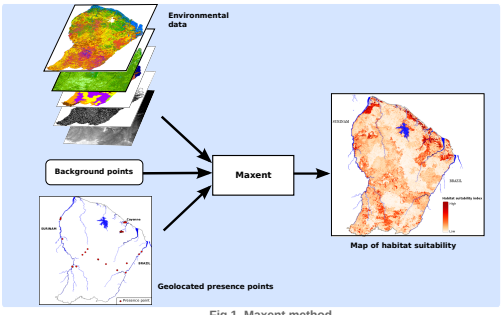


Fig 1. Maxent method

Scientific issues

- Presence data are often biased because most of effort sampling occurs in easily accessible areas.
  So that the model is not affected by a such bias, the selected background and the presence dataset should exhibit the same environmental bias (Phillips *et al*. 2009)
- Moreover, the number of available presence samples is often low.
- Some methods exist (Tab.1) to correct the effect of the sampling bias but : (1) they are mainly adapted to a large number of presence samples; (2) to our knowledge the biased selection of the background based on environmental criteria has not been studying yet.

Require a large number of samples

	Geographical criterion (geo)	Environmental criterion (env)
Presence dataset filtering (F)	Boria *et al.*, 2014	Fourcade *et al.*, 2014
Background selection (BG)	Elith *et al.*, 2010	Present study

Tab. 1. Existing correction methods

Objectives

- Propose a method to correct the effect of sampling bias :
  - based on the selection of the background with environmental criterion
  - efficient for a small number of samples
- Evaluate the proposed sampling bias correction method

Methodology

Proposed method

The sampling effort is used to bias the background selection (the higher the sampling effort is, the higher the chance to select a point is).

- For a given pixel *p*, **sampling effor**t is defined as the relative density of pixels where sampling occurred, in the environmental neighborhood of *p*.
- The **environmental neighborhood** of *p* is defined by a Gaussian-like membership function parametrized by the euclidean distance from *p* (see Fig. 2) within the environmental variable space (beforehand transformed by means of a Factorial Analysis of Mixed Data (FAMD)) .

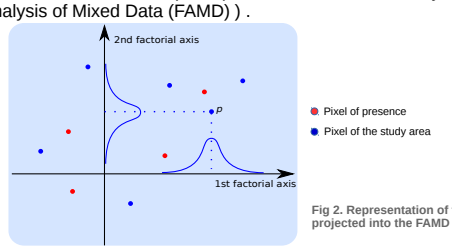


Fig 2. Representation of the Gaussian neighborhood projected into the FAMD first plane

Evaluation in the context of *An. darlingi* habitat spatialization (French Guiana)

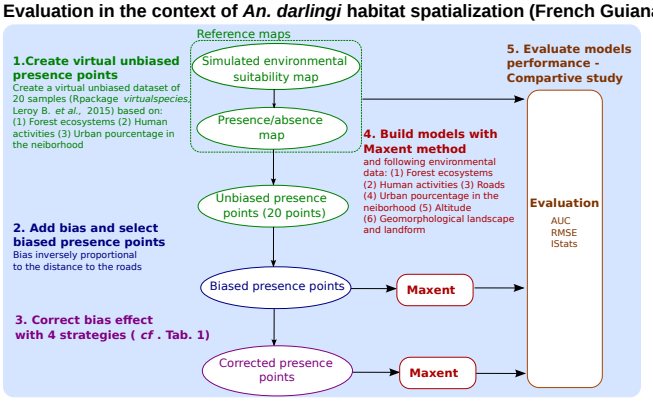


Fig 3. Illustration of the evaluation methodology

Results

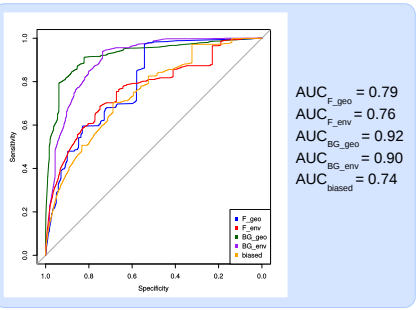


Fig 4. ROC curves and Areas Under the Curve (AUC) for the four correction strategies, evaluating the capacity of the model to predict the species presence.

	RMSE	Istat (*)
F_geo	0.365	0.849
F_env	0.389	0.777
BG_geo	0.343	0.888
BG_env	0.376	0.764
Biased	0.404	0.679

(*) Istat is the I similarity statistic for quantifying niche overlap. It measures the similarity between two probability distributions (Warren *et al.* 2008)
Value 0 : two distributions have no overlap
Value 1 : two distributions are identical

Tab 2. Root Mean Square Error (RMSE) and I similarity statistic (Istat) computed with the reference and the estimated environmental suitability maps

Discussion

- According to ROC curves and AUC results, background selection methods seem the most adapted to a low number of samples (20)
- The proposed correction method seems to be less efficient than the other according to RMSE and Istat results. However, the statistical significance of the result differences were not studied. Further investigations should be carried out by using simulated replicates
- The sensitivity of the proposed method to the number of presence samples should be studied in the future.

Conclusion

- The present study proposes a new method to correct the effect of sampling bias in ENM and especially in Maxent modeling approach, based on an environmental criterion and adapted to small sample datasets.
- Preliminary results show that the proposed method is very promising in predicting the species presence for small sample datasets (20).

Affiliations

[1]Espace Dev, UMR 228 UG/IRD/UM2/UR, Université de Guyane, Cayenne, French Guiana
[2]Espace Dev, UMR 228 IRD/UM2/UR/UG, Institut de Recherche pour le Développement, Montpellier, France
[3]Direction Interarmées du service de Santé en Guyane et Institut de Recherche biomédicale des Armées, France
[4]Laboratoire de parasitologir, Institut Pasteur de la Guyane, Cayenne, French Guiana

Acknowledgements

Guyamazon projets GAPAM-Sentinela and DS BIODIVA

References

WHO,World Health Organisation, *http://www.who.int*
Boria, R.A. *et al.* 2014. « Spatial Filtering to Reduce Sampling Bias Can Improve the Performance of Ecological Niche Models. » *Ecological Modelling* 275 (mars): 73-77.
Elith, J. *et al.*. 2010. « The Art of Modelling Range-Shifting Species: The Art of Modelling Range-Shifting Species. » *Methods in Ecology and Evolution* 1 (4): 330-42.
Elith, J., *et al.*. 2006. « Novel methods improve prediction of species' distributions from occurrence data. » *Ecography* 29: 129-51.
Fourcade, Y. *et al.* 2014. « Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. ». *PLoS ONE* 9 (5)
Leroy, B., *et al.* 2015. « Virtualspecies, an R Package to Generate Virtual Species Distributions. » *Ecography*
Phillips, S.J., *et al.* 2009. « Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. » *Ecological Applications* 19 (1): 181-97.
Varela, S. *et al.*. 2014. « Environmental Filters Reduce the Effects of Sampling Bias and Improve Predictions of Ecological Niche Models. » *Ecography*
Warren, D. *et al.* 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62 (11): 2868-83.
Environmental data : [1] Gond, V. *et al.* 2011. « Broad-Scale Spatial Pattern of Forest Landscape Types in the Guiana Shield. » *International Journal of Applied Earth Observation and Geoinformation* 13 (3): 357-67.**[2] [3] [4]** de Thoisy, B., *et al* 2010. « Rapid Evaluation of Threats to Biodiversity: Human Footprint Score and Large Vertebrate Species Responses in French Guiana. » *Biodiversity and Conservation* 19 (6): 1567-84. **[5]** NASA **[6]** Guitet, S., *et al.* 2013. « Landform and landscape mapping, French Guiana (South America). » *Journal of Maps* 9 (3): 325-35.

III. Moua Y., Roux E., Girod R., Dusfour I., de Thoisy B., Seyler F., and Briolant S. Distribution of the habitat suitability of the main malaria vector in French Guiana using Maximum Entropy modeling. *Journal of Medical Entomology*, accepté pour publication.

Journal of Medical Entomology

Distribution of the habitat suitability of the main malaria vector in French Guiana using Maximum Entropy modeling

Yi Moua[1], Emmanuel Roux[2], Romain Girod[3], Isabelle Dusfour[3], Benoit de Thoisy[4], Frédérique Seyler[2], Sébastien Briolant[3, 5, 6, 7]

[1] Université de Guyane, ESPACE-DEV, UMR 228 (IRD, UM, UR, UA, UG), Cayenne, French Guiana

[2] Institut de Recherche pour le Développement, ESPACE-DEV, UMR 228 (IRD, UM, UR, UA, UG), Montpellier, France

[3] Unité d'Entomologie Médicale, Institut Pasteur de la Guyane, Cayenne, French Guiana

[4] Laboratoire des Interactions Virus Hôtes, Institut Pasteur de la Guyane, Cayenne, French Guiana

[5] Direction Interarmées du Service de Santé en Guyane, Cayenne, French Guiana

[6] Institut de Recherche Biomédicales des Armées, Unité de Parasitologie et d'Entomologie Médicale, Marseille, France

[7] Unité de Recherche en Maladies Infectieuses Tropicales Emergentes, UMR 63, CNRS 7278, IRD 198, INSERM 1095, Faculté de Médecine La Timone, Marseille, France

Abstract

Malaria is an important health issue in French Guiana. Its principal mosquito vector in this region is *Anopheles darlingi*. Knowledge of the spatial distribution of this species is still very incomplete due to the extent of French Guiana and the difficulty to access most of the territory.

Species Distribution Modeling based on the maximal entropy procedure was used to predict the spatial distribution of *An. darlingi* using 39 presence sites.

The resulting model provided significantly high prediction performances (mean 10-fold cross-validated partial AUC and continuous Boyce index equal to, respectively, 1.11 – with a level of omission error of 20 % – and 0.42). The model also provided a habitat suitability map and environmental response curves in accordance with the known entomological situation.

Several environmental characteristics that had a positive correlation with the presence of *An. darlingi* were highlighted: non-permanent anthropogenic changes of the natural environment; the presence of roads and tracks; opening of the forest. Some geomorphological landforms and high altitude landscapes appear to be unsuitable for *An. darlingi*.

The Species Distribution Modeling was able to reliably predict the distribution of suitable habitats for *An. darlingi* in French Guiana. Results allowed completion of the knowledge of the spatial distribution of the principal malaria vector in this Amazonian region, and identification of the main factors that favor its presence. They should contribute to the definition of a necessary targeted vector control strategy in a malaria pre-elimination stage, and allow extrapolation of the acquired knowledge to other Amazonian or malaria-endemic contexts.

Keywords: Maxent, species distribution model, presence-only, *Anopheles darlingi*, sampling bias

Malaria is a public health issue in the Amazonian region, with major transmission foci depending on specific local characteristics associated with changing environmental and socio-demographic contexts. French Guiana is a French overseas territory with ~260,000 inhabitants. It remains one of the major malaria foci in the region, despite an improving epidemiological situation during the past ten years. The number of reported clinical cases has significantly dropped from 4,479 in 2005 to 434 in 2015 (Petit-Sinturel et al. 2016), and now corresponds to an incidence rate of two cases for 1,000 inhabitants for the whole territory, making it possible to target the pre-elimination of the disease in 2018 (Agence Régionale de Santé Guyane 2015). *Plasmodium vivax* is at present predominant and this species was responsible for 67% of the diagnosed cases of malaria in the territory in 2014, the others being mainly due to *Plasmodium falciparum* (Musset et al. 2014, Ardillon et al. 2015). However, this epidemiological situation is heterogeneous in space and time. In particular, a recrudescence of malaria cases is currently observed in the inland region (Saül, Cacao, and Régina) and eastern French Guiana (municipalities of Camopi and Saint-Georges-de-l'Oyapock), with a general incidence rate reaching 55.2 cases per 1,000 inhabitants in 2013 (Musset et al. 2014), likely due to the emergence and/or persistence of local foci of high malaria transmission (Berger et al. 2012, Musset et al. 2014). This Amazonian region, especially near the international borders, includes vulnerable populations. Some are hard-to-reach and have poor access to health services and treatment-seeking behaviors that may favor the development of resistance to antimalarial drugs (Musset et al. 2014, Wangdi et al. 2015). Uncontrolled areas of malaria transmission are also prevalent in illegal gold mining areas (Pommier de Santi et al. 2016a, Pommier de Santi et al. 2016b). The epidemiological situation remains quite unstable, and pre-elimination of malaria, corresponding to an incidence rate below one case for 1,000 inhabitants in any locality of French Guiana, remains a major challenge.

In this context, public health authorities must maintain control efforts while targeting them more precisely and objectively in time and space (Alimi et al. 2015). A map of malaria risk in French

Guiana is updated regularly by the regional unit of the French National Public Health Agency, based on the number of cases reported per locality and the data available on movements of human populations at risk, especially due to gold mining activities. This map is validated by the local expert committee of epidemic diseases (Comité d'Experts des Maladies à Caractère Épidémique, CEMCE), which brings together different experts of the disease in the region (from the Health Surveillance Agency, the Pasteur Institute of French Guiana, the Regional Unit of the National Public Health Agency, vector control services, hospitals, and other diagnosis and care centers, and the Defense Health Service in French Guiana). The lack of objective knowledge of several key factors, especially the spatiotemporal distribution of the main malaria vectors and human populations infected by *Plasmodium* and/or carrying gametocytes, makes such a map highly approximate.

