Fouille de données spatio-temporelles pour l'étude du risque de transmission résiduelle du paludisme à échelle paysagère en milieu rural ouest-africain

Paul Taconet

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Écologie et Biodiversité

École doctorale GAIA

Unité de recherche MIVEGEC (UMR 224)

Fouille de données spatio-temporelles pour l'étude du risque de transmission résiduelle du paludisme à échelle paysagère en milieu rural ouest-africain

Présentée par Paul TACONET
le 2 juin 2022

Sous la direction de Nicolas MOIROUX

Devant le jury composé de

Catherine LINARD, Professeure, Université de Namur	Rapportrice
Ibrahima DIA, Directeur de recherche, Institut Pasteur de Dakar	Rapporteur
Florence FOURNET, Directrice de recherche, IRD (UMR 224)	Examinatrice, Présidente
Jean GAUDART, Professeur, Aix Marseille Université (UMR 1252)	Examinateur
Emmanuel ROUX, Chargé de recherche, IRD (UMR 228)	Examinateur
Nicolas MOIROUX, Chargé de recherche, IRD (UMR 224)	Directeur de thèse
Annelise TRAN, Chargée de recherche, CIRAD (UMR 1470)	Invitée
Morgan MANGEAS, Directeur de Recherche, IRD (UMR 250)	Invité

UNIVERSITÉ DE MONTPELLIER

Remerciements

Il se dit en général qu'une thèse ne se fait pas seul... et bien, celle-ci ne dérogera pas à la règle ! Profitons donc de cette tribune qui nous est offerte en début de manuscrit de thèse pour remercier toutes les personnes qui, de près ou de loin, en m'accompagnant professionnellement, personnellement, ou les deux, ont contribué à ce travail :)

Mes premiers remerciements vont tout naturellement à Nicolas, mon directeur de thèse. Si l'on m'avait annoncé que notre entretien, il y a quatre ans, mènerait à cette thèse, je n'y aurais pas cru ! Merci donc, pour commencer, de m'avoir initialement recruté puis rapidement proposé ce sujet de thèse. Merci, également, pour ton suivi et tes conseils éclairés tout au long de ce travail. Et enfin et surtout, merci de m'avoir accordé en tous temps ta confiance tout au long de ces trois années et demi pour mener à bien ma thèse. Cette confiance et la liberté que tu m'as laissées m'ont permis d'explorer des domaines de recherche inconnus de moi jusqu'alors et qui se sont révélés passionnants, de produire un travail à mon image, et finalement je pense, de mieux appréhender le métier de chercheur !

En second lieu, je souhaite vivement remercier les personnes qui portent les projets de recherche sur lesquels ma thèse a été adossée. Merci à Cédric Pennetier et Nicolas Moiroux, porteurs des projets REACT 1 et REACT 2 ; et à Karine Mouline, porteuse du projet ANORHYTHM. Vos compétences et votre motivation pour ces sujets si passionnants sont un exemple à suivre pour les chercheuses et chercheurs en herbe comme moi ! Merci à l'unité MIVEGEC qui m'a hébergée pour mener à bien cette thèse. Merci également aux organismes publics qui ont financé, directement ou indirectement, ma thèse : l'Initiative 5%, le Ministère des Affaires Étrangères, l'Agence Nationale de la Recherche, l'Institut de Recherche pour le Développement.

La garantie d'une recherche de qualité est son évaluation par les pairs. Merci, ainsi, aux chercheurs et chercheuses qui ont pris le temps de suivre ou d'évaluer ma thèse, à

travers le comité de suivi individuel et le jury de thèse : Catherine Linard, Ibrahima Dia, Annelise Tran, Morgan Mangeas, Eric Daudé, Florence Fournet, Jean Gaudart, Emmanuel Roux, Carlo Costantini, Frédéric Simard. Un remerciement particulier à Catherine Linard et Ibrahima Dia, qui ont accepté la chronophage tâche de rapporter ma thèse.

Cette thèse se basant très largement sur les données du projet REACT 1, je souhaite remercier particulièrement toutes les personnes qui ont oeuvré à son accomplissement, sur le terrain, au laboratoire, ou au bureau ; ainsi que les habitants des villages du projet REACT. Ma thèse n'aurait pu voir le jour sans cet immense travail de collecte de données auquel vous avez contribué. Merci, au passage, aux collègues des organismes scientifiques partenaires de l'IRD avec lesquels j'ai échangé pendant ma thèse : l'Institut de Recherche en Sciences de la Santé au Burkina Faso et l'Institut Pierre Richet en Côte d'Ivoire.

Une mention spéciale à Aristide et Abou, mes accompagnateurs moto sur le terrain - je sais que j'ai parfois pu vous faire suer pour collecter mes parcelles d'occupation du sol :) ! Et tant qu'on est sur le terrain, merci à Hamidou Konaté, qui a pris grand soin de moi lors de mon accès palustre à Iolonioro. Grâce à toi, j'ai pu appréhender mon sujet de thèse - toute mesure gardée - au plus proche de la réalité terrain, en toute sécurité !

Ce travail a également été rendu possible grâce aux innombrables personnes travaillant, au niveau technique ou politique, à ouvrir de nombreuses données (satellitaires ou autres) et développer des logiciels à code source ouvert, pour que chacun puisse les utiliser facilement et librement. Je souhaite les remercier ici. Merci en particulier à Raffaele Gaetano, contributeur au développement du logiciel Orfeo Toolbox, pour ton aide sur l'utilisation de cet outil. De même, je remercie toutes les personnes anonymes qui font vivre les forums de discussion sur la science des données, en répondant aux questions techniques. Ces forums ont été largement utilisés dans mes travaux.

Avant le commencement de ce travail, de nombreuses personnes m'ont fait confiance au niveau professionnel. Je souhaite ici les en remercier : ce travail a aussi été rendu possible grâce à elles. Merci à Nathalie Molines de m'avoir initialement introduit au monde de la géomatique au cours de ma formation initiale, et de m'avoir fait confiance en m'autorisant - déjà à l'époque - de faire des stages 'alternatifs' m'ayant permis de

creuser cette voie précocément. Merci à Axel Falguier de m'avoir recruté, il y a tout juste neuf ans, pour ma première expérience professionnelle en tant que géomaticien pour la réserve naturelle des Terres Australes et Antarctiques Françaises - je conserve une belle nostalgie de cette époque. Enfin, merci à vous, Julien Barde et Emmanuel Chassot, de m'avoir introduit au monde de la recherche scientifique il y a sept ans (à l'époque, sur les pêcheries thonières) et en particulier, au sein de l'IRD. De la pêche au thon à la chasse aux moustiques, il n'y eu qu'un pas ! Ces années à l'IRD ont été pour moi extrêmement enrichissantes, grâce à tous ces projets portés par des personnes passionnées, dans un institut qui porte de belles valeurs humanistes. Pourvu que ça dure !

Vivre hors de son pays est toujours l'occasion de formidables rencontres et d'ouvrir un peu plus son esprit, et les deux premières années de ma thèse passées à Bobo-Dioulasso, au Burkina Faso, l'ont une fois de plus confirmé. Je salue donc bien bas toute l'équipe bobolaise ! Les innombrables Brakina savourées avec vous au Zoffi, les séances ciné Rembob's, les longues soirées dansantes à l'Entente ou encore les tournois de badminton au club Muraz ont rythmé une vie douce et remplie d'apprentissages de vie. Un Anitié particulier à la bande proche : Simon, Julia, Clément, Fleurance, Yara, Issouf, Hélène, Soizic, Farès.

Ces dix-huit derniers mois, les amis montpellierains ont bien pris le relais ! Merci à Camille et Bifi pour les innombrables soirées dans les mas et pour les tours de vélo, à Ben et Agnès pour les soirées dansantes dignes de celles de Bobo, à Marco (aaaaiie) et Coralie pour les apéros et virées dans l'arrière pays, à Rodaco pour un peu tout ça à la fois et même plus. Une pensée pour toi Martin, mon brolloc de la rue M2S, ça kif ! Merci à mes collègues et amies de bureau qui m'ont accompagné sur ma période montpelliéraine : Lison, Angélique et Adeline. Si c'est un plaisir de venir au bureau, c'est en bonne partie grâce à vous. Pourvu que l'on continue à travailler (et pas que) comme on le fait si bien ! Merci en particulier à toi, Angélique, pour le code Latex de la page de garde de ce manuscrit, mais surtout pour tes conseils avisés de statisticienne et les discussions enrichissantes sur notre rôle dans la recherche.

Bien sûr, je n'oublie pas les bases. La base de la base. Les *Vrais Gars* : Phil, Ninin, Adi, Jo, Seba, Tigrou, Flo, Lolo, Bouns. Merci pour vos messages de soutien tout au long de la thèse, pour les week-ends entre potes qui permettent de s'évader et oublier - un peu - la thèse. Longue vie aux VG et à bas les sapins ! Un merci spécial à Benoit

de Haas, auteur des dessins séparant les trois grandes parties du manuscrit (pages 5, 63, 210). Plus largement, merci à tous les amis de l'Université de Technologie de Compiègne : le réseau de l'UTC est une ancre pour moi, qui me permet d'appréhender plus sereinement mes divagations géographiques de vie !

Pierrot, mon cher cousin, une mention spéciale pour toi. Tu es pour moi un modèle, un exemple de motivation et de résilience. Obrigado !

Merci à toi, ma Clairette. Ta délicatesse, tes plans d'évasion pour le week-end, sorties au cinéma ou encore autres petites attentions culinaires, et nos longues soirées à refaire le monde, m'ont grandement aidé à tenir et avancer. Ton abnégation nous a permis de mener à bien en parallèle cette (longue) fin de thèse et notre début de relation ! Merci pour ton accompagnement, ton soutien, et ta patience. Le meilleur est devant nous ! M'bifé !

Enfin, merci à ma famille. Papa, Maman, Caroline, Marine, Julien, Alizée, Simon, Eliott, Noé. Je ne vous dirai jamais assez combien je vous aime. Vous êtes la moustiquaire qui me protège des piqûres de la vie ;) Et avec vous, pas de développement de résistances ! Papa et Maman, merci de m'avoir accompagné toutes ces années, de m'avoir toujours soutenu dans mes choix, de m'avoir guidé, en douceur, sans jamais interférer. Merci pour les valeurs que vous m'avez transmises. Je vous dédie cette thèse.

La maturité de l'homme, c'est d'avoir retrouvé le sérieux qu'on avait au jeu quand on était enfant

F. Nietzsche

Résumé de la thèse / Abstract

Français. La lutte contre la transmission du paludisme fait aujourd'hui face au ralentissement des progrès enregistrés au cours des quinze premières années du 21è siècle. Pour les redynamiser, il est nécessaire de passer d'une approche universelle de la prévention à une approche ciblée, adaptée au faciès local du risque de transmission. Ces stratégies nécessitent de décrire, comprendre et prédire le risque de transmission du paludisme à des échelles spatio-temporelles fines - adaptées à la prise de décision locale. Dans cette thèse, nous avons tenté d'expliquer et évaluer la prédictibilité dans l'espace et dans le temps, à échelle paysagère en milieu rural ouest-africain (au Burkina Faso et Côte d'Ivoire), de plusieurs indicateurs entomologiques du risque de transmission : présence et abondance des anophèles, résistances physiologiques et comportementales des anophèles aux insecticides. Nous avons pour cela utilisé des données entomologiques et environnementales hétérogènes, spatialisées et temporalisées, multi-source et multi-échelle (images satellitaires, capteurs micro-climatiques, enquêtes de comportement humains, etc.), et des méthodes de fouille de données avancées (notamment basées sur l'interprétation des modèles d'apprentissage automatique), dans des approches holistico-inductives de la génération de connaissances scientifiques. Nos résultats ont montré à l'échelle des villages une forte hétérogénéité spatio-temporelle des densités vectorielles, et une relative homogénéité de la prévalence des phénotypes / génotypes résistants. À partir des associations capturées par les modèles statistiques, nous avons émis de nombreuses hypothèses sur les déterminants environnementaux (climatiques, paysagers, socio-culturels, etc.) des différents indicateurs entomologiques étudiés ; autrement dit sur l'impact de l'environnement sur les traits de vie des vecteurs. Nos modèles étaient en mesure de prédire correctement et anticiper plusieurs semaines à l'avance les densités vectorielles à l'échelle du village, ce qui n'était pas le cas pour les résistances aux insecticides. À l'issue de ce travail, nous faisons des propositions pour l'amélioration (i) des méthodes actuelles de lutte anti-vectorielle, (ii) de l'utilisation de la science et ingénierie des (géo-)données en général, et de la modélisation statistique en particulier,

pour la recherche et le contrôle du paludisme, et (iii) des outils de surveillance et prévention du risque de transmission du paludisme à échelle locale en milieu rural ouest-africain.

English. The fight against malaria transmission is currently stalling. To reinvigorate progress, it is necessary to shift from an universal approach of prevention to a targeted one, adapted to the local transmission risk profile. Such strategy requires to characterise, understand, and predict the risk of transmission of malaria at fine spatial and temporal scales, i.e. scales suitable for local decision-making. In this thesis, we have attempted to explain and evaluate the spatial and temporal predictability of several entomological indicators of transmission risk at a landscape scale in rural West Africa (in Burkina Faso and Côte d'Ivoire) : presence, abundance of anopheles, physiological and behavioral resistance to insecticides. We used heterogeneous, spatio-temporal, multi-source and multi-scale entomological and environmental data, and data mining methods (among others, based on interpretable machine learning techniques), in a holistic-inductive approach to scientific knowledge generation. Our results showed strong spatio-temporal heterogeneities in vector abundances at the village scale, and relative homogeneities in the prevalence of vector resistances. Based on the associations captured by the statistical models, we made numerous hypotheses on the environmental determinants (climatic, landscape, socio-cultural, etc.) of the various entomological indicators studied ; in other words, on the impact of environmental conditions on the vectors' life traits. Our models were able to acurately forecast vector abundances at the village scale several weeks ahead, which was not the case for the prevalence of insecticide resistance. At the end of this work, we make proposals for the improvement of (i) current vector control methods, (ii) the use of (geo)data science and data engineering in general, and statistical modelling in particular, for malaria research and control, and (iii) tools for the surveillance and prevention of malaria transmission risk at the local scale in rural West Africa.

Table des matières

Table des figures

Liste des abréviations

BF : Burkina Faso
càd. : c'est-à-dire
CI : Côte d'Ivoire
CCM : Cross-Correlation Map
EDA : Exploratory data analysis
EIC : European Innovation Council
ERC : Essai Randomisé Contrôlé
FAIR : Findable, Accessible, Interoperable, Reusable
FOSS : Free and Open Source Software
GAM : Generalized Additive Model
GLMM : Generalized Linear Mixed Model
GPM : Global Precipitation Measurement
IEC : Information, Education, Communication
IPR : Institut Pierre Richet
IRD : Institut de Recherche pour le Développement
IRSS : Institut de Recherche en Sciences de la Sante
LAV : Lutte Anti-Vectorielle
LST : Land Surface Temperature
MAP : Malaria Atlas Project
MIILDA : Moustiquaire Imprégnée d'Insecticide à Longue Durée d'Action
MODIS : Moderate Resolution Imaging Spectroradiometer
MNT : Modèle Numérique de Terrain
NASA : National Aeronotics and Space Agency
NDVI : Normalized Difference Vegetation Index
OMS : Organisation Mondiale de la Santé
OPeNDAP : Open-source Project for a Network Data Acces Protocol
PDP : Partial Dependence Plot

PID : Pulvérisations Intra-Domiciliaires d'insecticide à effet rémanent
PNLP : Programme National de Lutte contre le Paludisme
SIG : Système d'Information Géographique
SMI : Synthetic Meteorological Indicator
SPOT : Satellite pour l'Observation de la Terre
SRTM : Shuttle Radar Topography Mission
TR : Transmission Résiduelle
VIIRS : Visible Infrared Imaging Radiometer Suite

Préface

Transmission résiduelle du paludisme

Les deux premières décennies du XXI^{ème} siècle ont représenté un âge d'or dans l'histoire de la lutte contre le paludisme. Au cours de ces années, grâce à un engagement scientifique, politique et financier sans précédent, le fardeau du paludisme a été significativement allégé à l'échelle mondiale. Plusieurs milliards de moustiquaires imprégnées d'insecticide, de tests de diagnostic rapide, de traitements à base d'artémisine, ont été distribués sur tous les continents touchés par la maladie (WHO, 2021). Ces efforts considérables effectués dans la prévention, le diagnostic et le traitement du paludisme, ont porté leurs fruits : au niveau mondial entre 2000 et 2019 (avant la pandémie de covid-19), l'incidence du paludisme a reculé de 29 % et la mortalité de 60 %, des niveaux jusqu'alors inédits (WHO, 2020). Sur cette même période, 1,5 milliards de cas de paludisme et 7,6 millions de décès associés ont été évités dans le monde (WHO, 2020).

Mais depuis quelques années, les progrès stagnent. L'incidence ne recule plus aussi rapidement : alors qu'elle diminuait de 4,25 % en moyenne par an sur la période 2000 – 2015, elle ne baissait plus que de 2 % par an sur la période 2015 - 2019 (WHO, 2020). De même, la mortalité associée à la maladie ne diminue plus aussi rapidement. En cause : des changements socio-économiques et environnementaux, mais surtout des menaces biologiques telles que la résistance des parasites responsables du paludisme aux antipaludiques ou encore celle des moustiques vecteurs aux insecticides utilisés dans la lutte anti-vectorielle. Aussi, dans des zones pourtant couvertes par les outils de lutte anti-vectorielle conventionnels, la transmission du paludisme reste soutenue, voire réaugmente : cette transmission qui échappe aux stratégies de lutte mises en place est appelée transmission *résiduelle* du paludisme (Gerry F. Killeen, 2014).

Comment redynamiser les progrès ? Dans un contexte de ressources (financières, humaines, matérielles) limitées et d'émergence de menaces à l'efficacité des outils de lutte actuels, la communauté scientifique propose de repenser certaines approches de gestion de la maladie. En particulier, elle incite à développer de nouveaux outils de lutte contre la maladie, et à passer d'une démarche « universelle » - où les efforts sont déployés uniformément sur un territoire donné – à une démarche localisée, où les interventions sont optimisées en fonction des spécificités locales et du niveau de risque, notamment de transmission résiduelle de la maladie.

La mise en œuvre de ces approches nécessite d'approfondir les connaissances sur le risque de transmission résiduelle et sur sa distribution spatio-temporelle. L'enjeu est donc double. D'une part, il s'agit d'approfondir les connaissances fondamentales sur les phénomènes entrant en jeu dans le risque de transmission résiduelle. Ces connaissances peuvent permettre de développer de nouveaux outils de lutte. D'autre part, pour gérer le risque au niveau local, il est nécessaire d'acquérir une connaissance fine de la distribution spatio-temporelle de ce risque sur un territoire d'intérêt : comprendre les déterminants de son hétérogénéité et être en mesure de la prédire. Une telle connaissance permet ensuite de définir puis déployer les interventions qui sont susceptibles d'avoir le plus d'impact au regard de la situation locale.

Comment acquérir ces connaissances ? Le risque de transmission résiduelle est directement lié à la probabilité qu'un moustique vecteur du paludisme pique un humain. Cette probabilité de contact entre l'homme et le vecteur dépend elle-même en grande partie de l'environnement dans lequel les populations, humaines et vectorielles, évoluent. Une compréhension fine et holistique des interactions complexes, sur le terrain, entre le vecteur, l'homme, le parasite, et l'environnement est donc nécéssaire pour élaborer et mettre en œuvre ces nouvelles approches de la gestion du risque. A cet égard, la fouille de données (*data mining*) constitue une approche intéressante et en plein essor permettant de mieux appréhender l'étude de ce genre de systèmes complexes.

Données volumineuses, fouille de données et méthodes statistiques pour l'étude des systèmes complexes

Le début du XXI^{ème} siècle a aussi représenté le point de basculement dans l'ère de la donnée numérique. Des paramètres biophysiques de notre planète à nos comportements individuels, les réalités physiques, matérielles, de notre univers, sont toujours plus mesurées et archivées. En parallèle et en lien direct, les ressources informatiques et les méthodes statistiques ont connu une véritable révolution afin d'offrir les capacités de traiter ce « déluge de données ». À l'interface entre informatique, statistiques et connaissances métiers, la science des données a émergé en tant que discipline scientifique visant à exploiter ces « données volumineuses » pour mieux comprendre et prédire le monde qui nous entoure.

Dans un premier temps cantonnée à des fins commerciales, la science des données est en train de prendre une part toujours plus importante dans l'ensemble des disciplines de la recherche scientifique ; avec une prise de conscience croissante que ces données volumineuses peuvent jouer un rôle dans le processus de vérification, voire création, de connaissance scientifique. Ainsi, les projets de recherche collectent toujours plus de données pour étudier des phénomènes d'intérêt. Mais dans certaines disciplines, si la démarche d'utilisation de données à des fins de vérification d'hypothèses scientifiques pré-établies paraît souvent bien connue et maîtrisée par les chercheurs, le processus de création de connaissances ou nouvelles hypothèses scientifiques à partir de données volumineuses et modèles statistiques semble généralement moins maîtrisé.

Comment, au-delà de la « simple » vérification d'hypothèses scientifiques pré-établies, les données et méthodes statistiques peuvent-elles nous aider à générer de nouvelles connaissances ou théories scientifiques ? En quoi la modélisation statistique et les données servent-elles les objectifs fondamentaux de la recherche scientifique (créer, consolider, vérifier des hypothèses) ? Comment exploiter pleinement leur potentiel ? En particulier, comment peuvent-elles appuyer l'étude des systèmes biologiques complexes, tel que le système environnement – hôte - vecteur ?

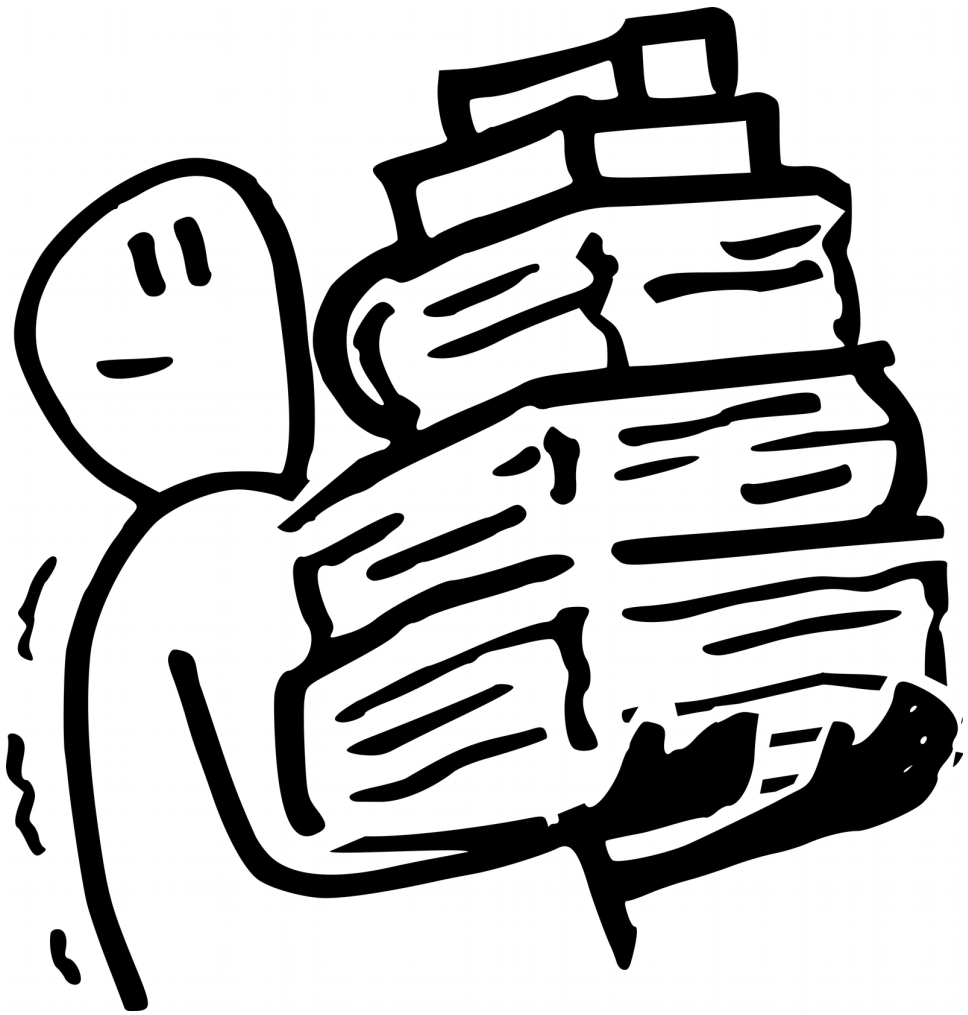
Présentation générale de la thèse et de l'organisation du manuscrit

Cette thèse étudie le risque de transmission résiduelle du paludisme, et en particulier les liens entre environnement et vecteurs du paludisme, en utilisant des méthodes avancées de modélisation statistique. Bien que les problématiques abordées concernent un grand nombre de pays africains, la présente étude se concentre sur deux zones rurales ouest-africaines endémiques du paludisme de la taille de districts sanitaires, situées au Burkina Faso et en Côte d'Ivoire. La thèse se veut à l'interface entre les sciences de l'entomologie médicale, de la géo-épidémiologie, et des données. L'objectif principal est d'améliorer certaines connaissances fondamentales sur le risque de transmission résiduelle du paludisme dans nos deux zones d'étude, en étudiant la bio-écologie des vecteurs. Le second objectif est d'offrir au lecteur un éclairage épistémologique et méthodologique sur le rôle de la modélisation statistique dans la recherche scientifique, et en particulier sur son potentiel pour l'étude des systèmes biologiques complexes ; les travaux de thèse en représentant des cas d'étude concrets.

Le manuscrit se divise en 7 chapitres répartis dans 3 grandes parties. La première partie (chapitres 1 et 2) a pour objectif de dresser un cadre théorique et bibliographique sur le paludisme et la modélisation statistique utile à la compréhension des travaux de thèse ; et de présenter les objectifs et enjeux de la thèse. La deuxième partie du manuscrit (chapitres 3 à 6) constitue le cœur du travail de thèse. Enfin, dans la troisième partie (chapitre 7), nous discutons l'ensemble résultats.

Partie 1

Bibliographie



Chapitre 1

Contexte scientifique : Paludisme, transmission résiduelle et enjeux de la thèse

Dans ce chapitre, nous introduisons tout d'abord certaines notions d'entomologie médicale essentielles à la compréhension de la thèse (cycle de la transmission du paludisme, cycle biologique et comportement trophique des vecteurs du paludisme, liens environnement-vecteur). Dans un second temps, nous présentons le concept et les outils de lutte anti-vectorielle, ainsi que leurs limites actuelles ; ce qui nous conduit finalement au concept de transmission résiduelle du paludisme. Dans une dernière partie, nous énonçons les enjeux de la thèse et précisons l'organisation générale du manuscrit.

1.1 Paludisme et lutte anti-vectorielle

1.1.1 Fardeau du paludisme dans le monde et en Afrique

Environ la moitié de la population mondiale est exposée au risque de paludisme (WHO, 2021) (figure 1.1). En 2020, le paludisme était endémique dans 85 pays (WHO, 2021) et était l'une des quatre maladies infectieuses (avec la tuberculose, le sida et le covid-19) ayant causé le plus de décès au niveau mondial (Le Monde, 2020). Cette année-là, l'OMS estimait le nombre de cas de paludisme à 241 millions (59 cas / 1000 personnes à risque), et le nombre de décès attribuables à la maladie à 627 000 (15 décès pour 100 000 personnes à risque).

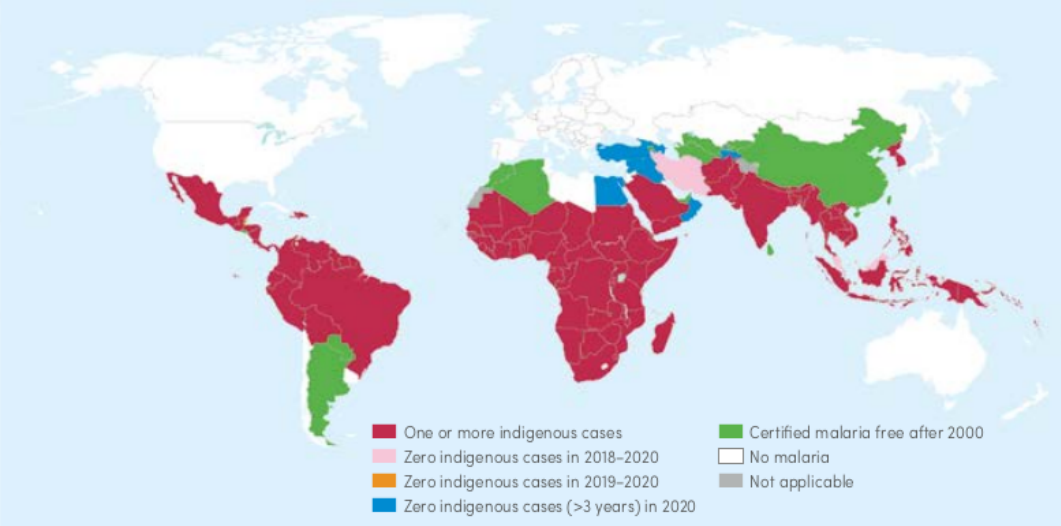


FIGURE 1.1: Pays d'endémicité palustre en 2000 et leur statut en 2020 (WHO, 2021)

Le poids du paludisme dans le monde est inégalement réparti, à la fois géographiquement et démographiquement. Ainsi par exemple, en 2020, 95 % des cas et 96 % de la mortalité étaient concentrés en Afrique sub-saharienne ; et les enfants de moins de 5 ans, tranche de la population la plus vulnérable, représentaient 77 % de la mortalité (WHO, 2021).

La morbidité et mortalité liées au paludisme au cours de ces vingt dernières années en Afrique et dans le monde a connu trois phases (figure 1.2). La période 2000 - 2015 a été marquée par une réduction forte et continue de la maladie. L'incidence du paludisme a diminué de 27 % sur cette période, et la mortalité de 52 %. Entre 2015 et 2019, les progrès ont stagné. Sur cette période, l'incidence n'a diminué que de 2 % entre 2015 et 2019 et la mortalité de 16 %. Enfin, l'année 2020 a été marquée par une augmentation significative des cas et de la mortalité, en partie liés à la pandémie de covid-19 qui a perturbé les services sanitaires (WHO, 2021).

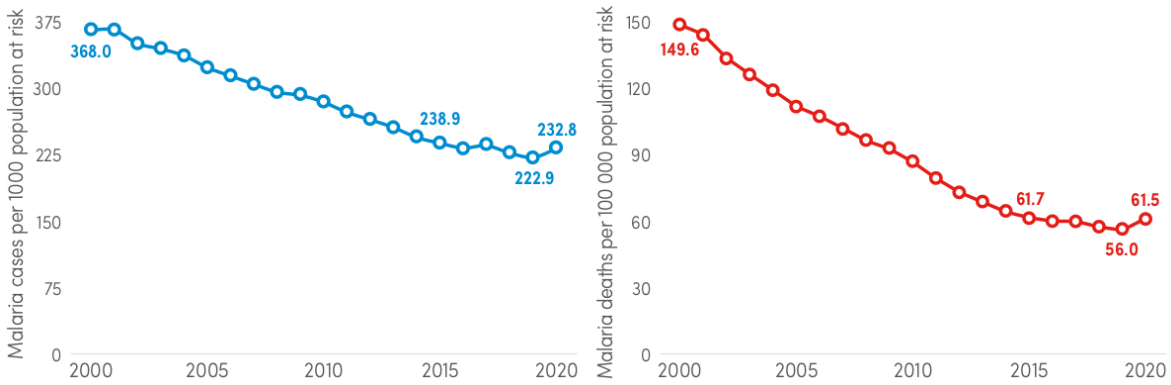


FIGURE 1.2: Morbité (à gauche) et mortalité (à droite) annuelle liées au paludisme dans la région Afrique de l'OMS entre 2000 et 2020 (WHO, 2021)

1.1.2 L'agent pathogène : *Plasmodium*

Le paludisme humain est une maladie infectieuse à transmission vectorielle faisant intervenir trois protagonistes : l'homme[1] (dit hôte) est infecté par un protozoaire parasite (dit agent infectieux) du genre *Plasmodium* qui lui a été transmis par un moustique (dit vecteur) du genre *Anopheles*. Les rôles de *Plasmodium* et *Anopheles* dans la maladie ont été découverts respectivement en 1880 et 1898 (Cox, 2010). Dans les prochaines sections, nous résumons le cycle biologique du parasite et du vecteur, et apportons quelques précisions supplémentaires, d'importance pour la thèse, sur les vecteurs.

Diversité et distribution

Parmi les 156 espèces de *Plasmodium* décrites, six causent le paludisme chez l'humain : *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*, *Plasmodium knowlesi* et *Plasmodium cynomolgi*. La plus pathogène des six espèces, *P. falciparum*, est à l'origine de la très grande majorité des cas de paludisme enregistrés en Afrique (WHO, 2021) (figure 1.3). Hors Afrique, c'est *P. vivax* qui prédomine (WHO, 2021).

1. dans ce manuscrit, nous employons le terme "homme" pour désigner l'être humain en général, selon la définition du dictionnaire Le Robert

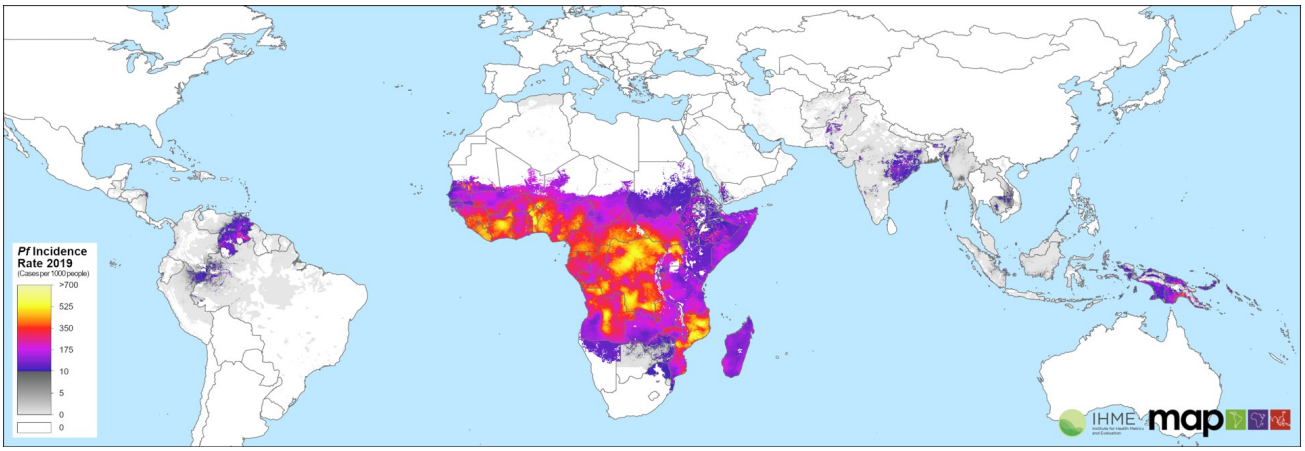


FIGURE 1.3: Taux d'incidence du paludisme à Plasmodium falciparum en 2019 (Weiss et al., 2019)

Cycle biologique

Le cycle biologique de *Plasmodium* fait intervenir deux hôtes : un moustique femelle du genre *Anopheles* et un être humain (figure 1.4). Le cycle chez l'anophèle commence lorsque celui-ci prend un repas sanguin sur un humain infecté, porteur de gamétocytes. Le parasite entame alors dans l'estomac de l'anophèle une phase de multiplication sexuée aboutissant à la migration des sporozoïtes jusqu'aux glandes salivaires du moustique. Ce premier cycle, chez l'anophèle, est appelé cycle sporogonique (ou extrinsèque) et dure environ une dizaine de jours selon l'espèce plasmodiale et la température (Baudon, Molez, & Guiguemde, 1984). Lors d'une prochaine piqûre une fois le cycle sporogonique effectué, l'anophèle alors infectieux injecte les sporozoïtes à l'homme. Ces derniers migrent dans le foie pour s'y multiplier (phase exo-érythrocytaire, 8 à 10 jours), puis sont libérés dans le sang sous forme de mérozoïtes qui pénètrent dans les globules rouges. S'ensuit une phase de multiplication des mérozoïtes dans les hématies, produisant de nouvelles cellules qui sont à leur tour libérées dans le sang et qui infecteront des érythrocytes sains (phase érythrocytaire). C'est cette libération qui entraîne les symptômes caractéristiques des accès palustres (frissons, chaleur et sueurs). Une partie des parasites peut également subir un processus de différenciation, aboutissant à la formation de gamétocytes mâles et femelles. Lors d'un repas de sang sur l'homme dès lors infectieux, un moustique anophèle femelle peut ingérers ces gamétocytes : le cyle recommence alors.

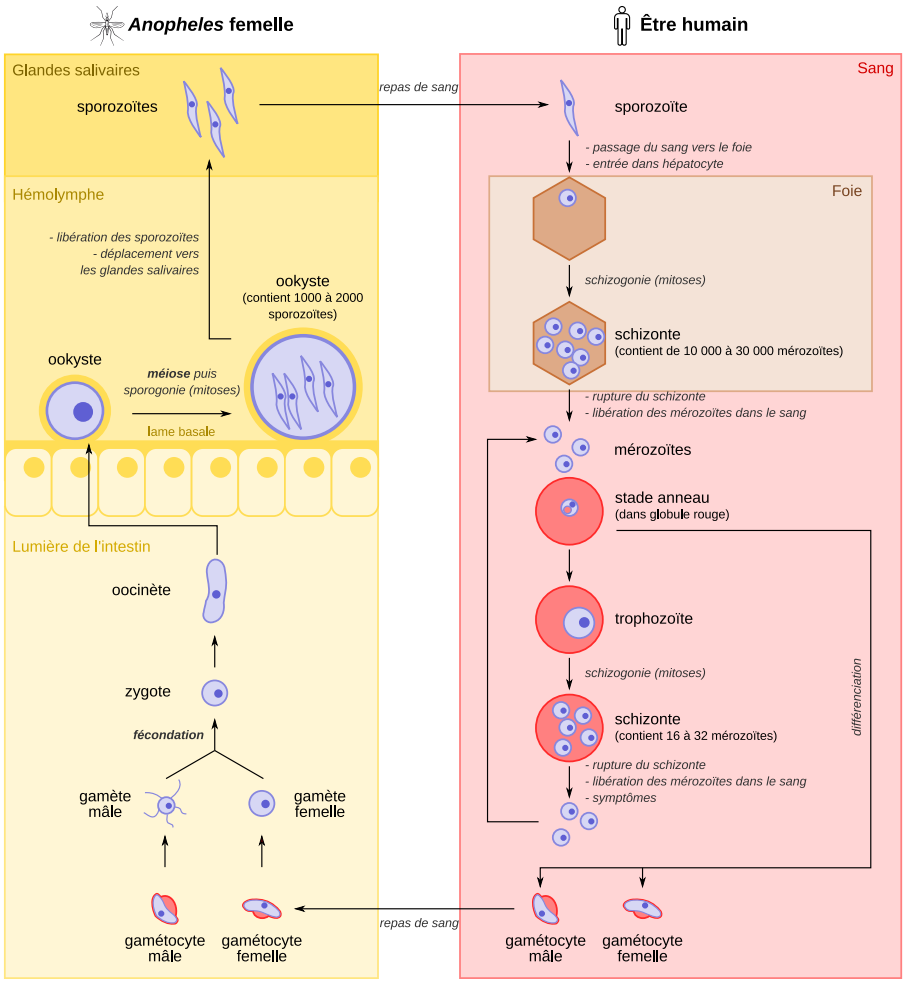


FIGURE 1.4: Cycle de développement et de reproduction des Plasmodium spp. (auteur : Pascal Combemorel, License : CC-BY-SA)

1.1.3 Le vecteur : *Anopheles*

Diversité et distribution

Sur les plus de 3000 espèces de moustiques (*Diptera : Culicidae*) recensées à ce jour, environ 500 font partie du genre *Anopheles*, dont une soixantaine est effectivement vectrice de la maladie. En Afrique sub-saharienne, on trouve au total environ 150 espèces d'anophèles, dont une trentaine vectrice de *Plasmodium*. Dans cette région, les principales espèces vectrices sont *An. arabiensis*, *An. gambiae s.s.* et *An. coluzzii* du com-

plexe Gambiae, et *An. funestus* du groupe Funestus (Sinka et al., 2010, 2012) (figure 1.5).

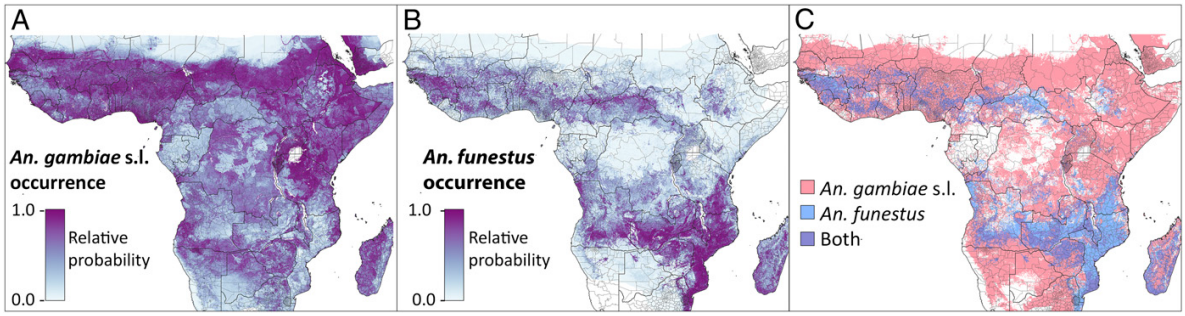


FIGURE 1.5: Distribution spatiale d'An. gambiae s.l. et An. funestus en Afrique sub-saharienne (Moyes et al., 2020)

Cycle biologique

Le cycle biologique de l'anophèle (figure 1.6) comprend quatre stades : oeuf, larve, nymphe et âge adulte. Le stade larvaire (œuf, larve, nymphe) se déroule en milieu aquatique et dure de 1 à 3 semaines en fonction de l'espèce et de la température ; quant au stade adulte, il se déroule en milieu aérien et la durée de vie de l'anophèle femelle peut aller jusqu'à 4 semaines (Holstein, 1952). Le stade adulte est marqué par une phase d'accouplement qui a lieu dans les 24 à 48 h suivant l'émergence. L'anophèle femelle ne copule en principe qu'une seule fois et stocke les spermatozoïdes dans une spermathèque jusqu'à sa mort. Une fois accouplée, l'anophèle femelle part à la recherche d'un hôte afin de prendre un repas de sang essentiel à la maturation des follicules ovariens. La piqûre est suivie d'une phase de repos au cours de laquelle la femelle digère le sang. Enfin, la femelle gravide cherche un site d'oviposition et y pond entre 40 et 100 œufs à la surface de l'eau. Après la ponte, la femelle cherche à prendre un nouveau repas sanguin afin d'effectuer une nouvelle oviposition ; et reproduit ce cycle (recherche de repas de sang, piqûre, repos, ponte) jusqu'à sa mort. Ce cycle est appelé gonotrophique et dure 2 à 5 jours en fonction, en particulier, de la température et de l'espèce (Gillies, 1953 ; Shapiro, Whitehead, & Thomas, 2017 ; Tchuinkam et al., 2010).