Anopheles (Nyssorhynchus) darlingi Root (Diptera: Culicidae) is one of the most efficient malaria vectors in South America and is considered to be the primary malaria vector in French Guiana because of its anthropophilic behavior, natural infectability, high density, and sensitivity to *P. falciparum* (Girod et al. 2008, Hiwat et al. 2010, Fouque et al. 2010). Used entomological data collection for the entire territory, for the mapping of entomological risk indicators at the regional scale, is not feasible. French Guiana occupies a large territory (84,000 km²) which is mostly covered by rain forest (more than 80%) and highly inaccessible. Knowledge of the recent geographical distribution of *An. darlingi* is thus restricted to coastal areas, some villages along the international border rivers, and some illegal gold mining sites (Figure 1).

Species Distribution Modeling (SDM) offers an efficient solution to geographically extrapolate such knowledge to the entire territory (Pearson et al. 2007). Species Distribution Modeling produces maps of species habitat suitability by using known presence locations of the species and relevant environmental data. The use of SDM is thus encouraged to “improve and facilitate the development of alternative vector control strategies” (Alimi et al. 2015). Numerous SDM approaches are

proposed in the literature. Some of them, such as Maximum Entropy (Maxent; Phillips et al. 2006), Genetic Algorithm for Rule-Set Prediction (GARP; Stockwell 1999), Boosted Regression Trees (BRT; Friedman et al. 2000), Generalized linear and additive models (GLM and GAM; Guisan, et al. 2002), and Multivariate adaptive regression splines (MARS; Leathwick et al. 2005), exploit only species presence information, offering a significant advantage over methods that also require absence data. Indeed, absence data are often difficult to obtain. According to Peterson (2007) and Hirzel et al. (2002), absence can result from (1) the non-detection of the species in a suitable habitat, even if it is present, (2) the actual absence of the species for historical reasons, whereas the habitat is suitable, and (3) the true absence of the species and the unsuitability of the habitat. Comparative studies (Elith et al. 2006, Tognelli et al. 2009, Pearson et al. 2007, Hernandez et al. 2006, Wisz et al. 2008) show that Maxent is able to fit complex functions between habitat suitability and predictor variables, is the least sensitive to the size of the presence dataset, and tends to outperform other comparable methods when the dataset is small.

In this study, the mapping of the habitat suitability of *An. darlingi* at the scale of all of French Guiana was performed using the Maxent SDM approach. This work aims to provide reliable maps for improving malaria transmission risk mapping in French Guiana, and to identify the environmental factors and associated mechanisms that favor the presence of *An. darlingi*.

Materials and Methods

Study area

French Guiana (84,000 km²), a French overseas region located in South America, is separated from Suriname by the Maroni River and from Brazil by the Oyapock River and the Tumuc-Humac mountains. More than 80% of the territory is covered by rain forest. The country has an equatorial climate characterized by two annual dry seasons, from mid-August to mid-November and in March, and two wet seasons, from mid-April to mid-August and mid-November to February. The average annual rainfall reaches 4,000 mm and 2,000 mm in the wettest (north-east) and driest (north-west)

areas, respectively (Hammond 2005). The average monthly rainfall is >100 mm for the entire territory throughout the year, except for the three driest months: September, October, and November (Héritier, 2011). The average humidity is between 80% and 90%. The temperature is homogeneous over the entire territory throughout the year, with an average annual temperature of 26°C. The difference between the minimum and maximum daily temperature is more important than the annual variations. For example, in Maripasoula (on the border with Surinam) and Camopi (on the border with Brazil), the annual ranges of the minimum and maximum temperatures were, 4.3°C and 9.6°C (averages over the period 2001-2008), respectively, whereas the mean daily thermal amplitude was 9.8°C (average over the period 2001-2008; Météo-France, 2016). The population of ~260,000 inhabitants is unequally distributed throughout the territory. Approximately 90% of the population lives in the coastal area and most of the rest lives along the Maroni and the Oyapock rivers (Amerindians and Bush-Negroes). However many people live and/or transit through inland and remote areas of the territory (forestry workers, gold miners, and soldiers). According to many studies (Berger et al. 2012, Verret et al. 2006, Queyriaux et al. 2011, Hustache et al. 2007, Stefani et al. 2011, Pommier de Santi et al. 2016a), Amerindians, gold miners, and soldiers may be highly infected by malaria, whereas the areas in which they live and/or transit are those with the poorest knowledge of the presence and density of malaria vectors. It is thus of potential interest to consider the malaria risk, study the distribution of malaria vectors, and implement prevention and control actions over the entire territory of French Guiana.

Species Records

Presence sites of *An. darlingi* were provided by surveys of the Medical Entomology Unit of the Pasteur Institute of French Guiana and the Defense Health Service in French Guiana. *Culicidae* collections were performed using either human landing catches or traps (light traps or odor baited traps). Human landing catches consisted of exposing collector's lower leg and collecting landing mosquito with a mouth aspirator. Collectors were members of the Pasteur Institute or Defense

Health Services, they were aware of the risks associated with the method and had given their free consent. Malaria prophylaxis was proposed and information on the medication was explained. Light trap catches were performed with Center for Disease Control and Prevention (CDC) light traps, and odor baited catches were performed with Mosquito Magnet ® traps (Woodstream Corporation, Lititz, PA) baited with Octenol, a combination considered to be the best candidate for *Anopheles* surveillance in the region (Vezenegho et al., 2015).

Anopheles species were morphologically identified using taxonomic keys specific for the region (Floch and Abonnenc 1951, Faran and Linthicum 1981, Forattini 1962). Only *Culicidae* collections performed since the year 2000 were precisely geolocated by GPS coordinates and were used for the study (Figure 1). These data correspond to 74 capture sites for the family C*ulicidae*, and to 48 presence sites for the species *An. darlingi*.

The difficulty in accessing most of the French Guiana territory, and the priority given to the areas at risk of malaria transmission where many people live, led to a significant sampling bias with oversampling of the anthropized region of the territory, notably those easily accessible by roads (Figure 1).

Ecological knowledge and hypotheses

The presence of *An. darlingi* is linked to compositional and configurational features of the land cover and land use, as they partially determine breeding, feeding, and resting sites of the vector (Stefani et al. 2013). The natural environment for this vector in the Amazonian region includes floodable savanna, swamps (Girod et al. 2011, Zeilhofer et al. 2007), and flooded forest (Rozendaal 1992). Larvae are found along river edges, on flooded riverbanks, creeks, and pools formed near river-beds (Rozendaal 1992, Hiwat et al. 2010). Breeding sites are generally situated at low altitude (Mouchet 2004) and solely in freshwater, as *An. darlingi* is sensitive to salinity (Deane et al. 1948). Hydrological and geomorphological factors are responsible for the formation and destruction of *Anopheles* breeding sites (Smith et al. 2013).

Human activities, comprising deforestation and fish farming, also contribute to the creation of active breeding sites (Patz et al. 2000, Richard 1987, Stefani et al. 2013, Takken et al. 2005, Terrazas et al. 2015, Vittor et al. 2006, Vittor et al. 2009). Unpaved roads, tracks, and culverts form ideal breeding sites for *An. darlingi* in the Amazon region (Singer and Castro 2001). The presence of *An. darlingi* is also maintained by regular human presence due to its strong anthropophilic behavior. However, the presence and density of *An. darlingi* can either be favored or restricted depending on the type and intensity of the anthropogenic impacts. Stefani et al. (2013) systematically reviewed the literature and showed that all the studies describe the same mechanisms linking deforestation, land use, and the degree of urbanization with malaria transmission risk in the Amazonian region: opening the forest and maintaining a high degree of interaction between forested and deforested areas decreases the distance between feeding, breeding, and resting sites of *An. darlingi,* favoring the presence and high density of the vector (as well as a high probability of contact between humans and vectors); in contrast, intensifying deforestation and creating large urbanized and/or cultivated surfaces tends to decrease suitable habitat for *An. darlingi.* These two antagonistic consequences of human activities were considered in the SDM described here, by explicitly separating favorable and unfavorable factors in the environmental characterization.

The optimum temperature range for *An. darlingi* is between 20 and 30°C with a humidity of above 60% (Martens et al. 1995). Several studies established a minimal monthly rainfall threshold to designate suitable breeding habitats for *Anopheles* (reviewed in Smith et al. 2013). These values vary between 10 and 80 mm and need to be maintained for three or four months.

Environmental Variables

Environmental variables chosen as SDM inputs must characterize the ecological factors that influence the presence of *An. darlingi*, previously described. These factors are separated into three types: 1) natural environment features, associated with land cover, land use, and geomorphology for which the impact on the presence of *An. darlingi* depends on specific values or categorical classes;

2) anthropogenic activities that non-permanently alter the natural environment on a highly local scale and favor the presence of *An. darlingi*; 3) urbanization, corresponding to human presence and activities that permanently alter the natural environment over large areas and hinders the presence of the vector. Meteorological variables were not included in the model, because the temperature, rainfall, and humidity fall within the optimal ranges for presence of the species in French Guiana. Thus, these variables cannot significantly explain differences in the time average habitat suitability distribution over the year (this point is extensively discussed in the Discussion section).

Raw Geographic Data. Variables chosen as SDM inputs were derived from the following raw geographic data:

- Geomorphological landscape (*GLS*) and Geomorphological landforms (*GLF*) from the French Forest Office (ONF) (Guitet et al. 2013);
- Landscape types (*LS*) from the French Agricultural Research Centre for International Development (CIRAD) (Gond et al. 2011). This provides the distribution of landscape types in French Guiana, most being forested landscapes;
- Altitude (*ATL*) derived from the Digital Elevation Model provided by the Shuttle Radar Topography Mission (SRTM, spatial resolution: 30 meters) of the United States Aeronautics and Space Administration (NASA);
- Human footprint (*HFP*): An integrated human activity index that gives a general measure of the extent of expected threats on biodiversity, by assigning a score depending on the nature of the disturbance. It combines sublayers spatializing human population density, urban areas, legal and illegal mining sites, agriculture, forest settlements and camps, tourist camps, logged areas (forest activities), and potential hunting areas corresponding to a zone of two kilometers around roads, tracks and rivers, likely to be used by humans. The total disturbance score is the sum of all human activity scores (de Thoisy et al. 2010);
- Roads and tracks from the BD TOPO® database of the French Institute of Geographical

and Forestry Information (IGN).

Table 1 summarizes the main features of these raw geographic data.

Definition of Environmental Variables Used as Inputs for SDM. Several variables were extracted from the previously described raw data to better reflect the ecological knowledge and hypotheses mentioned above. The reference spatial resolution (pixel size) permitting the integration of all environmental layers was set to 1 by 1 km, *i.e.*, the coarsest resolution of the available layers, associated with the *LS* map.

The length of roads and tracks outside of urban areas (*ROADS*) was computed in the 1 km-cell grid from the BD TOPO® database.

The sublayers composing the *HFP* were first rasterized into 30-m grid cells, the smaller polygon of the *HFP* having a size of approximately 40 by 10 m. Distinct attributes were then extracted:

- The percentage of urbanization (*PER_URB*) within the 1 km grid cells;
- The percentage of urbanization within the eight neighbor cells of each urban cell (*PER_URB_NEIGH*), which permits distinguishing small from large urban areas. This layer was obtained for each 1 km-cell considered to be urban (*i.e.*, with $PER\_URB \geq 50\%$), by averaging the *PER_URB* values for the eight (1 km side) neighbor pixels;
- The human activities which non-permanently alter the natural environment (*HA*), by first summing the scores of the following sublayers: tourist and forest camps, mining activities and logged areas, hunting areas nearby rivers, and then, by computing the minimum, median, and maximum values within the 1-km grid cells.

The agriculture sublayer from *HFP* was not used because it covers only the coastal area. The population density sublayer was also excluded because it did not have sufficient level of detail. The sublayer of potential hunting areas near roads and tracks were not used to avoid duplication of the length of roads and tracks outside of urban areas computed previously.

For each 1-km grid cell, the majority class of the categorical variables *GLS* and *GLF*, and the

minimum, median, and maximum altitude (*ALT*) values were computed.

Eventually, some corrections of the *LS* layer were performed as it did not identify urban areas and did not distinguish flooded forests associated with freshwater from those of the coastal strip associated with brackish water (mangroves): *LS* cells with an *PER_URB* value greater than or equal to 50% were reclassified into a new *LS* class referred to as *Urban*; *LS* cells classified as *Flooded forest* and corresponding to mangroves according to the coastal land use map provided by the ONF (Office National des Forêts Direction Régionale de Guyane, 2013) were reclassified as *Mangrove*. The variable *PER_URB* was excluded from the input SDM variables, as the urban areas were mapped, and their extent quantified, by the corrected *LS* and *PER_URB_NEIGH* layers, respectively.

Table 1 lists and describes the environmental variables used to build the model.

Maxent Model Principle

Maxent is an SDM which requires environmental variables and species presence-only data. It is based on the principle of maximum entropy to estimate an (*a priori*) unknown probability distribution over the entire study area. This probability distribution assigns a value that is proportional to the probability of the presence of the species to each pixel of the study area. It is therefore interpreted as a habitat suitability index (HSI) across the study area (Phillips et al. 2006). The Maximum Entropy principle consists of approximating the unknown probability distribution by finding the one that maximizes entropy and satisfies the constraints imposed by the environmental features at the known sites of presence. Environmental features are a set of input environmental variables chosen according to their expected relevance for the studied taxon (Phillips et al. 2006, Elith et al. 2011). The constraints ensure that the environmental values expected under the approximated probability distribution are consistent with environmental information observed at the presence points.