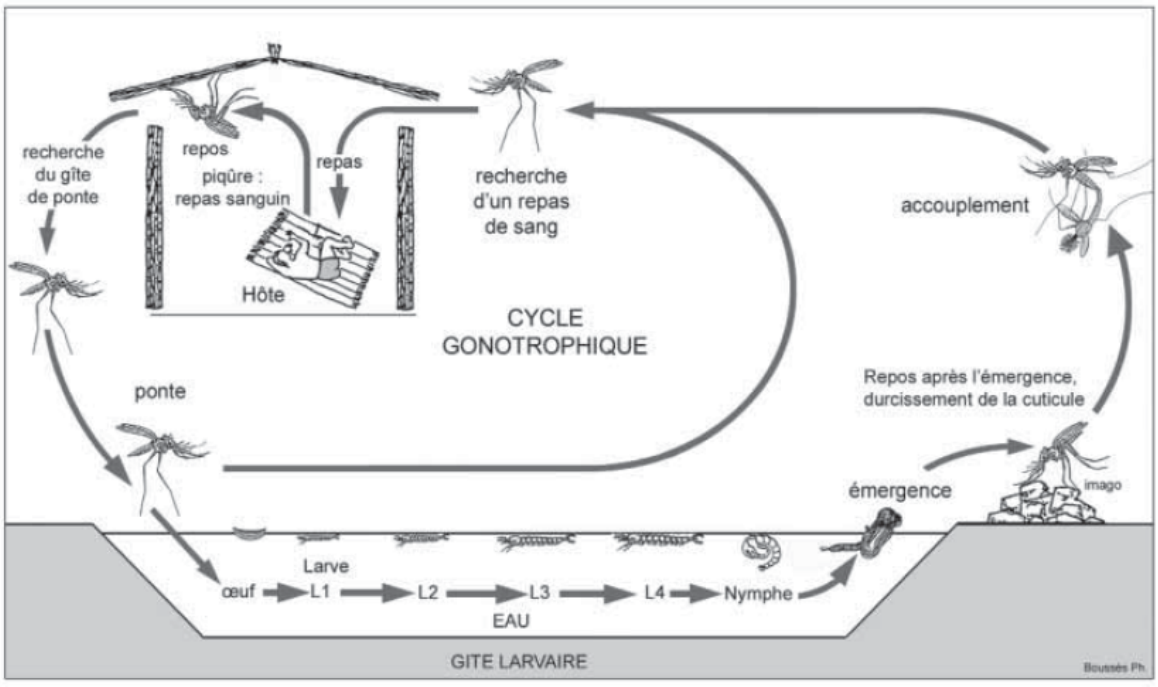


FIGURE 1.6: Cycle biologique de l'anophèle (Carnevale et al., 2009)

Comportement d'alimentation

La transmission de *Plasmodium* entre l'anophèle et l'homme s'effectue donc lorsque le vecteur prend un repas de sang sur l'hôte *via* la piqûre. Le comportement de piqûre (dit comportement trophique) de l'anophèle ainsi que son comportement de repos suivant la piqûre sont primordiaux dans l'étude de la transmission du paludisme. Quatre paramètres du comportement trophique et de repos des anophèles sont particulièrement déterminants :

— l'*anthropophilie* : propension d'un vecteur à piquer les humains (à l'opposé de zoophilie, désignant la propension à piquer les animaux) ;
— l'*endophagie* : propension d'un vecteur à piquer à l'intérieur des maisons (à l'opposé d'exophagie, désignant la propension à piquer à l'extérieur des habitations) ;
— l'*endophilie* : propension d'un vecteur à se reposer, après la piqûre, à l'intérieur des maisons (à l'opposé d'exophilie, désignant la propension à se reposer à l'extérieur des habitations) ;
— l'*activité précoce ou tardive* : propension d'un vecteur à piquer précocément ou tardivement dans la nuit, par rapport aux horaires habituellement observés (voir

ci-dessous).

Elements de bionomie

Bien que le cycle biologique de tous les anophèles soit identique, les différentes espèces (ainsi que les différents individus au sein d'une même espèce) exhibent souvent des préférences écologiques ou trophiques sensiblement différentes. Ci-après, nous donnons quelques éléments de bionomie (gites larvaires et comportements préférentiels) des 4 principales espèces d'anophèles vectrices en Afrique sub-saharienne. Ces éléments sont intégralement extraits de synthèses bibliographiques sur la bionomie des vecteurs effectuées à l'échelle de l'Afrique sub-saharienne, effectuées par Sinka et al. (2010) (gites larvaires préférentiels + comportements trophiques) et Sherrard-Smith et al. (2019) (comportement trophique).

An. gambiae s.s. et *An coluzzii* sont des espèces majoritairement associées aux gites larvaires respectivement temporaires et semi-permanents, globalement d'eaux douces peu profondes et ensoleillées. Ainsi, les larves d'*An. gambiae s.s.* sont typiquement retrouvées dans les gites se remplissant avec les précipitations, telles que les flaques d'eau, les empreintes de sabot ou les ornières ; et *An. coluzzii* pond typiquement dans les rizières ou zones inondées contenant de la végétation flottante et immergée, telles que les bas-fonds et zones marécageuses. Les deux espèces ont été décrites comme hautement anthropophiles et majoritairement endophages et endophiles. L'activité de piqûre de ces espèces se concentre au milieu de la nuit avec une tendance à la piqûre plutot tardive (figure 1.7). Malgré ces grandes tendances, ces espèces présentent une certaine plasticité phénotypique dans leur comportement de piqûre et de repos.

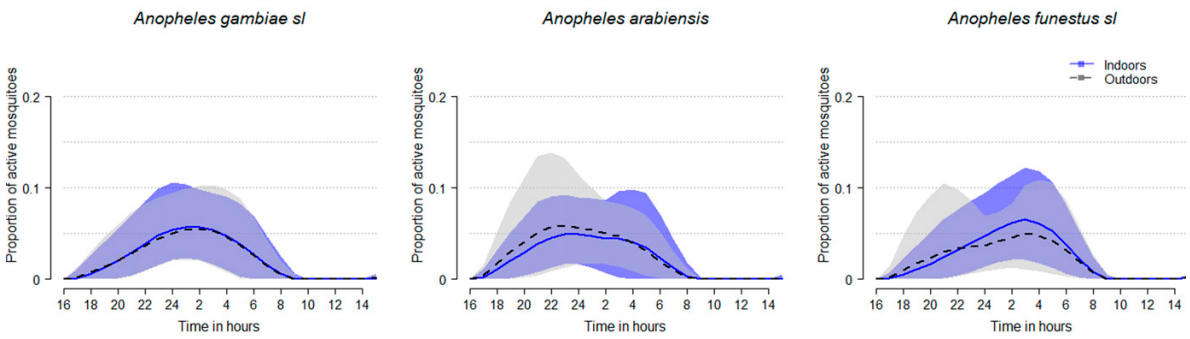


FIGURE 1.7: Courbes d'agressivité horaire nocturne pour trois espèces majeures d'anophèles en Afrique (Sherrard-Smith et al., 2019)

An. arabiensis est une espèce associée aux environnements de savanes et forêts clairsemées. Les gites larvaires d'*An. arabiensis* ressemblent à ceux d'*An. gambiae s.s.* : petits bassins d'eau douce temporaires, ensoleillés, clairs et peu profonds ; bien que l'espèce ait également été observée dans d'autres gites larvaires plus profonds ou turbides. *An. arabiensis* est considérée davantage zoophage, exophage et exophile qu'*An. gambiae s.s.*. Cependant, l'espèce montre un large éventail de comportements de piqûre et de repos en fonction des zones géographiques, constituant ainsi une espèce au comportement a piori plus plastique encore qu'*An. gambiae*. *An. arabiensis* pique en moyenne plus tôt dans la nuit qu'*An. gambiae*. Les courbes d'agressivité diffèrent cependant selon le site de piqûre, les vecteurs exophages étant actifs relativement plutôt précocément et les vecteurs endophages relativement plus tardivement (figure 1.7).

An. funestus, de son côté, est une espèce majoritairement associée aux grandes étendues d'eau douces, permanentes ou semi-permanentes, contenant une végétation émergente ou flottante, comme les marécages, les grands étangs et les bords de lacs. C'est une espèce décrite comme hautement adaptable, ce qui lui permet d'occuper et de maintenir une distribution spatiale large. *An. funestus* est considérée hautement anthropophile. L'espèce est majoritairement endophage, et présente un comportement de piqûre relativement tardif (figure 1.7). Par rapport aux autres espèces de vecteurs dominantes en Afrique, *An. funestus* présente des schémas comportementaux assez stables (généralement anthropophile et endophile) dans l'ensemble de son aire de répartition.

1.1.4 Le système vectoriel

Pour que la transmission du paludisme puisse s'effectuer, hôte, vecteur et agent infectieux doivent interagir dans un environnement favorable (Reisen, 2010). La triade vectorielle (système composé de l'agent pathogène, du vecteur et de l'hôte), complétée de l'environnement dans lequel elle évolue et de l'ensemble des interactions entre les acteurs et l'environnement, forme le *système vectoriel* (Rodhain, 2015) (figure 1.8). Comme nous allons le voir plus loin (section 1.3), l'analyse des intéractions dans le système vectoriel est au cœur de l'étude du risque de transmission du paludisme, et plus largement des maladies à transmission vectorielle.

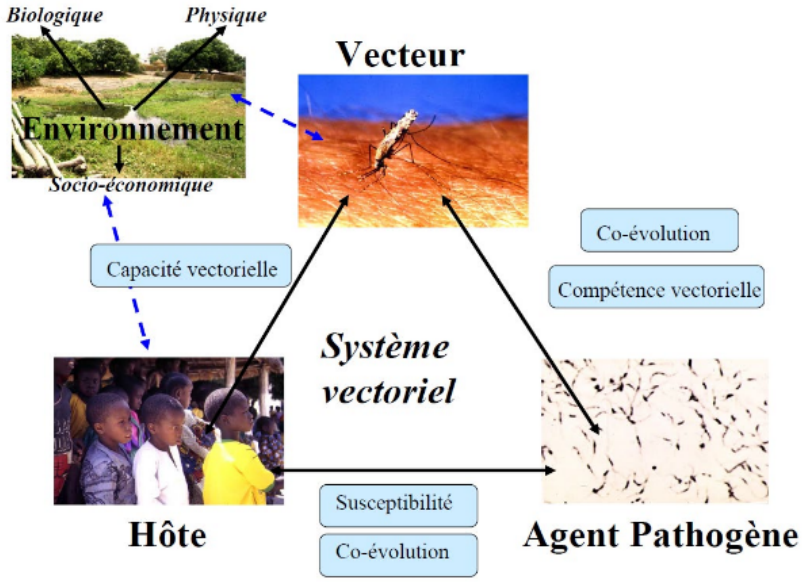


FIGURE 1.8: Système vectoriel : agent pathogène, vecteur, hôte, environnement (Fontenille, 2009)

De part son impact sur les traits de vie de chacun des protagonistes de la triade vectorielle, l'environnement (pris au sens large du terme : météorologie, paysage, facteurs socio-culturels, etc.) conditionne considérablement les dynamiques épidémiologiques, notamment spatio-temporelles, des maladies vectorielles (Reisen, 2010). Ainsi par exemple, des traits de vie des anophèles tels que l'émergence, la croissance, la survie, la dispersion, ou encore l'activité (notamment trophique) peuvent être impactés par des facteurs environnementaux météorologiques (températures, précipitations, humidité, etc.), paysagers (utilisation et occupation du sol, etc.), anthropiques (interventions de lutte contre

les vecteurs, etc.), etc.[2]. Il en découle que certains indicateurs entomologiques de la transmission tels que la densité agressive des vecteurs (nombre de piqûres / homme / nuit) sont, à priori, largement dépendants des conditions environnementales (Moiroux et al., 2013, 2014). La figure 1.9 expose par exemple un ensemble de facteurs ayant été identifiés, dans la bibliographie, comme pouvant impacter les densités agressives des anophèles ('m.a. vecteur'), ainsi que les relations à priori existantes entre ces facteurs (Moiroux, 2012).

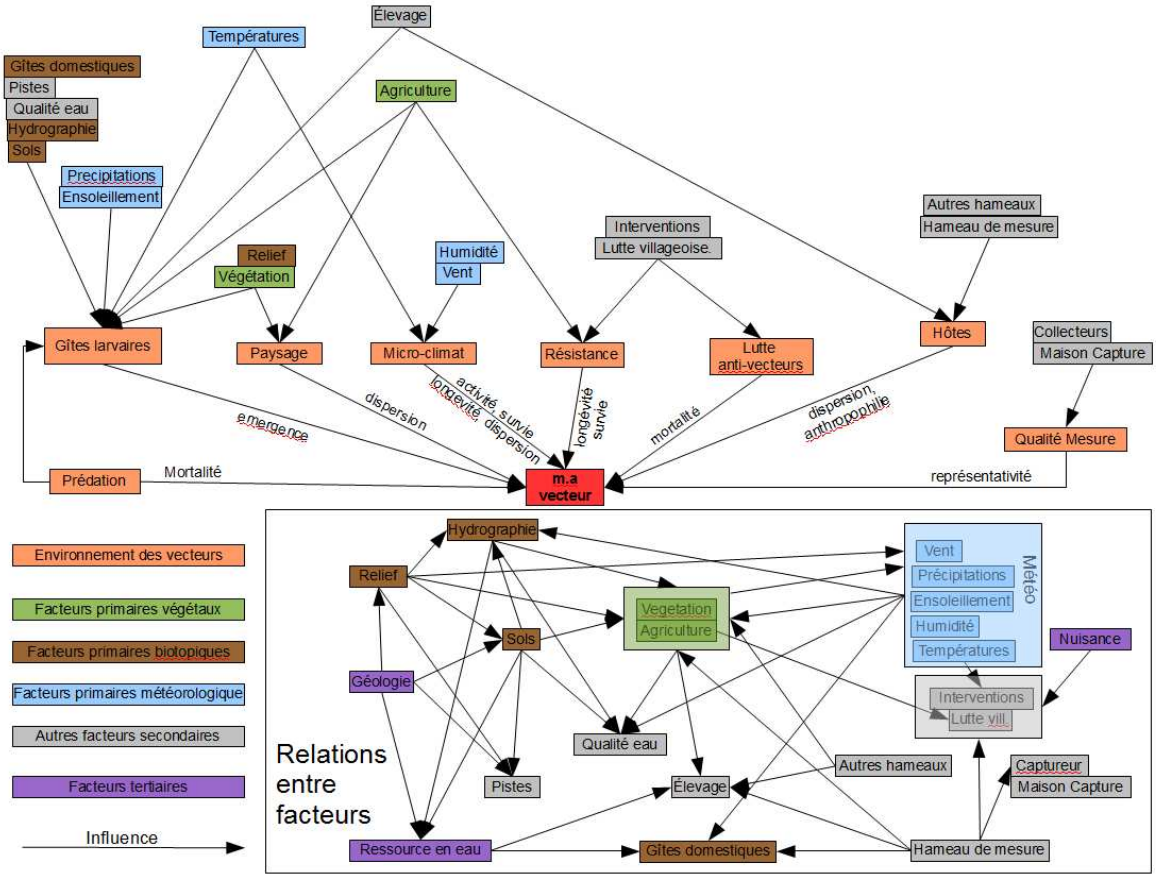


FIGURE 1.9: Modèle conceptuel du système biologique {densités agressives des anophèles - environnement} (Moiroux, 2012)

2. Les liens entre environnement et traits de vie des anophèles sont plus largement détaillés dans les chapitres 4 et 5 du manuscrit

1.1.5 Lutte anti-vectorielle

La lutte contre le paludisme s'oriente autour de trois axes : i) prévention, ii) diagnostic des cas suspects et iii) traitement des cas confirmés. Le diagnostic et le traitement visent respectivement à identifier la présence de *Plasmodium* dans le corps humain et à traiter les cas confirmés. La prévention, en amont, vise à réduire le risque de transmission ou les conséquences de la maladie si l'infection a lieu. Trois méthodes de prévention existent : la chimiothérapie préventive, la vaccination, et la lutte anti-vectorielle (WHO, 2021). Attardons-nous sur la lutte anti-vectorielle, seule des trois méthodes ciblant le moustique vecteur.

Concept

La lutte anti-vectorielle (LAV) consiste en un ensemble d'outils et de méthodes visant à empêcher la transmission des parasites depuis le vecteur vers l'humain. Dans leur immense majorité, ces outils ont pour objectif de limiter la probabilité (i) soit qu'un contact entre l'anophèle et l'humain se réalise (cad. la piqûre, ou encore l'intéraction homme-vecteur), (ii) soit qu'un moustique atteigne l'âge épidémiologiquement dangereux, (iii) soit les deux. Les leviers d'action sont multiples : réduire la densité des vecteurs, réduire leur longévité, ou empêcher physiquement le contact homme-vecteur. Pour cela, toutes sortes d'outils de lutte anti-vectorielle existent.

Principaux outils de LAV

Comme présenté précédemment, les principaux vecteurs du paludisme en Afrique ont été historiquement décrits comme majoritairement endophages, endophiles et nocturnes. C'est sur la base de ces traits comportementaux qu'ont été élaborés les deux principaux outils de lutte anti-vectorielle utilisés aujourd'hui dans la lutte contre la paludisme (WHO, 2021) : la moustiquaire Imprégnée d'Insecticide à Longue Durée d'Action (MIILDA) et les Pulvérisations Intra-Domiciliaires d'insecticide à effet rémanent (PID).

La MIILDA est un rideau de tulle imprégné d'insecticide dont on entoure en général les lits. La MIILDA offre une double barrière face aux vecteurs :

— barrière physique : le rideau de tulle offre une protection individuelle contre les vecteurs pour les personnes utilisant la moustiquaire, en empêchant le vecteur en

recherche de repas de sang d'entrer en contact avec l'hôte
— barrière chimique : l'insecticide dont la MIILDA est imprégnée a un effet à la fois repulsif (à distance) et létal (pour les vecteurs entrant en contact avec la moustiquaire).

Cette barrière chimique réduit donc la longévité des vecteurs sensibles à l'insecticide, et ainsi sa probabilité d'atteindre l'âge epidémiologiquement dangereux. Par ailleurs, en réduisant la longévité des vecteurs individuellement, les MIILDA réduisent leur densité de population globale, et protègent donc théoriquement également les non-utilisateurs de moustiquaires dans la communauté (Hawley et al., 2003 ; Gerry F. Killeen & Smith, 2007). Cet effet de protection, appelé "communautaire", se manifeste au delà d'un certain seuil d'utilisation des MIILDA dans une communauté donnée (les exercices de modélisation mathématique avancent un seuil situé entre 35 % et 65 % en fonction des spécificités écologiques locales (Gerry F. Killeen et al., 2007)).

La MIILDA a été l'outil de LAV phare du programme mondial de lutte contre le paludisme *Roll back malaria*, lancé par l'OMS en 2000. Ainsi, au niveau mondial, 2,3 milliards de moustiquaires imprégnées d'insecticide ont été vendues par les producteurs entre 2004 et et 2020 (WHO, 2021) ; et une grande partie de ces moustiquaires a été distribuée aux populations exposées au risque de paludisme par le biais des différents Programmes Nationaux de Lutte contre le Paludisme (PNLP) [3]. On estime en 2020 que 65% des maisons en Afrique sub-Saharienne étaient équipées d'au moins une moustiquaire et que 43 % de la population dormait sous une moustiquaire (WHO, 2021). L'efficacité des MIILDA a été largement documentée et prouvée : on estime qu'elle a permis d'éviter environ 450 millions de cas de paludisme et 1 million de décès associés entre 2000 et 2015 (Bhatt et al., 2015).

Les PID sont, après la MIILDA, le deuxième outil de LAV le plus communément utilisé (WHO, 2021). La méthode des PID consiste à pulvériser un insecticide sur les murs intérieurs d'une habitation, afin de réduire la longévité (et donc la densité) des vecteurs venant se reposer sur les murs intérieurs de l'habitation après la piqûre. Elle vise donc les vecteurs endophiles. En 2020, 5,3 % de la population africaine à risque était protégée par les PID (WHO, 2021).

3. Organismes nationaux chargés de mettre en oeuvre les politiques de lutte contre le paludisme

FIGURE 1.10: Installation d'une moustiquaire imprégnée (gauche) et pulvérisations intra-domiciliaire d'insecticide (droite) (crédit photo : Jean-Jacques Lemasson et Vincent Robert)

Une myriade d'outils et méthodes de lutte anti-vectorielle existent en sus de la MIILDA et des PID (Wilson et al., 2020) (la figure 1.11 en présente certains), mais restent à ce jour utilisés en proportion bien moindre : les répulsifs individuels, la lutte anti-larvaire, les pulvérisations spatiales extérieures, la lutte génétique, les grillages de fenêtres, etc.

Chemical	Immature	Chemical larvicides	Contact pesticides affecting insect nervous system (e.g., temephos) or endocrine system (insect growth regulators, e.g., pyriproxyfen)
	Adult	ITNs	Pyrethroid-treated ITNs or combination ITNs (e.g., pyrethroid plus synergist piperonyl butoxide) for malaria, LF, and leishmaniasis control
		Insecticide-treated materials for personal protection	Insecticide-treated clothing for workers and mobile populations
		IRS	Spraying of residual insecticides (typically either pyrethroids, carbamates, or organophosphates) indoors for malaria and *Aedes*-borne disease control
		Space spraying	Aircraft, vehicle or hand-held space spraying for dengue epidemic and other *Aedes*-borne disease control
		Insecticidal treatment of habitat	Focal, perifocal, ground, or aerial insecticide spraying
		Insecticide-treated cattle	Pour-on or spot-on pyrethroids for control of tsetse
		Insecticide-treated traps and targets	Targets for control of HAT and insecticide-treated adulticidal oviposition traps for *Aedes*-borne diseases
		Topical repellent	Chemicals (e.g., N,N-diethyl-meta-toluamide [DEET], picaridin) applied to the skin to reduce vector biting
		Spatial repellent	Transfluthrin/metafluthrin passive emanators or coils
Nonchemical	Immature	Microbial larvicides	*Bacillus thuringiensis* var. *israelensis*, *B. sphaericus*
		Predator species	Predatory fish or invertebrates
		Habitat modification, i.e., a permanent change of land and/or water	Drainage of surface water, land reclamation and filling, and coverage of large water storage containers (or complete coverage of water surfaces) with a material that is impenetrable to mosquitoes, such as expanded polystyrene beads
		Habitat manipulation, i.e., a recurrent activity	Water-level manipulation, exposing habitats to the sun (depending on the ecology of the vector), flushing of streams, drain clearance, and source reduction, including rubbish disposal and regular emptying and cleaning of domestic containers (e.g., flowerpots, animal drinking water troughs)
		Regulatory measures	Removal of man-made aquatic habitats and appropriate waste disposal
	Adult	House improvement and screening	Closing eaves, door and window screening
		Removal trapping	Solar-powered mosquito trapping system for malaria control and sticky adulticidal oviposition traps for *Aedes*-borne diseases

Abbreviations: HAT, human African trypanosomiasis; IRS, indoor residual spraying; ITN, insecticide-treated bed net; LF, lymphatic filariasis

https://doi.org/10.1371/journal.pntd.0007831.t002

FIGURE 1.11: Exemples d'outils de LAV utilisés contre la transmission des maladies vectorielles, triés par catégories (basés ou non sur les insecticides) et stade de développement du vecteur ciblé (Wilson et al., 2020)

1.2 Transmission résiduelle du paludisme : problématique, définition, enjeux

1.2.1 Limites actuelles de la LAV

Ces outils, MIILDA en tête, ont donc été et restent les principaux artisans de la diminution de l'incidence du paludisme à large échelle. Cependant, les niveaux toujours soutenus de transmission et la récente stagnation - voire augmentation dans certaines régions - du nombre de cas de paludisme dans des régions pourtant couvertes par ces outils, questionne : quelles en sont les limites ? Pourquoi cette stagnation ? Plusieurs hypothèses sont généralement avancées :

— *Fraction de la population de vecteurs ciblés.* Ces outils présentent certaines limites intrinsèques. Comme expliqué précédemment, les MIILDA et les PID ciblent, par définition, les vecteurs endophages, endophiles, et anthropophages. Le corollaire à cette observation est que tout vecteur exophage, exophile, ou zoophage leur échappe. Aussi, dans les zones où les vecteurs montrent de tels comportements, ces outils sont à priori peu efficaces (Gerry F. Killeen, 2014).

— *Taux de possession et d'utilisation.* Ces outils ne sont efficaces que s'ils sont disponibles et utilisés. Bien qu'assez triviale, cette assertion peut expliquer en partie les niveaux toujours élevés de transmission. Comme énoncé précedemment, la couverture en outils de LAV et leur utilisation est loin d'être universelle. Par ailleurs, les taux de possession et utilisation de moustiquaires sont très variables selon les sous-régions, pays, et à des échelles spatiales plus fines encore (Bertozzi-Villa et al., 2021) (figure 1.12). Notons, sur la figure 1.12, l'évolution de la possession et utilisation des moustiquaires à l'échelle de l'Afrique : les taux moyens ont fortement augmenté entre 2000 et 2017 mais déclinent depuis cette année-ci.

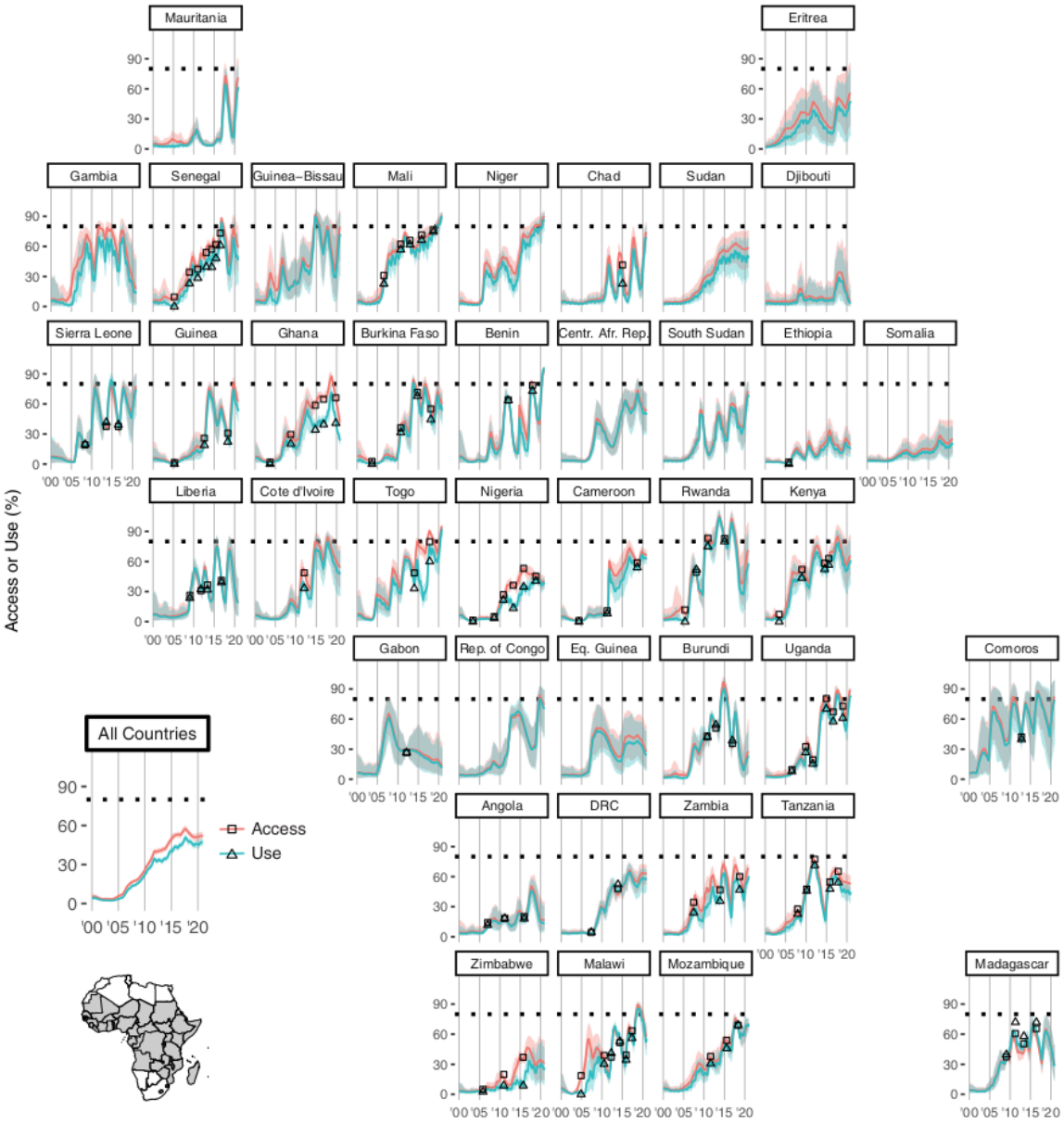


FIGURE 1.12: Evolution du taux de possession et d'utilisation de moustiquaires par pays en Afrique (Bertozzi-Villa et al., 2021)

— *Altération de la composition spécifique des vecteurs.* Ces outils sont susceptibles d'altérer la composition spécifique des vecteurs, en réduisant, à terme, la part des vecteurs endophiles, endophages, anthropophages, nocturnes et en favorisant les espèces exophages, exophiles, zoophages, et piquant précocément ou tardivement (Derua et al., 2012 ; Gatton et al., 2013 ; Mwangangi et al., 2013 ; Russell et al.,

2011 ; S. Sougoufara, Harry, Doucouré, Sembène, & Sokhna, 2016). Ces vecteurs, non ciblés par les MIILDA ou les PID, seront alors susceptibles de transmettre le paludisme.

— *Développement de mécanismes de résistance aux insecticides.* Enfin, on observe que les vecteurs au départ ciblés par ces outils développent des mécanismes de résistance aux insecticides leur permettant d'éviter ou de contourner leurs effets léthaux (Corbel & N'Guessan, 2013 ; Durnez & Coosemans, 2013 ; Gatton et al., 2013 ; Gerry F. Killeen, 2014 ; Riveron et al., 2018).

Parmi les différentes limites et problématiques sus-mentionnées, le développement de résistances aux insecticides dans les populations vectorielles est particulièrement important et probablement, impactant. En effet, les résistances menacent directement l'efficacité des principaux outils de lutte anti-vectorielle. La section suivante précise les différentes formes de résistance décrites dans la littérature, présente brièvement les mécanismes impliqués dans le développement des résistances, et décrit succinctement leur distribution spatio-temporelle en Afrique.

1.2.2 Résistances des anophèles aux insecticides

On reconnaît deux formes principales de résistances des vecteurs aux insecticides : les résistances physiologiques et les résistances comportementales (Lockwood, Sparks, & Story, 1984 ; Sokhna, Ndiath, & Rogier, 2013).

Résistances physiologiques

La résistance physiologique fait référence à un ensemble de mécanismes qui permettent au moustique de survivre à un contact avec l'insecticide (Davidson, 1957). Les bases moléculaires et génétiques de la résistance physiologique sont bien connues : sous la pression des insecticides, les mutations qui permettent aux vecteurs de survivre sont naturellement sélectionnées et se propagent ensuite dans les générations successives (Labbé et al., 2017 ; Martinez-Torres et al., 1998).

Plusieurs mécanismes de résistance physiologique aux insecticides ont été décrits, notamment biochimiques et morphologiques. Les modifications de la cible physiologique de l'insecticide - forme de résistance physiologique qui va être étudiée dans la suite

de cette thèse - provoquent une réduction de la sensibilité aux insecticides en raison de mutations ponctuelles sur les gènes codant pour les protéines cibles (Davies, Field, Usherwood, & Williamson, 2007 ; O'Reilly et al., 2006). La mutation la plus commune décrite chez les membres du complexe *Gambiae* est la mutation dite "kdr" ("knock-down resistance"). Cette mutation induit une résistance aux pyréthrinoïdes et aux organochlorés, insecticides les plus largement utilisés dans la lutte anti-vectorielle. On distingue 2 formes principales pour cette mutation : la mutation L1014F (ou "kdr-ouest") - historiquement détectée et largement répandue en Afrique de l'Ouest - et la mutation L1014S (ou "kdr-est") - historiquement détectée et largement répandue en Afrique de l'Est (Martinez-Torres et al., 1998 ; Ranson et al., 2000). Une autre mutation dite "ace-1" (G119S) induit chez les anophèles une résistance aux carbamates, et dans une moindre mesure, aux organochlorés (Weill et al., 2004).

Bien que les premières traces de résistances physiologiques chez les anophèles aient été observées bien avant les distributions massives des MIILDA (Corbel & N'Guessan, 2013), on note une corrélation temporelle forte entre le déploiement à large échelle des outils de lutte anti-vectorielle (figure 1.12) et la généralisation des résistances physiologiques chez les vecteurs du paludisme depuis les années 2000 (figure 1.13). La problématique des résistances physiologiques aux insecticides concerne maintenant la quasi-totalité des populations d'anophèles dans la sous-région ouest-africaine, et en particulier dans les zones d'étude de cette thèse - au Burkina Faso et en Côte d'Ivoire (Moyes et al., 2020) (figure 1.14).

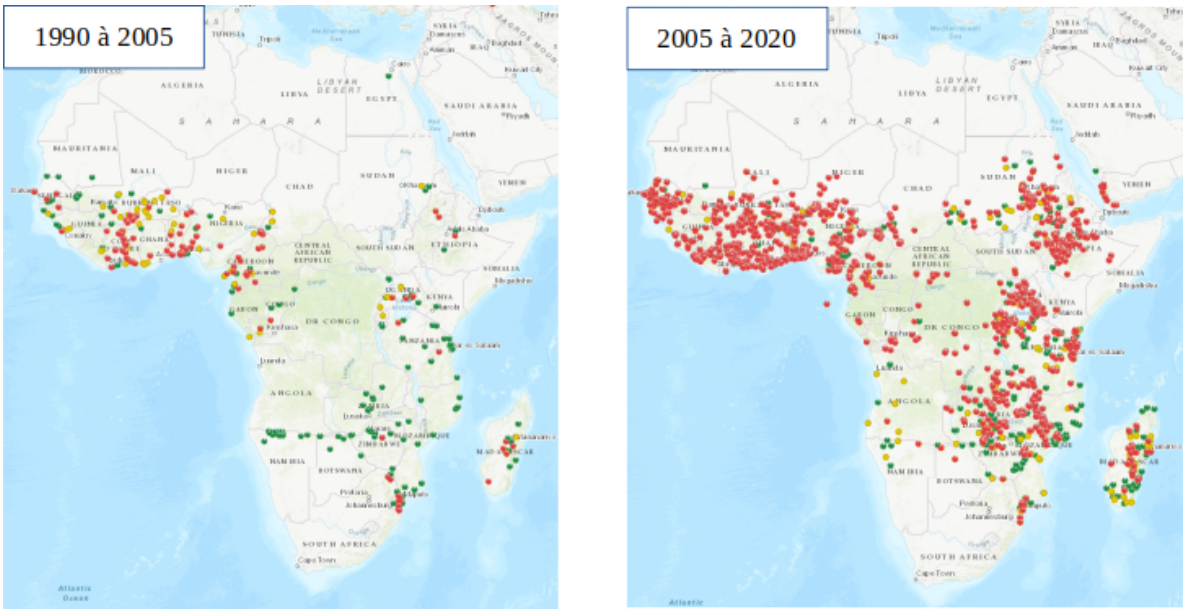


FIGURE 1.13: Distribution spatiale des études confirmant une résistance aux insecticides chez les anophèles entre 1990 et 2005 (à gauche) et entre 2005 et 2020 (à droite) (source : IR Mapper www.irmapper.com)

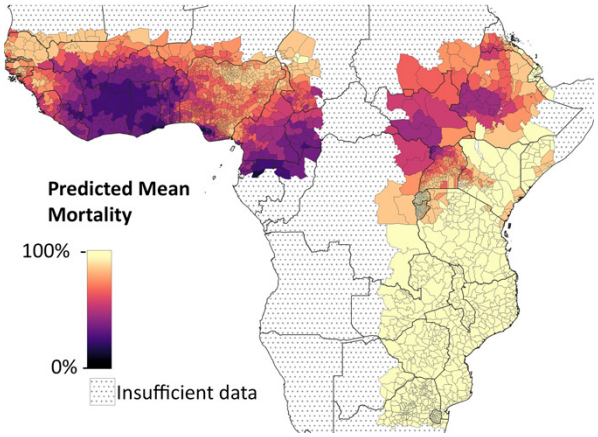


FIGURE 1.14: Distribution spatiale de la résistance à la deltaméthrine dans les populations d'An. gambiae s.l. (Moyes et al., 2020)

Résistances comportementales

La résistance comportementale consiste en des modifications dans le comportement du moustique lui permettant de prévenir ou réduire les conséquences négatives des insecticides (Carrasco et al., 2019). La résistance peut être qualitative (modifications

qui empêchent ou limitent le contact avec l'insecticide) ou quantitative (modifications qui arrêtent, limitent ou réduisent l'action de l'insecticide une fois le contact établi) (Carrasco et al., 2019). A ce jour, les mécanismes de résistance comportementale décrits dans la littérature sont principalement qualitatifs et consistent en des évitements spatiaux, temporels ou trophiques de l'insecticide. En particulier, dans les populations d'anophèles, les mécanismes de résistance qualitative comportementale suivants ont été décrits après la mise à l'échelle des outils de LAV à base d'insecticides (Durnez & Coosemans, 2013) : i) augmentation des comportements exophages ou exophiles (évitement spatial), ii) augmentation des comportements de piqûre précoce ou tardive (évitement temporel), iii) augmentation des comportements zoophages (évitement trophique).

Contrairement à la résistance physiologique, les mécanismes biologiques qui sous-tendent les résistances comportementales sont encore mal connus (Main et al., 2016 ; Carrasco et al., 2019 ; Durnez & Coosemans, 2013, ; Gerry F. Killeen, 2014). En particulier, il reste à comprendre si les changements de comportement reflètent des adaptations évolutives en réponse à la pression induite par les insecticides utilisés dans la LAV, comme pour les résistances physiologiques (*résistance constitutive*) ou sont des manifestations d'une plasticité phénotypique préexistante qui se déclenche face à l'insecticide ou en réponse à une variation environnementale qui réduit la disponibilité des hôtes humains (*résistance inductible*) ; ces mécanismes n'étant pas mutuellement exclusifs (Durnez & Coosemans, 2013). Ces considérations peuvent avoir des implications importantes en matière d'efficacité sur le long terme des outils de LAV actuels. En effet, la résistance inductible pourrait impliquer que les vecteurs retrouvent rapidement leurs comportements de base lorsque les interventions de LAV sont modifiées, tandis que la résistance constitutive (héréditaire), qui pourrait impliquer que les vecteurs sensibles soient peu à peu remplacés par des vecteurs résistants, pourrait éroder progressivement et durablement l'efficacité des outils de LAV actuels. Certaines études récentes tendent à montrer qu'il pourrait y avoir une composante héréditaire à ces comportements résistants chez *An. arabiensis* (Govella, Johnson, Killeen, & Ferguson, 2021).

Les résistances comportementales sont à ce jour, dans l'ensemble, moins étudiées que les résistances physiologiques (mécanismes biologiques sous-jacents moins compris, distributions spatio-temporelles moins rapportées, etc.) (Carrasco et al., 2019). A

notre connaissance, une seule revue systématique des données existantes à l'échelle de l'Afrique a été effectuée (Sherrard-Smith et al., 2019). Cette étude rapporte, entre autres, les variations spatiales et temporelles des taux d'endophagie (et donc exophagie) des vecteurs : elle montre notamment qu'à l'échelle de l'Afrique, la proportion des piqûres de moustiques effectuées à l'extérieur des habitations a augmenté de presque 10 % entre 2003 et 2018 (figure 1.15), et qu'une telle augmentation de l'exophagie pourrait résulter en un accroissement significatif de l'incidence du paludisme à l'échelle du continent (+ 10,6 millions de cas annuels).

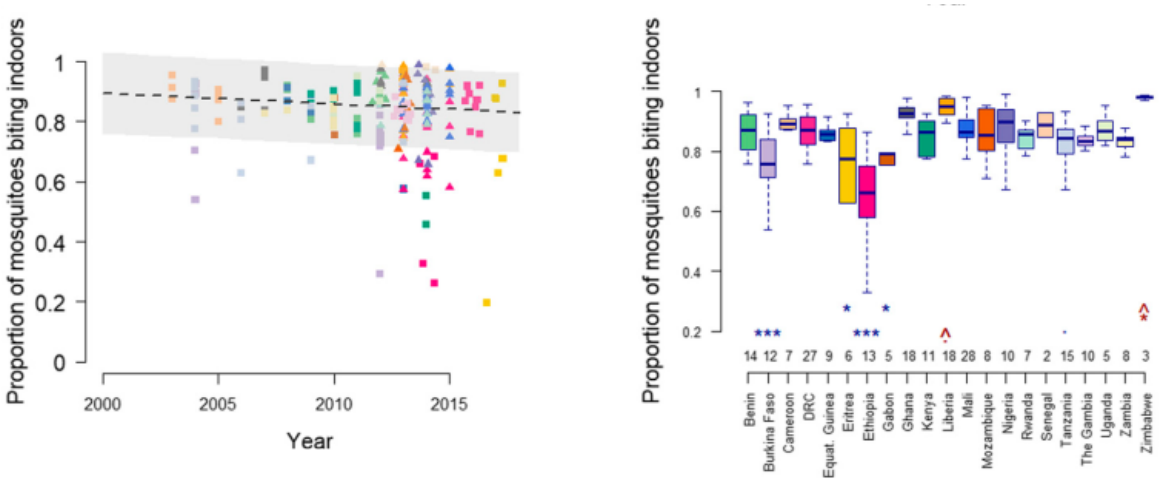


FIGURE 1.15: Distribution temporelle (en haut) et spatiale (en bas) du taux d'endophagie des anophèles en Afrique (Sherrard-Smith et al., 2019)

1.2.3 Transmission résiduelle du paludisme

Les différentes limites des principaux outils de LAV aujourd'hui utilisés expliquent donc que la transmission continue de s'effectuer malgré la mise en oeuvre de ces interventions. La transmission qui persiste après avoir atteint une couverture universelle complète en MIILDA et/ou PID est appelée *transmission résiduelle* du paludisme (Gerry F. Killeen, 2014).

On peut envisager deux scenarii d'évolution de l'intensité de transmission résiduelle suite à l'introduction de MIILDA ou PID (Gerry F. Killeen, 2014). Dans les deux scénarii, dans un premier temps l'intensité de la transmission diminue, jusqu'à atteindre

un palier bas, sans disparaitre totalement à cause des limites inhérentes aux outils. Dans un second temps :

- soit **l'intensité de la transmission reste stable** (scenario 1), car :
  - les outils de LAV restent largement utilisés ;
  - les résistances comportementales des vecteurs sont induites (la fraction de vecteurs échappant aux outils de LAV reste donc stable)
- soit **l'intensité de la transmission réaugmente** (rebond de la transmission) (scénario 2), à cause de :
  - une diminution progressive des taux d'utilisation des outils de LAV (par exemple, à cause d'une réduction des financements publics, ou de réticenses de la population à utiliser les interventions) ;
  - et/ou une augmentation progressive de la prévalence des vecteurs physiologiquement résistants ;
  - et/ou une augmentation progressive de la prévalence des vecteurs comportementalement résistants (dans ce scénario, les résistances comportementales sont donc constitutives) ;
  - et/ou une modification progressive de la composition spécifique des vecteurs, vers des espèces davantage exophages ou piquant précocement ou tardivement.

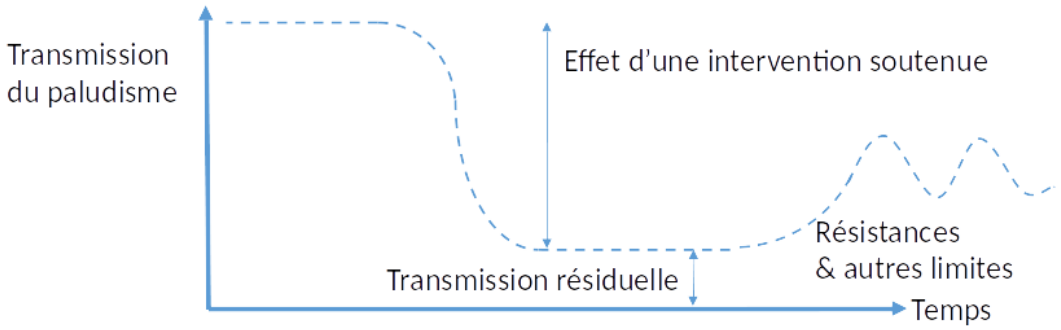


FIGURE 1.16: Concept de transmission résiduelle du paludisme (adapté de (Gerry F. Killeen, 2014))

Ces différentes problématiques concernant la transmission résiduelle du paludisme nous amènent ainsi à la présentation des enjeux et objectifs de la présente thèse.

1.3 Enjeux, objectifs, organisation de la thèse

1.3.1 Mesurer et caractériser, comprendre, prédire le risque le risque de transmission résiduelle du paludisme

Pour éviter les rebonds de transmission résiduelle, limiter leur impact, et plus largement redynamiser le progrès de la lutte contre le paludisme, plusieurs pistes sont proposées par la communauté des acteurs de la lutte contre le paludisme. En particulier, il est préconisé de concevoir de nouvelles interventions et méthodes de lutte, d'adapter les interventions au contexte local, et de cibler et prioriser le déploiement des interventions (WHO, 2020, 2021). Afin de tendre vers ces objectifs opérationnels, il est d'une part nécessaire d'appronfondir certaines connaissances fondamentales sur les déterminants de la transmission résiduelle (par exemple, les mécanismes biologiques sous-jacents aux résistances comportementales). D'autre part, pour optimer et prioriser les interventions sur un territoire d'intérêt, il est important d'acquérir une connaissance fine du risque de transmission résiduelle, en particulier, de ses composantes, de son intensité, et de sa distribution spatio-temporelle sur ce territoire. Nous proposons ci-après trois approches, constituant autant d'enjeux de la thèse, permettant de générer des connaissances essentielles à ces effets. Ces approches visent respectivement à i) mesurer et caractériser le risque de transmission résiduelle, ii) comprendre ce risque, et iii) prédire ce risque.

Approche n°1 : Mesurer et caractériser le risque

Le risque de transmission résiduelle peut être défini comme la probabilité qu'un anophèle entre en contact avec un humain (autrement dit, probabilité de contact homme-vecteur), en zone couverte par les MIILDA ou PID. Bien que le simple contact avec l'humain ne soit pas suffisant pour transmettre le parasite (il faut par exemple, en sus, que l'anophèle soit infectieux, que la piqûre soit suffisamment longue, etc.), nous admettrons cette définition du risque de TR pour la suite de ce manuscrit. Le contact homme-vecteur se produit lorsque les anophèles sont à la recherche d'un repas de sang et que simultanément les hommes ne sont pas protégés par les moustiquaires. La probabilité de ce contact dépend donc en partie du comportement de l'anophèle - ses horaires et sites d'activités de recherche de repas de sang - et de celui de l'humain - son utilisation ou absence d'utilisation de moustiquaire, ses horaires d'utilisation

de moustiquaires, ses habitudes nocturnes. En mesurant les comportements horaires anophéliens et humains au sein d'une même unité spatio-temporelle, il est possible de quantifier la probabilité de l'intéraction entre l'anophèle et l'humain : autrement dit, le risque de transmission résiduelle (Garrett-Jones & Organization, 1964 ; Gerry F. Killeen et al., 2006 ; Gerry F. Killeen, 2014).

L'approche descriptive du risque de transmission résiduelle (*mesurer et caractériser le risque*) consiste donc à quantifier la probabibilité de contact homme-vecteur et caractériser les sites et horaires où s'effectue ce contact, en zone couverte par les MIILDA.

Une méthode permettant l'étude des interactions comportementales entre les populations humaines et vectorielles en zone couverte par les MIILDA a été décrite par Gerry F. Killeen et al. (2006) et améliorée par Geissbühler et al. (2007). Cette méthode requiert la collecte de données fines sur les comportements humains (possession, utilisation, et horaires d'utilisation, de moustiquaires) et vectoriels (abondances horaires des piqûres de vecteurs). Ces données sont ensuite introduites dans un modèle mathématique calculant l'exposition humaine horaire à la piqûre d'anophèle.

Ces données et cette approche permet de dégager de précieuses informations concernant la transmission résiduelle sur un territoire d'intérêt : taux de possession et utilisation globale des moustiquaires par la population, niveau de protection conféré par les moustiquaires, site (intérieur ou extérieur des habitations) et horaire (soir, nuit, matin) où la transmission résiduelle s'effectue, hétérogénéité spatio-temporelle du risque de transmission résiduelle. Sur la base de ces informations et connaissances, il sera possible d'élaborer des plans de gestion et outils efficaces : par exemple, programmer une campagne de distribution de moustiquaires (si les taux de possession sont faibles) ou d'information, éducation, communication à leur utilisation (si les taux d'utilisation sont faibles), ou encore déployer des interventions de LAV complémentaires à la MIILDA qui ciblent la part résiduelle de la transmission (cad. qui visent les vecteurs impliqués dans le risque de transmission résiduelle).

Approche n°2 : Comprendre le risque

Si l'approche descriptive présentée dans la section précédente permet de mesurer la probabilité de contact hôte-vecteur, elle ne permet pas de comprendre les raisons sous-jacentes de l'intensité et de la variabilité spatio-temporelle de cette intéraction. **L'approche explicative du risque de transmission résiduelle proposée ici consiste ainsi à identifier les déterminants de l'intensité et de l'hétérogénéité spatio-temporelle de la probabibilité de contact homme-vecteur, en zone couverte par les MIILDA.**

En géographie, le risque est souvent défini comme «*la probabilité d'occurrence de dommage compte tenu des interactions entre facteurs d'endommagement (aléas) et facteurs de vulnérabilité. On peut ainsi résumer cette définition par une formule : risque = aléa × vulnérabilité*[4]». Le risque de transmission résiduelle du paludisme (selon la définition proposée dans la section précédente) fait intervenir deux protagonistes : l'anophèle et l'homme. Les facteurs d'aléa peuvent être définis comme ceux directement liés au vecteur, et les facteurs de vulnérabilité comme ceux directement dépendants de l'homme. Ainsi, les facteurs d'aléa sont par exemple :

— *Abondance journalière des vecteurs* : la densité environnante de vecteurs augmente à priori la probabilité pour un homme d'entrer en contact avec un vecteur ;
— *Résistances physiologiques des vecteurs* : les vecteurs résistants aux insecticides ont une longévité accrue, augmentant ainsi à priori à la fois la probabilité pour ces vecteurs (qui vivront plus longtemps) de transmettre le parasite, et la densité globale des vecteurs ;
— *Résistances comportementales des vecteurs* : les vecteurs exophages ou piquant précocement ou tardivement échappent à la protection conférée par les MIILDA et augmentent donc à priori la probabilité de contact homme-vecteur.

Les facteurs de vulnérabilité sont par exemple :

— *Possession et utilisation de moustiquaire* : La probabilité de posséder et utiliser une moustiquaire réduit à priori la probabilité de contact homme-vecteur ;
— *Horaires d'utilisation des moustiquaires* : Les horaires d'utilisation des moustiquaires modulent à priori la probabilité de contact homme-vecteur.
— *Qualité de l'habitat* : Des facteurs tels que l'utilisation de grillages au fenêtres

4. http://geoconfluences.ens-lyon.fr/glossaire/risque-s, consulté le 2022-01-12

empêchant les vecteurs d'entrer dans les maisons modulent à priori la probabilité de contact homme-vecteur.

Comprendre le risque de transmission résiduelle consiste à améliorer les connaissances sur les systèmes {environnement - vecteur} et {environnement - hôte} du système vectoriel : identifier les déterminants (environnementaux, génétiques, socio-économiques, etc.) de chacune de ces composantes du risque et la manière dont ils impactent la composante. En d'autre termes, il s'agit d'approfondir les connaissances fondamentales sur les déterminants du risque :

- Caractériser la niche écologique des vecteurs ;
- Caractériser la niche "d'activité" des vecteurs en recherche d'hôte ;
- Comprendre les conditions d'émergence et de développement des résistances aux insecticides au sein d'une population de vecteurs ;
- Comprendre les conditions de possession et d'utilisation des moustiquaires par la population ;
- etc.

Un des enjeux (et difficultés) de cette approche est son caractère holistique : pour chaque composante de risque, il ne s'agit pas uniquement d'étudier si et comment un facteur donné impacte cette composante, mais plutôt de comprendre de quelle manière l'ensemble des potentiels facteurs impacte de manière complexe, en conditions "réelles" (cad. de terrain), cette composante de risque. Cette connaissance holistique des déterminants des différentes composantes du risque peut permettre d'identifier les leviers d'actions les plus pertinents pour diminuer le risque de transmission résiduelle.

Cette analyse peut se réaliser à l'aide de données (sur les composantes de risque à expliquer et leurs déterminants potentiels) et de méthodes de modélisation statistique qui seront détaillées dans le chapitre 2 de ce manuscrit. De manière générale, les enjeux sont ici d'identifier, pour chaque composante du risque : i) ses déterminants (quels facteurs influencent les intensités observées et la variabilité spatio-temporelle ?), ii) l'importance relative de chaque déterminant dans le comportement de la composante de risque (quels sont les déterminants qui influencent le plus la composante ?), iii) l'effet respectif (relation fonctionnelle) de chaque déterminant sur la composante du risque (si la valeur d'un des déterminants change, comment va changer la valeur de la composante

du risque ?), et iv) l'existence, la nature et l'effet de potentielles intéractions entre les déterminants (comment les intéractions entre les déterminants impactent-elles la composante de risque ?).

Approche n°3 : Prédire le risque

Les deux approches précédentes (décrire et comprendre le risque) permettent d'accroitre les connaissances sur les intéractions (systèmes) hôte-vecteur, hôte-environnement, et vecteur-environnement ; ces connaissances permettant à leur tour de définir des caractéristiques pour de nouveaux outils de LAV et d'envisager des interventions adaptées au contexte local. Mais une fois ces interventions définies, comment savoir où et quand les déployer sur l'ensemble du territoire ? Prédire spatio-temporellement le risque de transmission résiduelle permet de cibler et prioriser les lieux et moments pour le déploiement des actions de gestion. Cette troisième approche consiste à estimer les valeurs des composantes du risque de transmission (abondance des vecteurs, taux de vecteurs résistants, etc.), voire du risque en lui-même (nombre de piqures / homme ou probabilité de contact homme-vecteur), en tout point de l'espace et du temps pour lesquels les observations « terrain » de ces composantes ne sont pas disponibles.

L'approche prédictive du risque de transmission résiduelle (*prédire le risque*) consiste donc à prédire ou anticiper la probabibilité de contact homme-vecteur en tout point de l'espace et du temps.

Concrètement, cette approche consiste à générer des cartes (éventuellement saisonnières) pour chaque composante du risque (par exemple, cartes de la distribution de l'abondance des vecteurs), ou un système d'alerte précoce qui permettra d'anticiper à courte ou moyenne échéance, dans l'espace et dans le temps, les composantes du risque. Ces outils permettront de prioriser les zones d'intervention pour d'éventuels outils de LAV complémentaires aux MIILDA, et d'anticiper précocement le besoin (éventuellement ponctuel) en ces interventions. Notons que les granularités (résolutions) spatiales et temporelles de la prédiction (ou les échéances d'anticipation) sont à définir en fonction de plusieurs caractéristiques et contraintes : dynamiques spatio-temporelles observées pour la composante de risque, échelles opérationnelles envisageables pour le déploiement des mesures, disponibilité et granularité des données environnementales, etc.

Les liens entre les différentes composantes du risque et l'environnement permettent leur prédiction spatio-temporelle. L'estimation de la valeur d'une composante de risque en tout point de l'espace et du temps se réalise en deux étapes : i) l'apprentissage des relations qui existent entre la composante de risque étudiée et l'environnement (sur le même principe que l'analyse explicative - à la différence près que, dans l'analyse prédictive, il n'est pas nécessaire d'expliciter ces liens) et ii) l'extrapolation de cet apprentissage en tous points de l'espace et du temps pour lesquels les données environnementales sont disponibles. Cette approche nécessite donc la collecte de données représentatives des composantes du risque à prédire (entomologiques et environnementales notamment) et peut se réaliser là aussi grâce à des méthodes de modélisation statistique.

Nous avons résumé dans le tableau ci-dessous les trois approches proposées (mesurer, comprendre, prédire le risque) : enjeux, questions, données nécessaires, approches d'analyse des données, exemple de connaissances ou outils potentiellement générés, exemple d'enjeux opérationnels.

	Mesurer et caractériser le risque de transmission résiduelle	**Expliquer** le risque de transmission résiduelle	**Prédire** le risque de transmission résiduelle
Enjeux	Quantifier l'intensité du contact homme-vecteur et caractériser les sites et horaires où s'effectue ce contact	Identifier les déterminants de l'intensité et de l'hétérogénéité spatio-temporelle de chaque composante du risque	Prédire les composantes du risque en tout point de l'espace et du temps (et in-fine le risque lui même)
Questions	- Quel niveau de protection à la piqûre d'anophèle les moustiquaires offrent-elles ? - A quel endroit (intérieur ou extérieur des maisons) et moment de la journée (soir, nuit ou matin) les populations sont-elles principalement exposées aux piqûres d'anophèles ? - L'exposition à la piqûre varie-t-elle selon les âges, les saisons, les villages ?	- Les composantes du risque sont-elles hétérogènes dans l'espace et dans le temps ? - Quels sont les déterminants de l'intensité et de la distribution spatio-temporelle de chaque composante du risque ?	- Peut-on prédire l'intensité des composantes du risque en tout point de l'espace et du temps ?
Données nécéssaires (échantillonage spatio-temporel sur la zone d'étude)	- Données entomologiques : Abondances horaires des piqûres d'anophèles - Données de comportement humain : Taux de possession, d'utilisation, et horaires d'utilisation des moustiquaires	- Données entomologiques : Abondances journalières et horaires des piqûres de vecteurs - Données de comportement humain : Taux de possession, d'utilisation, et horaires d'utilisation des moustiquaires - Données environnementales : environnement à proximité (spatiale et temporelle) des données entomologiques et de comportement humain	- Idem "Expliquer le risque"
Méthode d'analyse des données	Modélisation mathématique (modèle d'exposition humaine à la piqûre)	Modélisation statistique explicative & descriptive	Modélisation statistique prédictive
Exemples de connaissances ou outils générés	- Taux de possession et utilisation des moustiquaires par la population - Niveau de protection à la piqûre conféré par les moustiquaires - Sites (intérieur / extérieur) et horaires où la transmission résiduelle s'effectue - Niveau d'hétérogénéité spatio-temporelle du risque de TR	- Niche écologique des vecteurs - Niche "d'activité" des vecteurs en recherche d'hôte - Conditions d'émergence et de développement des résistances des vecteurs aux insecticides - Conditions d'utilisation des moustiquaires par les populations - Niveau d'hétérogénéité spatio-temporelle de chaque composante du risque de TR	- Cartes pour chaque composante du risque (éventuellement saisonnières) - Cartes du risque de transmission résiduelle (nb. piqûres / homme / nuit) - Système d'alerte précoce
Exemples de portées opérationnelles	- Évaluation de la nécessité d'une nouvelle distribution de moustiquaires - Évaluation de la nécessité d'une campagne d'information, éducation, communication à l'utilisation de la moustiquaire - Identification d'outils de LAV adaptés au contexte local (ciblant la part résiduelle de la transmission)	- Aide à la conception de nouveaux outils de LAV - Aide au ciblage des interventions de LAV	Aide au ciblage et à la priorisation du déploiement des interventions

FIGURE 1.17: Enjeux et objectifs des trois approches théoriques pour décrire, comprendre et prédire le risque de transmission résiduelle du paludisme

1.3.2 Enjeu de la thèse et organisation du manuscrit

Cette thèse propose d'étudier, en implémentant en partie les approches décrites précédemment, le risque de transmission résiduelle du paludisme dans deux zones d'étude situées au Burkina Faso et en Côte d'Ivoire. Chaque zone recouvre environ la surface d'un district sanitaire rural ouest-africain (environ 2500 km^2). Ainsi, l'échelle spatiale d'étude est dite "paysagère" : autrement dit, nous travaillons à l'échelle du *village* dans ces zones. L'enjeu d'ensemble est de montrer en quoi ces différentes approches apportent des éléments complémentaires permettant de proposer des stratégies de prévention adaptées aux contextes locaux.

Ce travail fera très largement appel à des méthodes avancées et non triviales issues de la science des données, en particulier la modélisation statistique. Aussi, à ces enjeux scientifiques s'en ajoute un davantage méthodologique, consistant à détailler les différentes manières dont la modélisation statistique peut servir la recherche scientifique ; et plus particulièrement à préciser son intérêt et potentiel dans l'étude des systèmes biologiques complexes tel que le système environnement-vecteur en conditions naturelles.

Les travaux de thèse s'articulent autour de quatre articles. Au total, deux de ces articles ont été rédigés en tant qu'auteur principal, et les deux autres ont été co-rédigés. Parmi les deux articles rédigés en auteur principal, l'un a été accepté et l'autre devrait être soumis prochainement. Les deux articles co-rédigés ont été acceptés. Tous les articles sont en anglais et sont donc préfacés dans ce manuscrit d'une introduction et d'un résumé en français.

Le manuscrit se compose de six chapitres faisant suite à ce premier chapitre introductif.

Le **chapitre 2** (***Contexte méthodologique : Étude des systèmes complexes et modélisation statistique***) a pour objectif de présenter et justifier la forme de raisonnement scientifique et l'approche méthodologique utilisée dans les principaux travaux de la thèse (chapitres 4 et 5). Nous y introduisons les différentes manières d'appréhender l'étude des systèmes biologiques complexes, et le rôle que peut tenir la modélisation statistique dans ce contexte. Nous élaborons sur les questions suivantes :

— Comment aborder l'étude des systèmes biologiques complexes tel que le système environnement-vecteur ? Quelles sont les deux approches existantes pour ce faire, et en quoi sont-elles complémentaires ?
— A quoi sert la modélisation statistique ? En quoi cet ensemble d'outils et de méthodes permet-il d'appréhender les systèmes complexes, et au sens plus large, certains enjeux majeurs de la recherche scientifique (tester, consolider, créer des connaissances scientifiques ; prédire) ?
— Quelles sont les différentes étapes d'un travail de modélisation statistique ?
— En quoi les développements récents en science des données offrent-ils de nouvelles perspectives pour approfondir la compréhension des liens et intéractions dans les systèmes biologiques complexes, tels que le système environnement-vecteur ?

Le **chapitre 3** (***Zones d'étude et préparation des données environnementales télédétectées***) présente le projet dans lequel s'inscrit la thèse, les zones d'étude, et les travaux de production de certaines données environnementales utilisées dans les chapitres 4 et 5 (données paysagères et météorologiques produites à partir d'images satellitaires d'observation de la Terre).

Les chapitres 4 à 6 constituent le coeur de la thèse.

Au **chapitre 4** (***Modélisation des dynamiques spatio-temporelles des abondances des vecteurs***) (article n°1, auteur principal, publié), nous étudions la composante du risque "Abondance journalière des vecteurs" (autrement dit, la niche écologique des vecteurs). Nous expliquons (approche n°2) et évaluons la prédictibilité (approche n°3) des dynamiques spatio-temporelles des abondances journalières des principales espèces d'anophèles présentes dans nos deux zones d'études, en les modélisant avec des données environnementales issues de produits satellitaires d'observation de la Terre. Nous apportons des éléments de réponse aux questions suivantes :

— Les densités agressives des vecteurs sont-elles hétérogènes dans l'espace et dans le temps dans nos zones d'étude ?
— Quels sont les déterminants des densités agressives pour chaque espèce majeure de vecteurs, et comment les impactent-ils ?
— Est-on en mesure de prédire les densités agressives dans l'espace et dans le temps ?
— Les déterminants considérés dans l'étude suffisent-ils à expliquer et prédire

l'abondance des vecteurs et leur hétérogénéité spatio-temporelle dans nos zones d'étude ? Quels facteurs additionnels, non considérés dans l'étude, peuvent expliquer l'hétérogénéité des abondances ?

Au **chapitre 5** (***Modélisation des dynamiques spatio-temporelles des résistances physiologiques et comportementales des vecteurs***) (article n°2, auteur principal, à soumettre), nous étudions les composantes du risque "Résistances physiologiques des vecteurs" et "Résistances comportementales des vecteurs" (autrement dit, les conditions d'émergence et de développement de résistances des vecteurs aux insecticides). Nous expliquons (approche n°2) et évaluons la prédictibilité (approche n°3) des dynamiques spatio-temporelles des résistances physiologiques et des comportements des anophèles dans nos deux zones d'étude. En utilisant un nombre important de variables environnementales potentiellement explicatives des résistances, nous modélisons la probabilité individuelle de résistance physiologique des vecteurs ainsi que certains traits de leur comportement de piqûre (exophagie, agressivité précoce, agressivité tardive), afin d'apporter des éléments de réponse aux questions suivantes :

— Les résistances physiologiques et comportementales des vecteurs sont-elles hétérogènes dans l'espace et dans le temps dans nos zones d'étude ?
— Quels sont les déterminants des résistances physiologiques et comportementales pour chaque espèce majeure de vecteurs, et comment les impactent-ils ?
— Est-on en mesure de prédire les résistances physiologiques et comportementales dans l'espace et dans le temps ?
— Les déterminants considérés dans l'étude suffisent-ils à expliquer et prédire les résistances des vecteurs et leur hétérogénéité spatio-temporelle dans nos zones d'étude ? Quels facteurs additionnels, non considérés dans l'étude, peuvent expliquer l'hétérogénéité des résistances ?

Le **chapitre 6** (***Etudes complémentaires : contributions à des travaux de modélisation liés à la transmission du paludisme***) (articles n°3 et 4, co-auteur, publiés) rassemble les deux études complémentaires auxquelles nous avons contribué dans le cadre de la thèse. Ces deux études concernent la zone d'étude située au Burkina Faso. Le premier article complémentaire (*Modélisation de l'exposition humaine à la piqûre d'anophèles*) vise à mesurer et caractériser la transmission résiduelle (approche n°1) dans la zone d'étude. Le deuxième article complémentaire (*Modélisation des*

dynamiques spatio-temporelles des cas de paludisme) présente une étude visant à expliquer et prédire la distribution spatio-temporelle des cas de paludisme dans la zone d'étude, en utilisant des produits satellitaires d'observation de la Terre - comme pour les chapitres 4 et 5. Cette étude ne traite donc pas directement d'entomologie médicale, mais complémentairement aux études précédentes, permet d'illustrer la diversité des utilisations possibles des données satellitaires et modèles statistiques pour la gestion du paludisme sur le terrain.

Enfin, au **chapitre 7** (***Discussion générale***), nous discutons l'ensemble des résultats. Nous proposons certaines stratégies pour la gestion du risque de transmission du paludisme sur nos deux zones d'étude. En particulier, nous faisons des propositions pour l'amélioration (i) des méthodes actuelles de lutte anti-vectorielle, (ii) de l'utilisation de la science et ingénierie des (géo-)données en général, et de la modélisation statistique en particulier, pour la recherche et le contrôle du paludisme, et (iii) des outils de surveillance et prévention du risque de transmission du paludisme à échelle locale en milieu rural ouest-africain.

Chapitre 2

Contexte méthodologique : Étude des systèmes complexes et modélisation statistique

L'enjeu des principales études de cette thèse (chapitres 4 et 5) est d'approfondir les connaissances sur certains traits bio-écologiques, comportementaux ou physiologiques des vecteurs du paludisme. A cette fin, nous utiliserons une forme particulière d'étude des systèmes complexes, nommée "holistico-inductive". L'approche holistico-inductive est différente, conceptuellement et pratiquement, de l'approche hypothético-déductive généralement mieux maitrisée des chercheurs.

Qu'est ce que l'approche holistico-inductive, et en quoi diffère-t-elle de l'approche hypothético-déductive ? Quel rôle peut jouer la modélisation statistique dans ces différentes approches ? En quoi la modélisation statistique peut-elle servir les différents grands objectifs de la recherche scientifique : tester, améliorer, ou construire des théories scientifiques ? Ce chapitre apporte des éléments de réponse à ces questions. Son enjeu principal, dans le cadre strict de la thèse, est de préciser le raisonnement scientifique et les choix méthodologiques effectués dans les travaux à suivre. Au sens plus large, l'objectif est de montrer en quoi la modélisation statistique peut servir les différents grands objectifs de la recherche scientifique : tester, améliorer, ou - de part ses récents développements - construire des théories scientifiques.

2.1 Considérations épistémologiques sur l'étude des systèmes complexes

2.1.1 Les deux formes d'inférence logique (inductif et déductif)

L'objectif principal de la recherche scientifique est de faire avancer les connaissances en construisant et testant des hypothèses scientifiques. Les nouvelles hypothèses scientifiques sont construites à partir d'hypothèses existantes : ce processus s'appelle l'inférence logique. On reconnait deux formes principales d'inférence logique (Johnson-Laird, 2013 ; Kell & Oliver, 2004) : la déduction et l'induction.

Le **raisonnement déductif** (ou *hypothesis-driven* (Kell & Oliver, 2004)) confirme l'hypothèse par le cas. Dans cette forme de raisonnement, l'hypothèse (ou la théorie) est le point de départ. Les observations (ou données) sont utilisées pour la tester dans des situations particulières et ainsi la vérifier ou l'infirmer.

Le **raisonnement inductif** (ou *data-driven* (Kell & Oliver, 2004)), à l'inverse, part du cas pour générer l'hypothèse. Dans cette forme de raisonnement, l'observation (ou la donnée) est le point de départ. Ces données sont utilisées pour formuler des hypothèses, des théories, plus générales. Autrement dit, dans ce cas, les données servent à générer l'hypothèse, qui est donc l'objectif et le point final du raisonnement.

Boite info n°1 : **Exemples de raisonnements déductif et inductif** (extrait de Kell & Oliver (2004))

Raisonnement déductif : Toutes les baleines sont bleues ; Georges est une baleine, donc Georges est bleu.

Raisonnement inductif : Georges est une baleine et est bleu ; Anne est une baleine et est bleue ; Percy est une baleine et est bleue ; etc. ; nous pouvons donc induire l'idée (hypothèse) que toutes les baleines sont bleues.

Ces deux formes de raisonnement, à priori complémentaires, s'alimentent et forment ainsi le cycle de la génération de connaissances (figure 2.1). En particulier, elles tiennent

chacun leur rôle dans l'étude des systèmes complexes.

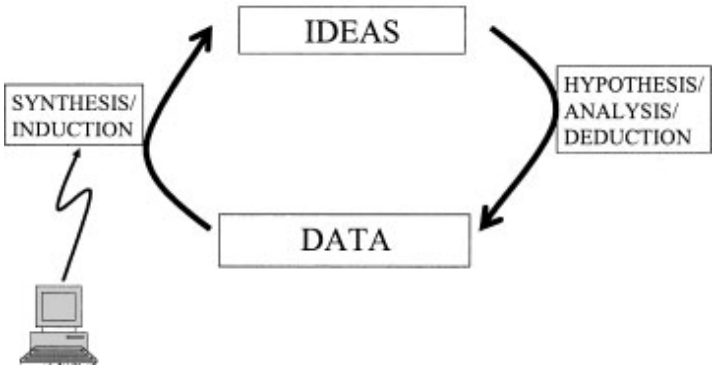


FIGURE 2.1: Le cycle de la connaissance (Kell & Oliver, 2004)

2.1.2 Etudier les systèmes complexes : approches holistique et réductionniste

Un système complexe est un sytème composé de nombreux éléments qui peuvent intéragir les uns avec les autres. S'il n'existe à priori pas de définition formelle largement acceptée du système complexe, ceux-ci sont définis, selon les cas et selon les auteurs, par l'existence d'effets et d'interactions non linéaires entre éléments du système, ou encore par l'existence de niveaux d'organisation différents (Bar-Yam, 2002). Ainsi, par exemple, nous pouvons qualifier le système {densités agressives des anophèles - environnement} de complexe (figure 1.9) : au sein de ce système, les effets de l'environnement sur le facteur étudié (la densités agressives des anophèles) peuvent être non-linéaires, de nombreuses interactions existent à priori (Stresman, 2010) ; l'ensemble provoquant un effet (les densités agressives) à priori difficilement prédictible. De manière générale, les sytèmes biologiques sont complexes (Bar-Yam, 2002).

Deux stratégies au moins peuvent être envisagées pour étudier et tenter d'approfondir la compréhension d'un sytème complexe : l'approche réductionniste et l'approche holistique (Amboise & Audet, 1996 ; Bar-Yam, 2002 ; Kell & Oliver, 2004). Nous résumons ces approches dans les prochains paragraphes, en nous basant sur ces trois références bibliographiques.

Dans l'approche réductionniste, le système complexe est considéré comme un ensemble de sous-sytèmes, moins complexes et ainsi plus simples à approcher, contenant un nombre réduit d'éléments, de relations et interactions. La compréhension de chacun de ces différents sous-systèmes permet ensuite de reconstruire le système complexe initial physiquement ou intellectuellement. Dans cette approche, le système complexe est tout d'abord décomposé en sous-systèmes pertinents en se basant sur les connaissances à priori du système complexe ; puis pour chaque sous-système, un nombre restreint de variables le caractérisant est sélectionné - là aussi en se basant sur les connaissances *à priori*. L'enjeu de l'étude est alors de vérifier si et comment ces variables parcimonieusement sélectionnées impactent le comportement du sous-sytème. L'approche réductionniste repose donc sur un certain niveau de connaissance à priori du système complexe : à la fois pour créer les sous-sytèmes et pour sélectionner des variables pour chacun d'entre eux. En ce sens, l'approche réductionniste de l'étude des systèmes complexes est associée au raisonnement hypothético-déductif : les données servent à valider des hypothèses préalablement construites.

L'approche holistique, à l'opposé, considère que la complexité théorique des relations et interactions dans le système complexe implique qu'il faille étudier, chercher à décrire et comprendre, le système en entier, dans sa complexité. La première étape dans cette approche consiste à recueillir un maximum d'observations (données), ayant un impact plus ou moins lointainement soupçonné, sur le sytème étudié, quitte à en écarter certaines par la suite si elles ne s'avèrent pas utiles. Dans un second temps, les relations et interactions entre ces observations sont décrites - en général à l'aide d'outils informatiques et statistiques au regard du volume d'observations - puis interprétées à la lumière des connaissances préalables existantes. Cette démarche, reposant donc principalement sur les données, peut permettre d'améliorer, raffiner, ou faire émerger de nouvelles hypothèses scientifiques sur le fonctionnement du système complexe. En ce sens, l'approche holistique de l'étude des systèmes complexes est associée au raisonnement inductif : les données sont la source de l'hypothèse. On parle ainsi d'approche holistico-inductive.

Boite info n°2 : **Exemple de questionnements de recherche et approches associées, en lien avec la thèse**

Répondre à la question *"Sur mon territoire d'étude, quel est l'impact de l'interaction entre les précipitations et les températures sur les densités agressives des moustiques ?"* relève d'une approche réductionniste hypothético-déductive. Parmi tous les déterminants potentiels des densités agressives, deux en particulier sont sélectionnés (températures et précipitations), que l'on sait impacter l'abondance. L'objectif de l'étude sera de quantifier précisement l'impact des précipitations, des températures, et de leur interaction sur les densités agressives des moustiques.

Répondre à la question *"Sur mon territoire d'étude, quels sont les déterminants des densités agressives des moustiques ?"* relève d'une approche holistico-inductive. L'objectif de l'étude sera de collecter un maximum d'observations sur l'ensemble des potentiels déterminants des densités agressives, puis de décrire les liens existant entre ces observations, afin d'élaborer des hypothèses sur les déterminants des densités agressives (facteurs déterminants, effets, interactions, etc.).

L'approche hypothético-déductive réductionniste nécessite donc un cadre établi, rigide : l'hypothèse de recherche, précise, est formellement énoncée puis vérifiée ou testée à l'aide d'expérimentations contrôlées et de méthodes statistiques rigoureuses. L'approche hostistico-inductive, de son côté, est de prime abord moins rigide : les hypothèses de recherche ne sont pas formellement énoncées (ou n'existent même pas nécéssairement), le nombre de variables est plus important, les méthodes d'analyse moins rigides, le tout afin de laisser place à la découverte potentielle d'informations intéressantes. L'enjeu de l'approche holistico-inductive n'est pas de produire des résultats généralisables mais de mieux comprendre un phénomène d'intérêt, en espérant que la connaissance du phénomène acquise au cours de la recherche permettra de raffiner et d'améliorer la théorie existante. L'approche holistico-inductive requiert donc beaucoup de données et des hypothèses de départ très ouvertes. Le différentiel de rigidité au moment de l'établissement d'hypothèses préalables est regagné dans les

étapes suivantes de l'analyse, qui nécéssitent rigueur, intégration du jugement subjectif et des connaissances humaines, afin d'interpréter les signaux (associations, etc.) révélés par l'analyse. En ce sens, ses défenseurs la considèrent comme une approche plus ouverte mais tout aussi rigoureuse que l'approche hypothético-déductive.

Ces deux approches de l'étude du système complexe, qui impliquent donc des démarches intrinsèquement différentes, sont pourtant complémentaires dans la compréhension des systèmes complexes. L'approche holistico-inductive, de part son caractère volontairement ouvert, est susceptible de soulever de nouvelles questions, ou hypothèses, potentiellement peu ou pas intuitionnées. La génération d'hypothèses est donc la fin du parcours. Ces nouvelles hypothèses pourront ensuite être testées expérimentalement, et validées, par raisonnement hypothético-déductif dans une approche réductionniste. Autrement dit, les approches hypothético-déductives (réductionnistes) et holistico-inductives sont itératives dans l'avancement des connaissances en général, et dans la compréhension des systèmes complexes en particulier (figure 2.2).

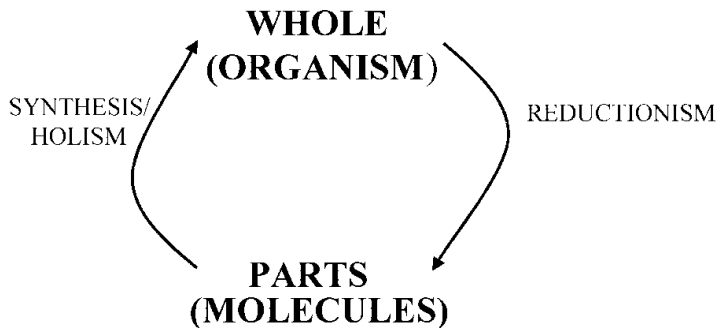


FIGURE 2.2: Holisme et réductionnisme en tant que stratégies complémentaires et itératives pour comprendre les systèmes complexes (Kell & Oliver, 2004)

Bien que ces approches soient donc à priori complémentaires, plusieurs auteurs constatent que les approches inductive en général, et holistico-inductive en particulier, sont aujourd'hui moins utilisées dans la recherche scientifique que les approches hypothético-déductives et réductionnistes (Amboise & Audet, 1996 ; Bar-Yam, 2002 ; Kell & Oliver, 2004 ; Shmueli & Koppius, 2010 ; Yanai & Lercher, 2019a). L'approche hypothético-déductive est considérée dans certains milieux comme la seule méthode valable et fiable pour faire avancer les connaissances. Ainsi, par exemple, les rejets

de projets ou idées scientifiques avec pour justification qu'ils "n'ont pas d'hypothèse testée", ou qu'ils consistent en des "fishing expeditions", sont fréquents : à l'extrême : "s'il n'y a pas d'hypothèse, ce n'est pas de la science" (Kell & Oliver, 2004 ; Yanai & Lercher, 2019a). Cette préférence de la déduction sur l'induction est probablement liée à une forme de sécurité psychologique qu'offre l'approche déductive (Kell & Oliver, 2004) (si l'axiome et l'observation sont correctes, l'inférence logique doit être correcte) mais pas l'approche inductive ; ainsi qu'à son cadre en apparence plus rigoureux, formel (Amboise & Audet, 1996). Par ailleurs, nous hypothétisons ici que la défection pour l'approche holistico-inductive pourrait venir - en sus de son caractère inductif - d'un déficit de maitrise de certains outils nécéssaires à la conduite de cette approche (comme nous allons le voir dans la section 2.2 à suivre) : l'analyse de données en général et les statistiques, mathématiques, informatique en particulier.

Quoi qu'il en soit, ces paragraphes ont montré que l'analyse des données est au coeur de la génération et validation d'hypothèses en général et de l'étude des systèmes complexes en particulier, quelle que soit l'approche et la forme d'inférence logique utilisée. La modélisation statistique est un puissant outil d'analyse de données, capable de servir, historiquement, l'approche hypothético-déductive mais aussi, de part ses développements récents, l'approche holistico-inductive - en faisant donc un outil essentiel du chercheur.

2.2 Les enjeux scientifiques de la modélisation statistique

2.2.1 Formalisation mathématique du modèle statistique

L'approche déterministe de la science, défendue entre autre par Karl Popper (Arnaud, 1986), veut que la structure du monde est telle que tout *évènement qui se produit* est déterminé par les *événements passés* conformément aux *lois de la nature*. Graphiquement et mathématiquement, cette approche peut être présentée par la figure 2.3 et l'équation *(1)* ci-dessous :

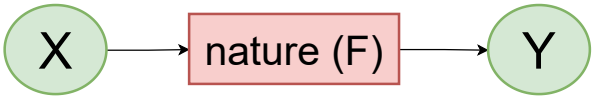


FIGURE 2.3: Illustration de l'approche déterministe de la science (adapté de (Breiman, 2001b))

$$(1)\ \mathrm{Y} = F(X)$$

où :

— Y est l'évènement qui se produit,
— X est l'ensemble des évènements passés causant Y,
— F est la loi de la nature (aussi appelé modèle causal théorique) reliant X à Y.

La modélisation statistique est un outil permettant d'approximer et de formaliser mathématiquement cette réalité. Dans un modèle statistique, une ou plusieurs variables dite(s) 'indépendante(s)', notées x et approximant X, sont associées à une autre variable dite 'dépendante', notée y et approximant Y, via une fonction (le modèle statistique) notée f. Mathématiquement, cela se traduit par l'équation suivante *(2)* :

$$(2)\ \mathrm{y} = f(x, \epsilon)$$

où :

— y est la variable dépendante, pendant de Y dans *(1)*,
— x est une ou plusieurs variables indépendante(s), pendant de X dans *(1)*,
— f est le modèle statistique associant y et x, pendant de F dans *(1)*,
— ϵ est un terme d'erreur regroupant la part non expliquée de y, puisque f ne fait qu'approximer F

Approximer et formaliser mathématiquement la réalité - cad. faire usage de la modélisation statistique - peut servir trois enjeux liés à *(1)* : i) expliquer (tester) une loi de la nature F, ii) décrire une loi de la nature F, iii) prédire un évènement Y (Fayyad, Piatetsky-Shapiro, & Smyth, 1996 ; Karpatne et al., 2017 ; Shmueli, 2010 ; Shmueli

& Koppius, 2010). En nous basant sur ces quatre références bibliographiques, nous résumons chacun de ces enjeux dans la prochaine section.

2.2.2 Les trois enjeux de la modélisation statistique (expliquer, prédire, décrire)

Note : au préalable, définissons dès maintenant les termes "modélisation statistique" et "fouille de données" tels qu'ils sont utilisés dans ce manuscrit. En effet, les définitions semblent varier selon les sources, au sein même de la famille des statisticiens. Nous définirons donc "modélisation statistique" comme l'ensemble du processus d'extraction de connaissances à partir de données et de modèle(s) statistique(s) (ledit processus est décrit dans la section 2.2.3). Cette définition équivaut à celle de "statistical modeling" dans Shmueli (2010). Nous définirons "fouille de données" comme l'ensemble du processus d'extraction de connaissances à partir de données. Cette définition équivaut à celle de "knowledge discovery in databases" dans Fayyad et al. (1996). À la différence de la modélisation statistique, la fouille de données n'implique pas qu'il soit spécifiquement fait usage d'un modèle statistique pour générer des connaissances à partir des données.

Modéliser pour expliquer (modélisation explicative)

Un modèle statistique peut être utilisé pour **tester ou vérifier un modèle causal théorique (cad. des hypothèses) pré-existant**. On parle dans ce cas de 'modélisation explicative' (Shmueli, 2010), 'd'objectif de vérification' (Fayyad et al., 1996), ou de 'theory-based model[ing]' (Karpatne et al., 2017).

Dans cette approche, le modèle causal (cad. la loi de la nature) existe au niveau conceptuel, préalablement à l'utilisation du modèle statistique. Le modèle statistique est utilisé pour tester ou vérifier ce modèle théorique. Pour cela, des variables x et y, représentant respectivement X et Y, sont contruites à partir de données judicieusement collectées. Un modèle statistique est ensuite utilisé pour associer les variables x et y. L'interprétation du modèle statistique permet finalement de générer des informations statistiques sur le modèle causal théorique.

Cette forme de modélisation sert donc l'approche hypothético-déductive : les

observations sont intégrées dans un modèle statistique dont l'objectif est de délivrer des informations statistiques sur les associations entre observations, permettant alors de vérifier et éventuellement préciser les hypothèses préalables. Dans l'étude des systèmes complexes, l'approche réductionniste fait donc ainsi usage de la modélisation explicative.

Ainsi, dans cette approche :

— **le modèle statistique f est l'objet d'intéret** de la modélisation : son analyse permet de **tester/vérifier** les hypothèses pré-existantes de F ;
— les données **x et y sont des outils** permettant d'estimer f.

Modéliser pour décrire (modélisation descriptive)

Un modèle statistique peut être utilisé pour **décrire un modèle causal**. Dans ce cas, l'enjeu principal du modèle étant de décrire des associations entre des évènements, on parle de 'modélisation descriptive' (Shmueli, 2010), d'objectif de 'description' (Fayyad et al., 1996) ou 'theory-guided data science model[ing]' (Karpatne et al., 2017).

Dans cette approche, le modèle causal n'existe pas nécéssairement, ou n'est pas formellement établi, au niveau conceptuel. Le modèle statistique f est utilisé pour décrire un modèle théorique F contenant des associations éventuellement peu ou pas hypothétisées. L'interprétation du modèle statistique permet finalement, éventuellement de mieux comprendre F.

Cette forme de modélisation sert donc l'approche holistico-inductive : les observations sont intégrées dans un modèle statistique dont l'objectif est de trouver des descriptions résumées et pertinentes expliquant les données. Le jugement subjectif et les connaissances préalables sur F permettent ensuite d'interpréter ces relations, afin d'améliorer les connaissances sur le modèle causal théorique.

Ainsi, dans cette approche :

— **le modèle statistique f est l'objet d'intéret** : son analyse permet éventuellement de **mieux comprendre** F ;
— les données **x et y sont des outils** permettant d'estimer f.

Modéliser pour prédire (modélisation prédictive)

Enfin, un modèle statistique peut être utilisé pour **prédire de nouvelles ou futures valeurs d'un évènement**. On parle dans ce cas de modélisation statistique prédictive (Shmueli, 2010) ou d'objectif de prédiction (Fayyad et al., 1996).

La modélisation prédictive s'effectue en deux étapes. Dans un premier temps, un modèle statistique, dit prédictif, est construit à partir de données x et y. En général, le pouvoir prédictif du modèle est évalué, à savoir sa capacité à générer des prédictions précises sur de nouvelles observations [1]. Dans un second temps, ce modèle est utilisé pour prédire y lorsque de nouvelles valeurs de x, pour lesquelles les valeurs de y sont inconnues, sont disponibles.

La modélisation prédictive a donc principalement une portée opérationnelle : les prédictions sont généralement utilisées à des fins pratiques. Cependant, elle peut également jouer un rôle dans la construction ou amélioration de théories scientifiques. Utiliser un modèle statistique à des fins de prédiction, et évaluer son pouvoir prédictif, peut par exemple permettre d'évaluer la pertinence d'une théorie (cad. un modèle causal F) (si le modèle prédit mal, la théorie est-elle réellement valable ?), d'évaluer la prédictibilité d'un phénomène empirique, de comparer plusieurs théories concurrentes (celle qui prédit le mieux a des chances d'être celle qu'il faut retenir), d'améliorer les théories existantes, etc. (Shmueli & Koppius, 2010). Utilisée dans ce contexte, la modélisation prédictive rejoint donc en partie les enjeux de la modélisation descriptive.

Ainsi, dans cette approche :

— les données **x et y sont les objets d'intérêt**, en particulier y ;
— **le modèle statistique f est un outil** permettant de générer des prédictions de y à partir de x.

Les rôles et fonctions du modèle statistique f et des données x et y diffèrent donc selon l'enjeu de la modélisation. Ces distinctions entre objets d'intérêt et outils sont importantes

1. ce point est détaillé dans la section 2.2.3 à suivre

car elles guident les choix durant tout le processus de modélisation statistique, dès la phase de définition de l'objectif de l'étude. Dans la prochaine section, nous détaillons les grandes étapes du processus de modélisation statistique, en précisant pour chacune d'elles les différences fondamentales entre les formes de modélisation.

2.2.3 Les étapes du processus de modélisation statistique

Quelle que soit l'approche, le travail de modélisation statistique est un processus complexe, itératif, constitué d'étapes bien définies. Chacune de ces étapes implique des choix, qui diffèrent suivant l'approche utilisée, et ces choix peuvent avoir un impact sur les étapes suivantes et sur l'information et la connaissance extraites en bout de processus (Fayyad et al., 1996 ; Shmueli, 2010). La figure 2.4 expose ces étapes. Dans cette section, nous énumérons et décrivons brièvement les principales d'entre elles, en exposant en quoi les choix diffèrent en fonction de l'approche empruntée. Sauf mention spécifique, l'ensemble de cette section est basée sur les travaux de Shmueli (2010), Shmueli & Koppius (2010) et Fayyad et al. (1996).