In practice, the Maxent distribution is defined on a set of points called background points. These

points should reflect the available environmental conditions of the study area and are chosen by uniform random sampling. This approach assumes that the presence data are not biased and that environmental conditions are uniformly sampled (Yackulic et al. 2013). However, in practice, some areas are more intensively sampled than others, and environmental conditions are not uniformly distributed and may imply a strong sampling bias. Phillips et al. (2009) proposed selecting the background points with the same environmental bias as the presence dataset to correct the effect of this sampling bias.

Model Building and Evaluation

Eleven environmental variables and 48 *An. darlingi* presence points (their coordinates were in the table in supplementary material S1) are used as inputs for Maxent. Only one presence site was selected to build the model when more than one occurred in the same pixel. As a result, only 39 presence sites were actually used for building the model. Hinge and categorical features were selected for the environmental variables. A hinge feature provides a good compromise between simplicity and the quality of the approximation of the species response curves (Elith et al. 2011, Phillips and Dudík 2008).

In this study, the distribution of the background points was biased so that the selection bias corresponds to that of the sampling. The sampling bias was defined as the relative sampling effort in the environmental space, and was estimated by considering the capture locations of *Culicidae*, obtained using the same capture techniques and supposed to be subjected to the same sampling bias as the *An. darlingi* species. The details of the method to create the relative sampling effort map are described in supplementary material S2.

The model was computed using version 3.3.3k of Maxent. The recommended values derived from Phillips and Dudík (2008) concerning the regularization parameters and the background set size, were applied. Regularization parameter values were set to 0.25 and 0.5 for categorical and hinge features, respectively, and the size of the background was set to 10,000. The extrapolation option

was not selected to avoid making predictions in environmental domains in which the model was not trained. The model was fitted using the full data set and evaluated using a 10-fold cross-validation procedure. The Receiver Operating Characteristic (ROC) curves and the associated Areas Under the ROC Curve (AUC) were computed. This was completed by computing the mean partial AUC ratios (Peterson et al., 2008), consisting of the ratios of the partial AUCs of the model over the null AUC (corresponding to random prediction), for omission errors (*E*) of 20, 10, and 5%. The Continuous Boyce Index (CBI), considered to better adapted to presence-only models than the AUC (Hirzel et al., 2006), was also computed. The gain (regularized training gain) was also used to evaluate the performance of the model prediction. It is a measure of the likelihood of the sample, and indicates how much better the estimated distribution fits the presence points than the uniform distribution, which corresponds to a null gain (Yost et al. 2008).

The importance of each variable was estimated using two methods, a heuristic method and the jackknife test. The heuristic method computes the percentage contribution of each variable to the model. During the training process, the increase of the gain is due to the adjustment of the feature weights and this increase is assigned to the environmental variable that the feature depends on. The sum of these increases in gain indicates the percentage contribution of each environmental variable. The jackknife test evaluates the individual contribution of each variable to the model by estimating the difference of the gain when removing each variable, one by one, and when considering the given variable alone to build the model.

Results

The mean AUC was 0.93, and the mean partial AUC ratios were 1.08, 1.03, and 1.01 for maximum omission errors sets to 20, 10, and 5% respectively. The mean CBI was 0.356 and the mean gain was 3.14. Three variables cumulatively contributed >80% (Table 2): the length of roads and tracks outside of urban areas (*ROADS*), the percentage of urbanization of neighboring pixels (*PER_URB_NEIGH*), and landscape (*LS*). The maximum value of the human activities which non-

297 permanently alter the natural environment (*HA_MAX*), geomorphological landscape (*GLS*),
298 minimum altitude (*ALT_MIN*), and geomorphological landform (*GLF)* contributed moderately to
299 the model, with contributions of 6.84%, 5.35%, 1.34%, and 1.19%, respectively. The following
300 input variables contributed very little to the model: minimum and median values of human activities
301 which non-permanently alter the natural environment (*HA_MIN* and *HA_MED*; 0.35 and 0.24%,
302 respectively); and median and maximum values of altitude (*ALT_MED* and *ATL_MAX*; 0.69 and
303 0.06%, respectively).
304 The results of the Jackknife test confirmed the non-significant contribution of the input variables
305 *HA_MIN*, *HA_MED*, *ALT_MED*, and *ALT_MAX* (Table 2).
306 A second model was built using only the most highly contributing environmental variables:
307 *ROADS*, *LS*, *PER_URB_NEIGH*, *HA_MAX*, *GLS*, *GLF*, and *ALT_MIN*. The overall performance of
308 this simpler model was very similar to the previous one, with the mean AUC and partial AUC ratios
309 equal to 0.93 and 1.11, 1.05, and 1.03, respectively. The mean gain was equal to 3.19 and the mean
310 CBI was 0.421. Relative contributions of the input variables were also very similar (Table 3).
311 The response curves of the environmental variables are represented in Figures 2 and 3. They show
312 that the HSI is maximal when the *PER_URB_NEIGH* is below 8%. Above this value, the HSI
313 decreases progressively towards 0. The HSI increases as *ROADS* increases up to 7,000 meters,
314 reaches a plateau value, and then tends to decrease above 10,000 meters. Among all *LS* classes,
315 *Woodland savanna/dry forest* and *Open forest* contribute the most to the high HSI values. The
316 geomorphological landscape classes *Coastal flat plain* and *Plain with residual relief* and the
317 geomorphological landform classes *Small-size and flat wet land*, *Small-size rounded hill,* and
318 *Lowered half-orange relief* – a tropical relief type corresponding to a hill with convex flanks giving
319 to it a roughly hemispherical shape (George, 1972) and usually linked to flat or swampy lowlands
320 drained by streams with meanders – are also associated with high HSI values. The HSI is maximal
321 when *ALT* is ~0, with a rapid decrease as altitudes increase. The *HA_MAX* response curve presents a

more complicated profile. The HSI increases for *HA_MAX* values between 0 and 8, decreases until *HA_MAX* reaches 24, and then again increases as values continue to climb above 24.

The map of habitat suitability for *An. darlingi*, based on all the presence data for modeling, shows six main areas (A – F) with a high HSI and a seventh area (G) corresponding to an epidemiological interest area (see Figure 4). A qualitative analysis was performed to determine the characteristics of the environmental variables of the areas with high HSI values (Table 4).

In the coastal area (A), where 90% of the Guyanese population lives, the HSI tends to be higher along the main road representing the main traffic route in French Guiana. Focusing on the main urban areas, represented in Figure 5, the HSI values within the highly urbanized districts of Cayenne and Kourou (rectangles in Figure 5) are lower than those of the surrounding pixels that are not considered to be highly urbanized. A very high HSI was predicted within the urban area of Saint-Laurent-du-Maroni. However, none of the pixels characterizing this city has a *PER_URB_NEIGH* value higher than or equal to 50%. The high HSI values in areas B, D, E, and G are characterized by the environmental variables *ROADS*, *HA_MAX*, the classes *Open fores*t and *Mixed high and open forest*, and flat or moderately hilly terrain. The high HSI in areas C and F is essentially linked to *Open forest* and flat terrain.

The areas for which the model did not predict the HSI, due to the choice to not extrapolate to environmental domains not used to train the model, correspond to areas with an altitude higher than 400 meters. They represent a small number of pixels of the study area.

Discussion

The prediction performances of the model are excellent and significantly greater than those of the null model. The following discussion focuses on the ecological interpretation of the results and the methodological choices and alternatives.

Environmental Factors Explaining the Habitat Suitability

The geographic distribution of habitat suitability is consistent with existing knowledge of the

entomological situation despite the small number of presence points. The high HSI values can be explained by different environmental contexts depending on the geographical locations. In most areas (A, B, D, and E), the HSI values depend on human presence and activities, characterized by the environmental variables *HA_MAX* and *ROADS* (in areas D, E, and B, most roads are not paved and correspond mostly to tracks). The significantly positive correlation between the variable *ROADS* and the HSI confirms that road and track opening, accompanied by deforestation and pooling of rainwater at the roadside, may favor breeding sites (Singer and Castro 2001). The response curve for the variable *ROADS* (Figure 3) reaches a plateau above 7,000 meters of road per square kilometer and decreases thereafter. The decrease of the HSI at values above 7,000 meters suggests that the density of the road network leads to an improvement of the road quality (paved road eliminating culverts, adding sidewalks), thus limiting the availability of breeding and/or resting sites, in the same way as urbanization. Indeed, the response curve of the *PER_URB_NEIGH* variable confirms that highly urbanized areas provide a poorly suitable habitat for *An. darlingi* (Figure 3). Intensive urbanization implies concrete paving, the decrease or removal of green areas and forests, and consequently, the destruction of breeding and resting sites for *An. darlingi* (Stefani et al. 2013). This phenomenon is observed in the highly urbanized areas of Cayenne and Kourou (Figure 5). In contrast, Saint-Laurent-du-Maroni, the second largest urban area of French Guiana in terms of urbanization size and density, has high HSI values. In fact, unlike Cayenne and Kourou, this area is not considered to be highly urbanized using the criterion of this study (*PER_URB_NEIGH* $\geq$ 50%). However, the result for Saint-Laurent-du-Maroni seems unlikely because the presence of *An. darlingi* has not yet been reported in an urban area. Further field works could confirm the presence of this species in the city. The sensitivity of the model for the criterion that defines a highly urbanized area may also merit further study.

The values of *HA_MAX* in areas D and E were essentially associated with mining activity. In French Guiana, this activity is responsible for forest loss reaching 2,000 hectares per year (Office National

372 des Forêts Direction Régionale de Guyane, 2014). Between 2001 and 2013, Alvarez-Berríos and
373 Aide (2015) estimated that the largest forest loss due to gold mining in the tropical and subtropical
374 moist forest in South America was situated in the Guianan region including French Guiana. This
375 suggests that this activity, resulting in deforestation and creating sources of standing water such as
376 mining pits, combined with the presence of a large number of people, creates suitable conditions for
377 *An. darlingi*. The high HSI in these two areas is also explained by the *Mixed high and open forest*
378 landscape which is associated with human disturbance (Gond et al. 2011). Indeed, this landscape is
379 described as a forest environment linked to young or unstable vegetation mostly due to first stages
380 of anthropization. These results confirm the important role of human presence in the creation of
381 suitable habitats for *An. darlingi*, which is also consistent with the strong anthropophilic behavior of
382 this vector.
383 Some landscape types which are not directly associated with human presence or activities were also
384 associated with a HSI. The *Woodland savanna/dry forest* class appears to highly contribute to high
385 HSI values (Figure 2). It corresponds to the driest landscape in French Guiana (Gond et al. 2011),
386 but can be seasonally inundated due to its poor drainage, creating breeding sites (Rosa-Freitas et al.
387 2007). The high HSI values in this area are in accordance with previous studies (Vezenegho et al.
388 2015, Dusfour et al. 2013), which reported finding *An. darlingi* in the coastal savanna environments
389 of French Guiana. In uninhabited areas (zones F and C in Figure 4), a high HSI is associated with
390 the *Open forest* class (LS layer) and flat terrain. This LS class can be associated with different land
391 cover types in French Guiana (Gond et al. 2011) depending on the geographical location.
392 Consequently, this *LS* class may differentially affect *An. darlingi* habitat suitability. The *Open forest*
393 in area C mainly corresponds to wetlands (classified as *Flooded forest* according to the coastal land
394 use map provided by the Office National des Forêts Direction Régionale de Guyane, 2013), whereas
395 in area F, it corresponds to *Large surfaces of bamboo thicket and forbs*. *Anopheles darlingi* was
396 found in flooded forest; however, to our knowledge, no information is available concerning its

presence in large areas of bamboo thicket and forb*s*. The prediction in these areas should be taken with precaution as a more precise description of the habitats within *Open forest* class is required. Overall, this information highlights that natural environment could form highly suitable habitats despite the high anthropophily of *An. darlingi*.

Meteorological Variables

In this study, meteorological variables were not used to build the model. Temperatures fall within the optimal range for the species presence, and were considered to be geographically and temporally too homogeneous to explain differences in the spatial distribution of habitat suitability. Such a hypothesis is common in the Amazonian context. Olson et al. (2009) report that in their study region (Amazon basin), "monthly temperatures were between 24.6°C and 29.4°C (well within the range for optimal malaria transmission) for 95% of the observations," and consequently did not include temperatures in their model. In French Guiana, several studies also used rainfall data to study the intra-annual variations in *An. darlingi* density (Hiwat et al., 2010, Girod et al., 2011). The exclusion of rainfall data is more debatable, as rainfall clearly influenced the intra-annual density of *An. darlingi* in the study region (Hiwat et al., 2010, Girod et al., 2011, Vezenegho et al., 2015) even if the relationship was not systematically observed (Girod et al., 2011). The evidence for this impact on densities is that *An. darlingi* habitat suitability varies at an intra-annual scale, due to the alternation of dry and wet seasons. However, the entire study area is subject to this alternation. Moreover, given the high density of the French Guiana hydrological network and that the driest area (north-west) still receives 2,000 mm a year, it can be reasonably assumed that *An. darlingi* can find suitable conditions within the entire territory throughout most of the year. In French Guiana, the geomorphological landscape highly influences the availability of breeding sites, and therefore their spatial distribution, whereas the rainfall quantities influence the intra-annual variations of *An. darlingi* densities. As a consequence, on an average over the year, we assume that the significant factor influencing the distribution of habitat suitability is not the quantity of rainfall, but the

capacity of the landscape to provide suitable breeding sites when it rains.