FIGURE 2.4: Etapes du processus de modélisation statistique (Shmueli, 2010)

Définition de l'objectif : **Cette étape consiste à définir l'objectif à priori du travail de modélisation (expliquer, décrire, ou prédire).** En effet, les différences conceptuelles entre les trois approches de modélisation impliquent, comme nous allons le voir ensuite, des choix différents dans les étapes du processus de modélisation à suivre ; même si les données utilisées peuvent être identiques.

Conceptualisation de l'étude et collecte des données : **Cette étape consiste à définir les caractéristiques de la collecte de données.** En fonction de l'approche, ces caractéristiques peuvent différer. Ainsi par exemple, en modélisation explicative, la puissance statistique est un critère majeur. Un certain nombre d'observation

est donc nécéssaire, mais au delà d'un certain volume, la puissance statistique n'augmente plus. En modélisation prédictive, en général, davantage d'observations sont nécéssaires. D'autres enjeux sont à considérer : plans d'échantillonage des données, conditions d'expérimentation (laboratoire ou terrain), instruments de collecte des données, etc.

Choix et construction des variables : **Cette étape consiste à construire des variables statistiques à partir des données.** Les critères pour construire les variables diffèrent largement selon l'approche. En modélisation explicative, l'objectif est la causalité : les variables x et y doivent donc représenter au plus proche les évènements X et Y que l'on cherche à vérifier. En modélisation prédictive, l'objectif est l'association : on ne cherche pas à comprendre le rôle de chaque variable en terme de relation de cause à effet. Les critères d'importance pour construire les variables sont donc principalement la qualité de l'association entre celles-ci et la disponibilité des variables prédictives (indépendantes), x, au moment des futures utilisations attendues du modèle (cad. quand il servira à prédire y à partir de nouveaux x).

Boite info n°3 : **Un exemple classique en géo-épidémiologie, le NDVI : variable explicative ou prédictive ?**

Une variable largement utilisée dans les travaux de modélisation statistique en géo-épidémiologie est le Normalized Difference Vegetation Index (NDVI) (Parselia et al., 2019), calculé à partir de valeurs de réflectance des sols mesurées par les capteurs embarqués dans les satellites ou les drones. Cette variable, adimensionnelle, permet de déterminer la santé de la végétation en mesurant la teneur en chlorophylle des plantes. Elle est donc à la fois représentative de la quantité de végétation et de la présence d'eau, deux paramètres environnementaux ayant à priori un impact sur les traits de vie des moustiques vecteurs (voir figure 1.9). Cette variable a donc un fort potentiel d'association avec la densité des moustiques, et à ce titre, peut être utilisée en modélisation prédictive de leur abondance. En revanche, en modélisation explicative, cette variable est peu pertinente : i) on ne peut discriminer l'effet de la présence d'eau et de la végétation et ii) quel que soit le sens de l'association, il est possible de fournir une explication (une association positive peut être expliquée par la présence d'eau, une association négative peut être expliquée par la densité de végétation impliquant

une réduction de la capacité de dispersion des moustiques (Le Goff, Carneval, & Robert, 1997)). On lui préfèrera ainsi, en modélisation explicative, des variables plus proches du modèle causal théorique : quantités de précipitations, suface occupée par la végétation, etc.

En sus de la pertinence des variables, une autre distinction de taille est la gestion de la multicollinéarité (collinéarité entre variables). En modélisation explicative, la multicollinéarité est problématique car elle peut conduire à des effets (par ex. coefficients de régression) ou intervalles de confiance biaisés, interférant avec l'inférence. En modélisation prédictive, l'interprétation du modèle n'étant pas nécéssaire, la multicollinéarité n'est en général pas problématique.

Choix du modèle statistique : **Cette étape consiste à sélectionner un modèle statistique, à savoir, une fonction mathématique ou un algorithme qui associe y à x.** Il existe de très nombreux modèles statistiques, et le choix dépend là encore de l'approche. En modélisation explicative et descriptive, où l'objectif est d'analyser f, le critère principal de Sélection est l'interprétabilité du modèle, c'est-à-dire la capacité à extraire les associations que le modèle a capturées. En modélisation explicative, le modèle doit être en mesure de délivrer des informations statistiques précises et chiffrées (intensité de l'effet, significativité de l'association, etc.). En modélisation descriptive, le modèle doit être en mesure de capturer au mieux les relations et interactions, potentiellement complexes (non-linéaires), entre variables. En modélisation prédictive, où f n'est que l'outil, l'enjeu est de sélectionner un modèle qui génèrera les meilleures prédictions possibles de y ; l'interprétabilité du modèle n'est donc pas un critère de choix. L'interprétabilité et interprétation des modèles statistiques sont intrinsèquement liés à leur nature et la manière dont chacun fonctionne pour associer les variables. Nous nous attardons sur les différentes philosophies d'associations entre variables et sur l'interprétation des modèles dans les sections 2.2.4 et 2.2.5.

Validation du modèle : **Cette étape consiste à vérifier certaines hypothèses permettant d'utiliser ou d'interpéter le modèle correctement.** En modélisation explicative, la validation du modèle consiste à vérifier que la forme de f représente adéquatement la relation à priori entre x et y (voir section 2.2.4). En modélisation prédictive, la validation consiste à évaluer la propension du modèle à généraliser

l'apprentissage, c'est à dire, à ne pas avoir sur-appris [2].

Evaluation du modèle : **Cette étape consiste à évaluer la puissance explicative ou prédictive du modèle.** En modélisation explicative, la puissance explicative du modèle est la force de la relation indiquée par f. Des mesures telles que le R^2, représentant la proportion de la variance d'une variable dépendante expliquée par les variables indépendantes, peuvent être rapportées. En modélisation prédictive, la puissance prédictive du modèle est la capacité du modèle à prédire sur de nouvelles données, non utilisées pour entrainer le modèle. Là aussi, des indicateurs mesurant l'écart entre les valeurs observées et prédites peuvent être utilisées (aire sous la courbe (AUC), erreur moyenne carrée (MSE), etc.). Le choix de l'indicateur de puissance prédictive dépend de la nature et de la distribution statistique des données. Une différence majeure entre les évaluations des modèles prédictifs et explicatifs est la nature des données sur lesquelles l'évaluation est effectuée : alors qu'en modélisation explicative l'évaluation est faite sur les données ayant servi à générer le modèle, en modélisation prédictive, l'évaluation doit être faite sur des données qui n'ont pas servi à générer le modèle, puisque l'objectif du modèle sera de prédire sur de nouvelles données. Par ailleurs, si les observations sont non-indépendantes (par exemple dans le cas des enquêtes transversales ou des données spatiales ou temporelles), le jeu de données de validation d'un modèle prédictif doit être judicieusement sélectionné afin d'être indépendant du jeu de données d'entrainement - ceci afin d'éviter des performances prédictives surévaluées dues au suraprentissage (Meyer, Reudenbach, Hengl, Katurji, & Nauss, 2018).

Sélection du modèle : **Cette étape consiste à sélectionner un modèle à interpréter ou utiliser parmi les différents modèles potentiellement valides.** En modélisation explicative, un des critères les plus cruciaux est l'importance théorique des variables dans le modèle causal F. Il est ainsi nécéssaire de retenir dans le modèle les variables qui ont un effet théorique important, même s'il s'avère que dans le modèle ces variables sortent non-significativement associées à la variable réponse (Shmueli, 2010) (par exemple, il est important de retenir la variable 'type de mesure de lutte anti-vectorielle implémentée' dans un modèle qui cherche à expliquer l'abondance des moustiques, même si cette variable n'est finalement pas statistiquement significativement associée à

2. le surraprentissage est la propension du modèle à s'ajuster trop proche des données qui ont été utilisées pour l'entraîner - provoquant ainsi son incapacité à prédire sur de nouvelles données x

l'abondance). En modélisation prédictive, le premier critère est la performance prédictive du modèle. Le choix se portera donc sur le modèle qui génère la meilleure prédiction, quitte à supprimer des variables théoriquement importantes au niveau conceptuel. De nombreuses méthodes de Sélection automatique de variables en modélisation prédictive existent à cet effet.

Interprétation du modèle et utilisation des résultats : **Cette étape consiste à finalement extraire de l'information ou de la connaissance pertinente à partir du modèle statistique.** En modélisation explicative, les informations d'intérêt sont les métriques statistiques renseignant sur l'effet des variables explicatives sur la variable à expliquer (coefficients directeurs par exemple), la significativité de l'effet (p-value par exemple), et la performances explicative du modèle (R^2 par exemple). En modélisation descriptive, il s'agit de rapporter les relations, potentiellement complexes, capturées par le modèle sous une forme compréhensible par l'humain (tableaux, graphiques, etc.). En modélisation prédictive, l'interprétation du modèle est secondaire. Les informations d'intérêt sont principalement celles issues des étapes de validation et évaluation du modèle. Les concepts et outils d'interprétation des modèles, notamment en modélisation descriptive, sont détaillées dans la section 2.2.5.

2.2.4 Les deux grandes familles de modèles statistiques (modèles paramétriques et non-paramétriques)

Le choix du modèle statistique est un des éléments primordial dans le processus de modélisation statistique. En effet, chaque modèle est défini de telle manière qu'intrinsèquement, il associe différemment les variables indépendantes (x) et dépendantes (y). En bout de chaine, cela peut avoir un impact considérable sur la nature de la connaissance qui est finalement extraite.

On peut distinguer deux grandes catégories de modèles statistiques, définies par deux philosophies d'association de y à x (cad. de construction de f) conceptuellement différentes (Breiman, 2001b) : les modèles paramétriques (que Breiman (2001b) appelle *data model(s)*) et les modèles non-paramétriques (que Breiman (2001b) appelle *algorithmic model(s)*).

Les modèles paramétriques simplifient la fonction f à une forme connue (par exemple : gaussienne, négative binomiale, etc.). Cette forme doit donc être spécifiée (par le modélisateur) dans le modèle. Le rôle du modèle statistique est ensuite d'estimer les coefficients de la fonction à partir des données. Des exemples de modèles paramétriques largement utilisés sont la régression linéaire et la régression logistique.

Les modèles non-paramétriques, de leur côté, ne font pas d'hypothèse concernant la forme de la fonction f. Ces modèles cherchent à s'ajuster au mieux aux données en construisant la fonction f à partir des données. Des exemples de modèles non-paramétriques largement utilisés sont les arbres de décision (et modèles dérivés, tels que les forêts aléatoires) et *Support Vector Machines*. Ces modèles sont parfois appelés modèles ou algorithmes d'apprentissage automatique (*machine learning*) (Bzdok, Altman, & Krzywinski, 2018) ou de fouille de données (*data mining*) (Shmueli, 2010).

Chacune de ces méthodes possède son lot d'avantages et d'inconvénients. Les modèles paramétriques sont transparents (les coefficients de f sont directement interprétables) et requièrent moins de données que les modèles non-paramétriques, puisque la forme de f est à priori determinée. Cependant, ils exigent que la forme de la fonction soit connue à l'avance, et sont ensuite contraints de se conformer à cette forme. Les modèles non-paramétriques, parce qu'ils doivent chercher de manière autonome la forme de f, requièrent plus de données, de puissance et de temps de calcul, sont davantage susceptibles de surapprendre et sont moins transparents que les modèles paramétriques. Cependant, ils ne requièrent pas d'hypothèse à priori sur la forme fonctionnelle et sont capables de s'adapter à une gamme bien plus large de formes, en faisant ainsi de bons candidats si les relations sont à priori inconnues ou soupçonnées complexes (relations entre variables non-linéaires ou interactions potentielles), ce qui est souvent le cas dans les processus naturels et en particulier biologiques (Breiman, 2001b).

Aussi, par définition, les modèles paramétriques sont à priori adaptés à la modélisation explicative (modèle causal théorique connu, besoin de résultats statistiques) et les modèles non-paramétriques à la modélisation prédictive (flexibilité, performance) (Bzdok et al., 2018 ; Shmueli, 2010). La modélisation descriptive, quand à elle, requiert à la fois un modèle flexible (puisque, par définition de cette approche de

modélisation, les relations ne sont pas nécéssairement connues) et interprétable (puisque l'objectif final est l'extraction de connaissances à partir des relations que le modèle a capturées), deux propriétés à priori difficilement conciliables au regard de ce qui est écrit ci-dessus. Consciente du potentiel des modèles non-paramétriques pour la génération de connaissances (au delà de leur potentiel prédictif indiscutable), la communauté des scientifiques des données a developpé un ensemble d'outils visant à interpréter les associations que ces modèles capturent. La prochaine section présente le concept et quelques outils d'interprétation des modèles statistiques non-paramétriques.

2.2.5 L'interprétation des modèles statistiques non-paramétriques

L'interprétation des modèles statistiques est un élément fondamental du processus de modélisation, en particulier en modélisation explicative et descriptive. L'interprétation des modèles peut être définie comme l'extraction de connaissances pertinentes à partir d'un modèle statistique concernant des relations soit contenues dans les données soit apprises par le modèle (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). Au niveau de l'interprétabilité, on distingue deux grandes catégories de modèles (Murdoch et al., 2019) : les modèles permettant de comprendre naturellement et directement les relations qu'ils ont capturées, et les modèles nécéssitant une phase supplémentaire d'interprétation à postériori de leur génération, avec des outils spécifiques, pour extraire de l'information sur les relations qu'ils ont capturées.

La première catégorie de modèles (modèles directement interprétables) est constituée dans l'ensemble des modèles paramétriques. Ces modèles sont transparents : les coefficients de la fonction (ainsi que d'autres métriques telles que les intervalles de confiance) sont les outils d'interprétation du modèle, à partir desquelles les connaissances sont extraites.

La deuxième catégorie de modèles (modèles nécéssitant une phase supplémentaire d'interprétation) est constituée dans l'ensemble des modèles non-paramétriques. Ces modèles ont l'avantage d'être en capacité de capturer des relations et interactions complexes mais l'inconvénient de ne pas délivrer directement les relations qu'ils ont capturées, à tel point qu'il sont souvent considérés comme des boites noires (Bzdok et

al., 2018). Aussi, un ensemble d'outils dont l'objectif est d'extraire des informations sur les relations que le modèle a capturées a été développé depuis une vingtaine d'années (Molnar, 2019).

Parmi les outils d'interprétation à postériori des modèles, on distingue deux grandes familles : les méthodes indépendantes du modèle interprété (dites "model-agnostic") et les méthodes spécifiques à un modèle donné (dites "model-specific") (Molnar, 2019). Les méthodes agnostiques peuvent elle-mêmes être subdivisées en deux classes : les méthodes globales et méthodes locales (Murdoch et al., 2019). Les méthodes globales décrivent la manière dont les variables indépendantes affectent la variable dépendante en moyenne, tandis que les méthodes locales visent à décrire l'effet des variables indépendante sur une observation individuelle (ou un groupe d'observations). Enonçons et expliquons le fonctionnement de deux des outils d'interprétation "model-agnostic" les plus anciens et utilisés (et utilisés en particulier dans les travaux de cette thèse) : l'importance des variables par permutation et les graphiques de dépendance partielle.

L'importance des variables par permutation (*permutation feature importance*) est une méthode introduite en 2001 par Breiman (Breiman, 2001a). Cette méthode renseigne sur "l'importance" de chaque variable indépendante dans le modèle en mesurant l'augmentation de l'erreur de prédiction du modèle après avoir permuté les valeurs de la variable. Une variable est "importante" si la permutation aléatoire de ses valeurs augmente l'erreur du modèle, car dans ce cas, le modèle s'est appuyé sur la variable pour la prédiction. A l'inverse, une variable est "sans importance" si la permutation aléatoire de ses valeurs ne modifie pas l'erreur de prédiction du modèle, car dans ce cas, le modèle n'a pas considéré la variable pour la prédiction. Un exemple de graphique d'importance des variables est fourni à la figure 2.5.

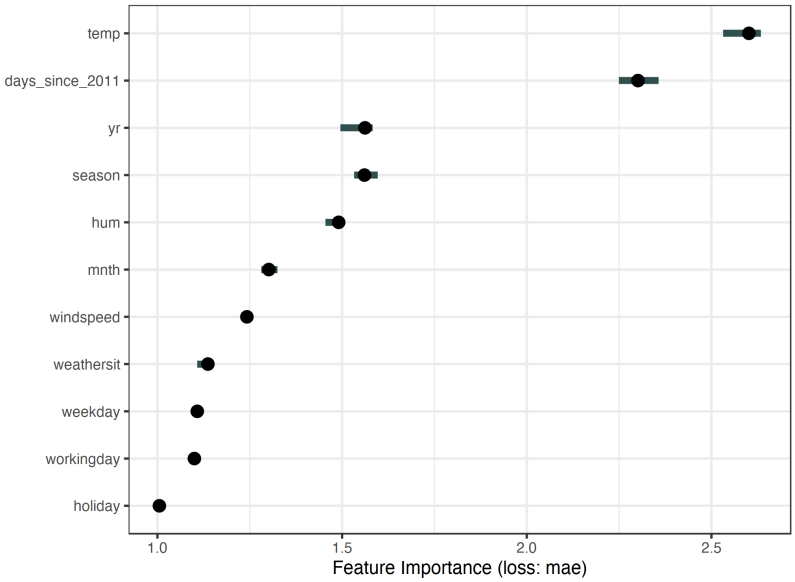


FIGURE 2.5: Exemple de graphique d'importance des variables extrait de (Molnar, 2019). Le modèle statistique sous-jacent prédit un nombre de vélos loués en fonction d'un ensemble de paramètres météorologiques et socio-économiques. Le graphique montre que la variable la plus importante est la température. Par extension, on peut donc émettre l'hypothèse que la température est le facteur principal impactant la location de vélos.

Les graphiques de dépendance partielle (*partial dependence plots* (PDP)), de leur côté, ont été introduits en 2001 par Friedman (Friedman, 2001). Cette méthode renseigne sur la relation fonctionnelle entre une variable indépendante et la variable dépendante. La relation fonctionnelle est calculée en fixant tour à tour chacune des valeurs de la variable indépendante d'intérêt pour toutes les observations, puis en calculant la valeur de la variable dépendante ainsi prédite par le modèle. Un graphique de dépendance partielle peut montrer si la relation entre la variable dépendante et indépendante est linéaire, monotone ou plus complexe. On peut utiliser la même méthode avec deux variables indépendantes : dans ce cas, le graphique renseigne sur l'effet de l'interaction entre ces deux variables indépendantes sur la variable dépendante. Un exemple de graphique de dépendance partielle est fourni à la figure 2.6.

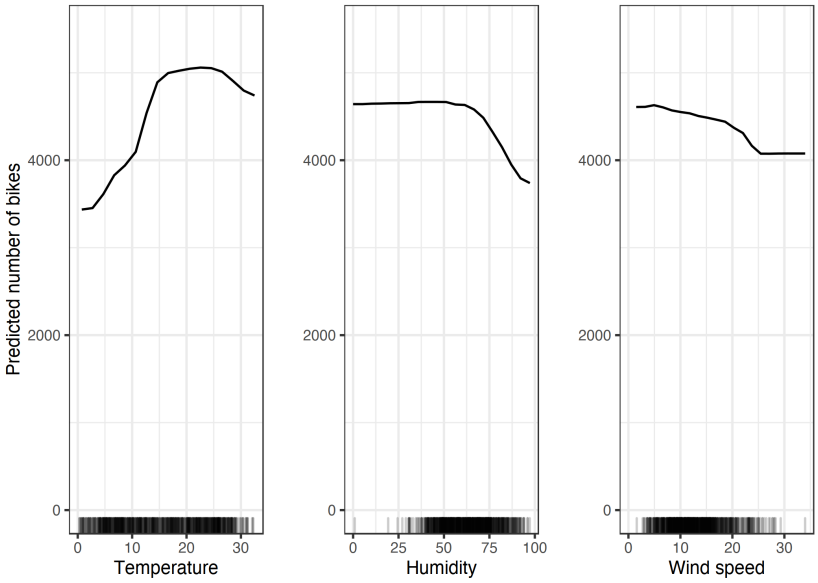


FIGURE 2.6: Exemple de graphiques de dépendance partielle des variables extrait de (Molnar, 2019). Le modèle statistique sous-jacent prédit un nombre de vélos loués en fonction d'un ensemble de paramètres météorologiques et socio-économiques. Le graphique montre que la relation capturée par le modèle entre le nombre de vélos prédits et respectivement la température (à gauche), l'humidité (au milieu) et la vitesse du vent (à droite) est non-linéaire

Au delà de ces deux exemples, il existe une myriade d'outils d'interprétation à posteriori des modèles statistiques (Molnar, 2019) ; et le secteur est en plein développement avec l'intérêt croissant pour l'interprétation des modèles non-paramétriques (Murdoch et al., 2019). Ces outils permettent d'interpréter un modèle ayant potentiellement capturé des relations complexes et non-hypothétisées, et donc d'étudier le comportement du sytème complexe sous toutes ses formes : contribution absolue et relative de ses différentes composantes, relations fonctionnelles, importance et effets des interactions, etc. Ces problématiques sont, typiquement, celles en jeu dans l'étude des systèmes biologiques (Yu et al., 2021).

Notons enfin que, au même titre que les modèles statistiques, chaque outil d'interprétation possède un lot d'hypothèses d'utilisation et de limites, et qu'il est ainsi important de bien en comprendre le fonctionnement intrinsèque afin de l'utiliser à bon escient et d'en extraire de l'information et de la connaissance pertinente (Molnar, 2019 ; Zhao & Hastie, 2021).

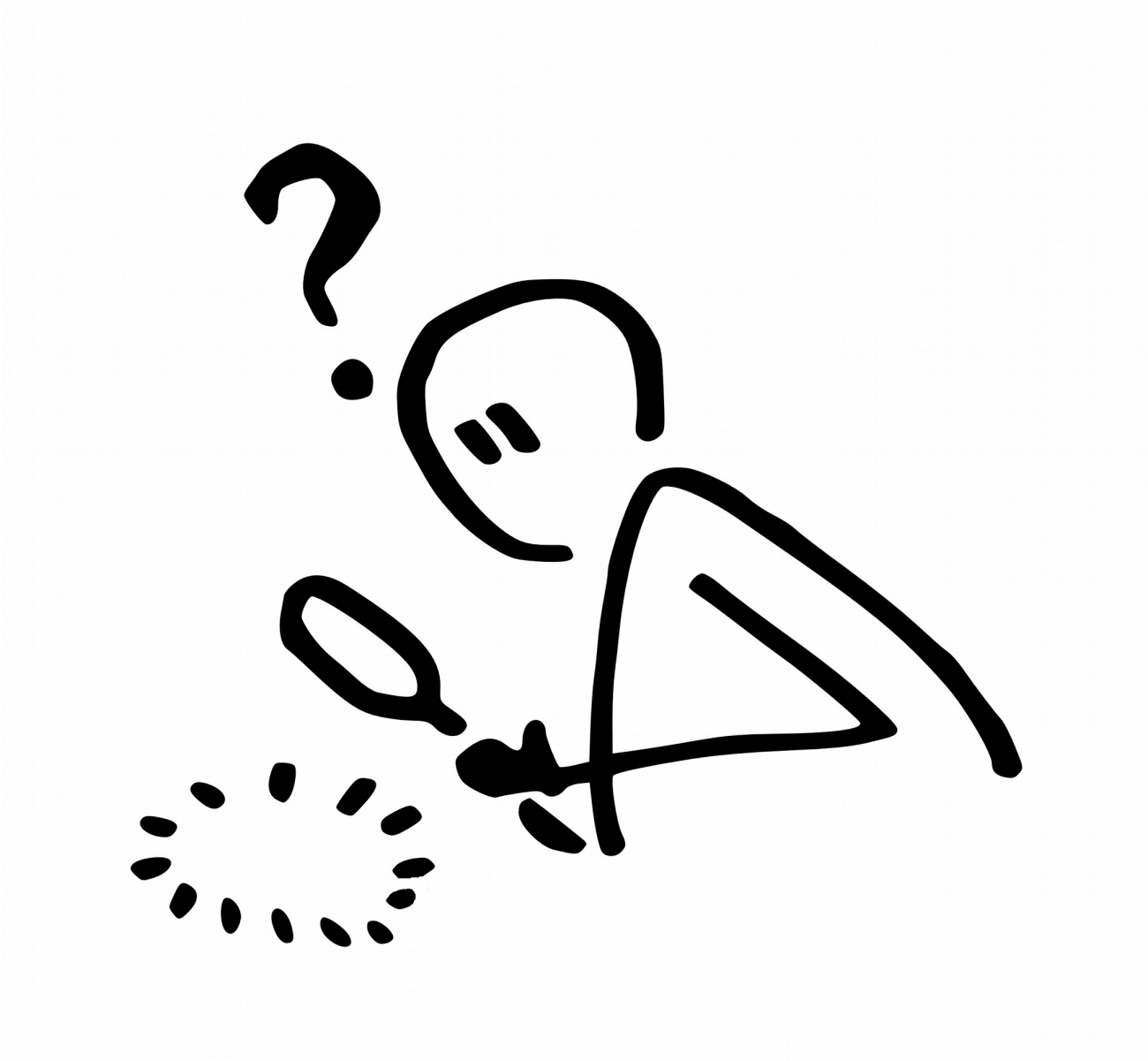
2.3 Notes conclusives

Un modèle statistique est un outil permettant d'associer des données, à savoir des informations mesurables du monde qui nous entoure. Associer des données peut servir différents enjeux scientifiques : tester une théorie scientifique (expliquer), mieux comprendre un phénomène d'intérêt (décrire), prédire de nouvelles valeurs d'un évènement (prédire). Selon l'enjeu, l'ensemble du processus de modélisation statistique diffère : pour un même système étudié, le "meilleur" modèle explicatif sera différent du "meilleur" modèle prédictif, lui-même différent du "meilleur" modèle descriptif (Shmueli, 2010).

Les premiers modèles statistiques, historiquement, servent principalement l'approche hypothético-déductive de la génération de connaissance scientifique - à savoir, tester une théorie scientifique. Aujourd'hui, l'avènement des données volumineuses, des méthodes statistiques non-paramétriques (qui sont en capacité de "trouver" de manière autonome des associations complexes entre variables, parfois difficilement hypothétisables) et des outils permettant d'interpréter les associations capturées par ces modèles, offrent des perspectives nouvelles pour mieux prédire, mais aussi comprendre, les systèmes complexes tels que le système vectoriel. Afin d'exploiter la puissance de la modélisation statistique - c'est à dire, exploiter pleinement son potentiel, sans pour autant le surévaluer (Molnar et al., 2022) - il est essentiel de maitriser les fondements mêmes de la génération de connaissance scientifique, et de connaitre les étapes du processus de modélisation statistique ainsi que les outils de science des données qui existent.

Partie 2

Travaux de thèse



Chapitre 3

Zones d'étude et préparation des données environnementales télédétectées

Les travaux à suivre s'inscrivent dans le cadre d'un projet plus large, nommé REACT ; et ont conduit à la production de nombreuses données. Dans ce chapitre, nous présentons dans un premier temps les objectifs du projet REACT et les deux zones d'études du projet et de la thèse. Dans un second temps, nous décrivons les travaux de génération des données utilisées pour les études présentées dans les chapitres suivants. Enfin, nous apportons quelques précisions sur les logiciels informatiques utilisés pour manipuler les données (recueil, génération, préparation, modélisation) dans le cadre de la thèse, et présentons rapidement les codes informatiques développés pour les besoins de ces travaux.

3.1 Présentation du projet REACT et des zones d'études de la thèse

Les travaux de cette thèse s'inscrivent dans le cadre du projet *REACT : Gestion de la résistance aux insecticides au Burkina Faso et en Côte d'Ivoire : recherche sur les stratégies de lutte anti-vectorielle*, mené en partenariat entre l'Institut de Recherche pour le Développement (IRD, France), l'Institut de Recherche en Sciences de la Santé (IRSS, Burkina Faso) et l'Institut Pierre Richet (IPR, Côte d'Ivoire). Ce projet était

financé par L'Initiative 5%. L'objectif principal de ce projet, dont la phase de terrain s'est déroulée entre les années 2016 et 2018, était d'évaluer l'impact de l'utilisation de mesures de lutte anti-vectorielles complémentaires à la MIILDA sur la transmission et l'épidémiologie du paludisme à travers un essai randomisé contrôlé (ERC). A cette fin, deux zones d'études ont été séléctionnées dans deux pays d'Afrique de l'ouest : le Burkina Faso (BF) et la Côte d'Ivoire (CI).

Ces deux pays sont situés en zone endémiques du paludisme à *P. falciparum*. Les courbes épidémiologiques des deux pays (morbidité et mortalité) suivent les tendances observées à l'échelle du continent (cf. figure 1.2). En 2019, avant la pandémie de covid-19, le nombre de cas de paludisme était estimé à 5,9 millions au Burkina Faso et autant en Côte d'Ivoire (WHO, 2021). Comme introduit dans le chapitre 1, les principales espèces d'anophèles dans ces pays sont *An. arabiensis*, *An. gambiae s.s.*, *An. coluzzii* et *An. funestus* ; et dans les deux pays les résistances physiologiques des anophèles aux insecticides y sont reportées depuis plusieurs décennies (voir figure 1.13).

Chaque zone d'étude du projet REACT couvre environ la surface d'un district sanitaire (˜2500 km^2). Il s'agit de zones principalement rurales. Pour le projet REACT, un total de 55 villages (27 au Burkina Faso, 28 en Côte d'Ivoire) a été séléctionné au sein de ces zones pour mener l'ERC selon les critères suivants : accessibilité pendant la saison des pluies, 200 à 500 habitants par village, et distance entre les villages supérieure à 2 km. La figure 3.1 présente la localisation géographique des zones et des villages séléctionnés ; ainsi que le chronogramme de collectes de données effectuées dans le cadre du projet REACT.

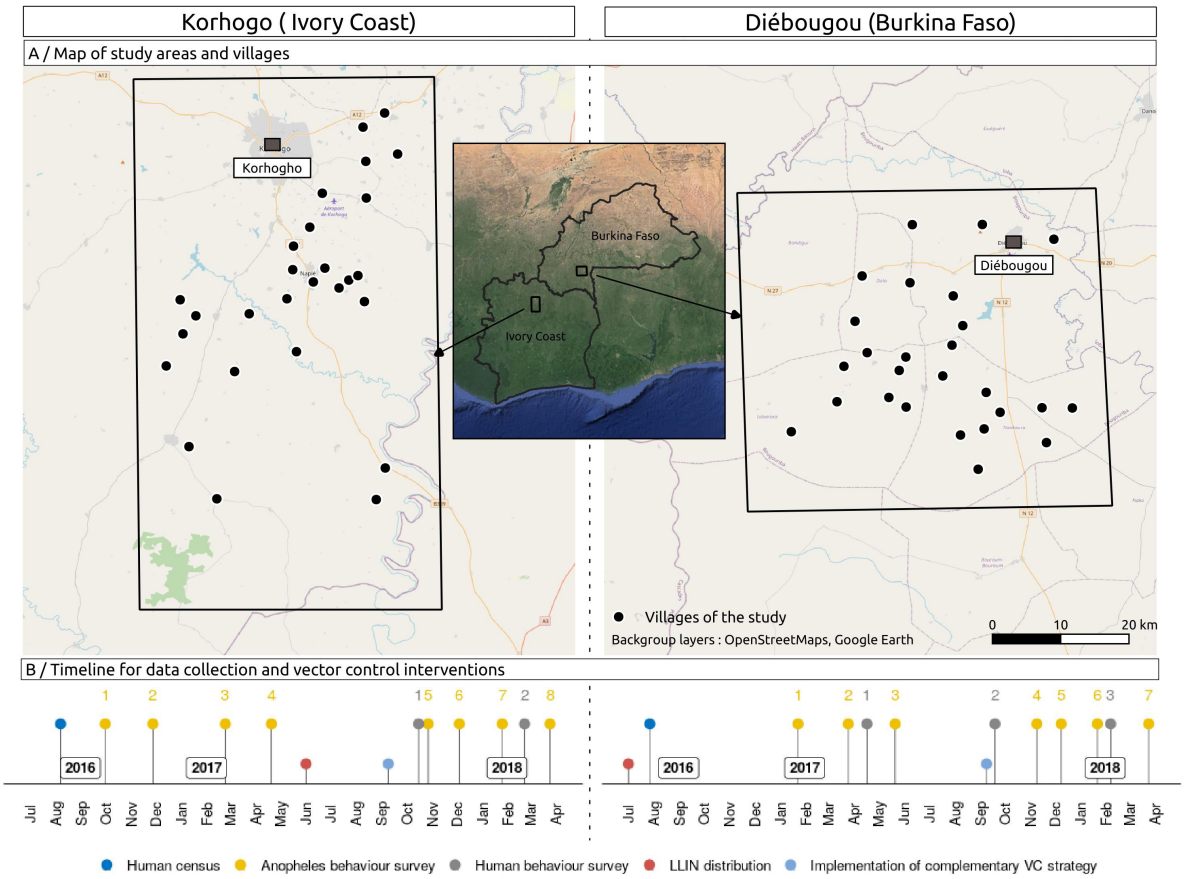


FIGURE 3.1: Projet REACT : zones d'étude, villages et dates de collectes des données

La zone d'étude burkinabé du projet REACT couvre la région de Diébougou, au sud-ouest du pays, en région bioclimatique soudanienne (CILSS, 2016). Le climat y est caractérisé par une saison sèche d'octobre à avril (incluant une période 'froide' de décembre à février et une période 'chaude' de mars à avril) et une saison pluvieuse de mai à septembre. Les amplitudes thermiques moyennes journalières sont 18-36 °C, 25-39 °C et 23-33 °C respectivement en saison sèche froide, sèche chaude et pluvieuse. Les précipitations annuelles moyennes sont de 1200 mm. Comme nous allons le voir à la section 3.2.2, la végétation naturelle est dominée par la savane arborée parsemée de forêts ripicoles. La principale activité économique est l'agriculture (culture des céréales) suivie par l'exploitation artisanale de l'or et la production de charbon et de bois (INSD, 2015, 2017). Le principal outil de lutte anti-vectorielle dans la région de Diébougou est la MIILDA, distribuée universellement par le gouvernement tous les 3-4 ans depuis 2010 (PNLP, 2014a). La dernière distribution avant le projet REACT datait de juillet

2016 (PNLP, 2014a), soit 6 mois avant la première mission de collecte de données entomologiques (voir ci-dessous).

La zone d'étude ivoirienne du projet REACT couvre la région de Korhogo, au nord du pays, elle aussi en région bioclimatique soudanienne (CILSS, 2016). La saisonalité de la climatologie y est relativement similaire à celle de Diébougou (voir section 3.2.1). Les précipitations annuelles varient de 1200 à 1400 mm, tandis que la température moyenne annuelle varie de 21 à 35 °C. La végétation naturelle est principalement un mélange de savane et de forêt ouverte (voir section 3.2.2). La région possède une forte densité de barrages hydrauliques qui permettent de pratiquer l'agriculture tout au long de l'année. Comme pour la région de Diébougou, la principale activité économique est l'agriculture (riz, maïs, coton). De même, Le principal outil de lutte anti-vectorielle est là aussi la MIILDA, distribuée universellement par le gouvernement, comme au Burkina Faso, tous les 3-4 ans depuis 2010 (PNLP, 2014b). La dernière distribution avant le projet REACT datait de 2014.

L'essai randomisé s'est déroulé en 3 phases. La phase pré-intervention a duré environ un an et a principalement consisté à i) établir un recensement exhaustif de la population et de la localisation géographique des ménages dans les villages, ii) recueillir des données entomologiques, épidémiologiques et de comportement humain dans ces villages, iii) distribuer des MIILDA dans les villages d'étude en Côte d'Ivoire (au Burkina Faso, une distribution universelle de moustiquaires a eu lieu en juillet 2016). Environ une année après le début du projet, la phase d'intervention a consisté à implémenter les mesures de LAV complémentaires à la MIILDA (détaillées ci-après) dans certains villages, tirés au sort dans le cadre de l'essai randomisé contrôlé. Enfin, en phase post-intervention (environ 1 an), plusieurs sessions de collecte de données ont été menées selon les mêmes protocoles qu'en phase de pré-intervention. Ainsi, en comparant les données entomologiques et épidémiologiques de pré-intervention et de post-intervention, il est possible de mesurer l'impact des mesures de lutte anti-vectorielles complémentaires à la MIILDA sur la transmission (taux d'inoculation entomologique) et l'épidémiologie (prévalence et incidence) du paludisme.

Boite info n°4 : **Mesures complémentaires de LAV déployées dans le cadre du projet REACT**

Les mesures complémentaires de lutte anti-vectorielle déployées dans le cadre du projet REACT étaient les suivantes :

- **Information, Education, Communication (IEC)** (testée dans les zones BF et CI). A travers des activités de sensibilisation des populations, l'objectif de cette intervention etait d'optimiser la mise en place des MILDA, l'adhésion des populations aux campagnes de lutte et l'utilisation correcte et régulière des MIILDA.
- **Pulvérisations intra-domiciliaires** (PID) de Pirimiphos-méthyle (Actellic) appliquées sur les murs des habitations (testées dans les zones BF et CI). En complément des MIILDA qui visent les vecteurs endophages uniquement, l'objectif des PID etait de tuer les vecteurs endophiles qui auraient résisté à l'insecticide utilisé dans les moustiquaires.
- **Lutte anti-larvaire** (Djènontin et al., 2014) (testée dans la zone CI uniquement). Cette intervention visait à diminuer la population générale de vecteurs, en tuant les moustiques à leur état larvaire par l'utilisation d'insecticides d'origine bactérienne.
- **Administration d'ivermectine** aux hôtes (testée dans la zone BF uniquement). L'ivermectine est une molécule administrée aux hôtes pour lutter contre les endoparasites. Elle diminue la longévité d'un moustique ayant pris un repas sanguin sur un hôte traité (Alout et al., 2014 ; Ouedraogo et al., 2015). Dans le projet REACT, l'ivermectine a été administrée aux populations animales péri-domestiques dans le but de cibler les populations de vecteurs présentant des comportements zoophages ou opportunistes.

Les travaux de thèse à suivre utilisent en grande partie des jeux de données recueillies sur le terrain dans le cadre du projet REACT, en particulier :

- les données **entomologiques**,
- les données de **recensement des villages** (population, localisation des habitations),
- un jeu de données **environnementales et climatiques au cours des collectes ento-**

mologiques,
— un jeu de données de **comportement humain** relatives à l'utilisation des MIILDA et aux habitudes horaires de sommeil.

Les données entomologiques constituent la source des variables à expliquer / prédire dans les travaux de modélisation, tandis que les autres données ont été utilisées pour caractériser l'environnement à proximité spatiale et temporelle des points de capture des vecteurs (variables explicatives / prédictives). Les protocoles de recueil de ces données sont détaillés dans les études qui les utilisent, ainsi qu'à l'annexe A de ce manuscrit.

Les conditions météorologiques (températures, précipitations) et paysagères (utilisation, occupation du sol) peuvent impacter l'abondance, le comportement, ou les résistances des vecteurs (voir les introductions des articles des chapitres 4 et 5 pour davantage de détails), objets d'étude de travaux de la thèse. Aussi, nous avons complété les données recueillies sur le terrain avec trois autres jeux de données environnementales, générées pour les deux zones d'étude :

— données **météorologiques** au cours des semaines précédant les collectes entomologiques,
— données d'**occupation et utilisation des sols**,
— données sur le **réseau hydrographique théorique**.

Ces données environnementales ont été générées à partir de produits satellitaires d'observation de la Terre. En effet, les satellites sont en mesure de capturer de nombreux paramètres environnementaux en surface ou dans l'atmosphère terrestre. Les capteurs embarqués sur ces satellites mesurent le rayonnement électromagnétique réfléchi ou émis par la surface terrestre, les océans ou l'atmosphère. Ces données brutes peuvent ensuite être traitées pour en extraire des informations environnementales telles que les précipitations, les températures au sol, l'altitude ou encore l'occupation du sol. Ces données sont particulièrement intéressantes et précieuses pour caractériser l'environnement dans des zones où les observatoires ou stations météorologiques au sol sont rares, telles que les zones rurales ouest-africaines. Aussi, les images satellitaires sont très largement utilisées en géo-épidémiologie, pour expliquer ou prédire des indicateurs entomologiques ou épidémiologiques (Ebhuoma & Gebreslasie, 2016 ; Parselia et al., 2019). Dans la suite de ce chapitre, nous détaillons les traitements qu'il a été nécessaire de réaliser pour produire ou exploiter ces données en vue des travaux de modélisation à suivre dans les prochains chapitres.

3.2 Production des données environnementales télédétéctées

3.2.1 Données de météorologie

Nous avons extrait les températures et les précipitations dans nos zones d'études sur les périodes précédant les collectes entomologiques à partir de produits satellitaires d'observation de la Terre. Pour les précipitations, nous avons utilisé les produits de la mission *Global Precipitation Measurement* (GPM) (voir point info ci-dessous). Pour les températures, nous avons utilisé les données recueillies par l'instrument *Moderate Resolution Imaging Spectroradiometer* (MODIS) embarqué à bord des satellites Terra et Aqua. En particulier, nous avons utilisé les collections GPM et MODIS suivantes :

— *MOD11A1.006* (Wan, Hook, & Hulley, 2015a) : Températures de surface terrestre diurnes et nocturnes extraites de l'instrument MODIS embarqué sur le satellite Terra (résolution spatiale : 1 km, résolution temporelle : 1 jour)
— *MYD11A1.006* (Wan, Hook, & Hulley, 2015b) : Températures de surface terrestre diurnes et nocturnes extraites de l'instrument MODIS embarqué sur le satellite Aqua (résolution spatiale : 1 km, résolution temporelle : 1 jour)
— *GPM_3IMERGDF.06* (NASA, 2019) : Précipitations extraites de GPM (résolution spatiale : 0.1 ° (˜ 10 km), résolution temporelle : 1 jour)

Boite info n°5 : **GPM et MODIS : des données météorologiques à l'échelle mondiale et à fine résolution spatio-temporelle**

Initiée par la NASA et la JAXA (agences spatiales respectivement états-uniennes et japonaises), la mission GPM est un projet international en cours depuis l'année 2014, comprenant une constellation de satellites appartenant à plusieurs agences spatiales nationales. À sa résolution spatio-temporelle la plus fine, elle fournit des estimations des précipitations à une résolution de 10 km / 30 minutes pour l'ensemble du globe en temps quasi réel (4 heures de latence entre l'acquisition du satellite et la mise à disposition des données). Les estimations sont ensuite consolidées via divers algorithmes de correction, pour créer des produits consolidés destinés à la recherche environ 3 mois après l'acquisition. Les

données GPM sont générées à différentes résolutions temporelles (30 minutes, 1 jour, 1 mois). Tous les produits sont gratuits et libres d'accès pour l'utilisateur.

L'instrument MODIS est embarqué à bord de Terra et Aqua de la NASA, deux satellites d'observation de la Terre lancés respectivement en 1999 et 2002. Les différents spectromètres de MODIS prennent une image complète de la Terre tous les 1 à 2 jours. Les satellites Terra et Aqua fonctionnant en phase et l'instrument MODIS capturant des données strictement identiques, les produits MODIS issus des deux satellites peuvent être combinés pour obtenir des produits à résolution temporelle très fine (jusqu'à 0.5 jour). Les observations brutes de MODIS sont traitées automatiquement par différents algorithmes de la NASA pour générer des produits dits de "haut niveau", directement utilisables par les différentes communautés scientifiques (océanographie, biologie, sciences de l'atmosphère, etc.). Les produits de haut niveau de MODIS comprennent, par exemple, la réflectance de la surface, la température de surface de la terre et de la mer, la couverture neigeuse, la concentration de chlorophylle-a dans l'océan, des indices de végétation, etc. Les résolutions spatiales et temporelles varient en fonction du produit. Toutes les données MODIS sont ouvertes et gratuites, et mises à disposition des utilisateurs finaux à différentes échéances temporelles après l'acquisition (quelques heures à une année, en fonction du produit).

Nous avons choisi d'extraire ces données de précipitations et températures sur une période de six semaines (soit 42 jours) précédant chaque collecte entomologique. Cette période permet en effet de couvrir largement la durée de vie d'un anophèle sur le terrain (incluant les phases aquatiques et larvaires) (Holstein, 1952). La quantité de données à extraire (3 collections de produits satellitaires, 2 zones d'études, plusieurs centaines de dates d'intérêt) nous a emmené à nous poser la question de la méthode à utiliser pour ce faire. Les données MODIS et GPM sont originellement stockées sur les serveurs de la NASA. Afin de s'adapter aux différents besoins, habitudes et compétences techniques des utilisateurs finaux des produits satellitaires, l'agence met à disposition de multiples outils pour les télécharger et les propose dans de nombreux formats numériques. Parmi les différents outils disponibles pour accéder aux données, un a particulièrement retenu notre attention : le protocole OPeNDAP.