Model Parametrization

The model was run by using the regularization parameter values and background set size recommended by Phillips and Dudík (2008), instead of those determined from specific experiments, as suggested by Merow et al. (2013). Phillips and Dudík (2008) tested a set of regularization parameter values with 48 species datasets that contained 11 to 13 environmental variables and a small number of categorical variables (1-3, as they considered discrete ordinal variables to be categorical). Nine of these datasets contained between 30 and 60 occurrences. The characteristics of the dataset exploited in our study (39 occurrence records; 13 and seven environmental variables including three categorical ones) are assumed to be quite similar of those of the datasets used by Phillips and Dudík (2008). We thus assumed that the pseudo-optimal parameters proposed by Phillips and Dudík (2008) could be confidently used in our study. Similarly, the background size was set to 10,000 based on the tests realized by Phillips and Dudík (2008), with 226 species and a median number of 57 presence sites. Better prediction performance may have been obtained by tuning the regularization values and background size and adding input environmental variables and features. However, the risk would have been to favor overfitting to the detriment of the bio-ecological interpretation of the model (see for example Merow et al., 2013). According to the entomologists who participated in the study, the model appears to be a good compromise between overfitting (that would have predicted suitable areas near occurrence points only) and being too general (that would have predicted suitable areas in too many environmental contexts for which the specialists have no species presence evidence).

Correction of the sampling bias effect

In this study, the effect of sampling bias was corrected by selecting background points with the same environmental bias as the sampled points. This approach appeared to be useful when applied to *An. darlingi* in French Guiana. Without a bias effect correction, the model predicted very high

447 HSI values in highly urbanized areas whereas these areas are known to be unsuitable for this vector
448 (see above). The biased background set is more concentrated around the sampled points (in the
449 environmental space) than the uniform random background, and is not likely to include
450 environmental conditions that are highly dissimilar to those encountered at the sampled points. As a
451 result, environmental conditions highly dissimilar to those of the sampled points can be subjected to
452 extrapolation, which may lead to erroneous habitat suitability predictions and bio-ecological
453 interpretations. This justifies not using the extrapolation option for modeling. The predicted HSI
454 map from the model with a biased background contains several excluded areas, whereas that of the
455 model with a uniform random background does not. Excluded areas correspond to high altitude
456 areas which are unsuitable for *An. darlingi* (Mouchet 2004).
457 When using a uniform random background, the three most contributive variables (cumulative
458 contribution equal to 85.5%) were all directly linked to human presence and territory accessibility
459 (*ROADS*: 38.1%, *PER_URB_NEIGH*: 34.9%, and *HA_MAX*: 12.5%). Thus, apart from urban areas,
460 high HSI values were associated with high *HA_MAX* and *ROADS* values. However, when
461 correcting the sampling bias effect, the *Landscape* (*LS*) variable was the second most contributive
462 variable (14.1%), the ROADS variable contribution increased to 62.6%, and the
463 PER_URB_NEIGH variable contribution decreased to 11.1% (see Table 3).
464 From a quantitative point of view, the two approaches (with and without applying the correction of
465 the sampling bias effect) resulted in identical AUC and partial AUC ratios. However, the regularized
466 gain and the CBI were lower without correction, with values equal to 2.81 (vs. 3.18) and 0.284 (vs.
467 0.421), respectively.
468 Thus, correction of the sampling bias effect gave better results: both more consistent with
469 knowledge from the field and more accurate in terms of prediction. The fact that the contributions
470 of the *LS* and *HA_MAX* variables respectively increased and decreased with the use of the biased
471 background, tends to show that the correction method actually manages to counterbalance the over-

representation of inhabited areas (cities, villages, and gold mining areas) in the sampled data. However, the very high contribution of the variable *ROADS* may be a residual effect of sampling bias, as sampling is essentially performed in the vicinity of accessible roads and tracks. Further studies are necessary to objectively and quantitatively assess the actual performance of the proposed methodology for correcting the effect of sampling bias.

Habitat Suitability and Malaria in French Guiana

Alimi et al. (2015) highlighted the utility of SDMs for gaining a better understanding of the geographical range and distribution of vectors for eliminating malaria and preventing outbreaks. The coastal strip in French Guiana is generally malaria free, although some cases resulting from local transmission are regularly diagnosed (Ardillon et al. 2015). This study, as well as that of Vezenegho et al. (2015), shows that the savanna in French Guiana may be highly suitable for *An. darlingi*. In the forest, Pommier de Santi et al. (2016c) found a link between mining, malaria cases, and the presence of *An. darlingi*. Indeed, >74% of malaria cases in French army soldiers were associated with operations to counteract illegal gold mining (Pommier de Santi et al. 2016a). According to the results of the present study, some areas associated with intense gold mining activity, known to be malaria transmission foci, are not necessarily associated with very high HSI values. In the village of Camopi, the annual malaria prevalence was 70% for children younger than seven years of age between 2000 and 2002 (Carme et al. 2005), reaching 100% in 2006 (Hustache et al. 2007). However, only some pixels on the border of the Camopi and Oyapock rivers have high values on the HSI map (area G in Figure 4). This is consistent with the study of Girod et al. (2011), which showed that the number of *An. darlingi* caught in this village was very low relative to the incidence of malaria cases. These findings collectively highlight two important points. First, the HSI map shown in Figure 4 does not correspond to a map of malaria transmission risk. Transmission risk depends on many factors that were not taken into account here, such as the parasitic charge and immunological status of the local population, compositional and

configurational features of the landscape (Stefani et al, 2013; Li et al, 2016), and behavioral factors. Second, this highlights that malaria transmission can occur in areas where there is a very low density of *An. darlingi.* This may be due to the presence of other *Anopheles* species such as *An. (Nys.) nuneztovar*i Galbaldón*, An. (Nys.) oswaldo*i Peryassú*, An. (Nys.) intermedi*us Peryassú, *An. (Nys.) marajoara* Galvão and Damasceno, or *An. (Nys.) ininii* Sénevet and Abonnenc (Diptera: Culicidae), already known to be naturally infected with *Plasmodium* species and/or described as efficient secondary malaria vectors (Dusfour et al. 2012, Pommier de Santi et al, 2016c).

Environmental Characterization

A significant limitation of this study was the spatial resolution of the environmental data. Capture campaigns are generally carried out at a local scale (villages or camps; Vezenegho et al. 2015, Dusfour et al. 2013). The spatial resolution of the study was not sufficient to take into account the heterogeneity of the environment at the capture scale. The use of environmental data with higher spatial resolution, such as the canopy height estimation from Fayad et al. (2014) or finer characterization of the land cover could improve future studies. However, these data are not consistently available across the entire territory.

In conclusion, the results of this study help to complete our knowledge on the spatial distribution of the principal malaria vector in this Amazonian region, and to identify the main factors that favor its presence. These results can be exploited to define the necessary targeted vector control strategies in a malaria pre-elimination context, and to extrapolate the acquired knowledge to other Amazonian contexts. They also suggest areas that need to be targeted to complete the field knowledge, validate the prediction and strengthen the model. Eventually, these proposed methodological developments can be applied to other species, including other disease vectors.

Acknowledgements

This study was funded by the Fonds Social Européen (FSE), Centre National d'Etudes Spatiales

(CNES), and Collectivité Territoriale de Guyane. Financial support was partially provided by the "Investissement d'Avenir" grants managed by the Agence Nationale de la Recherche (Center for the study of Biodiversity in Amazonia, ANR-10-LABX-0025) and by the GAPAM-Sentinela project of the Franco-brazilian scientific and academic cooperation program Guyamazon (funds: IRD, CIRAD, French Embassy in Brazil, Territorial Collectivity of French Guiana, Brazilian State-level research agencies of Amapá, Amazonas and Maranhão).

References cited

Agence Régionale de Santé Guyane. 2015. Plan de lutte contre le paludisme en Guyane 2015-2018. Agence Régionale de Santé Guyane, Cayenne, France.

Alimi, T.O., D. O. Fuller, M. L. Quinones, R-D. Xue, S. V. Herrera, M. Arevalo-Herrera, J. N. Ulrich, W. A. Qualls, and J. C. Beier. 2015. Prospects and recommendations for risk mapping to improve strategies for effective malaria vector control interventions in Latin America. Malar. J. 14:519.

Alvarez-Berríos, N. L., and T. M. Aide. 2015. Corrigendum: Global demand for gold is another threat for tropical forests (2014 Environ. Res. Lett. 10: 014006). Environ. Res. Lett. 10: 29501.

Ardillon, V., L. Carvalho, C. Prince, P. Abboud, and F. Djossou. 2015. Bilans 2013 et 2014 de la situation du paludisme en Guyane. Bulletin de Veille Sanitaire Antilles-Guyane. 1: 16-20.

Berger, F., C. Flamand, L. Musset, F. Djossou, J. Rosine, M.-A. Sanquer, I. Dusfour, E. Legrand, V. Ardillon, P. Rabarison, C. Grenier and R. Girod. 2012. Investigation of a sudden malaria outbreak in the isolated Amazonian village of Saul, French Guiana, January-April 2009. Am. J. Trop. Med. Hyg. 86: 591–597.

Carme, B., J. Lecat, and P. Lefebvre. 2005. [Malaria in an outbreak zone in Oyapock (French Guiana): incidence of malaria attacks in the American Indian population of Camopi] (in French) Médecine tropicale 65: 149–154.

Deane, L. M., O. R. Causey, and M. P. Deane. 1948. Notas sobre a distribuição ea biologia dos

547 anofelinos das regiões nordestina e Amazônica do Brasil. Revista do Serviço Especial de Saúde
548 Pública 1: 827–965.
549 **de Thoisy, B., C. Richard-Hansen, B. Goguillon, P. Joubert, J. Obstancias, P. Winterton, and**
550 **S. Brosse. 2010.** Rapid evaluation of threats to biodiversity: human footprint score and large
551 vertebrate species responses in French Guiana. Biodivers. Conserv. 19: 1567–84.
552 **Dusfour, I., J. Issaly, R. Carinci, P. Gaborit, and R. Girod. 2012.** Incrimination of *Anopheles*
553 *(Anopheles) intermedius* Peryassú, *An. (Nyssorhynchus) nuneztovari* Gabaldón, *An. (Nys.)*
554 *oswaldoi* Peryassú as natural vectors of *Plasmodium falciparum* in French Guiana. Memórias Do
555 Instituto Oswaldo Cruz 107: 429–432.
556 **Dusfour, I., R. Carinci, J. Issaly, P. Gaborit, and R. Girod. 2013.** A survey of adult *Anophelines*
557 in French Guiana: enhanced descriptions of species distribution and biting responses. J. Vector
558 Ecol. 38: 203–209.
559 **Elith, J., C. H. Graham, R. P Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F.**
560 **Huettmann, J. R. Leathwick, A. Lehmann, et al. 2006.** Novel methods improve prediction of
561 species' distributions from occurrence data. Ecography 29: 129–151.
562 **Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011**. A statistical
563 explanation of MaxEnt for ecologists: statistical explanation of MaxEnt. Divers. Distrib. 17: 43–
564 57.
565 **Faran, M. E., and K. J. Linthicum. 1981.** A handbook of the Amazonian species of *Anopheles*
566 *(Nyssorhynchus)* (Diptera: Culicidae). Mosq. Syst. 13: 1-81
567 **Fayad, I., N. Baghdadi, J-S. Bailly, N. Barbier, V. Gond, M. Hajj, F. Fabre, and B. Bourgine.**
568 **2014.** Canopy height estimation in French Guiana with LiDAR ICESat/GLAS data using
569 principal component analysis and Random Forest regressions. Rem. Sens. 6: 11883–11914.
570 **Floch, H., and E. Abonnenc. 1951.** Anophèles de la Guyane française. Archives de l'Institut
571 Pasteur de la Guyane et du territoire de l'Inini. 236:1-92

572 **Forattini, O. P. 1962.** Entomologia Médica : Vol. 1. Parte Geral, Diptera, Anophelini. Faculdade de
573 Higiene e Saúde Pública, Departamento de Parasitologia, Universidade de São Paulo, São Paulo,
574 Brazil.
575 **Fouque, F., P. Gaborit, R. Carinci, J. Issaly, and R. Girod. 2010.** Annual variations in the
576 number of malaria cases related to two different patterns of *Anopheles darlingi* transmission
577 potential in the Maroni area of French Guiana. Malar. J. 9: 80.
578 **Friedman, J., T. Hastie, and R. Tibshirani. 2000.** Additive Logistic Regression: a Statistical View
579 of Boosting. Ann. Stat. 28: 337–407.
580 **George, P. 1972.** Dictionnaire de la géographie. Presse Universitaire de France, Paris, France.
581 **Girod, R., P. Gaborit, R. Carinci, J. Issaly, and F. Fouque. 2008.** *Anopheles darlingi* bionomics
582 and transmission of *Plasmodium falciparum*, *Plasmodium vivax* and *Plasmodium malariae* in
583 Amerindian villages of the Upper-Maroni Amazonian forest, French Guiana. Memórias Do
584 Instituto Oswaldo Cruz 103: 702–710.
585 **Girod, R., E. Roux, F. Berger, A. Stefani, P. Gaborit, R. Carinci, J. Issaly, B. Carme, and I.**
586 **Dusfour. 2011.** Unravelling the relationships between *Anopheles darlingi* (Diptera: Culicidae)
587 densities, environmental factors and malaria incidence: understanding the variable patterns of
588 malarial transmission in French Guiana (South America). Ann. Trop. Med. Parasitol. 105: 107–
589 122.
590 **Gond, V., V. Freycon, J-F. Molino, O. Brunaux, F. Ingrassia, P. Joubert, J-F. Pekel, M-F.**
591 **Prévost, V. Thierron, P-J. Trombe, and D. Sabatier. 2011.** Broad-scale spatial pattern of forest
592 landscape types in the Guiana Shield. Int. J. Appl. Earth Obs. Geoinf. 13: 357–367.
593 **Guisan, A., T. C. Edwards, and T. Hastie. 2002.** Generalized linear and generalized additive
594 models in studies of species distributions: setting the scene. Ecol. Model. 157: 89–100.
595 **Guitet, S., J-F. Cornu, O. Brunaux, J. Betbeder, J-M. Carozza, and C. Richard-Hansen. 2013.**
596 Landform and landscape mapping, French Guiana (South America). J. Maps. 9: 325–35.