Boite info n°6 : **Le protocole OPeNDAP**

OPeNDAP est l'acronyme de "Open-source Project for a Network Data Access Protocol", un projet (et le nom du serveur) visant à faciliter l'accès à des données structurées (telles que les produits satellitaires) sur le Web. L'une des principales forces d'OPeNDAP est qu'il permet de filter les produits satellitaires dès la phase de téléchargement - spatialement, temporellement et dimensionnellement. Ainsi, seule la partie réellement utile des données pour l'utilisateur est téléchargée (et le volume de données téléchargées est donc limité au strict nécéssaire) ; ce qui contraste avec la plupart des interfaces 'clic-bouton' d'accès aux données - où de grands volumes de données sont généralement importés, quand bien même l'utilisateur n'en nécéssiterait qu'une petite partie. Notons aussi que le projet OPeNDAP est développé collaborativement par plusieurs institutions et entreprises, que le code source et ouvert et que le logiciel est gratuit.

Pour télécharger un produit satellitaire disponible sur un serveur OPeNDAP, il s'agit d'envoyer au serveur une URL dans laquelle les filtres (spatiaux, temporels, dimensionnels) sont spécifiés. Par exemple l'URL suivante :
https://opendap.cr.usgs.gov/opendap/hyrax/MOD11A1.006/h17v08.ncml.nc4?MODIS_Grid_Daily_1km_LST_eos_cf_projection,LST_Day_1km[6093 :6122][55 :140][512 :560],LST_Night_1km[6093 :6122][55 :140][512 :560],time[6093 :6122],YDim[55 :140],XDim[512 :560]
permet de télécharger les bandes LST_Day_1km et LST_Night_1km (filtre dimensionnel) du produit MOD11A1.006 entre le 1er et le 30 janvier 2017 (filtre temporel) sur la zone délimitée par les coordonnées géographiques suivantes (en WGS84) : xmin : -5.82 ymin : 8.84 xmax : -5.41 ymax : 9.55 (coordonnées de la zone CI du projet REACT).

La pluspart des données d'observation de la Terre produites par la NASA sont disponibles en accès OPeNDAP, mais utiliser ce protocole pour télécharger des données reste compliqué pour l'utilisateur néophyte (en particulier, la constitution de l'URL n'est pas triviale, comme le montre l'exemple précédent). Afin de faciliter l'extraction de ces données depuis les serveurs OPeNDAP, nous avons développé une librairie

dans le langage de programmation R (R Core Team, 2018) que nous avons nommée **opendapr**. La principale fonction de cette librairie, nommée `odr_get_url`, prend en argument une collection d'intérêt (par exemple `MOD11A1.006`), une période d'intérêt (sous forme d'une date de début et de fin), une aire géographique d'intérêt (sous forme des coordonnées géographiques qui la délimitent), et des bandes d'intérêts (par exemple `LST_Night_1km`), et contruit automatiquement l'URL qui permettra finalement de télécharger le produit satellitaire d'intérêt. Une seconde fonction, `odr_download_data`, permet ensuite de télécharger le produit en local. À ce jour, la librairie permet de télécharger 77 collections de produits satellitaires recueillis sur toute la surface terrestre (incluant les produits MODIS et GPM, mais aussi SMAP (humidité du sol) et VIIRS (successeur de MODIS)). Au delà de son utilisation pour les travaux de cette thèse, cette librairie présente l'intérêt de rendre l'accès à certains produits satellitaires plus aisé aux utilisateurs de R, en particulier si la connexion à internet de l'utilisateur est lente et/ou onéreuse, de promouvoir une forme de sobriété digitale dans les travaux de recherche scientifique, et de soutenir le mouvement des logiciels libres et ouverts (sur lesquels nous nous sommes exclusivement basés pour l'ensemble de nos travaux, cf. section 3.3.2).

La libraire est disponible à l'adresse suivante : https://github.com/ptaconet/opendapr, et une description plus détaillée de la librairie est disponible en annexe C de ce manuscrit.

Une fois les produits téléchargés localement, nous avons préparé les données dans l'objectif de constituer des variables statistiques exploitables dans des modèles. Nous avons rééchantillonné les produits GPM (précipitations) depuis leur résolution spatiale initiale (10 km) à une résolution d'un kilomètre, en utilisant une méthode d'interpolation bilinéaire. Nous avons également combiné les produits journaliers MODIS issus de Terra et Aqua, en conservant les valeurs de pixels les plus élevées (respectivement les plus basses) disponibles pour les températures diurnes (respectivement nocturnes). Nous avons finalement comblé les valeurs manquantes dans les pixels (principalement dues à la présence de nuages) en interpolant temporellement les valeurs disponibles aux dates les plus proches.

La figure 3.2 montre les séries temporelles hebdomadaires des précipitations et températures sur les deux zones d'études, extraites des données GPM et MODIS LST.

Le bandeau gris représente la variabilité spatiale autour des différents points de captures entomologiques. L'alternance des saisons sèches et pluvieuses dans les deux zones d'études est clairement visible sur les graphiques de précipitations. Notons que les températures diurnes sont plus élevées dans la zone de Diébougou que dans celle de Korhogo, que les températures nocturnes sont relativement similaires, et que les précipitations en saison pluvieuse sont plus abondantes à Korhogo qu'à Diébougou.

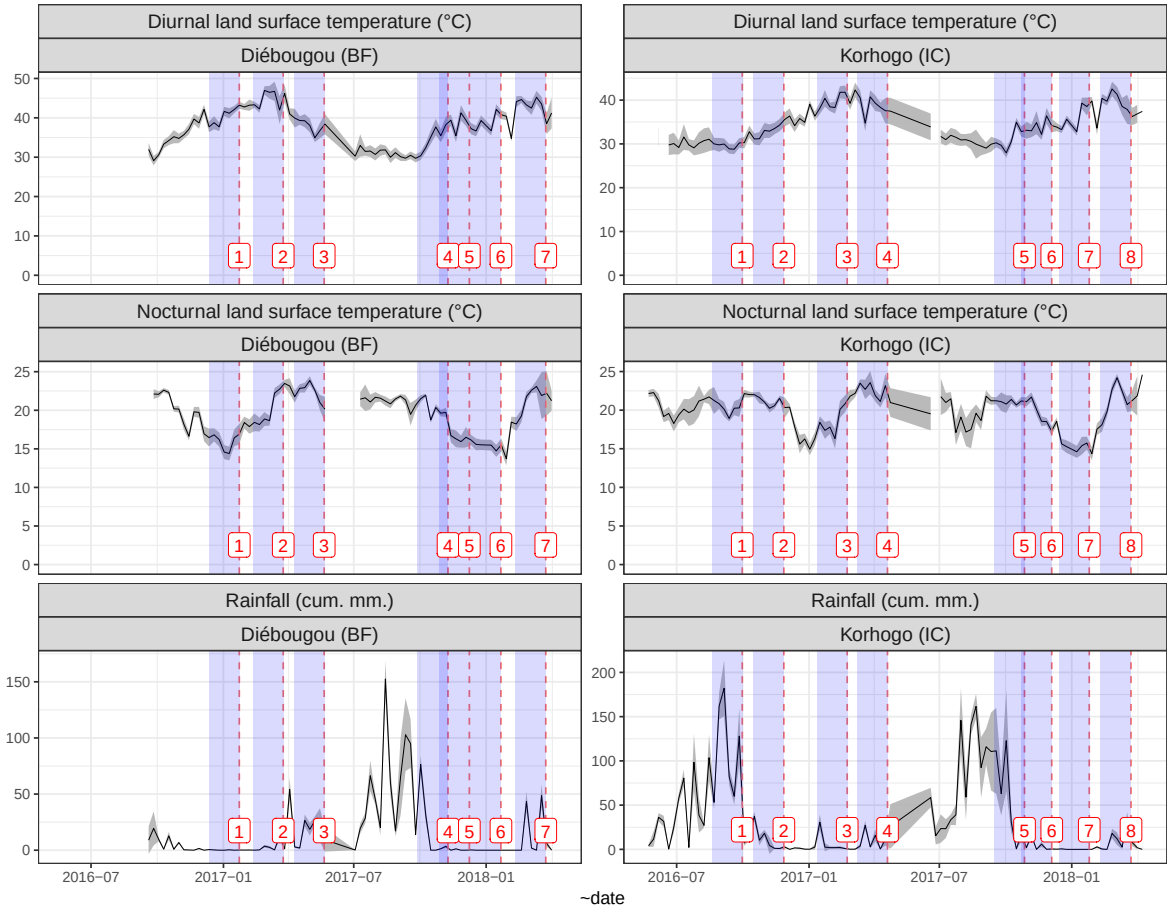


FIGURE 3.2: Courbes des conditions météorologiques sur les deux zones d'étude. Les lignes noires indiquent la moyenne de la variables météorologique pour tous les points de collectes entomologiques pour la semaine considérée. Les bandeaux gris indiquent la moyenne ± l'écart type (i.e. la variabilité spatiale pour la semaine considérée). Les lignes rouges verticales indiquent les dates des collectes entomologiques. Les bandeaux bleus indiquent une période de six semaines précédant chaque collecte entomologique. Source des données : GPM (précipitations), MODIS (températures diurnes et nocturnes)

3.2.2 Données d'occupation du sol

La caractérisation de l'occupation des sols sur un territoire peut être obtenue par classification d'images satellitaires (Anderson, Hardy, Roach, & Witmer, 1976) ou aériennes (Horning, Fleishman, Ersts, Fogarty, & Wohlfeil Zillig, 2020). Nous avons ainsi cartographié l'occupation du sol sur nos deux zones d'étude à l'aide d'une classification supervisée orientée objet de produits satellitaires d'observation de la Terre (G. J. Hay & Castilla, 2008).

Boite info n°7 : **Concept et principales étapes d'une classification supervisée orientée objet de produits satellitaires d'observation de la Terre**

Le principe de la cartographie de l'occupation du sol par télédétection spatiale est d'attribuer une classe d'occupation du sol à chaque pixel ou groupe de pixel d'une (ou plusieurs) image(s) géoréférencée(s). On peut distinguer deux grandes approches de classification d'images satellitaires à des fins de cartographie d'occupation du sol : i) l'approche orientée 'pixel', où chaque pixel de l'image satellitaire ou aérienne est classé individuellement sans tenir compte des pixels adjacents, et ii) l'approche orientée 'objet' (G. J. Hay & Castilla, 2008), où les pixels adjacents ayant des propriétés communes sont d'abord regroupés en 'objets', ces objets étant ensuite classés. L'approche orientée objet est particulièrement adaptée dans les cas où la résolution spatiale des pixels de l'image est largement inférieure à celle des entités constituant les classes d'occupation du sol que l'on cherche à extraire (G. J. Hay & Castilla, 2008) : par exemple, si la résolution du pixel est de quelques (centi)mètres mais que l'on cherche à extraire des informations type 'zones forestières'. Par ailleurs, l'approche orientée objet permet la classification non plus seulement sur les seules valeurs spectrales des pixels mais sur un ensemble de caractéristiques associées à l'objet : forme, relation avec les objets voisins, statistique sur les valeurs des pixels qui le compose, etc. Le qualificatif 'supervisé' fait référence, en modélisation statistique, au caractère connu *à priori* des classes (ici, d'occupation du sol) que l'on souhaite obtenir, par opposition à la classification non-supervisée où les classes sont automatiquement définies par un algorithme.

Les principales étapes d'une classification supervisée orientée objet sont les suivantes (G. J. Hay & Castilla, 2008) :

1. *Acquisition des produits satellitaires* : Il s'agit tout d'abord d'acquérir le(s) produit(s) satellitaire(s) qui sera(ont) utilisé(s) pour la classification. Notons que plusieurs produits satellitaires peuvent être utilisés, afin d'augmenter le volume et la diversité des informations capturées - et ainsi en théorie la qualité de la classification. Par exemple, il est possible d'utiliser une ou plusieurs images satellitaires optiques - qui donneront des informations sur la réfléctance des objets au sol dans plusieurs bandes spectrales - et un modèle numérique de terrain - qui donnera des informations sur le relief (altitude, pente, etc.).
2. *Constitution du jeu de données d'apprentissage et de validation* : Dans le cas d'une classification supervisée, il est nécessaire de constituer un jeu de donnée d'apprentissage / validation composé de plusieurs échantillons (parcelles) géoréférencés de chaque classe d'occupation du sol présentes dans la zone d'étude (les 'vérités terrain'). Ce travail nécessite donc i) de définir la liste des classes d'occupation du sol potentiellement présentes sur le territoire d'intérêt ; et ii) de constituer le jeu de données d'apprentissage / validation, par des enquêtes sur le terrain ou par photo-interprétation.
3. *Prétraitements des produits satellitaires* : Le ou les produits satellitaires utilisés pour la classification peuvent nécessiter un ensemble de prétraitements, selon leur degré de 'préparation' à l'étape d'acquisition. Parmi les prétraitements classiques, citons par exemple : la fusion des tuiles (dans le cas où la zone d'étude est composée de plusieurs tuiles satellitaires), la calibration optique (conversion des pixels dans le cas où les images satellitaires optiques ne sont pas prises sous les mêmes conditions atmosphériques), le traitement des pixels indisponibles (par exemple, à cause de la couverture nuageuse), ou encore l'orthorectification (afin d'améliorer le géoréférencement des images).
4. *Segmentation* : Cette étape est celle de la constitution des 'objets' par regroupement des pixels adjacents ayant des propriétés communes. Il existe plusieurs algorithmes de segmentation. Une des méthodes pour définir des objets consiste à agréger les pixels de proche en proche jusqu'à atteindre des seuils d'hétérogénéités fixés par l'utilisateur (liés à la taille des objets,

à leurs formes, et aux valeurs contenues dans les pixels), interrompant le processus et délimitant l'objet (Baatz & Schape, 2000).

5. *Constitution des variables prédictives* : Il s'agit ensuite de calculer pour chaque objet un ensemble d'attribut qui servira à entraîner le modèle sur le jeu d'entrainement, et à prédire sur les objets issus de la segmentation. Ces variables prédictives peuvent être basées sur des descripteurs statistiques des valeurs des pixels qui composent l'objet, sur la forme de l'objet, sa relation avec les objets voisins, etc.
6. *Classification* : Vient ensuite l'étape de la classification même : un modèle prédictif est d'abord entraîné sur les parcelles du jeu d'entraînement, puis est utilisé pour prédire la classe d'occupation du sol sur l'ensemble des objets de la zone d'étude.
7. *Evaluation de la qualité de la classification* : Enfin, la qualité de la classification est évaluée en prédisant l'occupation du sol sur le jeu de données de validation, puis en générant la matrice de confusion et les métriques de performances classiques afférentes (indices kappa, *accuracy*, etc.)

Dans notre cas, le détail des traitements ayant permis de génerer les produits d'occupation du sol à partir de classifications supervisées orientées objets de produits satellitaires est présenté ci-après.

1. Acquisition des produits satellitaires. Nous avons aquis les produits satellitaires suivants sur chaque zone d'étude :

— *images SPOT 6 et 7* (Satellite Pour l'Observation de la Terre) : images optiques à Très Haute Résolution Spatiale (THRS). Dates d'aquisition par le(s) satellite(s) : octobre 2017. Résolution spatiale : 1.6 m en panchromatique et 6.3 m en multispectral. Nombre de bandes spectrales : 4. Ces images ont été commandées via le dispositif Geosud, un projet (ANR-10-EQPX-20) du programme "Investissements d'Avenir" géré par le Centre National de la Recherche Scientifique.

— *images Sentinel-2* : images optiques à Haute Résolution Spatiale (HRS). Dates d'aquisition par le(s) satellite(s) : novembre / décembre 2018 (corresponsant aux

dates des campagnes d'acquisition de vérités terrain, voir ci-dessous). Résolution spatiale : 10 m à 60 m selon les bandes. Nombre de bandes spectrales : 10. Ces images libres d'accès ont été téléchargées sur le portail Copernicus SciHub de l'Agence spatiale européenne. L'intérêt d'utiliser des images Sentinel-2 en complément des images SPOT 6/7 est double : i) diversifier la nature et augmenter le nombre des variables prédictives (les images Sentinel-2 contiennent une information spectrale plus riche que les images SPOT-6 (10 bandes contre 4 bandes), et ii) intégrer des images acquises simultanément aux campagnes d'acquisition des vérités terrain.

— *Shuttle Radar Topography Mission (SRTM)* (JPL, 2013) : Modèle Numérique de Terrain (MNT) procurant la valeur de l'altitude en tout point du globe. Résolution spatiale : 30 m. Ce MNT libre d'accès a até téléchargé sur portail EarthExplorer (https://earthexplorer.usgs.gov/).

2. Constitution du jeu de données d'apprentissage et de validation. Nous avons mené une campagne d'acquisition de vérités terrain sur chacune des zones en novembre et décembre 2018 (10 jours sur la zone burkinabé, 14 jours sur la zone ivoirienne). Nous avons établi les classes d'occupation du sol dans chacune des zones sur la base de recherches bibliographiques sur les types de paysages potentiellement rencontrés dans nos zones (Aubréville, 1957 ; CILSS, 2016 ; OSS, 2015) et de nos observations du paysage sur le terrain. Nous avons collecté un minimum de 20 parcelles par classe, en tentant de les répartir au mieux sur l'étendue de chacune des zones. Sur le terrain, nous avons collecté les données à l'aide de l'application QField, compatible avec le logiciel de Système d'Information Géographique QGIS (QGIS Development Team, 2021). Nous avons ensuite complété le jeu de données en y ajoutant quelques parcelles par photo-interprétation d'images satellitaires très haute résolution (Google Earth et Spot 6/7).

3. Prétraitements des produits satellitaires. Les images SPOT ont été préparées selon la suite d'opérations suivante : fusion des tuiles de l'image panchromatique (PAN) ; conversion des images panchromatiques et multispectrales (MS) en valeurs de réflectance 'Top-of-Atmosphere' ; orthorectification des images PAN et MS en utilisant les informations disponibles dans les métadonnées des fichiers ainsi que le MNT SRTM (dont la résolution spatiale est suffisante pour une orthorectification de qualité au regard du profil de nos zones, peu accidentées) ; découpage des images sur l'étendue de nos zones d'études uniquement ; 'pansharpening' de l'image MS utilisant l'image PAN ; mosaïquage

des images (uniquement dans le cas de la zone de Diébougou, qui était constituée de deux images SPOT). Les produits Sentinel 2 et SRTM, de leur côté, ont nécéssité relativement peu de prétraitements. Nous avons reprojeté le MNT originellement fourni en WGS84 sur la zone UTM 30 Nord (zone UTM correspondant à nos zones d'étude), mosaiqué les images (dans le cas où nos zones d'études étaient couvertes par plusieurs tuiles) et découpé les produits afin de les conserver uniquement sur l'étendue de nos zones d'études.

4. Segmentation. Nous avons ensuite segmenté les images SPOT en utilisant un algorithme de croissance de région avec le critère d'homogénéité de Baatz and Shape (Baatz & Schape, 2000). Nous avons testé plusieurs paramétrisations, à la fois de l'algorithme (paramètres d'échelle, spectraux et de compacité) et des bandes spectrales utilisées pour la segmentation (bandes spectrales de l'image SPOT pan-sharpenées, bandes spectrales + indices spectraux type NDVI, etc.). Basé sur une approche visuelle des résultats de la segmentation (en les superposant à l'image SPOT 6/7), nous avons finalement retenu les paramètres suivants :

- Bandes spectrales utilisées pour la segmentation : les 4 bandes spectrales de l'image SPOT 6/7 pan-sharpenée, chacune avec un poids égal dans la segmentation ;
- Paramètres de segmentation pour la zone de Diébougou (BF) : seuil = 100 ; valeur pour le poids de forme = 0.1 ; valeur pour le poids spectral = 0.9
- Paramètres de segmentation pour la zone de Korhogo (CI) : seuil = 160 ; valeur pour le poids de forme = 0.1 ; valeur pour le poids spectral = 0.8

Nous avons vectorisé le jeu de données en sortie de l'algorithme afin d'avoir une version vecteur des objets segmentés. Puis, nous avons intersecté la base de données d'apprentissage (parcelles d'occupation du sol recueillies sur le terrain) avec la couche résultant de la segmentation, dans l'objectif d'obtenir des parcelles d'apprentissage plus homogènes du point de vue des critères de segmentation. Cette étape a sensiblement fait croître le nombre de parcelles d'entraînement - les objets issus de la segmentation étant globalement plus fragmentés et petits que les parcelles relevées sur le terrain. C'est cette couche de données d'apprentissage qui sera utilisée pour la suite du travail.

5. Constitution des variables prédictives. Afin de générer les variables prédictives, nous avons extrait ou calculé les couches géographiques suivantes (sous forme de fichiers raster) :

- Couches issues de l'image SPOT 6/7 :

  - chacune des 4 bandes spectrales de la THRS pan-sharpenée ;
  - image panchromatique ;
  - indices spectraux suivants : NDVI, NDWI2, BRI ;
  - indices de texture suivants, extraits de l'image PAN : energie, anthropie, correlation, inertie, haralick correlation, moyenne. Chacun des indices a été calculé sur 3 tailles de fenêtres glissantes : 5 pixels, 9 pixels, 17 pixels ;
- Couches issus de l'image Sentinel-2 :
  - chacune des 10 bandes spectrales ;
  - indices spectraux suivants : NDVI, NDWI, BRI, MNDWI, MNDVI, RNDVI (ces trois derniers indices sont des variantes des indices classiques NDWI et NDVI qui utilisent les bandes spectrales dans le moyen infra-rouge) ;
- Couches issues du MNT SRTM :
  - altitude ;
  - pente ;
  - réseau hydrographique théorique (couche vectorielle).

Au total, cela représentait ainsi 45 couches géographiques utilisables pour générer les variables prédictives. Nous avons calculé la moyenne et l'écart type des valeurs des pixels pour l'ensemble des indices préparées pour la classification sur chaque objet. Nous avons également calculé et ajouté les descripteurs contextuels suivants : i) la distance de chaque objet au réseau hydrographique théorique (calculé à partir du MNT, voir section 3.2.3), et ii) un ensemble d'indices liés à la forme des objets (aire, périmètre, etc.). La centaine de variables ainsi générée constituait les prédicteurs pour la classification à suivre.

6. Classification. Nous avons ensuite entrainé un modèle de forêts aléatoires sur le jeu de données d'entrainement. Nous avons généré la matrice de confusion en utilisant la procédure de validation interne aux forêts aléatoires (basée sur les 'out-of-bag' observations (Breiman, 1996)). En se basant sur cette matrice, nous avons ensuite regroupé, dans le jeu de données d'entrainement, les classes d'occupation du sol dont la confusion était importante (par exemple, zones de culture de mil et de sorgho) ; en prenant cependant soin de conserver la distinction entre les différentes classes à priori favorables à la présence de gîtes larvaires (par exemple, zones marécageuses et rizicoles) ou à la résistance. Nous avons entrainé un modèle de forêt

aléatoires sur cette nouvelle version du jeu de données d'entraînement puis l'avons utilisé pour prédire la classe d'occupation du sol sur chaque objet issu de la segmentation.

7. Evaluation de la qualité de la classification. Comme précédemment, nous avons généré la matrice de confusion puis en avons extrait un indice de qualité de la classification (*accuracy* (J. Cohen, 1960)) mesurant la proportion d'objets correctement classés.

Les différentes étapes de la classification sont résumées graphiquement dans la figure B.1 disponible en annexe B. Nous avons développé un script R implémentant l'ensemble des traitements (voir section 3.3.1).

Le tableau 3.1 présente les classes d'occupation du sol initialement définies et finalement retenues. La définition de chacune des classes ainsi qu'un ensemble de photographies représentatives des principales classes d'occupation du sol, prises lors des campagnes de terrain, sont disponibles en annexe B.

TABLE 3.1: Classes d'occupation du sol initialement définies et finalement retenues

Classes retenues	Classes initialement définies	Présence sur zone(s) d'étude
Eau permanente (Permanent water bodies)	Eau dormante	BF et CI
	Eau courante	BF et CI
Culture et jachère, hors coton et riz (Crops)	À maïs	BF et CI
	À poids de terre	BF et CI
	À arachide	BF et CI
	À mil	BF
	À sorgho	BF
	À haricot	BF
	À sésame	BF
	Jachère	CI
Culture cotonière (Cotton)	Culture cotonière	BF et CI
Rizière (Rice)	Rizière	BF et CI
Plantation (Plantation)	À mangue	CI
	À anacarde	CI
Savane ligneuse (Ligneous savannah)	Savane arbustive	BF
	Savane arborée	BF
	Savane boisée	BF
	Savane dégradée	CI
Prairie (Grassland)	Prairie	BF
Milieu forestier non humide (Woodland)	Forêt dense	CI
	Forêt claire	BF et CI
Forêt ripicole (Riparian forest)	Forêt ripicole	BF et CI
Prairie marécageuse (Marsh)	Prairie marécageuse	BF et CI
Bâti (Settlements)	Bâti	BF et CI
Sol nu (Bare soil)	Sol nu	BF et CI
Route principale (Main tracks)	Routes principales	BF et CI

Les matrices de confusion indiquaient que respectivement 84 % et 86 % des objets dans les zones de Diébougou et Korhogo étaient bien classés. Les cartes 3.3 et 3.4 présentent les produits finis d'occupation du sol dans les deux zones d'étude. Ces cartes incluent également le réseau hydrographique théorique, généré à partir du MNT SRTM (voir section 3.2.3).

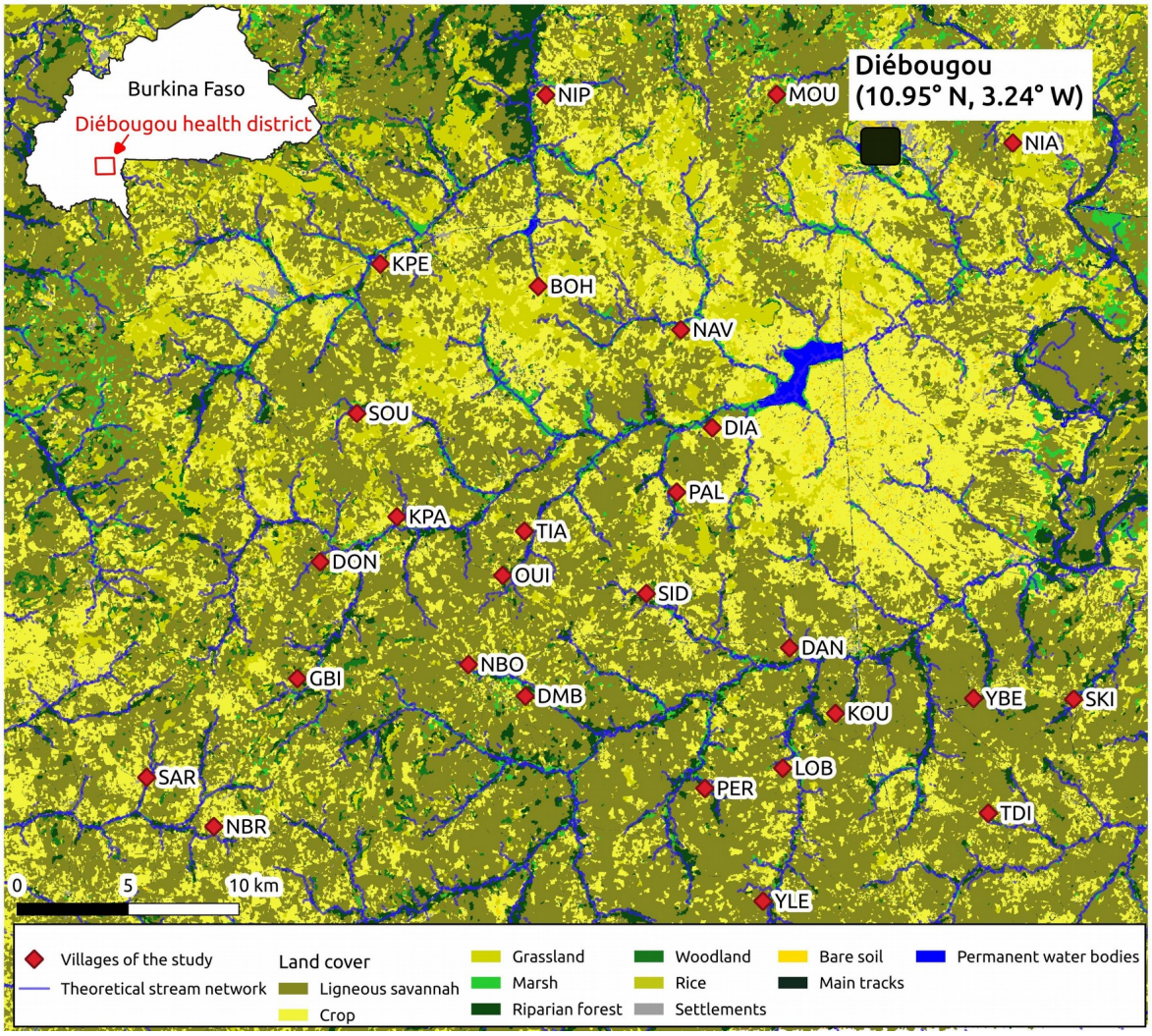


FIGURE 3.3: Carte d'occupation du sol résultante des travaux de classification dans la zone de Diébougou (BF) (résolution spatiale : 1,5 x 1,5 m)

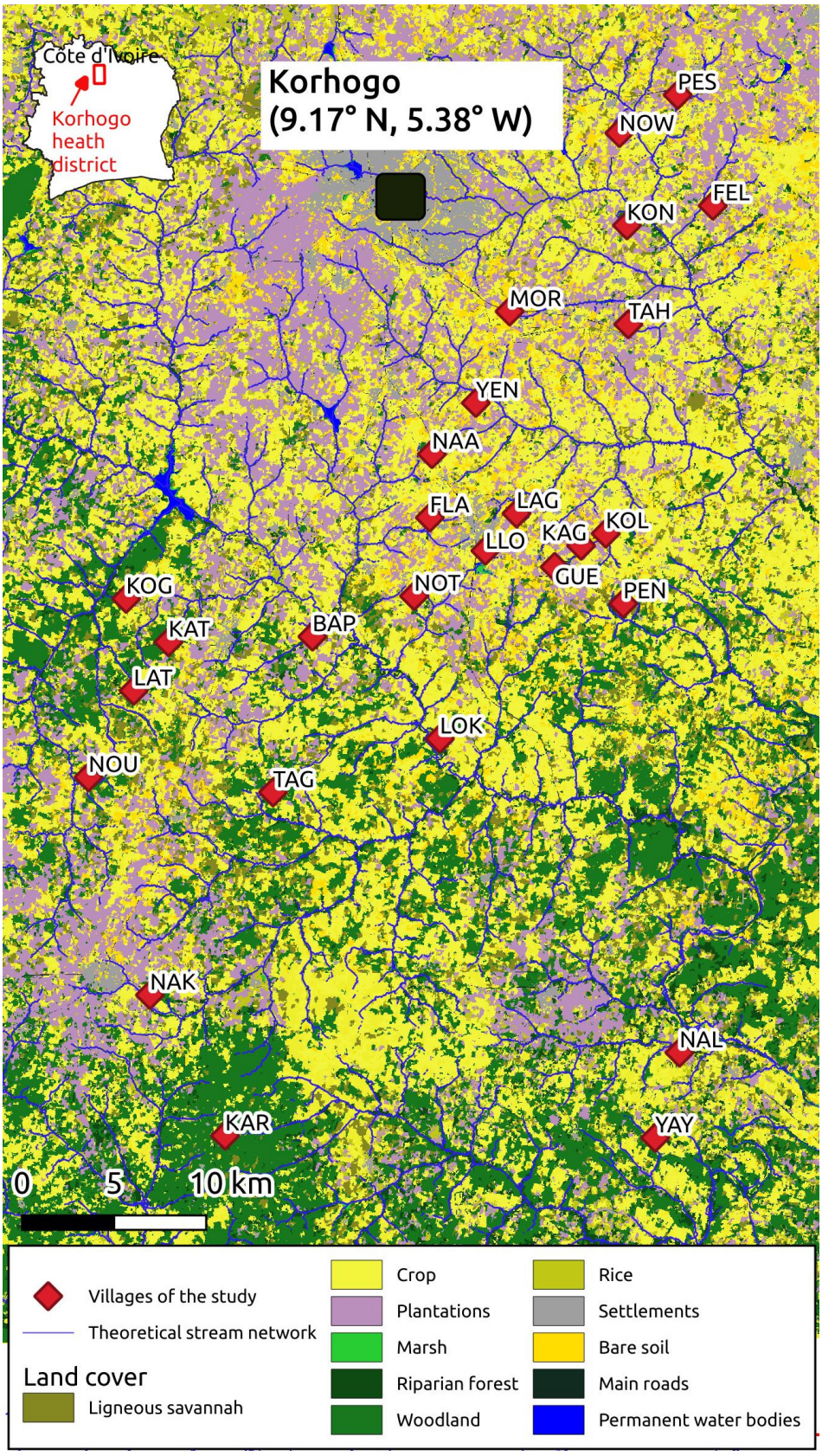


FIGURE 3.4: Carte d'occupation du sol résultante des travaux de classification dans la zone de Korhogo (CI) (résolution spatiale : 1,5 x 1,5 m)

Enfin, la figure 3.5 présente la proportion de surface occupé par chaque classe d'occupation du sol dans l'ensemble de la zone d'étude.

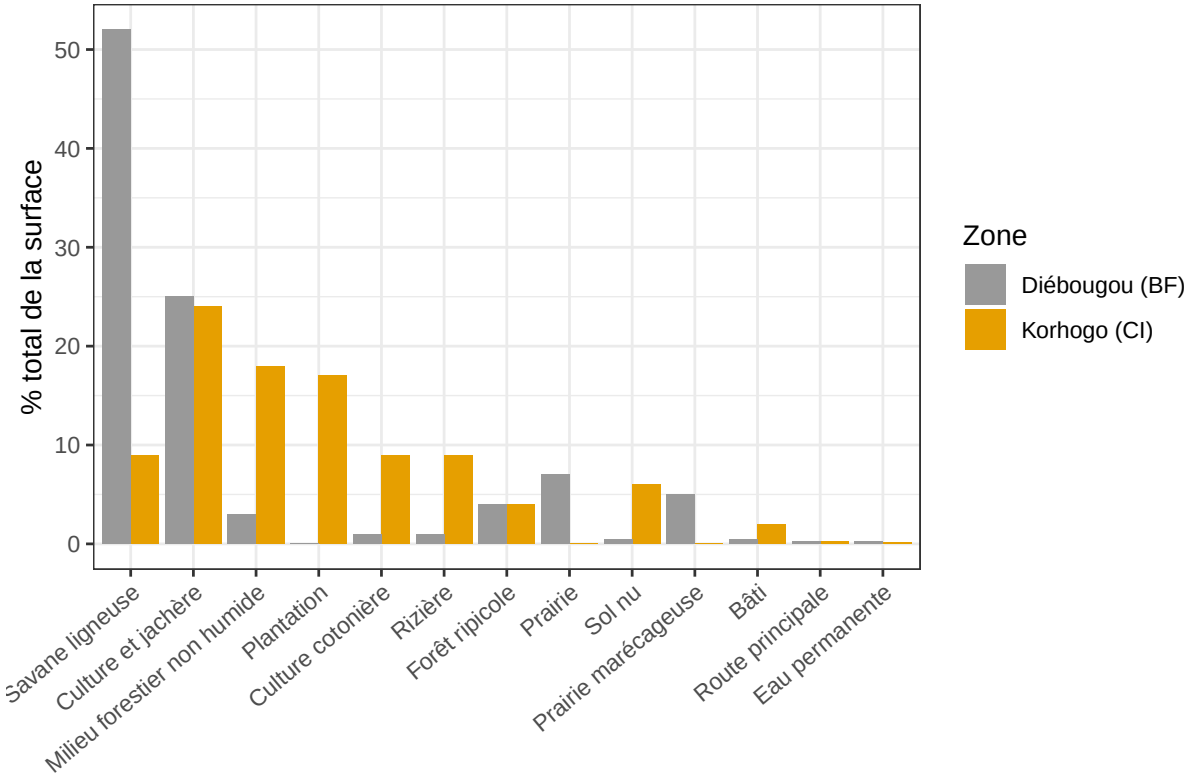


FIGURE 3.5: Proportion de surface occupée par chaque classe d'occupation du sol sur chaque zone d'étude

Nous pouvons noter que la zone de Diébougou était dominée par les savanes ligneuses (52% de la surface totale), les cultures non-inondées (25%) et les prairies non-inondées (7%). La zone de Korhogo, de son côté, était principalement composée de cultures non-inondées (24%), de milieux forestiers non-humides (18%) et de plantations d'anacardiers et mangues (17%). Les rizicultures et cultures de coton y représentaient chacune 9% de la surface totale.

3.2.3 Données sur le réseau hydrographique théorique

Le réseau hydrographique (cad. les rivières) est susceptible de produire des gîtes larvaires pour les anophèles. Nous avons utilisé le MNT SRTM pour produire le réseau hydrographique théorique dans nos zones d'étude. Nous avons tout d'abord produit une couche raster d'accumulation de flux à partir du MNT (Jenson & Domingue, 1988), puis

avons sélectionné tous les pixels dont la valeur d'accumulation de flux était supérieur à un seuil défini visuellement par superposition avec une image satellite THRS. Ces seuils étaient de 1000 pour la zone BF et 800 pour la zone CI. Les réseaux hydrographiques théoriques sont représentés sur les cartes 3.3 et 3.4.

3.3 Ressources informatiques : codes R développés et logiciels utilisés

3.3.1 Codes R développés

L'ensemble des travaux d'extraction et préparation des données décrits dans ce chapitre a été programmé, sous formes de codes informatiques, dans le language de programmation R (R Core Team, 2018) sous l'environnement RStudio (RStudio Team, 2020). Les travaux de modélisation de l'ensemble des articles à suivre ont eux aussi été scriptés en R.

Le niveau de description, généricité, réutilisation diffère selon les codes. Le plus abouti en ce sens est la librairie `opendapr`, puisqu'il s'agit d'une réelle libraire R. D'une manière générale, les codes de création, extraction et préparation des données ont un niveau acceptable de description et généricité (à savoir, ils peuvent être réutilisés tout ou partie à moindre coût par un utilisateur de R). Les codes de modélisation statistique sont moins décrits et reproductibles. Quel que soit le niveau de généricité et description, nous avons archivé l'ensemble de ces codes en leur état à l'issue des travaux de thèse, afin d'en conserver une copie pérenne et conforme aux travaux effectués. Le dossier contenant les codes est disponible à l'adresse suivante : https://doi.org/10.5281/zenodo.6334110 . L'architecture du dossier est décrite dans le tableau 3.2.

3.3.2 Logiciels et librairies utilisés

Les logiciels et librairies utilisées pour nos travaux de thèse étaient tous libres d'accès et à code source ouvert.

Les données d'occupation du sol (section 3.2.2) et du réseau hydrographique théorique (section 3.2.3) ont été générés en utilisant les librairies R suivantes : `RSAGA`

TABLE 3.2: Codes R développés au cours de la thèse

Nom du dossier	Description du contenu
data_creation extraction_preparation	Ce dossier contient les codes R développés pour créer, extraire, préparer les données environnementales : - le sous-dossier data_creation_extraction contient les codes pour i) (extraction_landcover_data) générer les données d'occupation du sol par classification supervisée orientée objet de produits satellitaires, ii) (extraction_meteo_data) extraire les produits météorologiques avec la librairie opendapr, iii) (extraction_miscellaneous_data) extraire des données environnementales diverses (les données de magnitude visuelle de la Lune (IMCCE), les données de vent (ERA-5), le MNT SRTM) - le sous-dossier data_preparation contient les codes développés pour extraire les données, les préparer, et calculer les variables explicatives pour les travaux de modélisation statistique
data_modeling	Ce dossier contient les codes R développés pour les travaux de modélisation statistique : - le sous-dossier modeling contient les codes pour générer les modèles statistiques des chapitres 4 et 5 - le sous-dossier models_analysis contient les codes pour analyser (interpréter) les modèles statistiques

(Brenning, Bangs, & Becker, 2018), `rgrass7` (Bivand, 2018), `raster` (Hijmans, 2020), `sf` (Pebesma, 2018), `rgdal` (Bivand, Keitt, & Rowlingson, 2019) et `randomForest` (Liaw & Wiener, 2002). Certaines de ces librairies utilisent en arrière-plan les logiciels libres SAGA GIS (Conrad et al., 2015) et GRASS GIS (GRASS Development Team, 2017). La segmentation a été réalisée grâce à l'algorithme 'Generic Region Merging Segmentation' implémenté dans le logiciel libre Orfeo Toolbox (Grizonnet et al., 2017).

Désireux de soutenir le mouvement du logiciel libre, nous avons rédigé au cours de la thèse un tutoriel d'initiation à la télédétéction spatiale (cartographie de l'occupation/utilisation du sol) sur logiciel libre (QGIS (QGIS Development Team, 2021) et SAGA GIS). Le tutoriel est disponible en annexe D. Nous l'avons utilisé au cours d'une formation en télédétéction dispensée à des étudiants de niveau master.

Les travaux de modélisation dans les études qui suivent ont eux aussi nécéssité l'utilisation de nombreuses librairies R, qui sont précisées dans les sections *Matériel et*

méthode des articles respectifs.

Le logiciel de gestion de la bibliographie utilisé pour cette thèse était Zotero. Enfin, l'ensemble des travaux de thèse a été réalisé sur un ordinateur équipé d'Ubuntu, un système d'exploitation à code source ouvert utilisant le noyau Linux ; et ce manuscrit de thèse a été rédigé en LaTeXen s'appuyant sur les librairies R `rmarkdown` (Allaire, Horner, Xie, Marti, & Porte, 2019), `knitr` (Xie, 2020), `bookdown` (Xie, 2019), `thesisdown` (Ismay & Solomon, n.d.), `kableExtra` (Zhu, 2019).

Chapitre 4

Article n°1 - Modélisation des dynamiques spatio-temporelles des abondances des vecteurs

Dans cette première étude, nous nous intéressons aux déterminants environnementaux de la présence et de l'abondance des espèces vectrices du paludisme dans nos zones d'étude. Comprendre de quelle manière l'environnement impacte la distribution et la densité des anophèles, et pouvoir prédire ces densités à fine échelle spatio-temporelle, peut en effet aider à concevoir et déployer des interventions de LAV efficaces (voir section 1.3.1). L'étude présentée dans ce chapitre avait ainsi pour objectif d'affiner les connaissances sur les liens entre environnement et nombre de contacts hommes-vecteurs et d'évaluer la prédictibilité spatio-temporelle des densités agressives, dans nos deux zones d'étude. Pour cela, nous étudions le système {environnement - abondance des vecteurs} dans une approche holistico-inductive, avec un processus de modélisation statistique en deux étapes. Nous exploitons pleinement la granularité spatio-temporelle des données environnementales à notre disposition et le potentiel descriptif et prédictif des modèles statistiques non-paramétriques pour expliquer et évaluer la prédictibilité des densités agressives des vecteurs dans nos zones d'étude. Pour la zone de Diébougou (BF), cette étude a fait l'objet d'une publication scientifique en tant qu'auteur principal (https://doi.org/10.1186/s13071-021-04851-x), que nous résumons puis intégrons dans ce chapitre. Dans une troisième partie (section 4.3), nous présentons et discutons les résultats dans la zone d'étude de Korhogo (CI).

4.1 Résumé de l'article

Les objectifs principaux de cette étude étaient i) d'approfondir les connaissances sur les déterminants environnementaux (en particulier météorologiques et paysagers) des densités agressives des anophèles dans la zone de Diébougou, et ii) d'évaluer la prédictibilité des densités agressives des anophèles dans l'espace et dans le temps.

Nous avons modélisé les densités agressives des vecteurs dans la zone de Diébougou en fonction des conditions météorologiques et paysagères à proximité des points de capture. La variable à expliquer était le nombre de contacts homme-vecteur par homme et par nuit de capture. Nous avons constitué des variables explicatives météorologiques à partir des données de météorologie précédant les collectes entomologiques, des variables paysagères à partir des données d'occupation du sol et du réseau hydrographique théorique, et des variables liées à l'attractivité et la pénétrabilité des habitations à partir des données de localisation des habitations. Les variables paysagères ont été calculées dans des zones tampon de rayon variable autour de chaque point de capture, et les variables météorologiques ont été calculées dans tous les intervalles de temps possibles entre 0 et 6 semaines avant les dates de collecte. Nous avons modélisé les densités agressives des vecteurs en deux étapes, apportant des informations complémentaires sur la bio-écologie des anophèles. Dans un premier temps, nous avons calculé les coefficients de corrélation de Spearman entre l'abondance des vecteurs et chaque variable environnementale prise aux différentes zones tampons (pour les variables paysagères) et intervalles de temps (pour les variables météorologiques) ; dans l'objectif d'identifier les périodes (précédant la capture) et espaces (autour du point de capture) pour/dans lesquels nos variables environnementales influençaient au plus les densités agressives. Dans un second temps, nous avons entraîné puis interprété des modèles multivariés d'apprentissage automatique (fôrets aléatoires) afin d'identifier i) l'importance relative des variables environnementales dans l'abondance des vecteurs et ii) l'effet respectif (relation fonctionnelle) de chaque variable environnementale sur l'abondance des vecteurs. Nous avons également évalué la capacité des modèles multivariés à prédire sur de nouveaux villages, par validation croisée. Pour des raisons à la fois d'ordre statistique et biologique, nous avons modélisé séparément la probabilité de présence des vecteurs (probabilité de contact homme-vecteur) et l'abondance des vecteurs (nombre de contacts homme-vecteurs dans les sessions de capture avec au moins un contact). De

plus, nous avons modélisé séparément les trois espèces majeures d'anophèles identifiées sur le terrain (*An. gambiae s.s.*, *An. coluzzii*, *An. funestus*), car les déterminants environnementaux de la présence / abondance de chacune peuvent différer.

Nous avons observé que les densités agressives étaient, sans surprise, très hétérogènes dans le temps (entre les saisons) et dans l'espace (entre les villages). Dans l'analyse bivariée (coefficients de corrélation), les variables météorologiques et paysagères étaient souvent statistiquement significativement corrélées avec la présence ou l'abondance des vecteurs ; et les coefficients de corrélation variaient fréquemment selon la distance, spatiale (pour les variables paysagères) ou temporelle (pour les variables climatiques), au point de capture. Dans l'analyse multivariée, de nombreux seuils et associations non-linéaires ont été révélés par les modèles d'apprentissage automatique, tant pour les variables météorologiques que pour les variables du paysage. Les modèles multivariés présentaient de bonnes puissances prédictives, indiquant que dans l'ensemble, les déterminants de l'abondance des vecteurs ont été identifiés.

L'interprétation des modèles nous a permis de formuler des hypothèses sur la bioécologie des principales espèces vectrices du paludisme dans la zone de Diébougou. Nous avons conjecturé que les conditions météorologiques (températures, précipitations) affectaient tous les stades de vie des moustiques capturés (larves, adultes) à des niveaux variables selon l'espèce et le paramètre météorologique. La météorologie avait parfois un impact plus important encore sur les périodes précédant la durée de vie de la génération échantillonnée. Des gites larvaires préférentiels pour chaque espèce ont été proposés : *An. funestus*, *An. coluzzii* et *An. gambiae s.s.* semblaient être distribués le long d'un gradient de persistance des sites de reproduction, de permanent à temporaire, confirmant la littérature. Par ailleurs, le niveau d'ouverture du paysage semblait impacter significativement les densités agressives des vecteurs (les milieux ouverts favorisant la probabilité et l'abondance des piqûres, et inversement), ce qui pourrait représenter un problème majeur au regard du rétrécissement progressif des surfaces de savane et de forêt au Burkina Faso.

Dans cette étude, nous avons approfondi nos connaissances sur les liens complexes entre environnement et abondance des vecteurs du paludisme, à fine échelle spatio-temporelle, en utilisant conjointement des données entomologiques issues de collectes sur le terrain et des données issues des images satellitaires d'observation de la Terre, dans un

cadre de modélisation statistique avancé. Ce travail pose les bases pour le développement d'outils opérationnels pour améliorer et optimiser la lutte contre la transmission du paludisme à l'échelle locale, tels que des plans d'action de lutte antivectorielle, des cartes saisonnières de la distribution des vecteurs à l'échelle du village, ou encore des systèmes d'alerte précoce pour la détection des épidémies de paludisme.

4.2 Texte intégral de l'article

Les figures additionnelles sont disponibles après l'article (section 4.2.1) ainsi que dans la version en ligne de l'article (https://doi.org/10.1186/s13071-021-04851-x)

Parasites & Vectors

RESEARCH Open Access

Data-driven and interpretable machine-learning modeling to explore the fine-scale environmental determinants of malaria vectors biting rates in rural Burkina Faso

Paul Taconet[1,2*], Angélique Porciani[1], Dieudonné Diloma Soma[1,2,3], Karine Mouline[1], Frédéric Simard[1], Alphonsine Amanan Koffi[4], Cedric Pennetier[1,2], Roch Kounbobr Dabiré[2], Morgan Mangeas[5] and Nicolas Moiroux[1,2]

Abstract

Background: Improving the knowledge and understanding of the environmental determinants of malaria vector abundance at fine spatiotemporal scales is essential to design locally tailored vector control intervention. This work is aimed at exploring the environmental tenets of human-biting activity in the main malaria vectors (*Anopheles gambiae s.s.*, *Anopheles coluzzii* and *Anopheles funestus*) in the health district of Diébougou, rural Burkina Faso.

Methods: *Anopheles* human-biting activity was monitored in 27 villages during 15 months (in 2017–2018), and environmental variables (meteorological and landscape) were extracted from high-resolution satellite imagery. A two-step data-driven modeling study was then carried out. Correlation coefficients between the biting rates of each vector species and the environmental variables taken at various temporal lags and spatial distances from the biting events were first calculated. Then, multivariate machine-learning models were generated and interpreted to (i) pinpoint primary and secondary environmental drivers of variation in the biting rates of each species and (ii) identify complex associations between the environmental conditions and the biting rates.

Results: Meteorological and landscape variables were often significantly correlated with the vectors' biting rates. Many nonlinear associations and thresholds were unveiled by the multivariate models, for both meteorological and landscape variables. From these results, several aspects of the bio-ecology of the main malaria vectors were identified or hypothesized for the Diébougou area, including breeding site typologies, development and survival rates in relation to weather, flight ranges from breeding sites and dispersal related to landscape openness.

Conclusions: Using high-resolution data in an interpretable machine-learning modeling framework proved to be an efficient way to enhance the knowledge of the complex links between the environment and the malaria vectors at a local scale. More broadly, the emerging field of interpretable machine learning has significant potential to help improve our understanding of the complex processes leading to malaria transmission, and to aid in developing

*Correspondence: paul.taconet@ird.fr
[1] MIVEGEC, Université de Montpellier, CNRS, IRD, Montpellier, France
Full list of author information is available at the end of the article

operational tools to support the fight against the disease (e.g. vector control intervention plans, seasonal maps of predicted biting rates, early warning systems).

Keywords: Malaria, Anopheles, Biting behavior, Abundance, Ecological niche, Earth observation data, Statistical modeling, Cross-correlation maps, Random forest, Interpretable machine learning, Africa

Background

Malaria is a vector-borne disease transmitted by *Anopheles* mosquitoes still affecting 229 million people and causing more than 400,000 deaths worldwide annually [1]. Malaria control efforts, mainly through the massive use of long-lasting insecticidal nets [2], led to a sustained decrease of the disease burden between 2000 and 2015 [1]. However, malaria cases have plateaued in the past 5 years or even increased in certain areas [1]. Vector resistance to insecticides, population growth and environmental changes are involved in such worrying trends [3, 4]. For effective and sustainable vector control (VC), locally tailored interventions, built on a thorough knowledge of the local determinants of malaria transmission, are needed [3–5]. To do so, it is of particular importance to decipher with vector bio-ecology at fine and operational spatiotemporal scales [3, 4, 6].

To develop efficient (i.e. species-, place- and time-specific) vector control strategies, important features of malaria vector ecology such as breeding site typologies, development and survival rates, flight ranges, or dispersal have to be considered. Meteorological conditions (temperature, precipitation) and land cover are major environmental factors frequently used to define the ecological niche of malaria mosquitoes [5] in complex, sometimes hardly hypothesizable, ways. Temperature affects the mosquito life history traits, nonlinearly, at each stage of its life cycle (e.g. larval growth, adult survival, biting rate). For example, the daily mortality rate of several adult *Anopheles* species follows a unimodal relationship with air temperature, with an optimal adult survival rate at around 25 °C [7–9]. Rainfall generates additional mosquito breeding sites and is therefore an important factor explaining the seasonality in species abundance. However, excessive rainfall can destroy developing larvae by flushing them out of their aquatic habitat [10, 11]. Land cover may affect mosquito population dynamics by creating breeding sites in hydromorphic areas or altering the dispersal ability of mosquitoes. A modification of land cover/use may therefore either increase or decrease vector abundance relative to species ecological preferences. As an example, deforestation can increase larval breeding sites of malaria vectors growing in sunny puddles, whereas it destroys habitats of some deep-forest *Anopheles* species [5, 12]. Moreover, even when found together, *Anopheles* species often exhibit specific ecological preferences [13, 14]. As an example, *Anopheles gambiae s.s.* was more frequently observed in temporary, rainfall-dependent breeding sites [15–17], whereas *Anopheles coluzzii* showed a preference for more permanent breeding sites [17–19]. Altogether, these examples illustrate that vector ecology is finely tuned with the environment. Using large-scale environmental indicators could therefore jeopardize the characterization of ecological niches of malaria vectors and consequently lead to suboptimal or even inappropriate VC intervention at a smaller scale [5]. To overcome this issue, we propose the use of high-resolution Earth observation (EO) data and develop novel statistical modeling approaches.

Indeed, in malariometric statistical modeling studies, "data" models [20] like linear or logistic regression are traditionally used with environmental variables extracted from EO data [21–23]. These models are well suited for testing pre-established hypotheses about theoretical constructs (e.g. to answer questions like "how much higher is mosquito abundance for each additional millimeter of rainfall?"); however, to explore hypotheses and extract knowledge in complex systems, machine-learning (ML) "algorithmic" models might be more suitable [20, 24]. In fact, these models are inherently able to capture complex patterns (such as nonlinear relationships and complex interactions between variables) contained in data. After a good predictive algorithm is fitted to a dataset, post hoc interpretation methods may uncover the complex relationships contained in the data and learned by the model, which in turn can be carefully linked to prior knowledge to identify meaningful—possibly unforeseen—cause–effect relationships, valuable thresholds or interactions [25–27]. This predict-then-explain modeling workflow is being increasingly used to generate knowledge from complex datasets [24, 26] and is commonly referred to as "interpretable machine learning" (IML) [25, 26].

The main objective of this study was to improve our overall understanding of the ecological niche and determinants of the biting rates of the main malaria vectors in a rural area of southwestern Burkina Faso. To do so, we used entomological collections and environmental variables extracted from high-resolution EO data, in a data-driven and IML modeling framework. In this research article, after a presentation of the methods and results, we discuss the environmental (landscape and meteorological) drivers of the human-biting activity of the main malaria

vectors in the Diébougou area. We also briefly present some potential practical uses of our results to support the conceptualization and deployment of locally tailored VC interventions. We conclude with methodological insight regarding the use of algorithmic models and IML for knowledge-building in the field of landscape entomology.

Methods

Data collection and preparation

Entomological data

Anopheles human-biting activity was monitored as part of a study carried out in the Diébougou rural health district located in southwest Burkina Faso [28]. Twenty-seven villages in this 2500 km^2 wide area were selected according to the following criteria: accessibility during the rainy season, 200–500 inhabitants per village, and distance between two villages greater than 2 km. Seven rounds of mosquito collection were conducted in each village between January 2017 and March 2018. The periods of the surveys span some of the typical climatic conditions of this tropical area (three surveys in the "dry-cold" season, two in the "dry-hot" season, one at each extremum of the rainy season) (see Additional file 1: Summary of the meteorological conditions around the sampling points).

Mosquitoes were collected using the human landing catch (HLC) technique from 17:00 to 09:00 both indoors and outdoors at four sites per village for one night during each survey. The procedure for conducting HLC was for a person to sit on a stool, and mosquitoes to alight on his exposed legs, where they were then collected using a hemolysis tube. Collectors were rotated hourly between collection sites and/or position (indoor/outdoor). Independent staff supervised rotations and regularly checked the quality of mosquito collections. Malaria vectors were identified using morphological keys [29, 30]. Individuals belonging to the *Anopheles gambiae* complex and the *Anopheles funestus* group were identified to species by PCR [31–33]. Mosquito collection design for this study has been described extensively elsewhere [28].

HLC enabled us to measure the presence and abundance of aggressive malaria vectors in time and space. In fact, landing on human legs is the behavioral event preceding the biting event. To avoid exposing mosquito collectors to infectious bites, we used landing as a proxy for biting, and in turn biting probability/rate as a proxy for the overall presence/abundance of aggressive vectors at the time and place of collection.

Landscape data

A land cover map of the study area was produced by carrying out a geographic object-based image analysis (GEOBIA) [34] using multisource very-high- and high-resolution satellite-derived products. The GEOBIA involved the following main steps: acquisition/collation of the satellite products (Satellite Pour l'Observation de la Terre (SPOT)-6 image acquired on 2017-10-11, Sentinel-2 image acquired on 2018-11-16, and a digital elevation model (DEM) from the Shuttle Radar Topography Mission [35]), acquisition of a ground-truth dataset composed of 420 known land cover samples by both fieldwork (held in November 2018) and photo-interpretation of satellite images, and classification of the land cover over the whole study area using a random forest (RF) algorithm [36]. The definitions of land cover classes were those proposed by the Permanent Interstate Committee for Drought Control in the Sahel [37]. The resulting dataset was a georeferenced raster image, where each 1.5×1.5 m pixel was assigned a land cover class. The confusion matrix was generated using the internal RF validation procedure based on the out-of-bag observations, and the quality of the final classification was assessed by calculating the overall accuracy from the confusion matrix [38].

Spatial buffers were then defined to characterize the environmental conditions at the neighborhood of each HLC collection point. Four buffer radii were considered: 250 m, 500 m, 1 km, 2 km. The distance of 2 km was chosen as the largest radius to minimize overlaps among buffers coming from different villages and because local dispersal of *Anopheles* beyond this distance can be considered negligible [29, 39–41]. We calculated the percentage of landscape occupied by each land cover class in each buffer zone around each collection site.

Additional indices related to the presence of water were calculated. The theoretical stream network was produced for the study area by first generating a flow accumulation raster dataset from the DEM and then applying a threshold value to select cells with accumulated flow greater than 1000 [42]. The quality of the product was assessed visually by overlaying it on the SPOT-6 satellite image. We then derived two indices for each collection site: the length of streams in each buffer zone, and the shortest distance to the streams.

In order to describe attractiveness and penetrability of households for malaria vectors, the geographical location of the households in the villages were recorded and two indices were computed. First, the Clark and Evans aggregation index [43] was calculated to describe the degree of clustering of the households in each village, as it has been suggested that scattered habitations in a village might increase the attractiveness for some vector species [44]. Second, we calculated the distance from each collection point to the edge of the village (defined as the convex hull polygon of each village—i.e. the minimum polygon that encompasses all the locations of the households), as it has been suggested elsewhere that living on the edge can increase biting rates [45].

Meteorological data

Daily rainfall estimates were extracted from the Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) Final products [46]. The raw satellite products were resampled from their original 10-km spatial resolution to a 1-km resolution using a bilinear interpolation method.

Daily diurnal and nocturnal temperatures were derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) Land Surface Temperature (LST) Terra and Aqua products [47, 48]. Terra and Aqua daily products were first combined, keeping the highest (or lowest) available pixel values for the diurnal (nocturnal) temperature. Missing values in pixels (mostly due to cloud presence) were then filled by temporally interpolating the values of the closest preceding and following available dates.

These meteorological data (daily rainfall, daily diurnal temperatures, daily nocturnal temperatures) were collected up to 42 days (i.e. 6 weeks) preceding each mosquito collection, so as to encompass largely the whole duration of the *Anopheles* life cycle in the field (including aquatic and aerial stages) [49]. They were then aggregated pixel-by-pixel on a weekly scale (cumulative 7-day rainfall and average 7-day diurnal and nocturnal LST). This temporal granularity represents a reasonable trade-off between the raw, daily information—which might overfit in the statistical models—and larger scales, which might prevent them from capturing fine-scale temporal relationships. Next, we calculated the cumulative rainfall and average temperatures for all possible intervals of time available in the data (e.g. b/w 0 and 1 week before the dates of collection, b/w 0 and 2 weeks, b/w 1 and 2 weeks, etc.). The data were finally averaged in the 2-km buffer zone only (considering the 1-km spatial resolution of the source data).

Statistical analyses

Overall approach

We used a two-step statistical modeling approach to study the relationships between the biting rates of each vector species and the environmental conditions. We first calculated correlation coefficients between the biting rates and the environmental variables at the various buffer sizes/time lags considered. The objectives of this bivariate analysis were twofold: (i) to better apprehend several aspects of the ecology of the vectors in the study area, and (ii) to screen out variables for the multivariate analysis. In a second stage, we integrated selected variables in multivariate algorithmic models that we further analyzed using interpretable machine-learning tools, to search for potential complex links (nonlinear relationships, relevant thresholds) between the environmental factors and the biting rates.

We ran the whole modeling framework separately for each species, as they might exhibit different ecological preferences.

From a statistical point of view, most algorithmic machine-learning models, although nonparametric, have difficulty coping with zero-inflated negative binomial response variables [50, 51], which are typically found in insect count data such as mosquito biting rates [52]. An alternative approach to model such data is the hurdle model that considers the data responding to two processes: one causing zero versus nonzero and the second process explaining the nonzero counts [53]. The hurdle methodology in the frame of a widely used algorithmic model (random forest) was proposed elsewhere to deal with such distributions of data [51]. Besides, this separation is biologically pertinent since it has been shown that the drivers of the presence might differ from those of the abundance [17, 44, 54]. Lastly, separate modeling of presence and abundance might enable us to identify distinct targets for vector control answering to, respectively, eradication (absence of bites) and control (reduction of the number of bites) [17].

We therefore separately modeled the probability of human–vector contact (called "presence" models in the rest of this article) and the positive counts of human–vector contact (called "abundance" models). Given that HLC data are used as a proxy for human-biting rate, presence models analyzed the probability of at least one individual biting a human during a night, while abundance models analyzed the number of bites received by one human in one night conditional on their presence (i.e. zero-truncated data). Hence, in our presence models, the dependent variable was the presence/absence of vectors (binarized as 1/0) collected during 1512 nights of HLC (27 villages $\times$ 4 collection sites $\times$ 2 places (indoors and outdoors) $\times$ 7 surveys), while in the abundance models, the dependent variable was the number of bites per human during the positive catch sessions—i.e. the sessions with at least one bite.

Bivariate analysis using correlation coefficients

The bivariate relationship between the presence/abundance of each vector species and the environmental variables was assessed using multilevel Spearman correlation coefficients [55] with the village entered as a random effect. Multilevel correlations, contrary to simple correlation, account for non-independency between observations in a dataset, by introducing a factor as a random effect in the correlation (on the same principle as random effects in mixed linear regressions).

Landscape variables: The correlation coefficient was calculated for each landscape variable (i.e. percentage of landscape occupied by each land cover class in each buffer zone).

Meteorological variables: Past weather is likely to influence the size of the sampled mosquito generation with varied delays. For example, (i) past weather in the week preceding the collection may influence adult survival rates of the collected generation, (ii) weather during 1 or 2 weeks preceding the collection may influence the development rates of the collected generation during the larval stages, and (iii) weather beyond the third week preceding the collection date may influence the development rates of parent generations. For the meteorological variables, cross-correlation maps (CCM) were hence computed [56] to assess the relationships between the biting rates and the precipitation and temperatures preceding the dates of collection. A CCM enables one to study the influence of environmental conditions during time intervals (instead of single time points) prior to the collection event. CCMs hence allowed us to account for the effects of cumulative precipitation and average temperature on the collected mosquito generation over intervals of weeks preceding the bites (e.g. average diurnal temperature between 1 and 3 weeks preceding the bite), instead of single weeks (e.g. average diurnal temperature during the third week preceding the bite).

Multivariate analysis using random forests and interpretable machine learning

We used the results of the bivariate analysis to select the environmental variables to include as predictors in the multivariate analysis. We first excluded variables that were poorly correlated with the response variable (i.e. correlation coefficients less than 0.1 or p-values greater than 0.2 at all time associations or buffer radii considered), except for variables related to the presence of water—i.e. possible breeding sites—which were all retained whatever their correlation. Then, for each meteorological (or landscape) variable, we retained the time lag interval (buffer radius) showing the higher absolute correlation coefficient value (see Additional file 6: Feature selection for the multivariate models). Because the entomological data used in this study were part of a trial, different vector control strategies were implemented in the villages of the study after the third survey. The implemented VC strategies and the place of collection (interior/exterior) were therefore introduced as adjustment variables in our models, but their effect on the biting rate was found to be negligible in our analysis (see "Results" section), and these results will not be discussed further.

Random forest classifiers were then trained for each species and response variable (presence and abundance models). Random forests are an ensemble machine-learning method that generates a multitude of random decision trees that are then aggregated to compute a classification or a regression [36]. They are known for their good predictive capacity, which is mainly due to their ability to inherently capture complex associations between the variables [20]. Binary classification RFs were generated for the presence models and regression RFs were generated for the abundance models. The modeling process involved the following steps:

- Feature collinearity: Collinear covariates (i.e. Pearson correlation coefficient > 0.7) were checked for and removed based on empirical knowledge.
- Feature engineering: In the classification models, data were up-sampled within the model resampling procedure to account for the imbalanced structure of the response variable [57]. In the regression models, the response variables were log-transformed prior to the model resampling procedure in order to reduce their overdispersion.
- Model training, tuning, selection: Model hyperparameters were optimized using a random 10-combination grid search [58]. For each set of hyperparameters tested, a leave-village-out cross-validation (LVO-CV) resampling method was used. The resampling method involved training the model using in turn the data from 26 of the 27 sampled villages, validating with the data from the remaining village using a predictive performance metric [for the presence model: the precision–recall area under the curve (PR-AUC); for the abundance model: the mean absolute error (MAE)] and averaging the metric across all hold-out predictions at the end of the procedure. The model retained was the one leading to the highest overall PR-AUC (lowest MAE). The retained model was then fit to all the observations and further used for the interpretation phase.
- Model evaluation: The predictive power of each model was assessed by LVO-CV. We hence evaluated the ability of the models to predict the presence or abundance of vectors on unseen nights of HLC, whilst excluding from the training sets all the observations belonging to the village of the evaluated observation. Doing so enables us to limit overfitting and over-optimistic performance metrics due to spatial autocorrelation [59]. For the presence models, precision–recall plots were then generated from the observed and predicted values, and the PR-AUC was calculated and compared to the baseline of PR curve (i.e. the PR-AUC of a random-guess classifier

for the dataset). The PR-AUC is a measure of predictive accuracy of a binary classification model particularly suitable for imbalanced classification problems [60]. It makes sense when compared to the baseline PR-AUC, which is the rate of "presence" observations in the dataset. For example, a model with a baseline of 0.01 and a PR-AUC of 0.2 performs 0.2/0.01 = 20 times better than a random-guess, or no-skill, classifier. We also calculated sensitivity and specificity at the optimal probability threshold (i.e. the one maximizing the AUC). For the abundance models, a visual evaluation (i.e. graphical comparison between observed and predicted values) was preferred to a numerical one because performance metrics were expected to be low given the overdispersion of the response data and the type of model used [51]. Evaluation plots for the abundance models included (i) the distribution of MAEs and (ii) observed versus predicted values for each out-of-sample village.

To interpret the models, we further generated permutation-based variable importance plots (VIPs) [36] and partial dependence plots (PDPs) [61] including standard deviation bands (that can be interpreted as confidence intervals). These plots, part of the interpretable machine-learning toolbox, enable us to study the effects of one predictor on the response variable while accounting for the effect of the other predictors in the model [25]. Variable importance measures a feature's importance by calculating the degradation of the predictive accuracy of the model after randomly permuting the values of the feature: the higher a variable's importance, the more that variable contributes to the prediction. Partial dependence plots, on their side, show the marginal effect that one feature has on the predicted outcome [25]. PDPs hence help visualize the relationship, learned by a model, between a feature and the response. A PDP is likely to reveal complex (nonlinear, nonmonotonic) effects when a model has learned such relationships. Importantly, the information provided by these tools should be trusted only if the underlying model has good predictive power [27].

We finally identified primary and secondary predictors for each model, according to the following criteria. Primary predictors were the top three most important predictors of the VIP. Secondary predictors were variables either presenting marked variations in their PDP (e.g. thresholds, significant slopes) or known to influence the bio-ecology of the vector.

Software used

The software packages used in this work were all free and open-source. The R programming language [62] and the RStudio environment [63] were used as the main programming tools. An R package was developed [64] to extract the NASA meteorological data (MODIS and GPM). The land cover layer was generated using the following R packages: "RSAGA" [65], "rgrass7" [66], "raster" [67], "sf" [68], "rgdal" [69] and "randomForest" [70]. The "spatstat" [71] package was used to compute the Clark and Evans aggregation index. The QGIS software [72] was used to create the map of the study area. The "landscapemetrics" package [73] was used to calculate the percentage of landscape occupied by each land cover class in the buffer areas. The "correlation" [55] package was used for the correlation analysis. The "caret" [74] and "ranger" [75] packages were used to fit the random forest models in the statistical analysis. The "CAST" [76] package was used to create the temporal folds for cross-validation. The "MLmetrics" [77] package was used to calculate the model evaluation metrics. The "iml" [78] and "pdp" [79] packages were used to generate the partial dependence plots. The "patchwork" [80] package was used to create various plot compositions. The "ggmap" [81] package was used to generate the map of the vector biting rates. The "precrec" [82] package was used to generate the precision–recall plots for the presence models. The "tidyverse" meta-package [83] was used throughout the entire analysis.

Results

Entomological data

A total of 1512 nights of HLC were conducted among the 27 villages during the seven entomological surveys. Altogether 3056 vectors belonging to the *Anopheles* genus were collected: 1322 *An. coluzzii*, 708 *An. funestus*, 616 *An. gambiae s.s.* and 410 from other species. *An. funestus* was present in 12% of the nights of HLC (182 times), while both *An. coluzzii* and *An. gambiae s.s.* appeared on 20% of the nights of HLC (respectively 297 and 302 times). The distribution of the biting rates in the positive sessions (i.e. sessions with at least one bite) was highly left-skewed (for *An. funestus*: median = 2, SD = 4.7, max. = 36; for *An. gambiae s.s.*: median = 2, SD = 1.5, max. = 10; for *An. coluzzii.*: median = 2, SD = 6.8, max. = 50).

Figure 1 shows the distribution of the biting rates of the three main vector species by village and survey. Overall, the map reveals heterogeneous spatiotemporal patterns of biting rates for the three main species. *An. funestus* was found in a few villages only, mainly at the end of the rainy season (November) and in the dry-cold season (December, January) (see Additional file 1: Summary of the meteorological conditions around the sampling point). It almost disappeared during the dry-hot season (March) and the beginning of the rainy season (May). *An. gambiae s.s.* and *An. coluzzii* were found

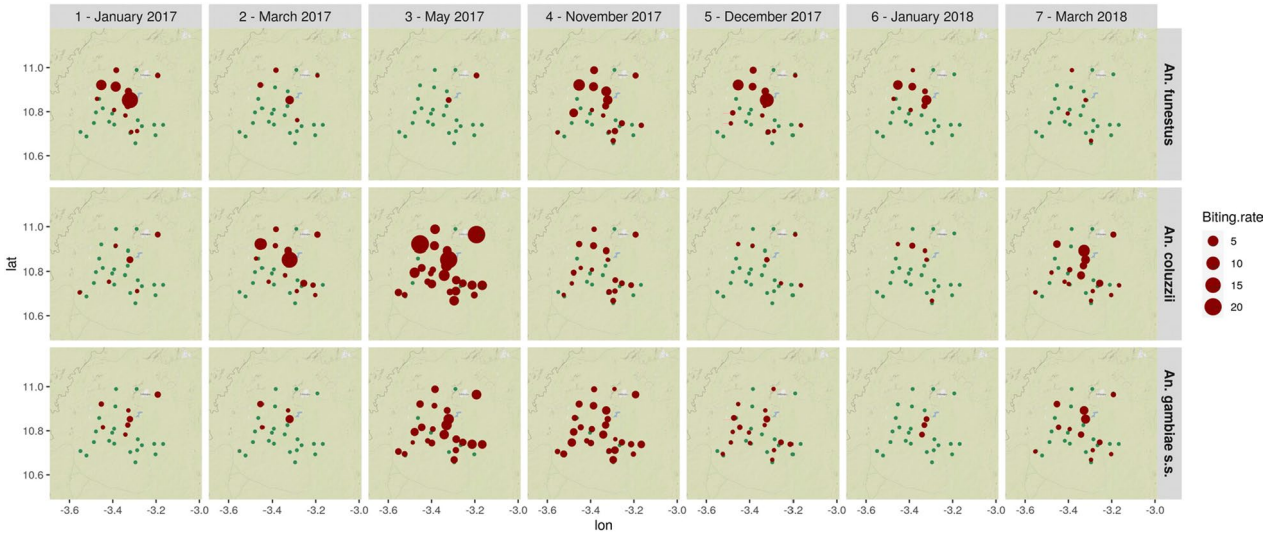


Fig. 1 Map of the biting rates of the three main vector species for each village and entomological survey. Unit: average number of bites/human/night. Blue dots indicate absence of bites in the village for the considered survey. Background layer: OpenStreetMapers

in almost all the villages at the beginning and the end of the rainy season (May, November). They were also found year-round in some villages, in particular, those located close to dams or to the Bougouriba River. *An. coluzzii* was particularly abundant at the beginning of the rainy season (May), while *An. gambiae s.s.* was found in similar abundance at the beginning and the end of the rainy season (May, November).

Land cover map

Eleven land cover classes were discriminated: ligneous savanna (52% of the total surface), crop (25%), grassland (7%), marsh (5%), riparian forest (4%), woodland (3%), rice (1%), settlements (0.5%), bare soil (0.5%), main roads (0.3%) and permanent water bodies (0.3%). In the buffer areas considered for the modeling study (250 m, 500 m, 1 km, 2 km radii), similar trends were observed regarding the percentage of area occupied by each land cover class (see Additional file 2: Summary of the landscape conditions around the sampling points). Ligneous savanna included shrub savanna, tree savanna and wooded savanna. Grassland included herbaceous savanna and Sahelian short grass savanna. Permanent water bodies included dams and the Bougouriba River. Marshlands included wetland–floodplain and agriculture in shallows and recessions. The overall accuracy of the classification was 0.84. The resulting land cover map of the study area, including the geographical position of the study villages, is presented in Fig. 2. Pictures representative of the main land cover classes are provided in Additional file 3: Pictures representative of the main land cover classes in the Diébougou area.

Bivariate analysis

Figure 3 shows the landscape variables that were significantly correlated [multilevel Spearman's correlation coefficient (cc) > 0.1 and p-value < 0.2] with the presence or abundance of each of the studied vector species. The presence or abundance of *An. funestus* was correlated to two to six landscape variables depending to the buffer radius considered, one to five variables for *An. gambiae s.s.* and two to five for *An. coluzzii*. Overall, among the three species, the highest correlation coefficients with the landscape variables were observed for *An. funestus*.

Both the presence and the abundance of *An. funestus* were positively correlated with the % of surface occupied by permanent water bodies in the 2-km radius buffer zone. They were also positively correlated with the % of surface occupied by marshlands in all buffer zones with radius ≥ 500 m, with increasing correlation coefficients as the buffer zone radii increased. The presence and the abundance of *An. funestus* were also positively correlated with the % of surface occupied by grasslands, and negatively correlated with the % of surface occupied by ligneous savannas, for all buffer radii. The correlation between the abundance and the % of surface occupied by grasslands and ligneous savannas increased (both in absolute value and significance) with smaller buffer radii. The presence of *An. funestus* was positively correlated with the % of surface occupied by crops in all buffer zones

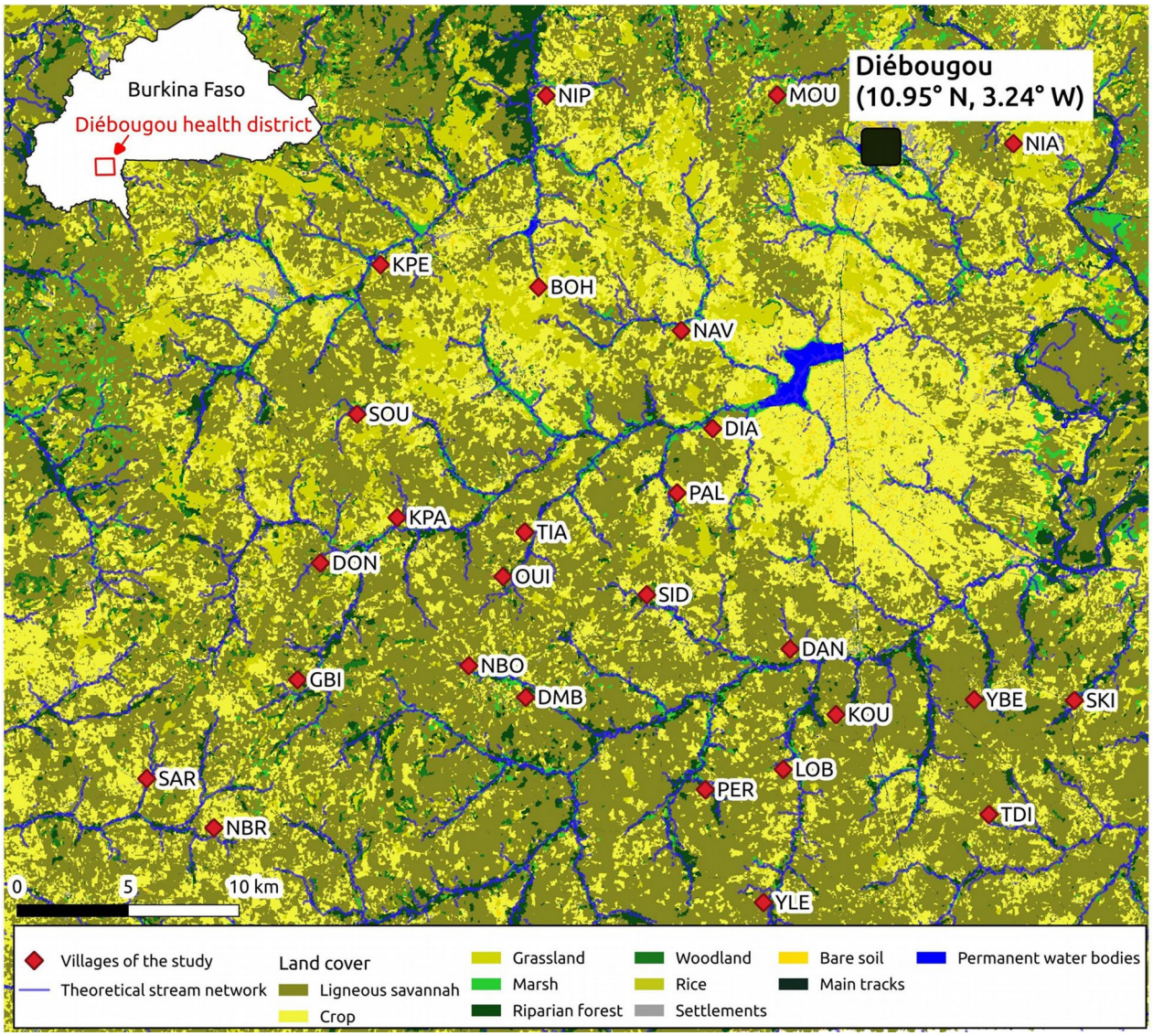


Fig. 2 Map of the study area. The map includes the villages of the study, the land cover derived from geographic object-based image analysis of a SPOT-6 satellite image acquired on 2017-10-11, and the theoretical stream network derived from the SRTM DEM

except for the 1-km radius. The abundance of *An. funestus* was positively correlated with the length of streams in the 250-m and the 2-km radii buffer zones.

Both the presence and the abundance of *An. gambiae s.s.* were positively correlated with the % of surface occupied by permanent water bodies in the 2-km radius buffer zone. Its presence was also positively correlated with the % of surface occupied by marshlands in all buffer zones with radius ≥ 500 m. The presence and abundance of *An. gambiae s.s.* were also positively correlated with the % of surface occupied by grasslands (in all buffer zones for the presence, and in the buffer zones with radius ≥ 1 km for the abundance), and negatively correlated with the % of surface occupied by ligneous savannas (in all buffer zones for the presence, and in the 2-km radius buffer zone for the abundance). The presence and abundance of *An. gambiae s.s.* were also positively correlated with the % of surface occupied by riparian forests, only in the 500-m radius buffer zone for the presence, and in buffer zones with radius ≤ 500 m for the abundance.

The presence of *An. coluzzii* was positively correlated with the % of surface occupied by permanent water bodies in the 2-km radius buffer zone. The presence and the abundance of that species were positively correlated with the % of surface occupied by marshlands in all buffer zones with radius ≥ 1 km for the presence, and in

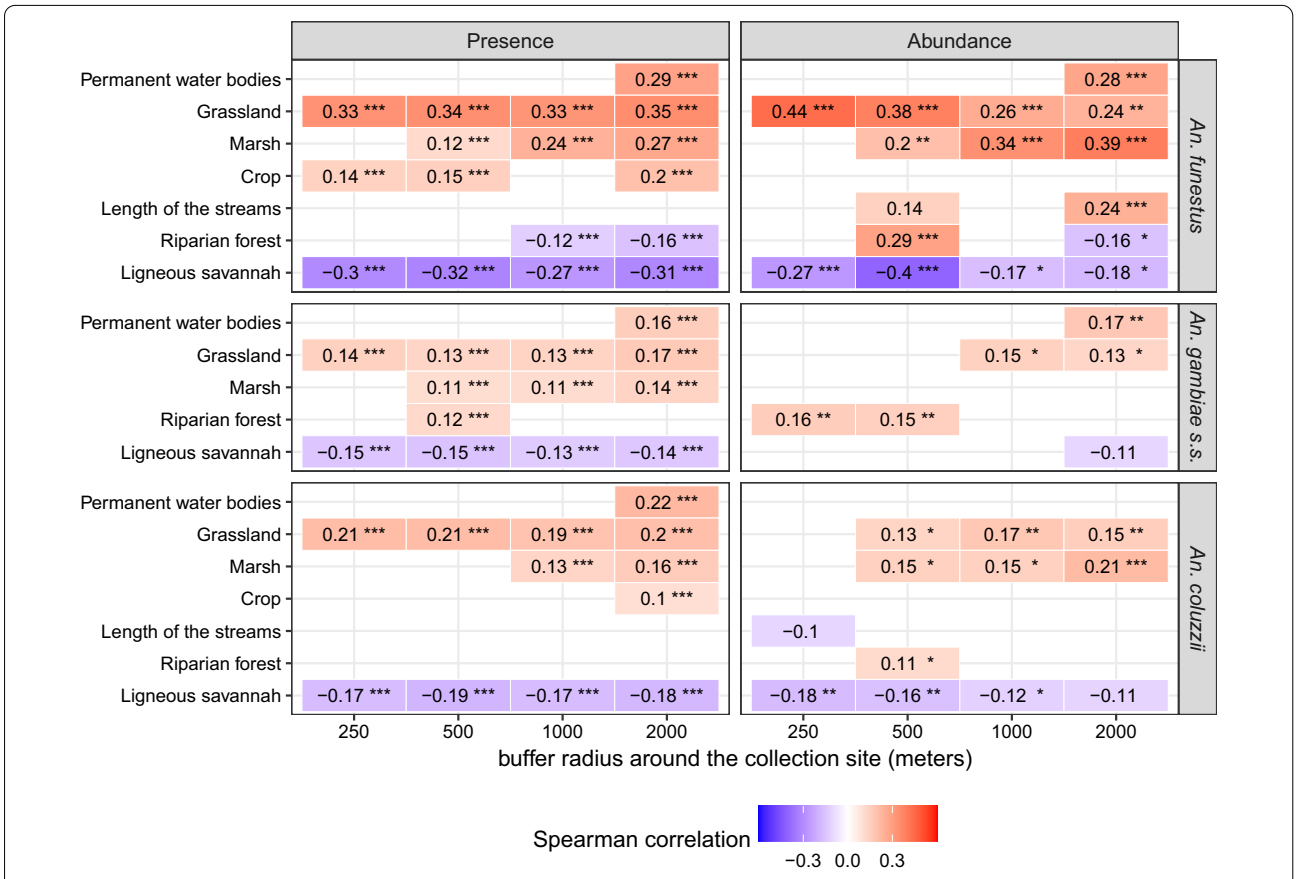


Fig. 3 Multilevel Spearman's correlation between the vectors' biting rates and the landscape variables. Biting rates were separated into presence/absence of bites (left) and abundance of bites (i.e. positive counts only) (right). Unit of biting rates: number of landings on human/person/night. Unit of landscape variables: % of landscape occupied by each land cover class. Landscape variables were extracted in four spatial buffer zones around the sampling locations (250 m radius, 500 m, 1 km, 2 km) for each main vector species. Only correlations with coefficient > 0.1 and *p*-values < 0.2 are displayed. Stars indicate the range of the *p*-value: *** *p*-value ∈ [0, 0.001]; ** *p*-value ∈ [0.001, 0.01]; * *p*-value ∈ [0.01, 0.05]; absence of stars: *p*-value ∈ [0.05, 0.2]

all buffer zones with radius ≥ 500 m for the abundance. Presence and abundance of *An. coluzzii* were also positively correlated with the % of surface occupied by grasslands, and negatively correlated with the % of surface occupied by ligneous savannas, in all the buffer zones (except in the 250-m radius buffer zone for the abundance). The correlation between the abundance of *An. coluzzii* and the % of surface occupied by ligneous savannas increased (in both absolute value and significance) with smaller buffer zones.

Figure 4 shows the meteorological variables that were significantly correlated [multilevel Spearman correlation coefficient (cc) > 0.1 and *p*-value < 0.2] with the presence or abundance of bites for each of the studied vector species (in the form of cross-correlation maps). Overall, among the three species, the highest correlation coefficients with the meteorological variables were observed for *An. coluzzii,* closely followed by *An. gambiae s.s.*

The presence and abundance of *An. funestus* showed quite weak correlations with the meteorological variables (Spearman's correlation coefficient always < 0.25, with all the meteorological variables at all time frames) when compared to *An. gambiae s.s.* or *An. coluzzii*. Correlations between both response variables (presence and abundance) and the three meteorological variables (cumulative rainfall, diurnal LST, nocturnal LST), when significant, were negative. The maximum correlation coefficients between each meteorological variable and both the presence and abundance were found for the following: cumulative rainfall recorded b/w 1–2 and 3 weeks before the date of collection, diurnal temperatures recorded b/w 3–4 and 6 weeks before the date of collection, and nocturnal temperatures recorded b/w 0–1 and 3 weeks before the date of collection.