Hammond, D. S. 2005. Tropical forests of the Guiana Shield: ancient forests in a modern world. CABI publishing, Wallingford, England.

Héritier, P. 2011. Le climat guyanais ; petit atlas climatique de la Guyane française. Météo France, Cayenne, France.

Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29: 773–785. **Hirzel, A. H., J. Hausser, D. Chessel, and N. Perrin. 2002.** Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83: 2027–2036.

Hirzel, A. H., G. Le Lay, V. Helfer, C. Randin, and A. Guisan. 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Model. 199: 142–152.

Hiwat, H., J. Issaly, P. Gaborit, A. Somai, A. Samjhawan, P. Sardjoe, T. Soekhoe, and R. Girod. 2010. Behavioral heterogeneity of *Anopheles darlingi* (Diptera: Culicidae) and malaria transmission dynamics along the Maroni River, Suriname, French Guiana. Trans. R. Soc. Trop. Med. Hyg. 104: 207–213.

Hustache, S., M. Nacher, F. Djossou, and B. Carme. 2007. Malaria risk factors in Amerindian children in French Guiana. Am. J. Trop. Med. Hyg. 76: 619–625.

Leathwick, J. R., D. Rowe, J. Richardson, J. Elith, and T. Hastie. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshw. Biol. 50: 2034–2052.

Li, Z., E. Roux, N. Dessay, R. Girod, A. Stefani, M. Nacher, A. Moiret, and F. Seyler. 2016. Mapping a knowledge-based malaria hazard index related to landscape using remote sensing: application to the cross-border area between French Guiana and Brazil. Remote Sens. 8: 1-22.

Martens, W. J., L. W. Niessen, J. Rotmans, T. H. Jetten, and A. J. McMichael. 1995. Potential impact of global climate change on malaria risk. Environ. Health Perspect. 103: 458–64.

622 **Météo-France. 2016**. Données pluviométriques disponibles au 01/01/2016
623 (http://pluiesextremes.meteo.fr/guyane/IMG/sipex_pdf/carte_reseau_dep973.pdf). Last accessed
624 the 7 Oct. 2016.
625 **Merow, C., M. J. Smith, and J. A. Silander. 2013.** A Practical guide to MaxEnt for modeling
626 species' distributions: what it does, and why inputs and settings matter. Ecography 36: 1058–
627 1069.
628 **Musset, L., S. Pelleau, R. Girod, V. Ardillon, L. Carvalho, I. Dusfour, M. SM Gomes, F.**
629 **Djossou, and E. Legrand. 2014.** Malaria on the Guiana Shield: a review of the situation in
630 French Guiana. Memórias Do Instituto Oswaldo Cruz 109: 525–33.
631 **Mouchet, J. 2004.** Biodiversité du paludisme dans le monde. John Libbey Eurotext, Montrouge,
632 France
633 **Office National des Forêts Direction Régionale de Guyane. 2013.** PROJET "EXPERTISE
634 LITTORAL 2011" Occupation du sol et sa dynamique sur la bande côtière de la Guyane de 2005
635 à 2011. Office National des Forêts et le Ministère de l'Agriculture, de l'Agroalimentaire et de la
636 Forêt, Cayenne, France.
637 **Office National des Forêts Direction Régionale de Guyane. 2014.** Rapport d'activité 2013.
638 Office National des Forêts et le Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt,
639 Cayenne, France.
640 **Olson, S. H., R. Gangnon, E. Elguero, L. Durieux, J. F. Guégan, J. A. Foley, and J. A. Patz.**
641 **2009.** Links between climate, malaria, and wetlands in the Amazon Basin. Emerg. Infect. Dis.15:
642 659-662.
643 **Patz, J. A., T. K. Graczyk, N. Geller, and A. Y. Vittor. 2000.** Effects of environmental change on
644 emerging parasitic diseases. Int. J. Parasitol. 30: 1395–1405.
645 **Pearson, R. G., C. J. Raxworthy, M. Nakamura, and A. Townsend Peterson. 2007.** Predicting
646 species distributions from small numbers of occurrence records: a test case using cryptic geckos

in Madagascar: Predicting species distributions with low sample sizes. J. Biogeogr. 34: 102–117.

Peterson, A. T. 2007. Ecological niche modelling and understanding the geography of disease transmission. Vet. Ital. 43: 393–400.

Peterson, A. T., M. Papeş, and J. Soberón. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol. Model. 213: 63–72.

Petit-Sinturel, M., L. Carvalho, A. Andrieu, C. Prince, P. Abboud, F. Djossou, and V. Ardillon. 2016. Situation du paludisme en Guyane française en 2015. Bulletin de Veille Sanitaire Antilles-Guyane. 2: 6-10.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190: 231–259.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31: 161–175.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19: 181–197.

Pommier de Santi, V., A. Dia, A. Adde, G. Hyvert, J. Galant, M. Mazevet, C. Nguyen, S. B. Vezenegho, I. Dusfour, R. Girod, and S. Briolant. 2016a. Malaria in French Guiana Linked to Illegal Gold Mining. Emerging Infect. Dis. 22: 344–346.

Pommier de Santi, V., F. Djossou, N. Barthes, H. Bogreau, G. Hyvert, C. Nguyen, S. Pelleau, E. Legrand, L. Musset, M. Nacher, and S. Briolant. 2016b. Malaria hyperendemicity and risk for Artemisinin resistance among illegal gold miners, French Guiana. Emerging Infect. Dis. 22: 903–906

Pommier de Santi, V., R. Girod, M. Mura, A. Dia, S. Briolant, F. Djossou, I. Dusfour, A. Mendebil, F. Simon, X. Deparis, and F. Pagès. 2016c. Epidemiological and entomological studies of a malaria outbreak among French armed forces deployed at illegal gold mining sites

reveal new aspects of the disease's transmission in French Guiana. Malar. J. 15: 35.

Queyriaux, B., G. Texier, L. Ollivier, L. Galoisy-Guibal, R. Michel, and J. B. Meynard. 2011. *Plasmodium vivax* malaria among military personnel, French Guiana, 1998–2008. Emerging Infect. Dis. 17 : 1280–1282.

Richard, A. 1987. Le Paludisme en forêt, pp. 249–250. *In* Connaissance Du Milieu Amazonien, 15-16 October 1985, Paris, France. ORSTOM, Paris, France. **Rosa-Freitas, M. G., P. Tsouris, A. T. Peterson, N. A. Honório, F. S. M. de Barros, D. B. de Aguiar, H. C. Gurgel, M. de Arruda, S. D. Vasconcelos, and J. F. Luitgards-Moura. 2007.** An ecoregional classification for the state of Roraima, Brazil. The importance of landscape in malaria biology. Memórias Do Instituto Oswaldo Cruz. 102: 349–358.

Rozendaal, J. A. 1992. Relations between *Anopheles darlingi* breeding habitats, rainfall, river level and malaria transmission rates in the rain forest of Suriname. Med. Vet. Entomol. 6: 16–22.

Singer, B. H., and M. C. Castro. 2001. Agricultural colonization and malaria on the Amazon frontier. Ann. N. Y. Acad. Sci. 954: 184–222.

Smith, M. W., M. G. Macklin, and C. J. Thomas. 2013. Hydrological and geomorphological controls of malaria transmission. Earth-Sci. Rev. 116: 109–127.

Stefani, A., E. Roux, J. M. Fotsing, and B. Carme. 2011. Studying relationships between environment and malaria incidence in Camopi (French Guiana) through the objective selection of buffer-based landscape characterizations. Int. J. Health Geogr. 10: 1-13.

Stefani, A., I. Dusfour, A. P. S. A. Corrêa, M. C. B. Cruz, N. Dessay, A. K. R. Galardo, C. D. Galardo, R. Girod, M. S. M. Gomes, and H. Gurgel. 2013. Land cover, land use and malaria in the Amazon: a systematic literature review of studies using remotely sensed data. Malar. J. 12: 1–8.

Stockwell, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Sci. 13: 143–158.

Takken, W., P. D. R. Vilarinhos, P. Schneider, and F. Dos Santos. 2005. Effects of environmental change on malaria in the Amazon region of Brazil. Frontis 9: 113–123.

Terrazas, W., V. Sampaio, D. de Castro, R. C. Pinto, B. C. de Albuquerque, M. Sadahiro, R. dos Passos, and J. U. Braga. 2015. Deforestation, drainage network, indigenous status, and geographical differences of malaria in the state of Amazonas. Malar. J. 14: 379.

Tognelli, M. F., S. A. Roig-Juñent, A. E. Marvaldi, G. E. Flores, and J. M. Lobo. 2009. An evaluation of methods for modelling distribution of Patagonian insects. Revista Chilena de Historia Natural 82: 347-360.

Verret, C., B. Cabianca, R. Haus-Cheymol, J-J. Lafille, G. Loran-Haranqui, and A. Spiegel. 2006. Malaria outbreak in troops returning from French Guiana. Emerg. Infect. Dis. 12: 1794–1795.

Vezenegho S.B., R. Carinci, P. Gaborit, J. Issaly, I. Dusfour, S. Briolant and R. Girod. 2015. *Anopheles darlingi* (Diptera: Culicidae) dynamics in relation to meteorological data in a cattle farm located in the coastal region of French Guiana: advantage of Mosquito Magnet trap. Environ. Entomol. 44: 454-462.

Vittor, A. Y., R. H. Gilman, J. Tielsch, G. Glass, T. I. M. Shields, W. Lozano, V. Pinedo-Cancino, and J. A. Patz. 2006. The effect of deforestation on the human-biting rate of *Anopheles darlingi*, the primary vector of falciparum malaria in the Peruvian Amazon. Am. J. Trop. Med. Hyg. 74: 3–11.

Vittor, A. Y., W. Pan, R. H. Gilman, J. Tielsch, G. Glass, T. Shields, W. Sánchez-Lozano, V. V. Pinedo, E. Salas-Cobos, and S. Flores. 2009. Linking deforestation to malaria in the Amazon: characterization of the breeding habitat of the principal malaria vector, *Anopheles darlingi*. Am. J. Trop. Med. Hyg. 81: 5-12.

Wangdi, K., M. L. Gatton, G. C. Kelly, and A. C. A. Clements. 2015. Cross-Border Malaria: A Major Obstacle for Malaria Elimination. Adv. Parasitol., 89:79–107.

722 **Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, A. Guisan, and NCEAS**
723 **Predicting Species Distributions Working Group. 2008.** Effects of sample size on the
724 performance of species distribution models. Divers. Distrib. 14: 763–773.
725 **Yackulic, C. B., R. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H. Campbell Grant,**
726 **and S. Veran. 2013.** Presence-only modelling using MAXENT: when can we trust the
727 inferences? Methods Ecol. Evol. 4: 236–243.
728 **Yost, A. C., S. L. Petersen, M. Gregg, and R. Miller. 2008.** Predictive modeling and mapping
729 sage grouse (*Centrocercus Urophasianus*) nesting habitat using Maximum Entropy and a long-
730 term dataset from Southern Oregon. Ecol. Inform. 3: 375-386.
731 **Zeilhofer, P., E. Santos, A. L. M. Ribeiro, R. D. Miyazaki, and M. Santos. 2007.** Habitat
732 suitability mapping of *Anopheles darlingi* in the surroundings of the Manso hydropower plant
733 reservoir, Mato Grosso, Central Brazil. Int. J. Health Geogr. 6: 1-14.