The presence and abundance of *An. gambiae s.s.* were positively correlated with cumulative rainfall, at all time lags. The maximum correlation coefficients both

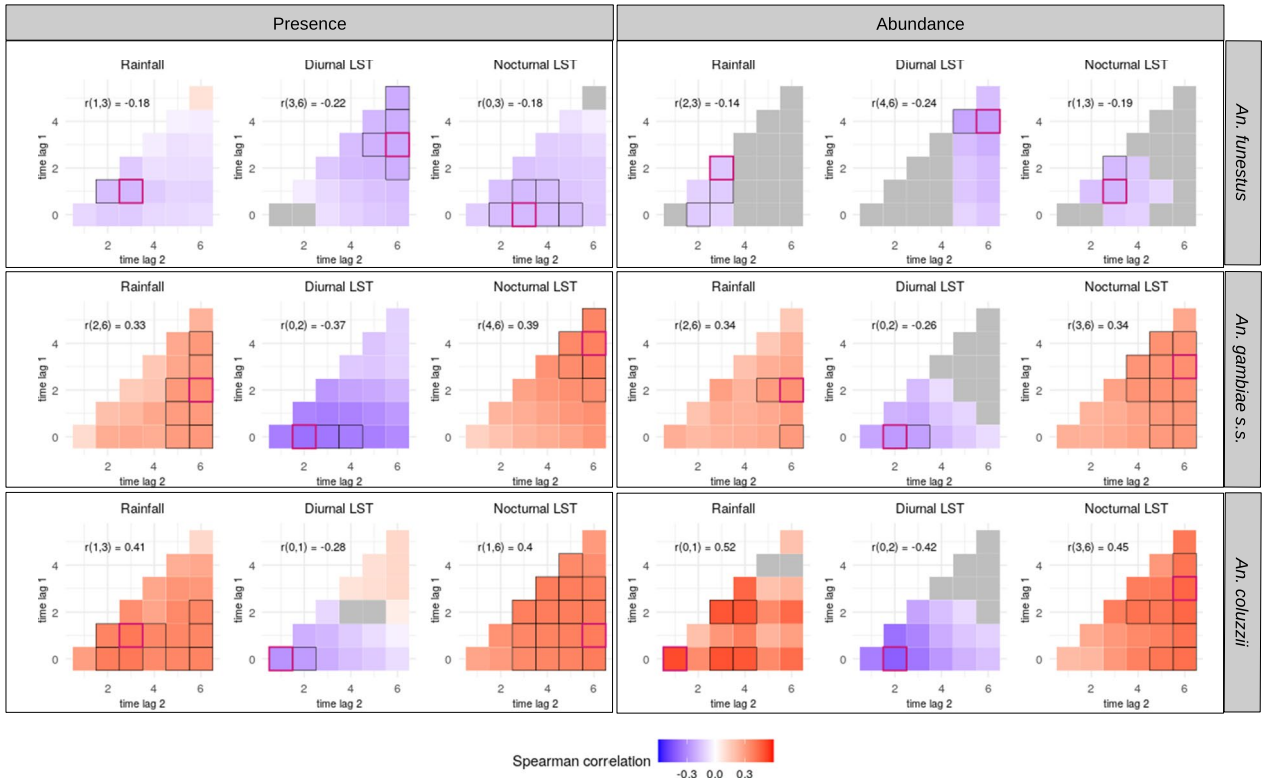


Fig. 4 Multilevel Spearman's correlation between the vectors' biting rates and the meteorological variables (as cross-correlation maps). Biting rates were separated into presence/absence of bites (left) and abundance of bites (i.e. positive counts only) (right). Unit of biting rates: number of landings on human/person/night. Unit of meteorological variables: °C for land surface temperatures (LST), cumulative millimeters for rainfall. Meteorological variables were extracted on a weekly scale up to 6 weeks before the dates of collection for each main vector species. In each CCM, time lags are expressed in week(s) before the date of collection. The red-bordered square indicates the time lag interval that showed the highest correlation coefficient (absolute value) with the meteorological variable (the associated time lag interval and correlation coefficient are reported on the top-left corner of the CCM). The black-bordered squares indicate correlations close to the highest observed correlation (i.e. less than 10% of difference). Gray-filled squares indicate correlations with *p*-value > 0.2 or coefficient > 0.1

presence and abundance were found for cumulative rainfall recorded b/w 2 and 6 weeks before the date of collection. The presence and abundance of *An. gambiae s.s.* were also positively correlated with nocturnal temperatures at all time lags, and the maximum correlation coefficients with both response variables were found for temperatures recorded b/w 3–4 and 6 weeks before the date of collection. The presence and the abundance of *An. gambiae s.s.* were negatively correlated with diurnal temperatures preceding the date of collection at almost all time lags. The maximum correlation coefficient with both response variables was found for temperatures recorded b/w 0 and 2 weeks before the date of collection.

The correlations between meteorological variables and both the presence and abundance of *An. coluzzii* exhibited similar trends as *An. gambiae s.s.*, with few notable differences. The presence and abundance of *An. coluzzii* were positively correlated with cumulative rainfall preceding the date of collection at all time lags. The maximum correlation coefficient with cumulative rainfall was found b/w 1 and 3 weeks before the date of collection for presence, and b/w 0 and 1 week before the date of collection for abundance (the correlation coefficient b/w 1 and 3 weeks was also among the highest for the abundance). The presence and abundance of *An. coluzzii* were positively correlated with nocturnal temperatures at all time lags with maximum correlation coefficients found b/w 1–3 and 6 weeks before the date of collection for both response variables. The presence and abundance of *An. coluzzii* were, overall, negatively correlated with diurnal temperatures preceding the date of collection. The maximum correlation coefficient between diurnal temperatures and both response variables was found b/w 0 and 1–2 weeks before the date of collection.

Multivariate analysis

The PR-AUC of the presence models were 0.56 (baseline = 0.12), 0.46 (baseline = 0.20) and 0.60

(See figure on next page.)

Fig. 5 Interpretation plots of the random forest models for *An. funestus*. Biting rates were separated into presence/absence of bites and abundance of bites (i.e. positive counts only), and two models were therefore generated [presence (top) and abundance (bottom)]. For each model, the top-left corner plot is the variable importance plot. The other plots are partial dependence plots (PDPs) for each variable included in the models (1 plot/variable). The y-axis in the PDPs represents: in the presence models, the probability of at least one individual biting a human during a night; in the abundance models, the log-transformed number of bites received by one human in one night conditional on their presence. The dashed lines represent the partial dependence function ± one standard deviation (i.e. variability estimates). The range of values in the x-axis represents the range of values available in the data for the considered variable. The rugs above the x-axis represent the actual values available in the data for the variable. LST = land surface temperature, b/w = between

(baseline = 0.20) for *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*, respectively. The specificity and sensitivity of the models at the optimal probability thresholds were respectively 80% and 73% for *An. funestus*, 75% and 76% for *An. gambiae s.s.*, and 79% and 75% for *An. coluzzii.* Overall, these results indicate good predictive accuracy of the presence models. The abundance models reflected the trends well for the three species, although they often underestimated high counts. The model evaluation plots are available in Additional file 4: Model evaluation plots for the presence models and Additional file 5: Model evaluation plots for the abundance models (for the presence models: precision–recall plots and observed versus predicted values for each out-of-sample village; for the abundance models: distribution of MAE and observed versus predicted values for each out-of-sample village). Figures 5, 6 and 7 show the model interpretation plots (variable importance plot and partial dependence plots) for *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*, respectively.

The most important predictors of the presence and abundance of *An. funestus* were landscape-based, including % of surface occupied by marshlands (in the 2–km radius buffer zone), grasslands (in the buffer zone radii ≤ 500 m) and ligneous savannas (in the 500-m radius buffer zone). The probability of the presence of *An. funestus* increased linearly with surface occupied by marshlands in the range available in the data (0–10%), while the abundance was constant in the range of 0–3%, increased approximately linearly from 3 to 6%, and finally stabilized in the range of 6–10%. Both the probability of presence and the abundance increased linearly with surface occupied by grasslands in the range of 0–20%, and stabilized above that threshold. Conversely, they decreased linearly with surface occupied by ligneous savannas in the range of 0–50%, and above that threshold stabilized for abundance and tended to diminish (with a lower trend though) for presence.

Secondary predictors of the presence and abundance of *An. funestus* were as follows: % of surface occupied by permanent water bodies in the 2-km radius buffer zone (increase in the range 0–0.1%, stable in the range 0.1–1%), diurnal LST b/w 3 to 4 weeks and 6 weeks before the date of collection (negative, approximately linear, association), and nocturnal LST b/w 0 to 1 and 3 weeks before the date of collection (stable in the range 14–19 °C, decrease in the range 19–25 °C).

The most important predictors of the presence of *An. gambiae s.s.* were meteorological variables, namely nocturnal LST (b/w 4 and 6 weeks, i.e. 28 and 42 days, before the date of collection), rainfall (b/w 2 and 6 weeks before the date of collection), and diurnal LST (b/w 0 and 2 weeks before the date of collection). The probability of presence of *An. gambiae s.s.* increased slowly for nocturnal LSTs in the range of 14–20 °C and more rapidly above that threshold. It increased linearly with the cumulative rainfall in the range of 0–10 mm, and was stable above that threshold. It decreased for diurnal LSTs in the range of 35–41 °C, and stabilized above that threshold. Secondary predictors of the presence of *An. gambiae s.s.* were as follows: % of surface occupied by grasslands in the 2-km radius buffer zone (increase in the range 0–15%, stable above), % of surface occupied by ligneous savannas in the 500-m radius buffer zone (negative linear association), and % of surface occupied by marshlands in the 2-km radius buffer zone (positive linear association).

When *An. gambiae s.s.* was present, cumulative rainfall (b/w 2 and 6 weeks before the date of collection) was, by far, the most important predictor of its abundance. Other primary predictors were diurnal temperatures (b/w 0 and 2 weeks before the date of collection) and % of surface occupied by marshlands (in the 2-km radius buffer zone). The abundance of *An. gambiae s.s.* increased linearly in the range of 0–50 mm cumulative rainfall and stabilized above that threshold (range 50–80 mm). It slowly increased with the % of surface occupied by marshlands. Secondary predictors of the abundance of *An. gambiae s.s.* included the following: % of surface occupied by ligneous savannas in the 2-km radius buffer zone (negative linear association), % of surface occupied by permanent water bodies in the 2-km radius buffer zone (increase in the range 0–0.1%, stable in the range 0.1–1%), % of surface occupied by riparian forests in the 250-m radius buffer zone (positive linear association), and % of surface

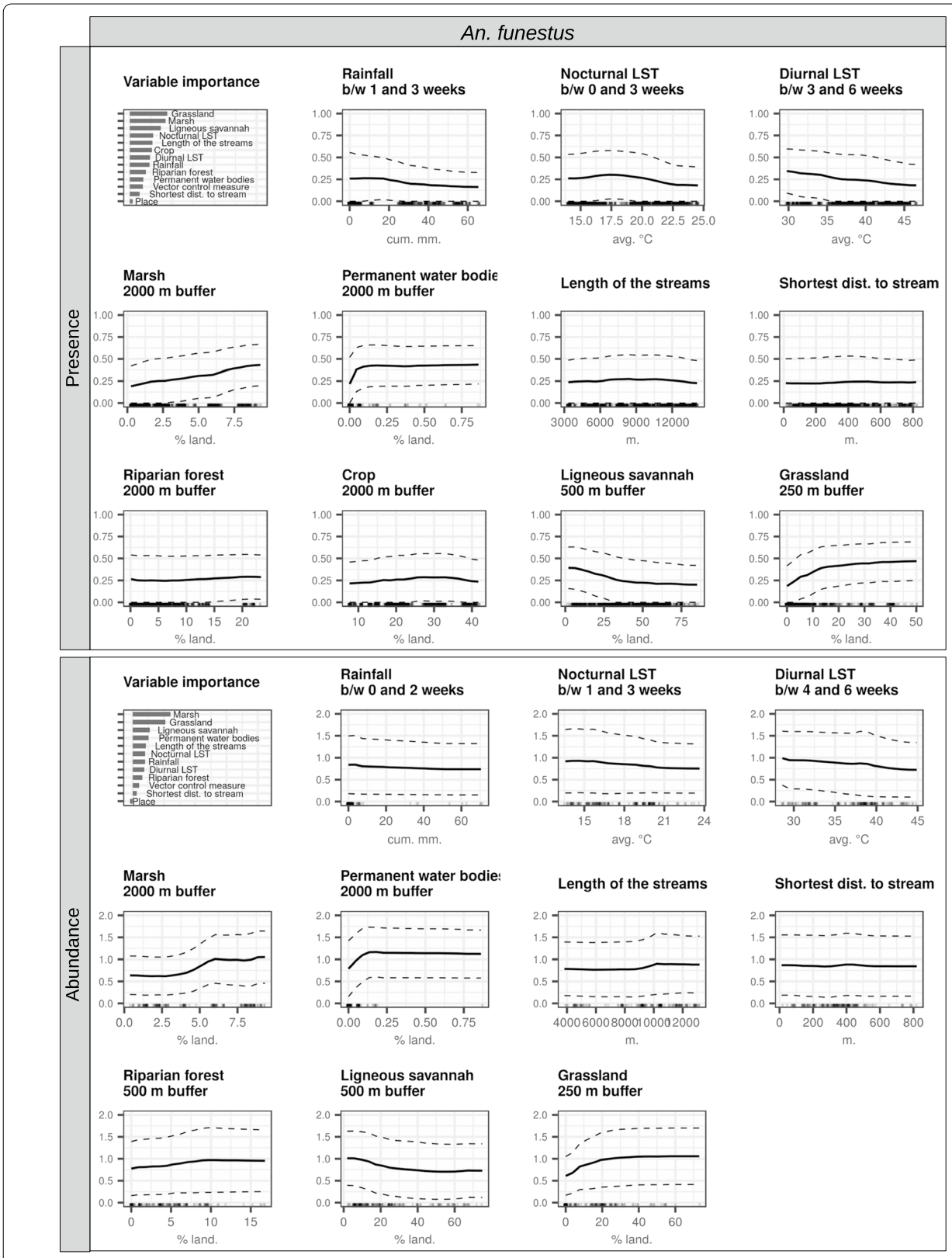


Fig. 5 (See legend on previous page.)

(See figure on next page.)

Fig. 6 Interpretation plots of the random forest models for *An. gambiae s.s.* Biting rates were separated into presence/absence of bites and abundance of bites (i.e. positive counts only), and two models were therefore generated [presence (top) and abundance (bottom)]. For each model, the top-left corner plot is the variable importance plot. The other plots are partial dependence plots (PDPs) for each variable included in the models (1 plot/variable). The y-axis in the PDPs represents: in the presence models, the probability of at least one individual biting a human during a night; in the abundance models, the log-transformed number of bites received by one human in one night conditional on their presence. The dashed lines represent the partial dependence function ± one standard deviation (i.e. variability estimates). The range of values in the x-axis represents the range of values available in the data for the considered variable. The rugs above the x-axis represent the actual values available in the data for the variable. LST = land surface temperature, b/w = between

occupied by grasslands in the 1-km radius buffer zone (positive linear association).

The most important predictors of the presence of *An. coluzzii* were as follows: cumulative rainfall (b/w 1 and 3 weeks before the date of collection), nocturnal LST (b/w 1 and 6 weeks before the date of collection), and % of surface occupied by marshlands (in the 2-km radius buffer zone). A total of approximately 40 mm of rainfall b/w 1 and 3 weeks before the date of collection was enough to double the probability of presence of *An. coluzzii* (from an average 0.25 without rainfall to 0.55). Beyond that amount of rainfall, the probability of presence tended to diminish (range 50–65 mm). The probability of presence of *An. coluzzii* increased linearly with nocturnal LSTs and % of surface occupied by marshlands. Secondary predictors of the presence of *An. coluzzii* included the following: diurnal LST b/w 0 and 1 week before the date of collection (decrease in the range 34 °C–40%, stable in the range 40–50 °C), % of surface occupied by ligneous savannas in the 500-m radius buffer zone (decrease in the range 0–40%, stable above), % of surface occupied by grasslands in the 250-m radius buffer zone (increase in the range 0–30%, stable above), and % of surface occupied by permanent water bodies in the 2-km radius buffer zone (increase in the range 0–0.1%, stable in the range 0.1–1%).

When *An. coluzzii* was present, primary predictors of its abundance were nocturnal LST (b/w 3 and 6 weeks before the date of collection), cumulative rainfall (b/w 0 and 1 week before the date of collection), and % of surface occupied by grasslands (in the 1-km radius buffer zone). The abundance of *An. coluzzii* was constant for nocturnal LSTs under 22 °C and strongly increased above, until 23 °C. The association between abundance and cumulative rainfall was quite weak, but overall positive. The abundance increased linearly with the surface of grasslands in the range of 0–20%, and stabilized above that threshold. Secondary predictors of the abundance of *An. coluzzii* were as follows: % of surface occupied by marshlands in the 2-km radius buffer zone (positive linear association), % of surface occupied by riparian forests in the 500-m radius buffer zone (positive linear association), % of surface occupied by ligneous savannas in the 250-m radius buffer zone (decrease in the range 0–40%, stable above), and distance to the closest stream (decrease in the range 0–100 m, stable above).

Notably, the confidence intervals of the partial dependence functions were overall high for all species and variables and, with a few exceptions, no variable emerged as much more predictive than others (in the VIPs) nor had signals outstandingly strong (in the PDPs).

Discussion

In this modeling study, we linked the biting rates of three major malaria vector species with environmental conditions at vicinities of places and periods of time of biting events to better understand their bio-ecology at fine spatiotemporal scales and identify important factors leading to increased biting risk. First, we correlated the biting rates of the vector species with (i) each meteorological variable at various time lags before the mosquito collection (using cross-correlation maps) and (ii) each landscape variable in various buffer zones around the HLC locations. Then, for selected time lags or spatial radii (the ones with the highest correlation coefficients), we generated multivariate models to study (i) the contribution of each environmental variable in predicting the biting rates and (ii) the nature of the relationship between each environmental variable and the biting rates (all other environmental conditions considered).

In this section, we first discuss the relationships between the biting rates of the malaria vectors and the meteorological and landscape conditions in the Diébougou area, and link them to the bio-ecology of the species. We then discuss how the results of our study could concretely support the conceptualization and deployment of locally tailored VC interventions. Next, we briefly summarize some of the advantages of the modeling method used for knowledge generation in the field of landscape entomology. We conclude the discussion with some limitations of this study and directions for future research.

Effects of meteorological variables

The cross-correlation maps enable us to study how meteorological conditions affect the various stages of the mosquito life cycle [56]. Here, we found that weather

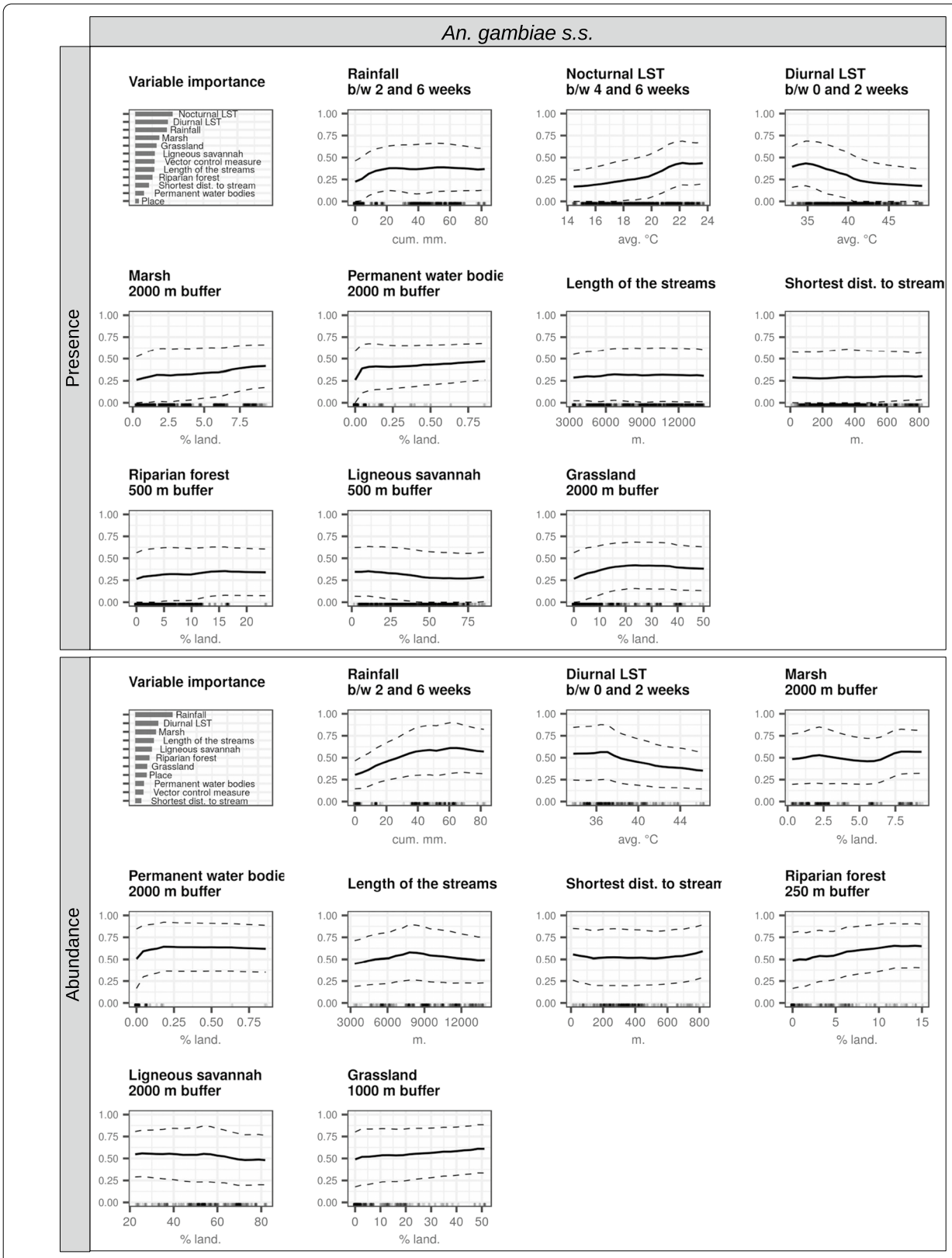


Fig. 6 (See legend on previous page.)

(See figure on next page.)

Fig. 7 Interpretation plots of the random forest models for *An. coluzzii*. Biting rates were separated into presence/absence of bites and abundance of bites (i.e. positive counts only), and two models were therefore generated [presence (top) and abundance (bottom)]. For each model, the top-left corner plot is the variable importance plot. The other plots are partial dependence plots (PDPs) for each variable included in the models (1 plot/variable). The y-axis in the PDPs represents: in the presence models, the probability of at least one individual biting a human during a night; in the abundance models, the log-transformed number of bites received by one human in one night conditional on their presence. The dashed lines represent the partial dependence function ± one standard deviation (i.e. variability estimates). The range of values in the x-axis represents the range of values available in the data for the considered variable. The rugs above the x-axis represent the actual values available in the data for the variable. LST = land surface temperature, b/w = between

conditions (rainfall, nocturnal LST and diurnal LST) were significantly correlated with, and almost always primary predictors of, the presence and abundance of the species of the *Anopheles gambiae* complex. Stronger effects of these meteorological variables were found at various time lags in the studied range (from 0 to 6 weeks before collections). As discussed by Lebl and colleagues [84], weather-dependent life expectancy and development rates make it difficult to link time lags (of weather recordings) influencing mosquito abundance to different development stages. Given the mean life span and larval development duration of the *Anopheles* species collected in our area [49, 85, 86], weather during the first week (i.e. b/w 0 and 1 week) before collection was expected to influence the adult lifetime of collected mosquitoes, and weather during weeks 1–3 (i.e. b/w 1 and 2–3) before collection was expected to influence the larval lifetime of collected mosquitoes. Weather during preceding weeks (i.e. beyond 3 weeks before the date of collection) might affect observed densities by influencing the survival and development rates of (i) parent generations through mechanical effects on the population dynamic [84], (ii) the current/sampled generation through maternal/paternal effects [87, 88], or (iii) the current generation by preparing different biotic and abiotic conditions (for instance, by filling suitable larval development sites with water or by enabling the development of food sources, competitors or predators of *Anopheles* larvae).

In the spatiotemporal frame of our study, nocturnal LST ranged from 14 to 24 °C, and diurnal LST from 33 to 50 °C. Both nocturnal and diurnal LST were important predictors of the presence and abundance of *An. gambiae s.s.* and *An. coluzzii*, often with marked thresholds. Indeed, for *An. gambiae s.s.* we were able to identify a threshold of minimal LST over which the probability started to increase (20 °C), and for both species we also identified a threshold of maximum LST over which the probability reached a minimum (40 °C). For both species, diurnal temperature had the strongest effect during the 2 weeks preceding the dates of collection. This indicates that increasing diurnal temperatures probably reduced adult survival and larval development rates of the sampled generation of mosquitoes, leading to lower observed abundance. Indeed, high temperatures are known to inhibit development of anopheline larvae [89] and to reduce adult survival [90]. Regarding nocturnal temperatures, the time period with the strongest effect on the presence and abundance of both *An. gambiae s.s.* and *An. coluzzii* was between 3 and 6 weeks before collection. This indicates that nocturnal (i.e. minimal) temperatures had their strongest impact by either affecting previous generations (low temperatures are known to reduce adult survival and inhibit larval development [89, 90]) or modifying habitats (with a delay) for the collected generation (for instance, low temperatures may inhibit the development of algae [91], whose biomass has been found to be associated with larval densities [92, 93]).

The high correlation coefficients between cumulative rainfall and both the presence and abundance of *An. gambiae s.s.* and *An. coluzzii*, and the fact that rainfall was systematically an important predictor of these species, might indicate that in our area *An. gambiae s.s.* and *An. coluzzii* are preferably attached to rainfall-dependent breeding sites, confirming the results of other studies [19, 94, 95] and explaining their seasonality. The time period with the strongest effects of rainfall on the presence and abundance of *An. gambiae s.s.* was between 2 and 6 weeks before collection, suggesting an effect on parental generations (as observed by Lebl and colleagues [84] for other mosquito species) or by modifying habitats (abiotic and/or biotic conditions) for the collected generation of these species. Conversely, rainfall had one of its highest correlation coefficients with the presence and abundance of *An. coluzzii* during weeks 1–3 before the dates of collection, indicating that rainfall might have had the greatest influence on the larval stages of the sampled generation of mosquitoes.

Different amounts of rainfall were needed for *An. coluzzii* and *An. gambiae s.s.* to be present or abundant, suggesting different breeding habitat preferences. Minimal rainfall was needed for *An. gambiae s.s.* to increase its probability of being present; additionally, rainfall was by far the most important predictor of its abundance (with a strong positive and approximately linear association). This could indicate that *An. gambiae s.s.* was more attached to breeding sites that quickly appear (presence)

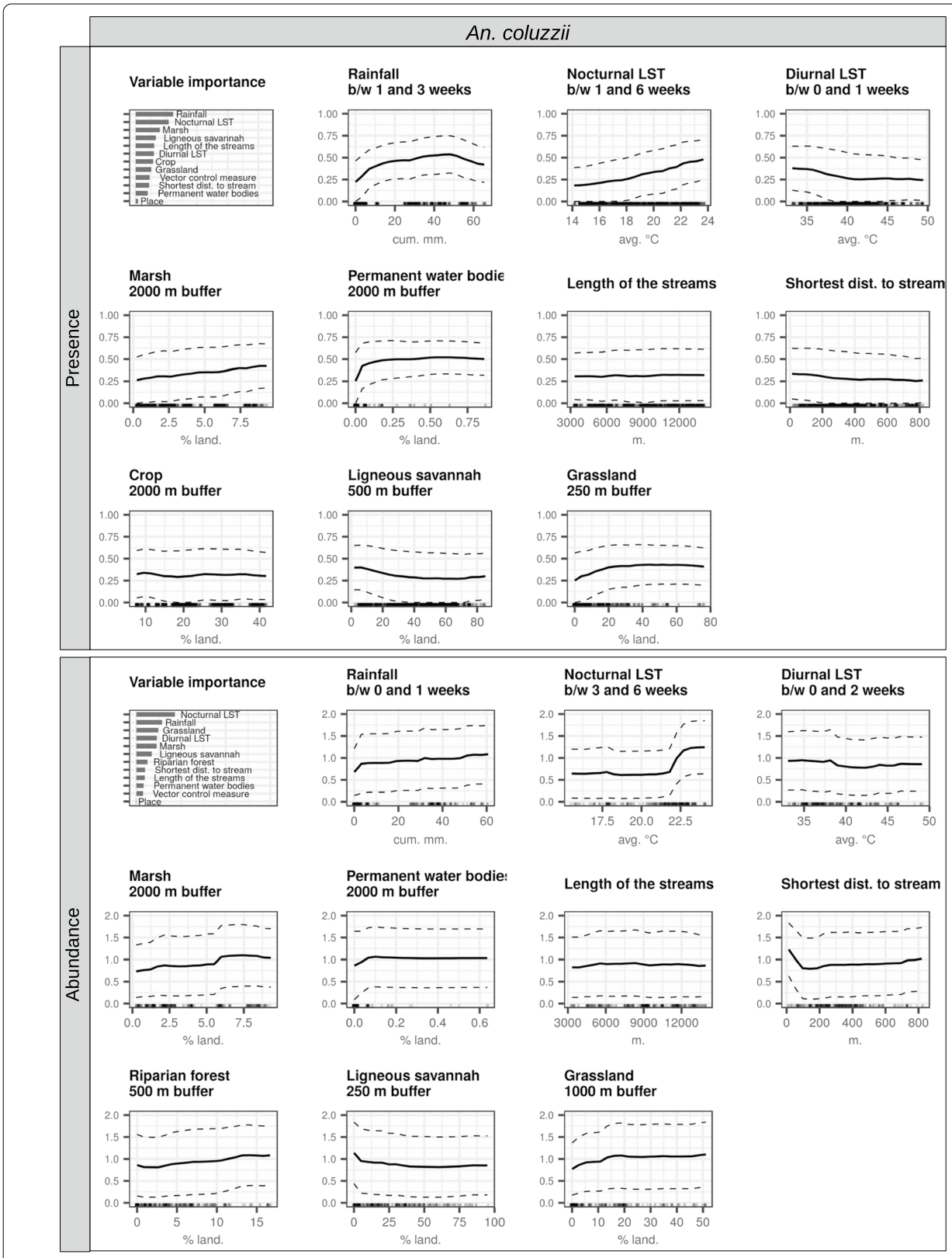


Fig. 7 (See legend on previous page.)

and abound (abundance) when limited rain falls and disappear after it stops, i.e. temporary breeding sites like puddles. *An. coluzzii* needed more rainfall to be present, suggesting preferences for breeding sites that require more water to be flooded, i.e. semipermanent surface water collections like marshlands or streams, which are usually filled in by rainfall throughout the rainy season and shortly after. Indeed, the % of surface occupied by marshlands was significantly correlated with, and the fourth most important predictor of, the abundance of *An. coluzzii*. Overall, these hypotheses about the preferred breeding habitats of *An. gambiae s.s.* and *An. coluzzii* confirm the literature reports [15–19].

Effects of landscape variables

The biting rates of the three species were significantly correlated with several landscape variables, at varying distances from the collection points, and with fluctuating correlation coefficients. In addition, primary predictors of the presence and abundance of *An. funestus* were systematically landscape-based, and some were also primary predictors of the abundance of *An. gambiae s.s.* and *An. coluzzii*. Overall, this indicates that local landscape conditions are important drivers of the bio-ecology of the malaria vectors in rural areas, confirming the literature [5, 12, 17].

The mere presence of permanent water bodies (irrespective of the surface that they occupied) was sufficient to increase (even moderately) the probability of presence and the abundance of the three species. Permanent water bodies, where available, are likely to form breeding habitats for the *Anopheles* species [17, 44, 96–98], and explain why the few study villages located close to the dams and the main river are exposed to year-round bites of high densities of the three species [28]. The % of surface occupied by marshlands at the vicinities of the biting sites was the most important predictor of the presence and abundance models of *An. funestus*. In our study area, marshlands, a semipermanent aquatic environment, hence seemed to be one of the preferred breeding habitats of *An. funestus*, as it has been observed in other places [96, 97]. Notably, the correlation coefficients between the presence/abundance of bites and the % of surface occupied by breeding habitat land cover types (marshlands and permanent water bodies) increased as buffer sizes increased. This might indicate that mosquitoes are able to fly over quite large distances to reach their biting site from these breeding habitats (≥ 2 km), as observed elsewhere in similar landscapes [99, 100]. Proximity to the streams (<100 m) and % of landscape occupied by the riparian forests (≤ 500 m) were secondary predictors of the presence and/or abundance of *An. gambiae s.s.* and *An. coluzzii*. Streams and riparian forests (which are spatially interrelated, i.e. streams flow under riparian forests) might hence form secondary, semipermanent breeding sites for the species of the *Anopheles gambiae* complex in the Diébougou area.

Grasslands—a very "open" landscape—and ligneous savannas—the most "closed" landscape in our study area—were alone or together significantly correlated with, and important predictors of, the presence and/or the abundance of the three malaria vectors studied. Increasing surfaces of grassland areas were associated with increasing probabilities of presence or abundance, while increasing surfaces of savannas were associated with decreasing probabilities of presence or abundance. With some rare exceptions, these landscape indicators were most highly correlated in small-radii (≤ 500 m) buffer areas around the collection sites. Although grassland may provide suitable breeding sites for, at least, *An. gambiae s.s.* and *An. coluzzii* [101], these results seem to indicate that the degree of openness of the landscape some hectometers around villages had a great impact on malaria mosquito biting rates. Our observations are supported by the hypothesis of the lower dispersal of *Anopheles* mosquitoes in closed landscape (in comparison to open landscape) [102] leading to shorter gonotrophic cycle durations and therefore increased biting frequencies and higher biting rates [103]. A similar observation was previously made with *An. coluzzii* in Benin [17]. In the Diebougou area, ligneous savannas seemed to act as natural protective barriers against the malaria vectors and, conversely, grasslands as an aggravating biting risk factor. In a country which is increasingly replacing its closed landscapes (savannas) with opened ones [37], this observation is worrying for malaria transmission. Removal of savannas may significantly increase biting densities. This concern may however be mitigated for our study area, as savannas are usually mainly replaced by crops [37], which themselves did not seem to be an aggravating risk factor (i.e. crops did not emerge as an important variable in our models).

Back to the field: how can these models and knowledge concretely support the fight against malaria transmission at the local scale?

An important question is how these results can ultimately help build locally tailored VC interventions (i.e. deploy the right VC tool at the right time and the right place) to support prevention and reduction of malaria transmission. The knowledge and models generated in this study could support at least three actions: (i) conceptualization of tailored vector control intervention plans, (ii) decisions regarding the places and times where recurrent (long-term) interventions should be deployed, in the form of seasonal maps of predicted biting rates, and

(iii) decisions regarding the places and times where occasional (short-term) interventions should be deployed, in the form of an early warning system.

Support conceptualization of tailored vector control intervention plans

The scientific knowledge confirmed, clarified, or gained through the interpretation of the models could help conceptualize tailored (i.e. species-, time- and place-specific) VC intervention plans. For example, management of temporary breeding sites (through e.g. larval control, or information education, and communication) during the rainy season is likely to be impactful, given the biting densities of *An. coluzzii* and *An. gambiae s.s.* in these seasons and their preferred breeding habitats. Similarly, larval source management in semi-temporary breeding sites (marshlands) would be an interesting option to reduce the presence and abundance of *An. funestus* during the dry-cold season, and if done, should cover quite large buffer zones (at least 2 km) around the households. Beyond these few examples, efficacious and cost-effective VC action plans could be designed on the basis of our characterization of the vectors' local bio-ecology. Importantly, however, several important traits of the vectors (e.g. physiological resistance, behavioral resistance) remain to be characterized in order to design highly efficient action plans.

Support decisions regarding the places and times where recurrent interventions should be deployed through seasonal maps of predicted distribution of biting rates

Once VC interventions have been conceptualized, one must choose the places and times they should be deployed. Here, the multivariate models could be used to generate maps of the predicted distribution of biting rates for each species over the whole study area, at fine spatial resolutions (village or household), and spanning the typical meteorological conditions in the area (for instance, three maps could be generated for each species: one for the dry-cold, one for the dry-hot, and one for the rainy season). These maps could help target and possibly prioritize the places and times for the deployment of recurrent, long-term VC interventions [4, 17, 21].

Support decisions regarding the places and times where occasional interventions should be deployed through an early warning system

A limitation of these maps is that they would only consider "typical," i.e. average, meteorological conditions within a given season. However, different-than-expected events, such as rainfall episodes in the dry season or longer/shorter rainy season, could possibly lead to higher-than-expected biting rates and consequently peaks of infectiousness and transmission of malaria. Our study has shown that meteorological conditions several weeks prior to the mosquito collections can accurately predict future biting rates. To help identify these potential "hot spots" of transmission in a timely manner, an early warning system (EWS) based on these predictive models could be built. Such an EWS, in the form of an automated algorithm, would routinely extract the up-to-date meteorological data and use the models to generate high-resolution maps of short-term forecasts (1 week ahead, 2 weeks ahead, etc.) of the biting rates. The potential hot spots of malaria transmission identified in the maps could then benefit from special interventions that remain to be conceptualized (e.g. increased vector control, special prevention or curation actions).

Methodological bonus: on the use of algorithmic models and interpretable machine learning in landscape entomology

In our study, we have shown how complex statistical models and IML can be used to enhance the fundamental knowledge and understanding of the complex links between the environment and the malaria vectors. Advantages of this modeling workflow over more traditional modeling methods (e.g. linear of logistic regressions) include the ability to (i) inherently capture and unveil complex patterns such as nonlinear or nonmonotonic relationships (e.g. effect of temperature and rainfall) and interactions and (ii) easily include more variables [20] and hence capture small—yet relevant—effects (e.g. effects from riparian forests or distance to streams). Necessary conditions to perform causal interpretation from "black-box" models are to (i) generate a good predictive model and (ii) have some prior domain knowledge about the causal structure of the system under study [27]. Both conditions were met in our work.

Using machine-learning black-box models for scientific discovery, i.e. to generate new knowledge from data, is an emerging trend in many disciplines [26, 104] that has been made possible by the recent development of both IML tools [78] and the know-how to interpret these complex models [25–27, 104, 105]. ML models enable us to integrate knowledge from existing theory in a less formal way than "data" models, and as such can be useful for theory development, provided that a careful linkage to existing knowledge is made [104, 106]. New theoretical insights generated from data and models may then in turn lead to unforeseen experimental research questions. We believe that the fields of landscape epidemiology and entomology still need to fully embrace the potential offered by these methods, in support not only of prediction or forecasting, but also explanation, i.e. to improve

our understanding of the complex processes leading to malaria transmission.

Limitations and directions for future research

An important limitation of our work is linked to the spatiotemporal sampling distribution of mosquito collection. First, no collection was conducted during the high rainy season (July to October) at the known mosquito abundance and malaria transmission peaks. Second, all the collection points were less than 800 m away from the theoretical hydrographic network (which is spatially interrelated with many breeding habitats such as marshlands, streams, riparian forests), meaning that our study could not identify potential differences in the drivers of vector abundance for households further than this distance. Year-round longitudinal collections, including sites further away from permanent or semipermanent breeding habitats, may enable a better understanding of the overall malaria mosquito spatiotemporal dynamics in the area. Meanwhile, these limitations must be accounted for if our models are used to generate predictive maps of the spatiotemporal distribution of vector abundance in the study area.

Similarly, predicting vector abundance outside the study area using the models generated in this study would be of high interest, to extend the operational tools previously mentioned (maps, EWS) to other areas than the Diebougou health district. However, careful attention should be given because of well-known problems linked to predicting beyond the model sampling locations or range of values (e.g. overfitting) [107, 108]. The scalability of our models to places with similar landscape and weather dynamics as the Diebougou area could be tested by collecting similar entomological data in another health district and comparing these ground-truth data with predictions generated by the models. In any case, these models should not be used to predict in remote ecoregions or urban settings, or at higher/lower spatial resolution.

Another limitation is the nature and diversity of the variables introduced in the models. Very fine-scale potential important drivers of mosquito abundance, such as the presence of alternative sources of blood meal (e.g. cattle), of domestic breeding sites, market gardening, or the micro-climatic conditions on the night of collection, have not been investigated. These variables were significant drivers elsewhere in West African rural settings [17, 44, 109]. Yet, the good predictive accuracy of the models suggest that, most probably, the most important drivers of vector abundance in our study area have been identified.

The absence of a strong signal from single variables in the models and the large confidence intervals in the PDPs suggest that the models might have learned important interactions between variables. IML tools such as the *H*-statistic [110] might help reveal such interactions, and others like the two-variable PDP [78] might help explore their effect on malaria vectors abundance. Other tools can be used to analyze individual, or a target set of, predictions made by the models (these tools are called local interpretation methods, e.g. LIME [111] or Shapley values [112]). Local interpretation could be useful, for example, to precisely determine the environmental drivers of vector presence/abundance for a village of interest, or to better identify the drivers of the spatial heterogeneity of biting rates within a season of interest. Altogether, these IML tools might enable us to dig deeper into the models and, hence, the complexity of the ecological niche of malaria vectors.

This work has revealed how landscape can influence the biting rates of vectors, either directly (by impacting vector dispersal) or indirectly (by providing suitable breeding sites and hence increasing vector densities and consequently biting rates). Further investigations on the role played by the level of openness of the landscape are needed to confirm the various hypotheses that we have previously enumerated. A finer-grained land cover classification (e.g. discriminating shrub, tree and wooded savanna) could help test some of our hypotheses: for instance, is there a correlation between the gradient of closedness of the savannas and the abundance of vectors in our study area? Moreover, additional fieldwork could help identify the potential cause–effect relationship between surface of grasslands and malaria vector presence/abundance (i.e. breeding habitat or open landscape favoring dispersal).

Lastly, as stated previously, the scientific knowledge confirmed or acquired in this study and the good predictive accuracy of our models lay the ground for the development of operational tools to support vector control and improve forecasting of epidemic outbreaks at the local scale (e.g. locally tailored VC intervention plans, seasonal maps of the spatiotemporal distribution of vector abundance, EWS).

Conclusion

In this study, several aspects of the bio-ecology of the main malaria vectors in the Diébougou area were explored using field mosquito collections and high-resolution EO data (reflecting both meteorological and landscape local conditions) in a state-of-the-art statistical modeling framework. Overall, the spatiotemporal distributions of biting rates of *An. coluzzii* and *An. gambiae s.s.* were closely associated with meteorological conditions (temperature, precipitation), while those of *An. funestus* were more closely linked to landscape conditions. Meteorological conditions (temperatures, rainfall) putatively

affected all developmental stages of the mosquitoes (larval, adult) at varying levels according to the species and the meteorological variable. Weather occasionally had an even greater impact on time periods preceding the life span of the sampled generation. Primary and possible secondary breeding habitats of each vector species were proposed: *An. funestus*, *An. coluzzii* and *An. gambiae s.s.* seemed to be distributed along a gradient of persistence of the breeding sites, from permanent to temporary, confirming the literature reports. The rate of openness of the landscape seemed to play a major role in the biting rates, which could represent a major concern in a context of progressive shrinkage of the savanna and forest surfaces in Burkina Faso. This work lays the foundation for the development of operational tools to enhance and optimize the fight against malaria transmission at the local scale, such as vector control action plans, seasonal maps of predicted distribution of biting rates or early warning systems for the detection of malaria outbreaks.

Abbreviations

IML: Interpretable machine learning; ML: Machine learning; HLC: Human landing catch; VC: Vector control; RF: Random forest; cc: Correlation coefficient; EO: Earth observation; GEOBIA: Geographic object-based image analysis; SPOT: Satellite Pour l'Observation de la Terre; SRTM: Shuttle Radar Topography Mission; DEM: Digital elevation model; LST: Land surface temperature; CCM: Cross-correlation map; PCR: Polymerase chain reaction; LVO-CV: Leave-village-out cross-validation; PR-AUC: Precision–recall area under the curve; MAE: Mean absolute error; VIP(s): Variable importance plot(s); PDP(s): Partial dependence plot(s); b/w: Between; SD: Standard deviation; LIME: Local interpretable model-agnostic explanations; EWS: Early warning system.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13071-021-04851-x.

Additional file 1: Figure S1. Summary of the meteorological conditions around the sampling points. Average meteorological conditions in a 2 km radius buffer zone around the collection points (weekly aggregation). Vertical red lines indicate the dates of the entomological surveys. Ribbons indicate the mean ± one standard deviation considering all the sampling points for the week.

Additional file 2: Figure S2. Summary of the landscape conditions around the sampling points. Average percentage of surface occupied by each land cover class in the various buffer zones (250 m, 500 m, 1 km, 2 km radii) around the collection points. Error bars indicate the mean ± one standard deviation considering all the sampling points.

Additional file 3: Figure S3. Pictures representative of the main land cover classes in the Diébougou area. Pictures were taken in November 2018.

Additional file 4: Figure S4. Model evaluation plots for the presence models. A1, A2, A3 are precision–recall curves for the presence models of respectively *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*. Precision–recall curves show the precision and the recall of the models for different probability thresholds of the "presence" class. Precision is the proportion of presence identifications that was actually correct, while recall is the proportion of actual presence observations that were identified correctly. The horizontal dashed line represents the baseline (i.e. random or no-skill) classifier. A precision–recall curve above the horizontal line indicates a better-than-no-skill classifier. The higher the area between the precision–recall curve and the horizontal line, the better the classifier. Plots B1, B2, B3 are observed vs. predicted presence probabilities for each out-of-sample village. The y-axis represents the sum over the 8 sampling points/village/survey (4 points by village * 2 places (interior and exterior)). Overall, the plots A1, A2, A3 show that the models had good predictive accuracies (precision–recall curves are higher than the baseline curve, particularly for *An. funestus* and *An. coluzzii*). The plots B1, B2, B3 show that the models predicted well the spatiotemporal trends of presence/absence of bites (lines of predicted presence probabilities are generally close to lines of observed probabilities), although they usually slightly overestimated the probabilities of being bitten (predicted presence probability > observed presence probability).