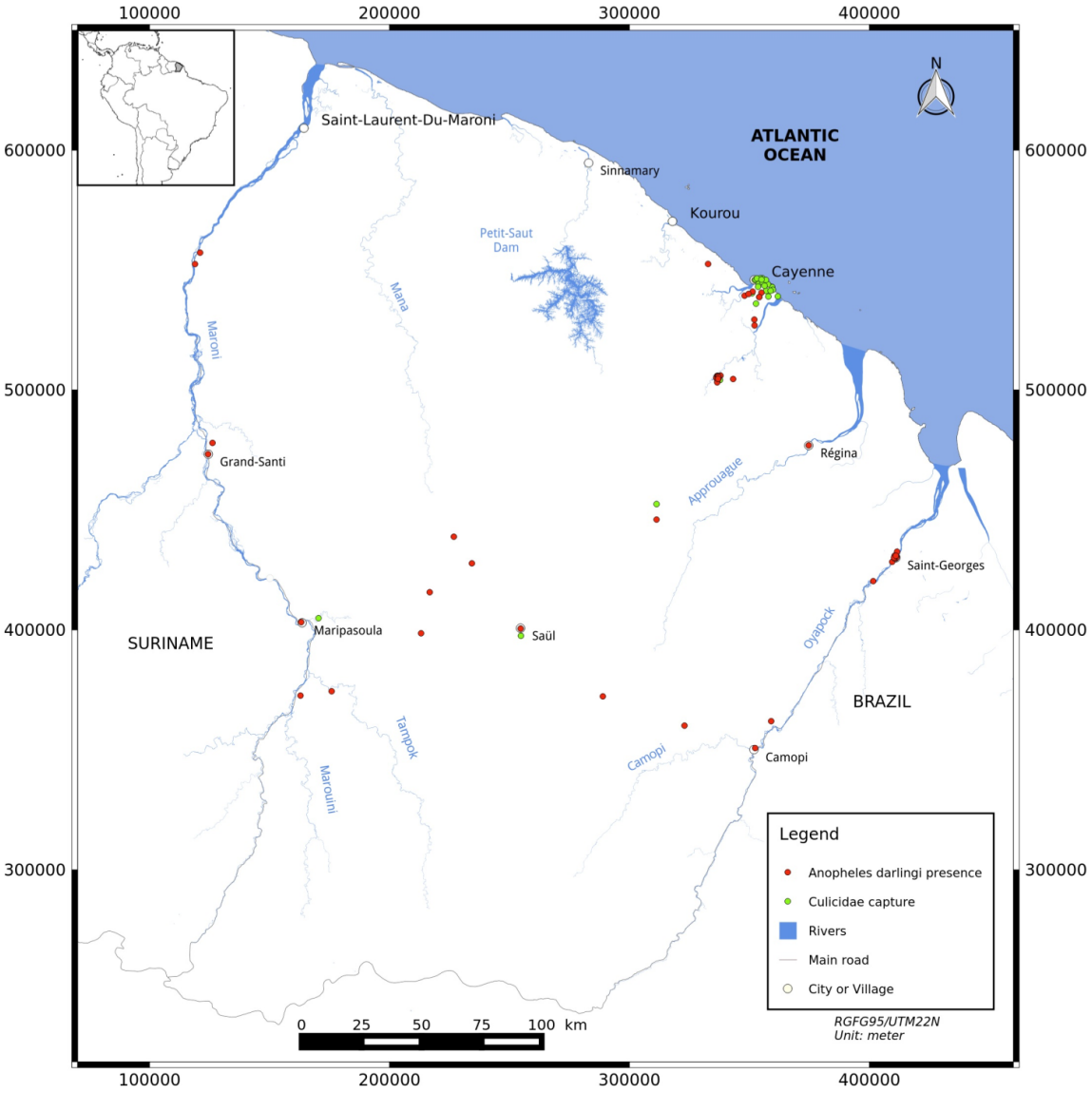


Figure 1. *Culicidae* capture points and *Anopheles darlingi* presence points (from 2000 to 2013).

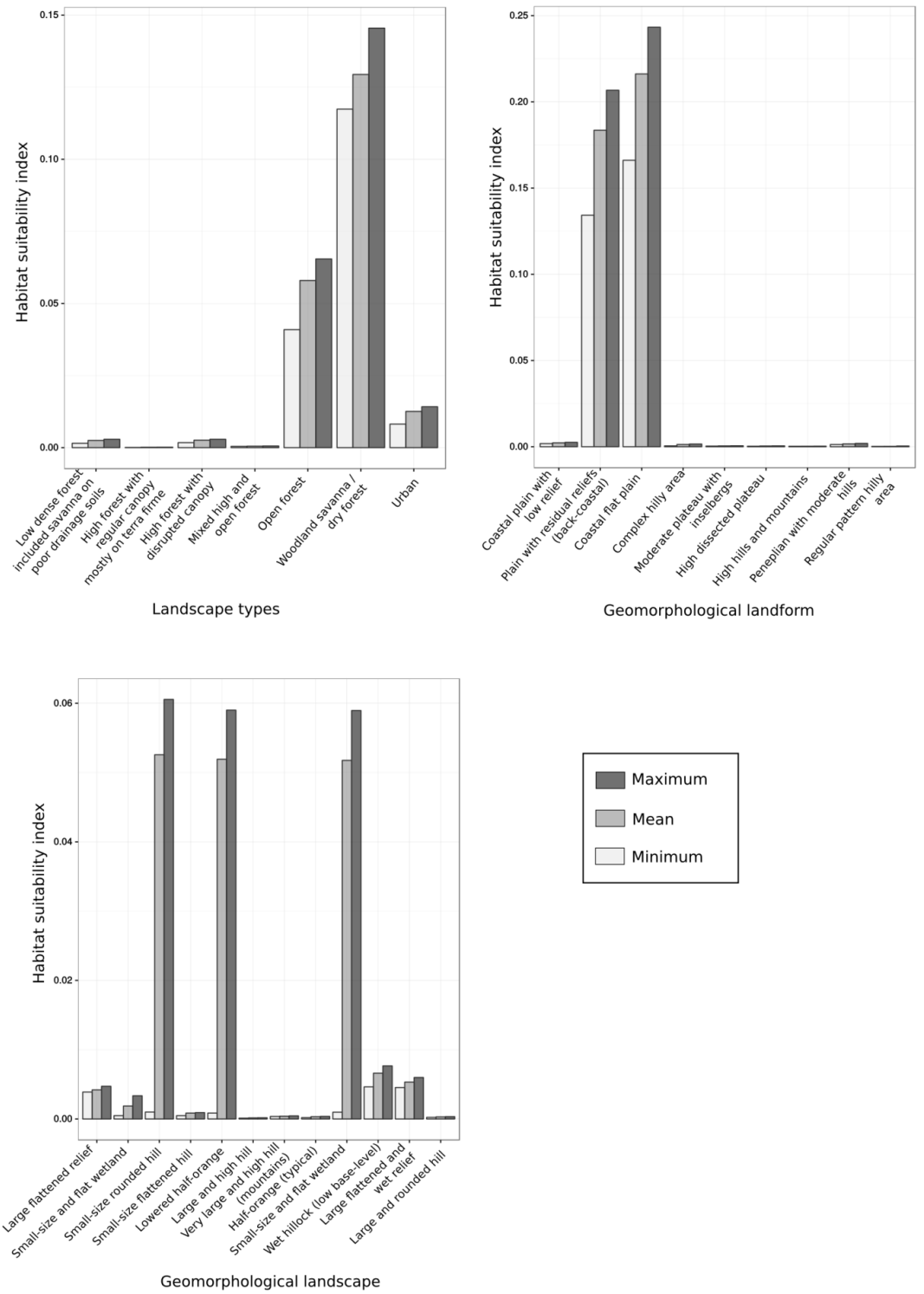


Figure 2. Response curves of categorical environmental variables.

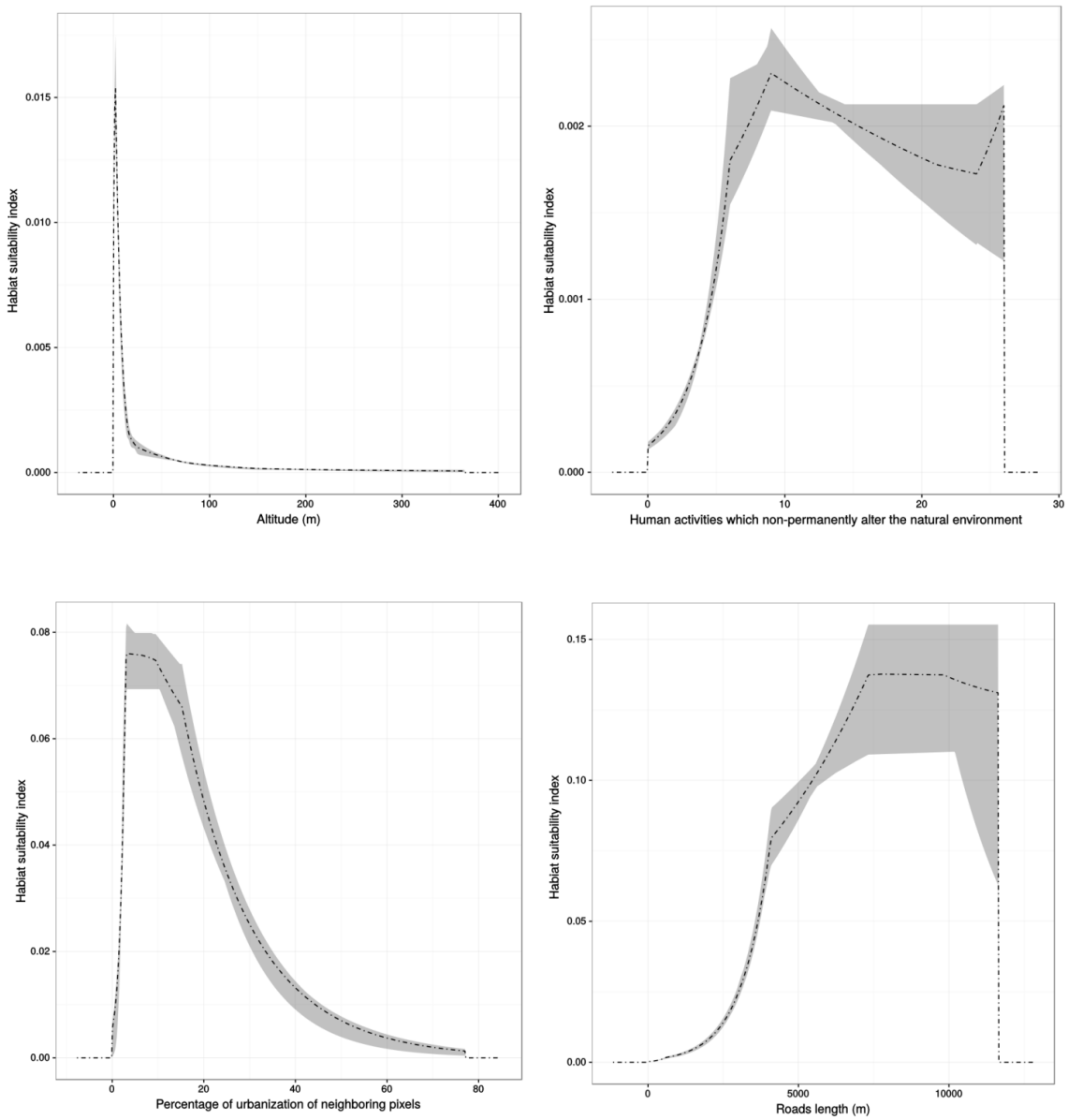


Figure 3. Response curves of numerical environmental variables. Dashed lines show the mean values and the grey regions represent the interval between the maximum and minimum values.

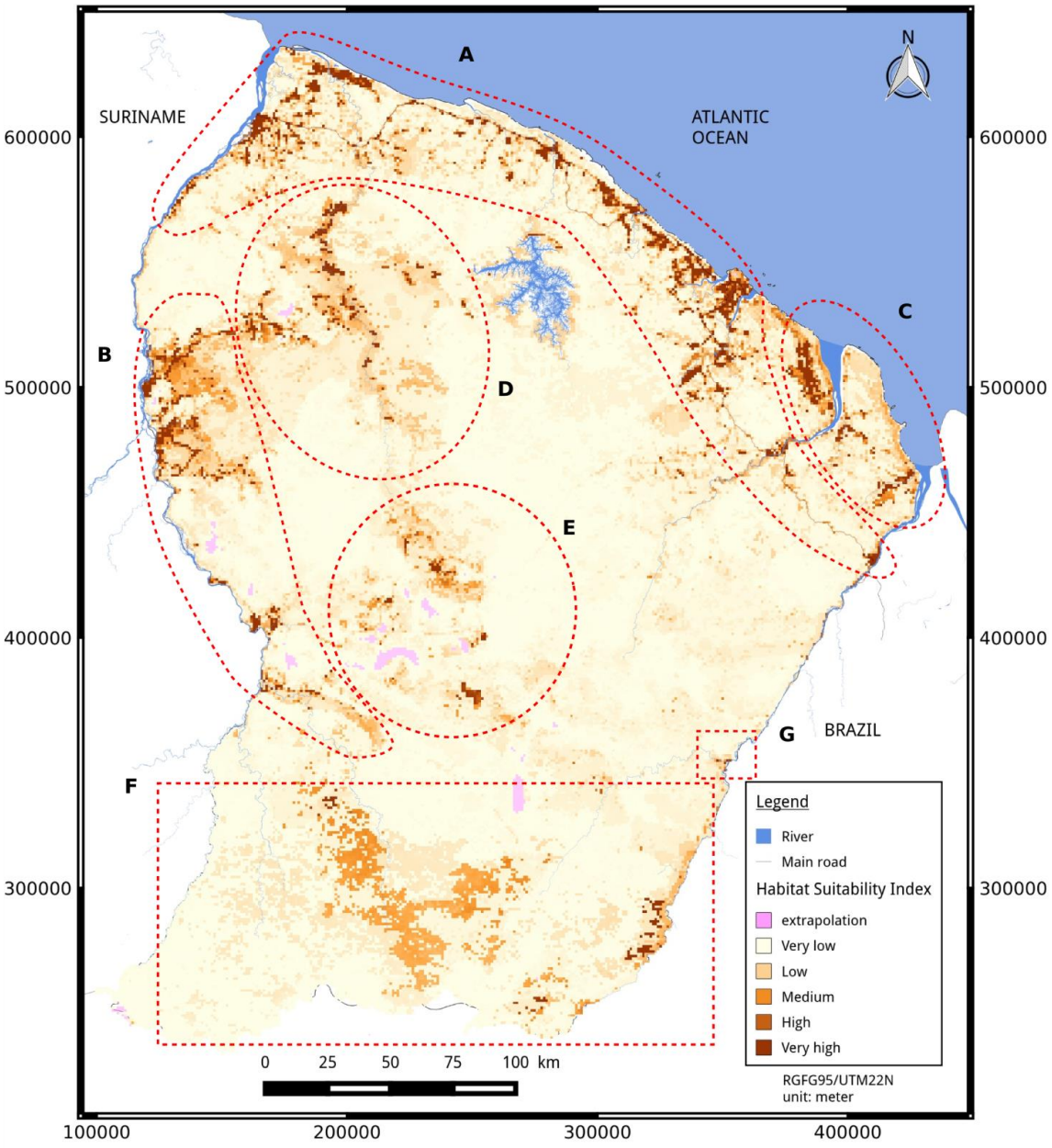


Figure 4. Habitat suitability index map. Six main areas with a high habitat suitability index (A to F) and Camopi village (G) are circumscribed by the red circles and rectangles.

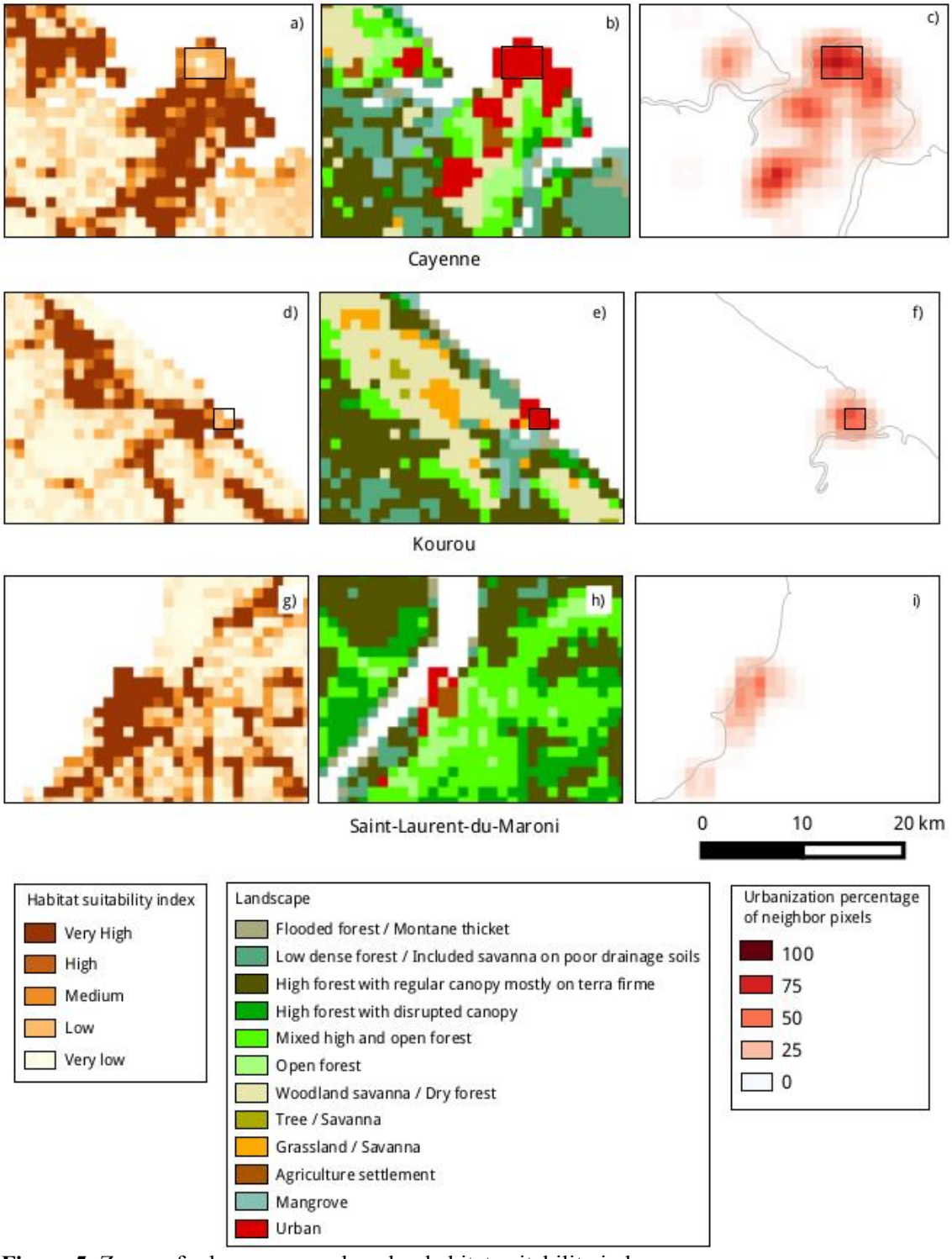


Figure 5. Zoom of urban areas. a, d, and g: habitat suitability index maps. b, e, and h: landscape type. c, f, and i: percentage urbanization of neighbor pixels. Rectangles correspond to highly urbanized areas (*LS* class is *Urban* and *PER_URB_NEIGH* $\geq$ 50%). Cayenne and Kourou include highly urbanized areas, but Saint-Laurent-du-Maroni does not.

Table 1. Raw environmental data and derived variables used to build the model.

Number of input variable	Producer, reference	Raw environmental data	Derived from (information source)	Date(s)	Original spatial resolution or interpretation scale	Derived SDM input variable(s)	Type of feature extraction for each 1x1 km pixel	Classes or range of values and units	Environment types [2]	*A priori* effect on An. darlingi presence [3] and bibliographic references	Input variable(s) type
1	Forest National Office (ONF), (Guitet et al. 2013)	Geomorphological landscape (*GLS*)	SRTM	2000	≥ 5000 m	Geomorphological landscape (*GLS*)	Majority class	12 classes	Natural environment	(/) Smith et al. (2013)	Categorical
2	Forest National Office (ONF), (Guitet et al. 2013)	Geomorphological landform (*GLF*)	SRTM	2000	≥ 200 m	Geomorphological landform (*GLF*)	Majority class	15 classes	Natural environment	(/) Smith et al. (2013)	Categorical
3	Agricultural Research Centre for International Development (CIRAD), (Gond et al. 2011)	Landscape types (*LS*)	Spot-Vegetation	2000	1000 m	Landscape types (*LS*)	Correction of pixels corresponding to urban areas and mangroves	14 classes	Natural environment and urbanization	(/) Stefani et al. (2013) Girod et al. (2011) Zeilhofer et al. (2007) Rozendaal (1992) Hiwat et al. (2010) Vittor et al. (2006) Vittor et al. (2009)	Categorical
4, 5, 6	National Aeronautics and Space Administration (NASA)	Altitude (*ALT*)	SRTM	2000	30 m	Altitude - minimum (*ALT_MIN*), - maximum (*ALT_MAX*) - median (*ALT_MED*)	Statistical computation	0 – 832 m	Natural environment	(-) Zeilhofer et al. (2007)	Continuous

Number of input variable	Producer, reference	Raw environmental data	Derived from (information source)	Date(s)	Original spatial resolution or interpretation scale	Derived SDM input variable(s)	Type of feature extraction for each 1x1 km pixel	Classes or range of values and units	Environment types [2]	*A priori* effect on An. darlingi presence [3] and bibliographic references	Input variable(s) type
7	National Institute of Geographic and Forestry Information (IGN)	Road and track network	BD TOPO®	2011	≥ 1000 m	Length of roads and tracks outside of urban areas (*ROADS*)	Computation of road/track lengths	0 – 12545 m	Non-permanent anthropogenic changes	(+) Singer and Castro (2001)	Continuous
8	Association Kwata 'Study and Conservation of French Guianan Wildlife ' (de Thoisy et al. 2010)	Human footprint (HFP)	From various sources [a]	2005	≥ 1000 m	Percentage of urbanization of neighboring pixels (*PER_URB_NEIGH*)	Percentage of urbanization within the eight neighbor cells	0-100%	Urbanization	(-) Stefani et al. (2013)	Continuous
9, 10, 11	Association Kwata 'Study and Conservation of French Guianan Wildlife ' (de Thoisy et al. 2010)	Human footprint (HFP)	From various sources [a]	2005	≥ 1000m	Human activities which non-permanently alter natural environment (*HA*) - minimum (*HA_MIN*) - maximum (*HA_MAX*) - median (*HA_MED*)	Statistical computation	0-30	Non-permanent anthropogenic changes	(+) Vittor et al. (2009)	Continuous

[a] French Institute for Statistical and Economic studies (INSEE); Regional Departments for Food, Agriculture and the Forest (DAAF); ONF; Regional Equipment, Habitat and Planning Authority (DDE) and Hammond et al. (2007).

[b] See the section on environmental variables in Materials and Methods.

[c] *A priori* effect on *An. darlingi* presence: (+) favorable; (-) unfavorable; (/) depends on categorical variable values.

Table 2. Mean contributions and jackknife results of the eleven input environmental variables.

Environmental variables	Contribution (%)	Cumulative contribution (%)	Gain with the variable only	Decrease of the gain without the variable (%)
ROADS	51.45	51.45	2.20	-7.98
PER_URB_NEIGH	17.17	68.62	1.86	-0.41
LS	15.32	83.94	2.23	-4.67
HA	7.43 (min: 0.35; median: 0.24; max: 6.84)	91.37	min: 0.02 median: 0.15 max: 0.43	min:-0.06 median: -0.22 max: -2.10
GLS	5.35	96.72	1.40	-2.59
ALT	2.09 (min: 1.34; median: 0.69; max: 0.06)	98.81	min: 1.12 median: 1.04 max: 0.76	min: - 1.04 median: -0.39 max: -0.03
GLF	1.19	100	0.80	-0.21

Table 3. Mean contributions and jackknife results of the seven input environmental variables of the simpler model.

Environmental variables	Contribution (%)	Cumulative contribution (%)	Gain with the variable only	Decrease of the gain without the variable (%)
ROADS	62.61	62.61	2.31	-8.61
LS	14.10	76.71	2.35	-6.23
PER_URB_NEIGH	11.15	87.86	2.05	-0.58
HA_MAX	5.39	93.25	0.37	-1.74
GLS	3.84	97.09	1.44	-1.90
GLF	2.1	99.19	1.01	-0.32
ALT_MIN	0.88	100	1.27	-1.29

Table 4. Characterization of areas with a high HSI

ns. signifies that the high HSI of the concerned area was not driven by that environmental variable, (+) signifies that when the value of the variable increases, the HSI also increases also, (-) signifies that when the value of the variable decreases, the HSI increases, and cells with classes name signifies that the presence of the given class implies a high HSI.

Area	*ROADS*	*LS* classes	*PER_URB_NEIGH*	*HA_MAX*	*GLS* classes	*GLF* classes	*ALT*
A	(+)	- Woodland savanna / Dry forest - Mixed high and open forest	(-)	(+)	- Coastal plain with low relief - Plain with residual reliefs (back coastal)	- Small size and flat wetland - Large flattened and wet relief - Wet hillock (low base-level)	(-)
B	(+)	- Open forest - Mixed high and open forest	*ns.*	(+)	- Peneplain with moderate hills	-Wet hillock (low base-level) - Large flattened relief	(-)
C	*ns.*	- Open forest	*ns.*	*ns.*	- Coastal flat plain	- Large flattened and wet relief	(-)
D	(+)	- Mixed high and open forest	*ns.*	(+)	*ns.*	*ns.*	(-)
E	(+)	- Mixed high and open forest	*ns.*	(+)	- Peneplain with moderate hills	- Large flattened relief	(-)
F	*ns.*	- Open forest	*ns.*	*ns.*	- Peneplain with moderate hills	- Lowered half-orange	(-)
G	(+)	Mixed high and open forest	*ns.*	(+)	*ns.*	*ns.*	(-)

S1. Coordinates of *Anopheles darlingi* presence sites (Coordinate system: RGFG95/UTM22N)

Number of sites	Locality	Longitude	Latitude
1	Cayodé	175866.785122426	374453.1653833
2	Taluène	162884.684665956	372648.888139721
3	Cacao	336895.086300153	505521.800163598
4	Midenangalanti	121041.015986301	557342.72581239
5	Grand Santi	124374.105534813	473261.338110987
6	Bois Martin	118927.138507594	552587.496776593
7	Flavien Campou	126212.611562883	477960.065899793
8	Régina	374698.048396995	476928.730177523
9	Camopi	352395.046055802	350764.216480451
10	Alikéné	322929.978284572	360135.527835386
11	Mine Boulanger	343189.582357019	504590.723307373
12	Carbet Légion crique Sikini	359046.553479	361996.097871
13	Cogneau - 23. Lot. Aquavilla	354091.053713437	538837.801956188
14	41. rue des Ixoras - Lot. Cogneau Larivot	351322.960529913	541116.10235073
15	Attila Cabassou	354987.967015231	540678.87835953
16	La Chaumière	347961.540051952	539402.864452834
17	1228. Ch. de La Chaumière	349626.610158034	540136.658859129
18	Saint-Georges	411052.118458134	429833.606330522
19	Quartier Espérance - Saint-Georges	411113.960068634	430709.665437749
20	Village Martin - Saint-Georges	411519.804866871	432628.481492214
21	Boulangerie – Saül	254644.262333224	400558.641871421
22	Chemin Mogès	352032.329517247	529456.717653458
23	Dorlin	216742.863044	415732.862195
24	Maripasoula	163199.637191727	403271.422431979
25	Repentir	234327.380239	427769.677349

Number of sites	Locality	Longitude	Latitude
26	Stoupan	352127.057631364	527017.037636427
27	Camp Pararé – Nouragues	311267.546406115	446010.858172732
28	Village Blondin - Saint-Georges	409492.529993713	428392.815435875
29	Quartier Adimo - Saint-Georges	410525.149504269	430989.662548913
30	Camp Bernet/ Légion étrangère	410437.343147834	429860.701524782
31	La ferme de Lait-Quateur	332757.212761399	552628.404024885
32	Grand Usine	288898.531939378	372330.073251345
33	Dagobert	226797.35968	438864.200286
34	Cacao	336515.199831583	505801.969900718
35	Saut-Maripa – Camp militaire	401579.145107468	420354.782186144
36	Eau-Claire	213145.824694	398626.838699
37	Cacao	336956.199791048	504651.969979654
38	Cacao	337909.199733996	506008.969874801
39	Cacao	336576.199808207	503413.970071365
40	Cacao	336555.199807874	503184.970087929
41	Impasse de la rafinerie – Cogneau	354091.053713437	538837.801956188
42	Quartier Bambou	411277.041477492	430178.275876772
43	Quartier Maripa - Saint-Georges	410252.974346512	430188.463058511
44	Quartier Savane - Saint-Georges	411024.737954334	430940.054473617
45	Cacao	336826.199809151	505764.969900894
46	Cacao	337091.199787723	505441.969921949
47	Cacao	336186.199848992	505049.969957266
48	Quartier Onozo- Saint-Georges	411339.040785181	430641.896115415

S2. Creation of the relative sampling effort map

Capture data of *Culicidae* (74 capture sites) were used to estimate the sampling effort of *An. darlingi*. The collection methods were identical and the sampling bias for the family was assumed to be representative of that for the focal species.

The sampling bias was defined as the relative sampling effort in the environmental space. For a pixel *i*, it corresponds to the ratio of the number of sampled pixels over the total number of pixels, within the *environmental neighborhood* of *i*.

First, all pixels of the study area were represented in the environmental variable space. This was accomplished by performing a Factorial Analysis of Mixed Data (FAMD) (Pagès, 2004). This analysis jointly takes into account numerical and categorical variables and makes it possible to represent the pixels within an Euclidean, orthonormal space defined from the whole set of environmental variables.

The membership degree of a pixel *j* to the neighborhood of pixel *i*, denoted w_{ij}, was defined by a Gaussian-like membership function:

$$w_{ij} = 0.5^{\left(d_{ij}/D_{min}\right)^2} \qquad (1)$$

with d_{ij} the euclidean distance between *i* and *j* in the factorial space, and D_{min} the threshold distance over which *j* does not significantly belong to the environmental neighborhood of *i*, *i.e.* over which $w_{ij} < 0.5$. The membership degree w_{ij} has the following properties:

- $w_{ij} \in]0,1]$;

- $w_{ij} = 1$ if $d_{ij} = 0$;

- $w_{ij} < 0.5$ if $d_{ij} > D_{\min}$.

The parameter D_{min} was set from *a priori* knowledge of *An. darlingi* bio-ecology. As highly urbanized areas are not suitable for *An. darlingi* (see § I.3), we stated that a pixel associated with *An. darlingi* presence cannot belong to a highly urbanized pixel. Reciprocally, a pixel considered to be highly urbanized cannot belong to the environmental neighborhood of a pixel where *An. darlingi* was observed.

Consequently, given *P*, the set of pixels where the species was observed and *U*, the set of pixels belonging to highly urbanized areas, D_{min} was defined as follows:

$$D_{min} = min\left(d_{pu}\right)_{p\in P, u\in U} \qquad (2)$$

A pixel is considered to be highly urbanized if it belongs to the *LC* class *Urban* and if its eight neighboring pixels present an average urbanization percentage (*PER_URB_NEIGH*) higher than or equal to 50%.

The concepts of environmental space and neighborhood, as well as the key method parameters are schematically represented in Figure S1.

Given *X*, the set of pixels of the study area, and $c = \{c_i\}_{i\in X}$, a vector such that $c_i = 1$ if *i* is sampled and $c_i = 0$ otherwise, the relative sampling effort at pixel *i*, z_i, is then defined as:

$$z_i = \sum_{j\in X} w_{ij} \cdot c \,/ \sum_{j\in X} w_{ij} \qquad (3)$$

The relative sampling effort was computed for each pixel of the study area. The resulting map was used to bias the random selection of background points. Consequently, for a given pixel, the greater the relative sampling effort, the higher the chance of selecting the pixel as a background point.

Reference

Pagès, J. 2014. Multiple Factor Analysis by Example Using R. Chapman & Hall, CRC Press.

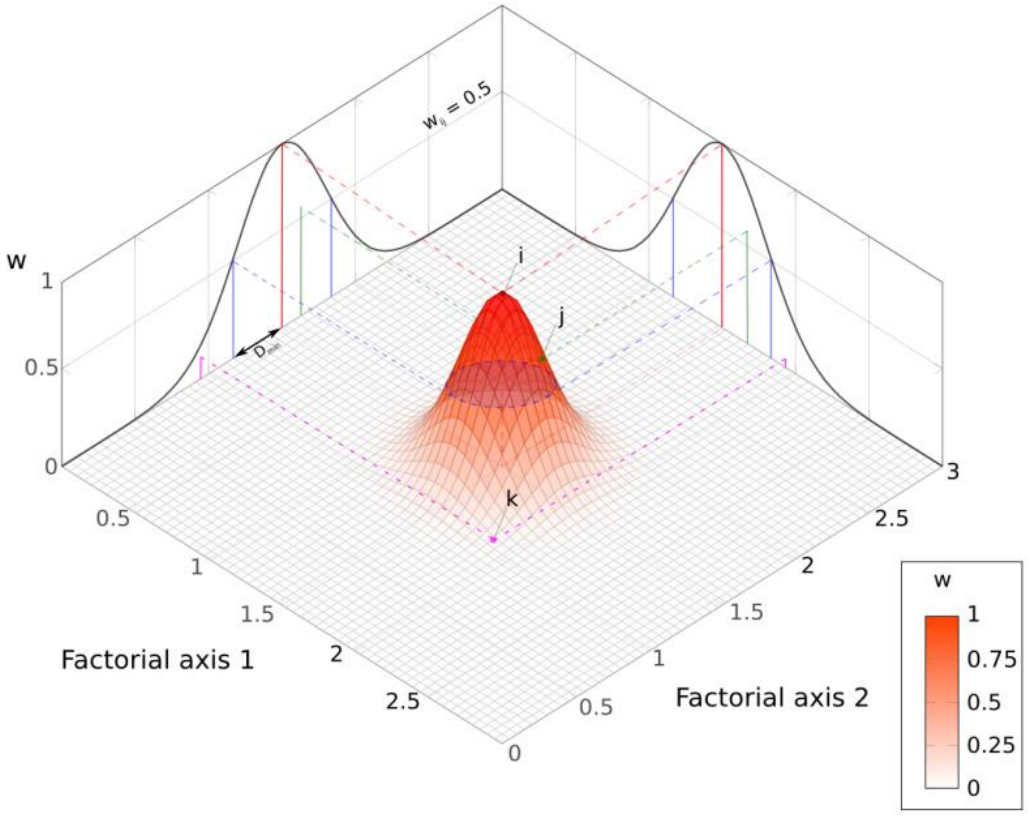


Figure S1. Neighborhood of a pixel *i* in the environmental space represented by the first and second factorial axes. The environmental neighborhood of point *i* is represented by the Gaussian function. The blue lines define the limit of the neighborhood of *i*. Only point *j* is situated above these lines. Thus *j* is in the neighborhood of *i* in the first factorial plane.

Résumé

En modélisation écologique, les modèles de distribution d'espèces sont très couramment utilisés pour des objectifs variés. Ils ont notamment été identifiés comme particulièrement pertinents pour cartographier et caractériser la qualité d'habitat des moustiques du genre *Anopheles*, vecteurs du paludisme, et ainsi participer à l'estimation du risque de transmission de cette maladie et à la définition de stratégies de lutte anti-vectorielle ciblées. Le paludisme est en effet un problème sanitaire majeur au niveau mondial. Sa transmission dépend directement de la présence et de la distribution des vecteurs, qui dépendent elles-mêmes des conditions environnementales contribuant à définir la qualité des habitats écologiques des *Anopheles*.

Cependant, dans certaines régions, les données de captures d'*Anopheles* restent rares, rendant difficile la modélisation de ces habitats. De plus, le recueil de ces données est très souvent soumis à des biais d'échantillonnage importants compte tenu, notamment, d'une accessibilité inégale à l'ensemble de la zone d'étude.

Cette thèse vise ainsi à fournir une solution à la cartographie des vecteurs du paludisme, en considérant explicitement deux aspects très peu étudiés dans la modélisation : le faible nombre de sites de présence disponibles et l'existence d'un biais d'échantillonnage important.

Ainsi, une méthode originale de correction de l'effet du biais d'échantillonnage est proposée. Elle est appliquée à des données de présence d'*Anopheles darlingi* – le principal vecteur du paludisme en Amérique du Sud – en Guyane française, où le paludisme est endémique. Un modèle de distribution d'*An. darlingi* a ensuite été construit, permettant d'obtenir une carte de qualité d'habitat en cohérence avec la connaissance actuelle des entomologistes et fournissant des performances de prédiction élevées (AUC et AUC partiel avec un taux d'omission de 20% égaux, respectivement, à 0,93 et 1,11).

La méthode de correction proposée a ensuite été comparée aux méthodes existantes dans un contexte applicatif caractérisé par la rareté des données d'occurrence de l'espèce et la présence d'un biais d'échantillonnage significatif. Pour cela, des données de présence virtuelles ont été simulées afin de bénéficier d'un jeu de données de référence non biaisé. Les résultats montrent que la méthode développée dans ce travail est adaptée aux cas où le nombre de sites de présence est faible. En revanche, lorsque ce nombre augmente, d'autres méthodes procurent de meilleurs résultats.

Cette thèse contribue, d'une part, à combler les lacunes théoriques et d'applicabilité des méthodes actuelles visant à corriger l'effet des biais d'échantillonnage et, d'autre part, à compléter la connaissance sur la distribution spatiale et la bio-écologie du principal vecteur du paludisme en Guyane française. Elle prend ainsi part aux efforts engagés par la Guyane française afin d'atteindre la pré-élimination du paludisme à l'horizon 2018. Les méthodes développées étant applicables à toute espèce animale ou végétale, elle contribue plus largement, dans le domaine de la modélisation écologique, à améliorer l'applicabilité et la fiabilité des modèles de distribution spatiale des espèces.

Abstract

In ecological modeling, species distribution models are very commonly used for various purposes. In particular, they have been identified as relevant to map and characterize the habitat quality of *Anopheles* genus mosquitoes, transmitting malaria, and thus to both participate in the estimation of the transmission risk of this disease and in the definition of targeted vector control actions. Malaria is indeed a major health problem worldwide. Its transmission depends directly on the presence and distribution of the vectors, which are themselves dependent on the environmental conditions that contribute to define the quality of the ecological habitats of the *Anopheles*.

However, in some areas, *Anopheles* collection data remain scarce, making it difficult to model these habitats. In addition, the collection of these data is very often subjected to significant sampling biases, due, in particular, to unequal accessibility to the entire study area.

This thesis aims at providing a solution to the mapping of malaria vectors, by explicitly considering two very little studied aspects in modeling : the low number of available presence sites and the existence of a significant sampling bias. Thus, an original method for correcting the effect of the sampling bias is proposed. It is applied to presence data of *Anopheles darlingi* species - the main vector of malaria in South America - in French Guiana, where malaria is endemic. Then, a distribution model of *An. darlingi* was built to obtain a map of habitat quality consistent with current entomologists' knowledge and providing high prediction performances (AUC and partial AUC with an omission rate of 20% equal to, respectively, 0.93 and 1.11).

The proposed correction method was then compared to existing methods in an application context characterized by the scarcity of the species occurence data and the presence of a significant sampling bias. For this purpose, virtual presence data were simulated in order to benefit from an unbiased reference dataset. The results show that the method developed in this work is adapted to cases where the number of sites of presence is low. On the other hand, when this number increases, other methods yield better results.

This thesis contributes, on the one hand, to fill theoretical and applicability lacuna of current methods intended to correct the effect of the sampling bias and, on the other hand, to supplement the knowledge on both the spatial distribution and the bio-ecology of the main malaria vector in French Guiana. It thus takes part in the efforts undertaken by French Guiana to achieve the pre-elimination of malaria by 2018. The methods developed in this thesis are applicable to any animal or plant species and contribute, more generally in the field of ecological modeling, to improve the applicability and reliability of species distribution models.