Additional file 5: Figure S5. Model evaluation plots for the abundance models. A1, A2, A3 are violin plots of the distribution of the residuals for the abundance models of respectively *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*, by observed counts of bites (4 classes: 1 bite, 2–3 bites, 4–10 bites, > 10 bites). Black dots indicate the median value. B1, B2, B3 are observed vs. predicted number of bites/village/entomological surveys. The y-axis represents the sum of bites over the 8 sampling points/village/survey (4 points by village * 2 places (interior and exterior)) on a logarithmic scale. The absence of a dot indicates that no vector was collected. MAE = mean absolute error; *n* = number of observations. Overall, the plots A1, A2, A3 show that the models predicted well small observed counts of bites (1 bite, 2–3 bites) (cf. small MAEs, small residuals), which represent the vast majority of observations (high *n*). Larger counts (4–10 bites, > 10 bites) tended to be underestimated by the models, especially for *An. funestus* and *An. gambiae s.s.* However, large counts (> 10 bites) represented few observations (small *n*). The plots B1, B2, B3 confirm these observations, and additionally show that general trends of biting rates over time were well predicted by the models (lines of predicted abundance are generally close to lines of observed abundance).

Additional file 6: Figure S6. Feature selection for the multivariate models. The figure shows the Spearman correlation coefficient between the explanatory variables and each response variable (presence and abundance of *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*). Based on these results, variables were retained for the multivariate models according to the following criteria: we first excluded variables that were poorly correlated with the response variable (i.e. correlation coefficients less than 0.1 or *p*-values greater than 0.2 at all time associations or buffer radii considered), except for variables related to the presence of water—i.e. possible breeding sites—that were all retained whatever their correlation. Then, for each meteorological (resp. landscape) variable, we retained the time lag interval (resp. buffer radius) showing the higher absolute correlation coefficient value. We finally excluded collinear variables (i.e. Pearson correlation coefficient between the variables > 0.7) based on empirical knowledge.

Acknowledgements

We thank populations of the villages for their kind support and collaboration. We also thank all the field and laboratory staff for their strong commitment to the REACT project. We thank all the anonymous persons that make data-science web forums alive by asking and answering technical questions; these forums were extensively used to elaborate this modeling work.

Map data copyrighted by OpenStreetMap contributors and available from https://www.openstreetmap.org. This work contains modified Copernicus Sentinel data (2018).

Authors' contributions

PT, CP and NM conceived and designed the study. DDS and PT collected and prepared the data. NM, KM and RKD supervised fieldwork. PT analyzed the data, helped by AP and MM. PT and NM drafted the manuscript. KM, CP, DDS, AAK, RKD, MM and AP reviewed the manuscript. All authors read and approved the final manuscript.

Funding

This work was part of the REACT project, funded by the French Initiative 5%—Expertise France (no. 15SANIN213). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the

manuscript. PT was supported by the French Institute of Research for Sustainable Development (IRD) through an international volunteer fellowship and the French National Research Agency (ANR) through the ANORHYTHM project (ANR-16-CE35-008). This work was supported by public funds received in the framework of GEOSUD, a project (ANR-10-EQPX-20) of the program "Investissements d'Avenir" managed by the French National Research Agency.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The protocol of this study was reviewed and approved by the Institutional Ethics Committee of the Institut de Recherche en Sciences de la Santé (IEC-IRSS) and registered as N°A06/2016/CEIRES. Mosquito collectors and supervisors gave their written informed consent. They received a vaccine against yellow fever as a prophylactic measure. Collectors were treated free of charge for malaria according to WHO recommendations.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]MIVEGEC, Université de Montpellier, CNRS, IRD, Montpellier, France. [2]Institut de Recherche en Sciences de La Santé (IRSS), Bobo-Dioulasso, Burkina Faso. [3]Université Nazi Boni, Bobo-Dioulasso, Burkina Faso. [4]Institut Pierre Richet (IPR), Bouaké, Côte d'Ivoire. [5]ESPACE DEV, Université Montpellier, IRD, Université Antilles, Université Guyane, Université Réunion, Montpellier, France.

Received: 14 April 2021 Accepted: 12 June 2021
Published online: 29 June 2021

References

1. WHO. World malaria report 2020: 20 years of global progress and challenges. Licence: CC BY-NC-SA 3.0 IGO; 2020.
2. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. Nature. 2015;526(7572):207–11.
3. Wilson AL, Courtenay O, Kelly-Hope LA, Scott TW, Takken W, Torr SJ, et al. The importance of vector control for the control and elimination of vector-borne diseases. PLoS Negl Trop Dis. 2020;14(1):e0007831.
4. WHO. Global vector control response 2017–2030. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO; 2017.
5. Stresman GH. Beyond temperature and precipitation: Ecological risk factors that modify malaria transmission. Acta Trop. 2010;116(3):167–72.
6. Ferguson HM, Dornhaus A, Beeche A, Borgemeister C, Gottlieb M, Mulla MS, et al. Ecology: a prerequisite for malaria elimination and eradication. PLoS Med. 2010;7(8):e1000303.
7. Beck-Johnson LM, Nelson WA, Paaijmans KP, Read AF, Thomas MB, Bjørnstad ON. The effect of temperature on *Anopheles* mosquito population dynamics and the potential for malaria transmission. PLoS ONE. 2013;8(11):e79276.
8. Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, Ben-Horin T, de Moor E, et al. Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecol Lett. 2013;16(1):22–30.
9. Shapiro LLM, Whitehead SA, Thomas MB. Quantifying the effects of temperature on mosquito and parasite traits that determine the transmission potential of human malaria. PLoS Biol. 2017;15(10):e2003489.
10. Shaman J, Stieglitz M, Stark C, Le Blancq S, Cane M. Using a dynamic hydrology model to predict mosquito abundances in flood and swamp water. Emerg Infect Dis. 2002;8(1):6–13.
11. Paaijmans KP, Takken W, Githeko AK, Jacobs AFG. The effect of water turbidity on the near-surface water temperature of larval habitats of the malaria mosquito *Anopheles gambiae*. Int J Biometeorol. 2008;52(8):747–53.
12. Fornace KM, Diaz AV, Lines J, Drakeley CJ. Achieving global malaria eradication in changing landscapes. Malar J. 2021;20(1):69.
13. Hamon Jacques, Mouchet Jean. (1961). Les vecteurs secondaires du paludisme humain en Afrique. In : Etudes sur le paludisme en Afrique. *Médecine Tropicale*, 21 (No spécial), p. 643–60. ISSN 0025-682X
14. Delmont J. Paludisme et variations climatiques saisonnières en savane soudanienne d'Afrique de l'Ouest. Cah DÉtudes Afr. 1982;22(85):117–33.
15. Simard F, Ayala D, Kamdem G, Pombi M, Etouna J, Ose K, et al. Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation. BMC Ecol. 2009;9(1):17.
16. Costantini C, Ayala D, Guelbeogo WM, Pombi M, Some CY, Bassole IH, et al. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*. BMC Ecol. 2009;9(1):16.
17. Moiroux N, Djènontin A, Bio-Bangana AS, Chandre F, Corbel V, Guis H. Spatio-temporal analysis of abundances of three malaria vector species in southern Benin using zero-truncated models. Parasit Vectors. 2014;7(1):103.
18. Diabaté A, Dabiré RK, Heidenberger K, Crawford J, Lamp WO, Culler LE, et al. Evidence for divergent selection between the molecular forms of *Anopheles gambiae*: role of predation. BMC Evol Biol. 2008;8(1):5.
19. Diabaté A, Dabire RK, Kim EH, Dalton R, Millogo N, Baldet T, et al. Larval development of the molecular forms of *Anopheles gambiae (Diptera: Culicidae)* in different habitats: a transplantation experiment. J Med Entomol. 2005;42(4):548–53.
20. Breiman L. Statistical modeling: the two cultures. Stat Sci. 2001;16(3):199–215.
21. Machault V, Vignolles C, Borchi F, Vounatsou P, Pages F, Briolant S, et al. The use of remotely sensed environmental data in the study of malaria. Geospat Health. 2011;5:151–68.
22. Ebhuoma O, Gebreslasie M. Remote sensing-driven climatic/environmental variables for modelling malaria transmission in sub-Saharan Africa. Int J Environ Res Public Health. 2016;13(6):584.
23. Parselia E, Kontoes C, Tsouni A, Hadjichristodoulou C, Kioutsioukis I, Magiorkinis G, et al. Satellite earth observation data in epidemiological modeling of malaria, dengue and west Nile virus: a scoping review. Remote Sens. 2019;11(16):1862.
24. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AIMag. 1996;17(3):37.
25. Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
26. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci. 2019;116(44):22071–80.
27. Zhao Q, Hastie T. Causal interpretations of black-box models. J Bus Econ Stat. 2021;39(1):272–81.
28. Soma DD, Zogo BM, Somé A, Tchiekoi BN, de Hien DFS, Pooda HS, et al. *Anopheles* bionomics, insecticide resistance and malaria transmission in southwest Burkina Faso: a pre-intervention study. PLoS ONE. 2020;15(8):e0236920.
29. Gillies MT, B. De Meillon. The Anophelinae of Africa south of the Sahara (Ethiopian Zoogeographical Region). Publications of the South African Institute for Medical Research. 1968;54.
30. Gillies MT, Coetzee M. A supplement to the Anophelinae of Africa South of the Sahara. Publ Afr Inst Med Res. 1987;55:1–143.
31. Koekemoer LL, Kamau L, Hunt RH, Coetzee M. A cocktail polymerase chain reaction assay to identify members of the *Anopheles funestus (Diptera: Culicidae)* group. Am J Trop Med Hyg. 2002;66(6):804–11.
32. Cohuet A, Simard F, Berthomieu A, Raymond M, Fontenille D, Weill M. Isolation and characterization of microsatellite DNA markers in the malaria vector *Anopheles funestus*. Mol Ecol Notes. 2002;2(4):498–500.
33. Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, della Torre A. . Insertion polymorphisms of SINE200 retrotransposons within speciation islands of *Anopheles gambiae* molecular forms. Malar J. 2008;7(1):1–10.
34. Hay GJ, Castilla G. Geographic Object-Based Image Analysis (GEOBIA): A new name for a new discipline. In: Blaschke T, Lang S, Hay GJ, editors. Object-based image analysis. Lecture notes in geoinformation

and cartography. Berlin: Springer; 2008. https://doi.org/10.1007/978-3-540-77058-9_4.
35. NASA JPL. NASA shuttle radar topography mission global 1 arc second. NASA EOSDIS land processes DAAC; 2013. https://lpdaac.usgs.gov/products/srtmgl1v003/. Accessed 12 Apr 2021
36. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
37. CILSS, 2016. Landscapes of West Africa—A window on a changing world: Ouagadougou, Burkina Faso, CILSS, 219 p. (Comité Permanent Inter-états de Lutte contre la Sécheresse dans le Sahel)
38. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
39. Takken W, Charlwood JD, Billingsley PF, Gort G. Dispersal and survival of *Anopheles funestus* and *A. gambiae* s.l. (Diptera: Culicidae) during the rainy season in southeast Tanzania. Bull Entomol Res. 1998;88(5):561–6.
40. Service MW. Mosquito (Diptera: Culicidae) dispersal—the long and short of it. J Med Entomol. 1997;34(6):579–88.
41. Clarke SE, Bøgh C, Brown RC, Walraven GEL, Thomas CJ, Lindsay SW. Risk of malaria attacks in Gambian children is greater away from malaria vector breeding sites. Trans R Soc Trop Med Hyg. 2002;96(5):499–506.
42. Jenson SK, Domingue JO. Extracting topographic structure from digital elevation data for geographic information-system analysis. Photogramm Eng Remote Sens. 1988;54(11):1593–600.
43. Clark PJ, Evans FC. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology. 1954;35(4):445–53.
44. Moiroux N, Bio-Bangana AS, Djènontin A, Chandre F, Corbel V, Guis H. Modelling the risk of being bitten by malaria vectors in a vector control area in southern Benin, west Africa. Parasit Vectors. 2013;6(1):71.
45. Debebe Y, Hill SR, Tekie H, Dugassa S, Hopkins RJ, Ignell R. Malaria hotspots explained from the perspective of ecological theory underlying insect foraging. Sci Rep. 2020;10(1):21449.
46. NASA Goddard Earth Sciences Data And Information Services Center. GPM IMERG final precipitation L3 1 day 0.1 degree x 0.1 degree V06. NASA Goddard Earth Sciences Data and Information Services Center; 2019. https://disc.gsfc.nasa.gov/datacollection/GPM_3IMERGDF_06.html. Accessed 11 Feb 2021.
47. Wan, Zhengming, Hook, Simon, Hulley, Glynn. MOD11A1 MODIS/terra land surface temperature/emissivity daily L3 global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC; 2015. https://lpdaac.usgs.gov/products/mod11a1v006/. Accessed 11 Feb 2021.
48. Wan, Zhengming, Hook, Simon, Hulley, Glynn. MYD11A1 MODIS/aqua land surface temperature/emissivity daily L3 global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC; 2015. https://lpdaac.usgs.gov/products/myd11a1v006/. Accessed 11 Feb 2021.
49. Holstein M. Biologie d'*Anopheles gambiae* : recherches en Afrique-Occidentale Française. Genève: OMS; 1952. (Monographies - OMS). http://www.documentation.ird.fr/hor/fdi:42581. Accessed 2 Dec 2020.
50. Lee S-K, Jin S. Decision tree approaches for zero-inflated count data. J Appl Stat. 2006;33(8):853–65.
51. Mathlouthi W, Larocque D, Fredette M. Random forests for homogeneous and non-homogeneous Poisson processes with excess zeros. Stat Methods Med Res. 2020;29(8):2217–37.
52. Boussari O, Moiroux N, Iwaz J, Djènontin A, Bio-Bangana S, Corbel V, et al. Use of a mixture statistical model in studying malaria vectors density. PLoS ONE. 2012;7(11):e50452.
53. Cragg JG. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica. 1971;39(5):829.
54. Gadiaga L, Machault V, Pagès F, Gaye A, Jarjaval F, Godefroy L, et al. Conditions of malaria transmission in Dakar from 2007 to 2010. Malar J. 2011;10(1):312.
55. Makowski D, Ben-Shachar MS, Patil I, Lüdecke D. Methods for correlation analysis. CRAN; 2020. https://github.com/easystats/correlation.
56. Curriero FC, Shone SM, Glass GE. Cross correlation maps: a tool for visualizing and modeling time lagged associations. Vector Borne Zoonotic Dis Larchmt N. 2005;5(3):267–75.
57. Tyagi S, Mittal S. Sampling approaches for imbalanced data classification problem in machine learning. In: Singh P, Kar A, Singh Y, Kolekar M, Tanwar S, editors. Proceedings of ICRIC 2019. Lecture notes in electrical engineering, vol. 597. Cham: Springer; 2020. https://doi.org/10.1007/978-3-030-29407-6_17.
58. Ten CD. quick tips for machine learning in computational biology. BioData Mining. 2017;10(1):35.
59. Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ Model Softw. 2018;101:1–9.
60. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.
61. Friedman JH. Machine. Ann Stat. 2001;29(5):1189–232.
62. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018. https://www.R-project.org/. Accessed 15 Oct 2020.
63. RStudio Team. RStudio: integrated development environment for R. Boston: RStudio, PBC; 2020. http://www.rstudio.com/. Accessed 15 Oct 2020.
64. opendapr. Fast download of many earth observation data in R using the OPeNDAP Capacities. https://github.com/ptaconet/opendapr. Accessed 01 Apr 2021
65. Brenning A, Bangs D, Becker M. RSAGA: SAGA geoprocessing and terrain analysis; 2018. https://CRAN.R-project.org/package=RSAGA. Accessed 15 Oct 2020.
66. Bivand R. rgrass7: interface between GRASS 7 geographical information system and R; 2018. https://CRAN.R-project.org/package=rgrass7. Accessed 15 Oct 2020.
67. Hijmans RJ. raster: geographic data analysis and modeling; 2020. https://CRAN.R-project.org/package=raster. Accessed 15 Oct 2020.
68. Pebesma E. Simple features for R: standardized support for spatial vector data. R J. 2018;10(1):439–46.
69. Bivand R, Keitt T, Rowlingson B. rgdal: bindings for the "Geospatial" data abstraction library; 2019. https://CRAN.R-project.org/package=rgdal. Accessed 15 Oct 2020.
70. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.
71. Baddeley A, Rubak E, Turner R. Spatial point patterns: methodology and applications with R. London: Chapman and Hall/CRC Press; 2015. http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/. Accessed 15 Oct 2020.
72. QGIS Development Team. QGIS geographic information system. QGIS Association; 2021. https://www.qgis.org. Accessed 15 Oct 2020.
73. Hesselbarth MHK, Sciaini M, With KA, Wiegand K, Nowosad J. landscapemetrics: an open-source R tool to calculate landscape metrics. Ecography. 2019;42:1–10.
74. Wing MKC from J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. caret: classification and regression training; 2018. https://CRAN.R-project.org/package=caret. Accessed 15 Oct 2020.
75. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. 2017;77(1):1–17.
76. Meyer H. CAST: "caret" applications for spatial-temporal models; 2020. https://CRAN.R-project.org/package=CAST. Accessed 15 Oct 2020.
77. Yan Y. MLmetrics: machine learning evaluation metrics; 2016. https://CRAN.R-project.org/package=MLmetrics. Accessed 15 Oct 2020.
78. Molnar C, Bischl B, Casalicchio G. iml: an R package for interpretable machine learning. JOSS. 2018;3(26):786.
79. Greenwell BM. pdp: an R package for constructing partial dependence plots. R J. 2017;9(1):421–36.
80. Pedersen TL. patchwork: the composer of plots; 2019. https://CRAN.R-project.org/package=patchwork. Accessed 15 Oct 2020.
81. Kahle D, Wickham H. ggmap: spatial visualization with ggplot2. R J. 2013;5(1):144–61.
82. Saito T, Rehmsmeier M. Precrec: fast and accurate precision-recall and ROC curve calculations in R. Bioinformatics. 2017;33(1):145–7.
83. Wickham H. tidyverse: Easily Install and Load the "Tidyverse"; 2017. https://CRAN.R-project.org/package=tidyverse. Accessed 15 Oct 2020.
84. Lebl K, Brugger K, Rubel F. Predicting *Culex pipiens/restuans* population dynamics by interval lagged weather data. Parasit Vectors. 2013;6(1):129.
85. Townson H. The biology of mosquitoes. Volume 1. Development, nutrition and reproduction. By A.N. Clements. (London: Chapman

& Hall, 1992).viii 509 pp. ISBN 0-412-40180-0. Bull Entomol Res. 1993;83(2):307–308.
86. Carnevale P, Robert V, Manguin S, Corbel V, Fontenille D, Garros C, et al. Les anophèles : biologie, transmission du Plasmodium et lutte antivectorielle. IRD; 2009. 391 p. (Didactiques). http://www.documentation.ird.fr/hor/fdi:010047862
87. Zirbel K, Eastmond B, Alto BW. Parental and offspring larval diets interact to influence life-history traits and infection with dengue virus in *Aedes aegypti*. R Soc Open Sci. 2018;5(7):180539.
88. Zirbel KE, Alto BW. Maternal and paternal nutrition in a mosquito influences offspring life histories but not infection with an arbovirus. Ecosphere. 2018;9(10):e02469.
89. Lyons CL, Coetzee M, Chown SL. Stable and fluctuating temperature effects on the development rate and survival of two malaria vectors, *Anopheles arabiensis* and *Anopheles funestus*. Parasit Vectors. 2013;6(1):104.
90. Lyons CL, Coetzee M, Terblanche JS, Chown SL. Desiccation tolerance as a function of age, sex, humidity and temperature in adults of the African malaria vectors *Anopheles arabiensis* and *Anopheles funestus*. J Exp Biol. 2014;217(21):3823–33.
91. Raven JA, Geider RJ. Temperature and algal growth. New Phytol. 1988;110(4):441–61.
92. Kweka EJ, Zhou G, Munga S, Lee M-C, Atieli HE, Nyindo M, et al. Anopheline larval habitats seasonality and species distribution: a prerequisite for effective targeted larval habitats control programmes. PLoS ONE. 2012;7(12):e52084.
93. Kaufman MG, Wanja E, Maknojia S, Bayoh MN, Vulule JM, Walker ED. Importance of algal biomass to growth and development of *Anopheles gambiae* larvae. J Med Entomol. 2006;43(4):669–76.
94. Gimonneau G, Pombi M, Choisy M, Morand S, Dabiré RK, Simard F. Larval habitat segregation between the molecular forms of the mosquito *Anopheles gambiae* in a rice field area of Burkina Faso, West Africa. Med Vet Entomol. 2012;26(1):9–17.
95. Gimnig JE, Ombok M, Kamau L, Hawley WA. Characteristics of larval anopheline *(Diptera: Culicidae)* habitats in Western Kenya. J Med Entomol. 2001;38(2):282–8.
96. Pages F, Orlandipradines E, Corbel V. Vecteurs du paludisme: biologie, diversité, contrôle et protection individuelle. Médecine Mal Infect. 2007;37(3):153–61.
97. Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J, et al. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. Parasit Vectors. 2010;3(1):117.
98. Zogo B, Koffi AA, Alou LPA, Fournet F, Dahounto A, Dabiré RK, et al. Identification and characterization of *Anopheles* spp. breeding habitats in the Korhogo area in northern Côte d'Ivoire: a study prior to a Bti-based larviciding intervention. Parasit Vectors. 2019;12(1):146.
99. Verdonschot PFM, Besse-Lototskaya AA. Flight distance of mosquitoes *(Culicidae)*: A metadata analysis to support the management of barrier zones around rewetted and newly constructed wetlands. Limnologica. 2014;45:69–79.
100. Thomas CJ, Cross DE, Bøgh C. Landscape movements of *Anopheles gambiae* Malaria vector mosquitoes in rural Gambia. PLoS ONE. 2013;8(7):e68679.
101. Fillinger U, Majambere S, Lindsay SW, Green C, Sayer DR. Spatial distribution of mosquito larvae and the potential for targeted larval control in the Gambia. Am J Trop Med Hyg. 2008;79(1):19–27.
102. Le Goff G, Carneval P, Robert V. Low dispersion of anopheline malaria vectors in the African equatorial forest. Parasite. 1997;4(2):187–9.
103. Afrane YA, Lawson BW, Githeko AK, Yan G. Effects of microclimatic changes caused by land use and land cover on duration of gonotrophic cycles of *Anopheles gambiae (Diptera: Culicidae)* in western Kenya highlands. J Med Entomol. 2005;42(6):974–80.
104. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, et al. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans Knowl Data Eng. 2017;29(10):2318–31.
105. Shmueli G. To explain or to predict? Stat Sci. 2010;25(3):289–310.
106. Shmueli G, Koppius O. Predictive Analytics in Information Systems Research. SSRN Electron J. 2010. http://www.ssrn.com/abstract=1606674. Accessed 18 Dec 2020.
107. Wardrop NA, Geary M, Osborne PE, Atkinson PM. Interpreting predictive maps of disease: highlighting the pitfalls of distribution models in epidemiology. Geospat Health. 2014;9:237–46.
108. Meyer H, Pebesma E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in ecology and evolution; 2021. https://doi.org/10.1111/2041-210X.13650. Accessed 11 Jun 2021.
109. Ngowo HS, Kaindoa EW, Matthiopoulos J, Ferguson HM, Okumu FO. Variations in household microclimate affect outdoor-biting behaviour of malaria vectors. Wellcome Open Res. 2017;2:102.
110. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. 2008;2(3):916–54.
111. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco California USA: ACM; 2016. p. 1135–44. https://doi.org/10.1145/2939672.2939778. Accessed 12 Apr 2021.
112. Strumbelj KI. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2013;41:647–65.

Publisher's Note

4.2.1 Figures additionnelles

Voir la version en ligne de l'article pour les figures tailles réelles (https://doi.org/10.1186/s13071-021-04851-x)

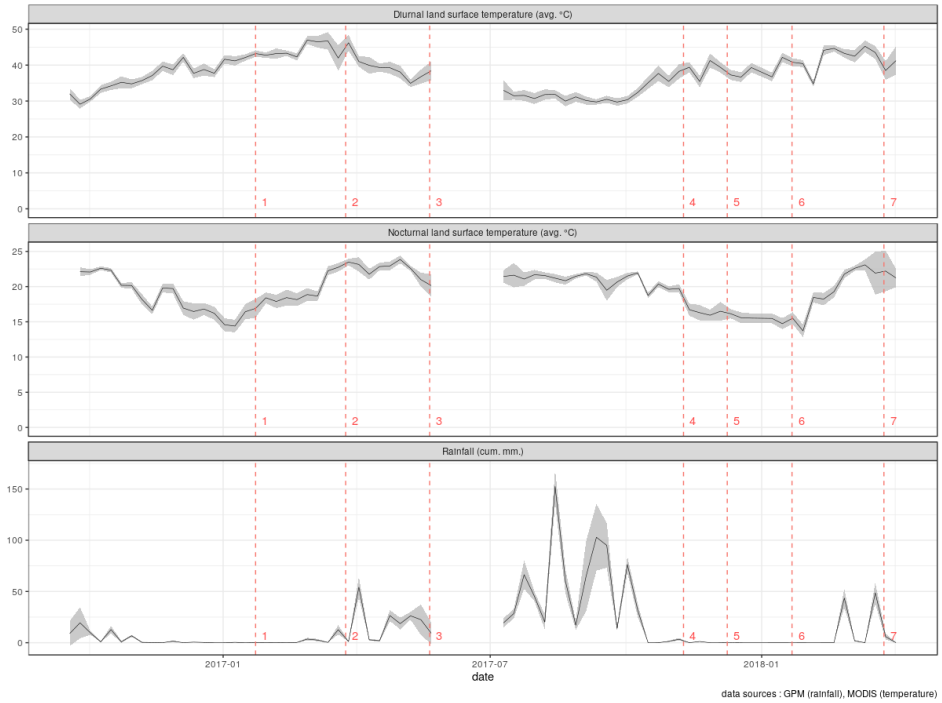


FIGURE 4.1: Figure additionnelle n°1 : Summary of the meteorological conditions around the sampling points. Average meteorological conditions in a 2 km radius buffer zone around the collection points (weekly aggregation). Vertical red lines indicate the dates of the entomological surveys. Ribbons indicate the mean ± one standard deviation considering all the sampling points.

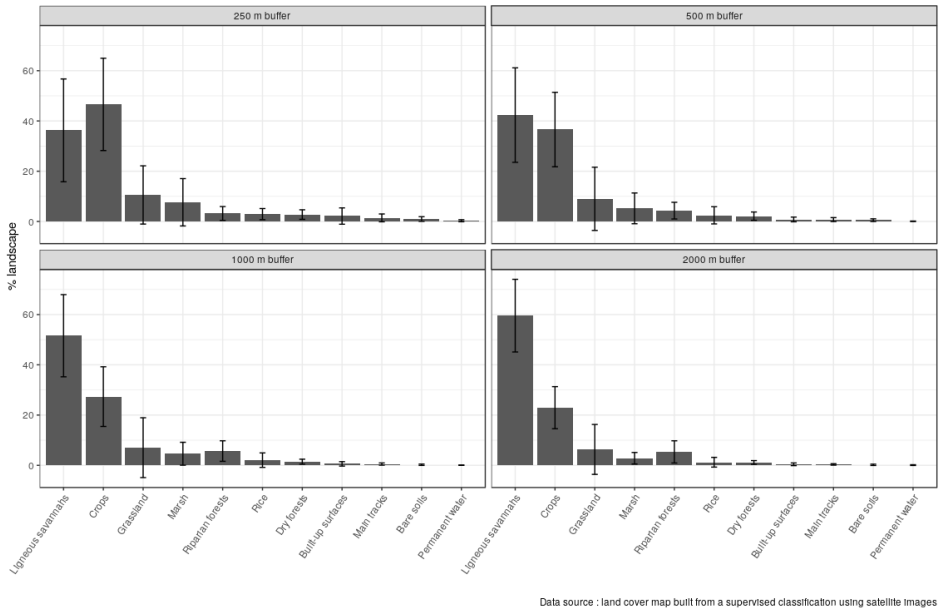


FIGURE 4.2: Figure additionnelle n°2 : Summary of the landscape conditions around the sampling points. Average percentage of surface occupied by each land cover class in the various buffer zones (250 m, 500 m, 1 km, 2 km radii) around the collection points. Error bars indicate the mean ± one standard deviation considering all the sampling points.

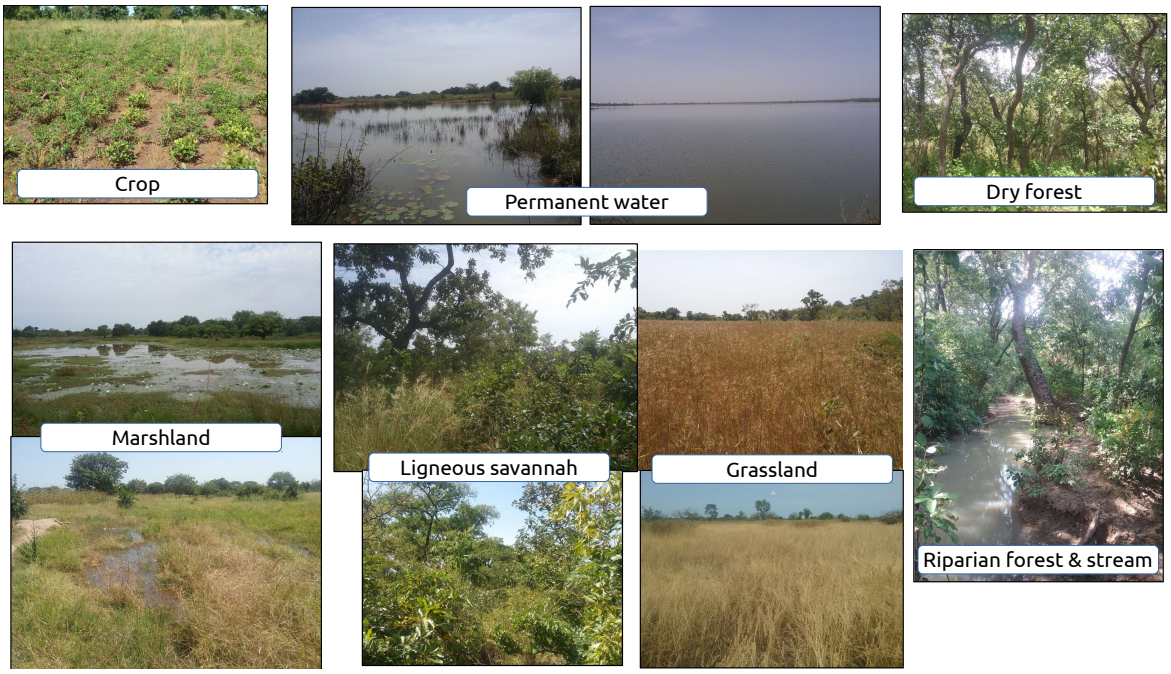


FIGURE 4.3: Figure additionnelle n°3 : Pictures representative of the main land cover classes in the Diébougou area. Pictures were taken in November 2018.

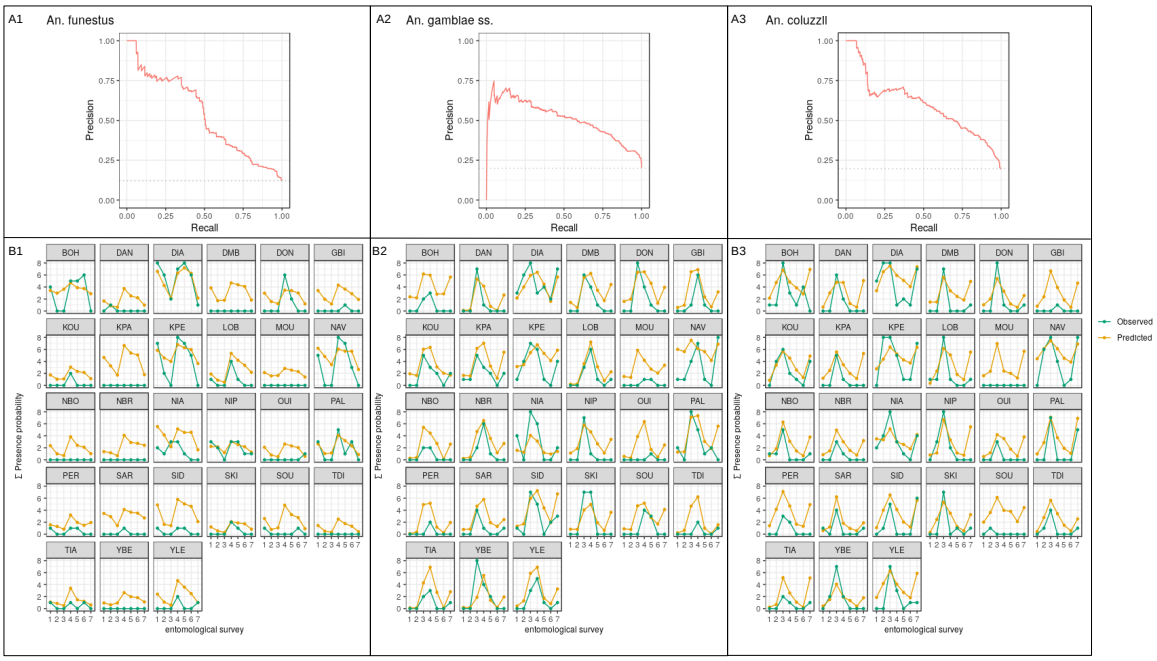


FIGURE 4.4: Figure additionnelle n°4 : Model evaluation plots for the presence models. A1, A2, A3 are precision–recall curves for the presence models of respectively An. funestus, An. gambiae s.s. and An. coluzzii. Precision–recall curves show the precision and the recall of the models for different probability thresholds of the "presence" class. Precision is the proportion of presence identifications that was actually correct, while recall is the proportion of actual presence observations that were identified correctly. The horizontal dashed line represents the baseline (i.e. random or no-skill) classifier. A precision–recall curve above the horizontal line indicates a better-than-no-skill classifier. The higher the area between the precision–recall curve and the horizontal line, the better the classifier. Plots B1, B2, B3 are observed vs. predicted presence probabilities for each out-of-sample village. The y-axis represents the sum over the 8 sampling points/village/ survey (4 points by village * 2 places (interior and exterior)). Overall, the plots A1, A2, A3 show that the models had good predictive accuracies (precision–recall curves are higher than the baseline curve, particularly for An. funestus and An. coluzzii). The plots B1, B2, B3 show that the models predicted well the spatiotemporal trends of presence/absence of bites (lines of predicted presence probabilities are generally close to lines of observed probabilities), although they usually slightly overestimated the probabilities of being bitten (predicted presence probability > observed presence probability).

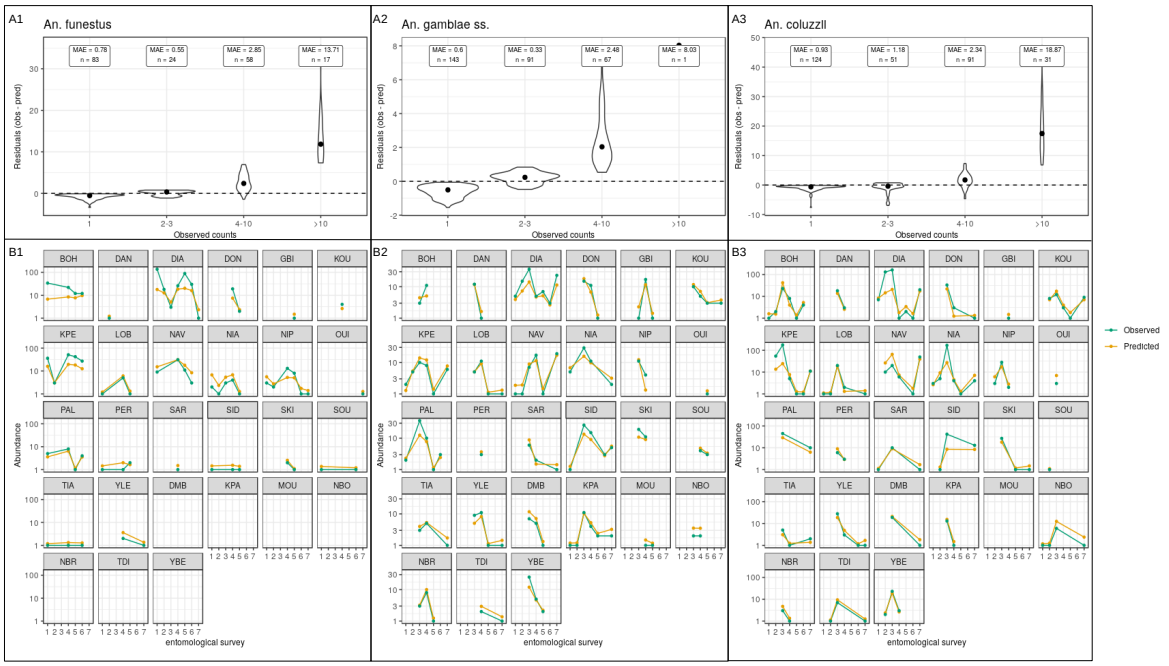
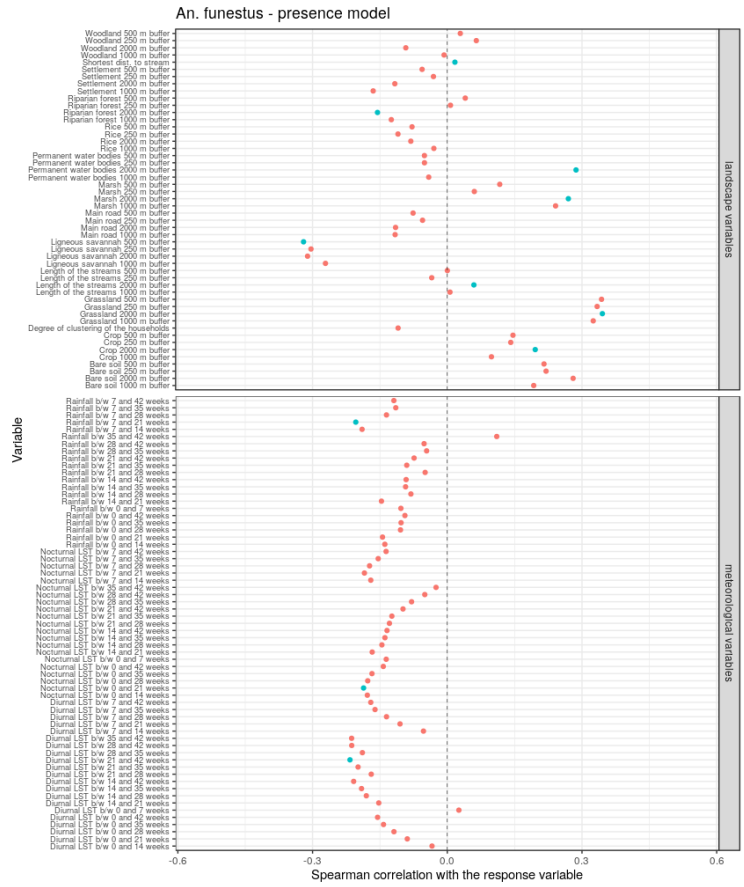
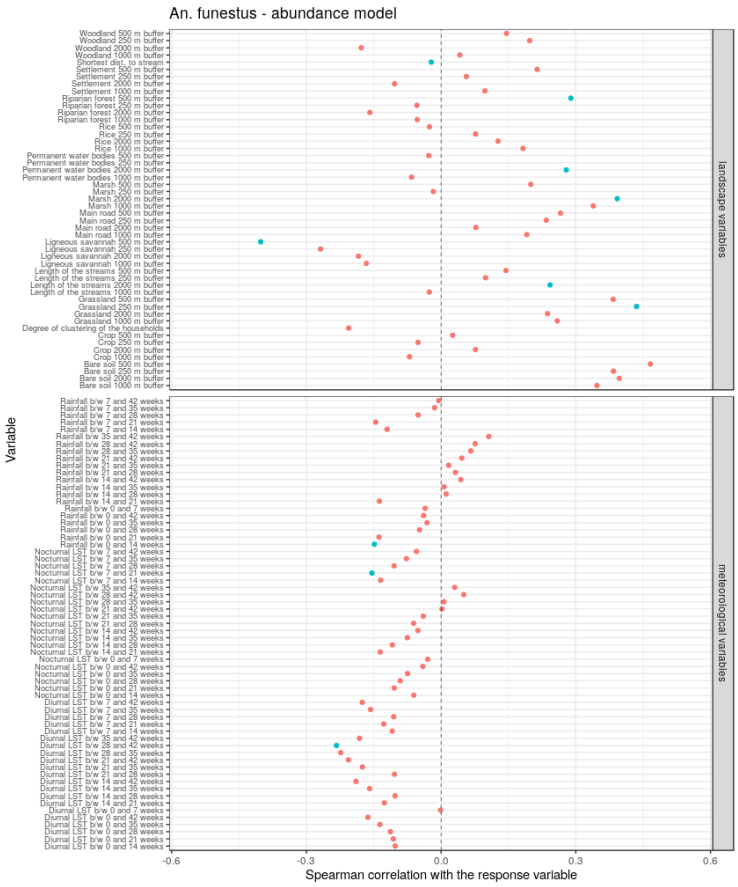


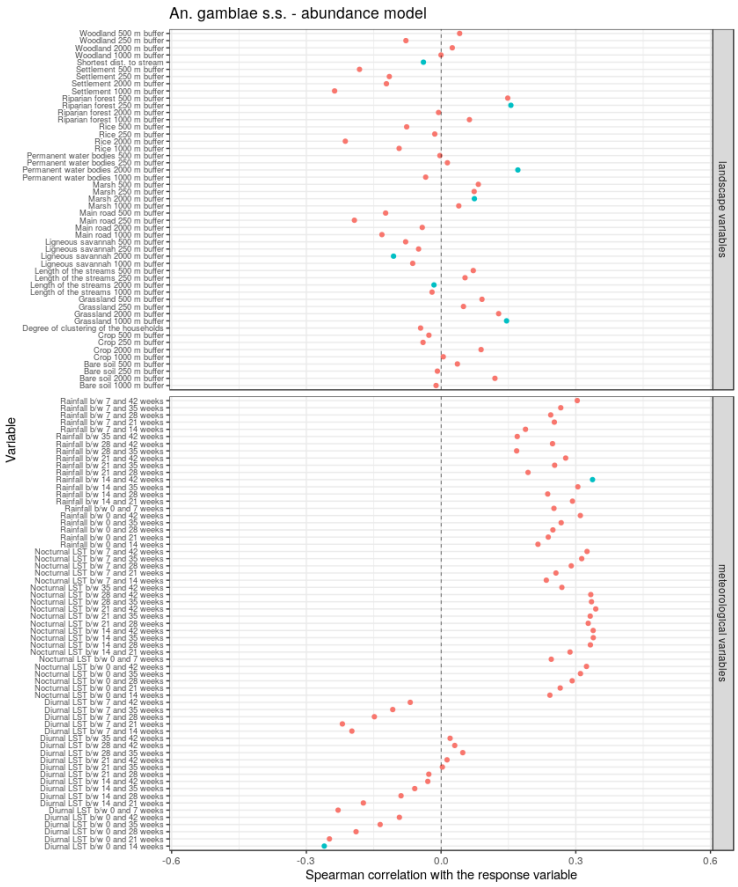
FIGURE 4.5: Figure additionnelle n°5 : Model evaluation plots for the abundance models. A1, A2, A3 are violin plots of the distribution of the residuals for the abundance models of respectively An. funestus, An. gambiae s.s. and An. coluzzii, by observed counts of bites (4 classes : 1 bite, 2–3 bites, 4–10 bites, > 10 bites). Black dots indicate the median value. B1, B2, B3 are observed vs. predicted number of bites/village/entomological surveys. The y-axis represents the sum of bites over the 8 sampling points/village/survey (4 points by village * 2 places (interior and exterior)) on a logarithmic scale. The absence of a dot indicates that no vector was collected. MAE = mean absolute error ; n = number of observations. Overall, the plots A1, A2, A3 show that the models predicted well small observed counts of bites (1 bite, 2–3 bites) (cf. small MAEs, small residuals), which represent the vast majority of observations (high n). Larger counts (4–10 bites, > 10 bites) tended to be underestimated by the models, especially for An. funestus and An. gambiae s.s. However, large counts (> 10 bites) represented few observations (small n). The plots B1, B2, B3 confirm these observations, and additionally show that general trends of biting rates over time were well predicted by the models (lines of predicted abundance are generally close to lines of observed abundance).

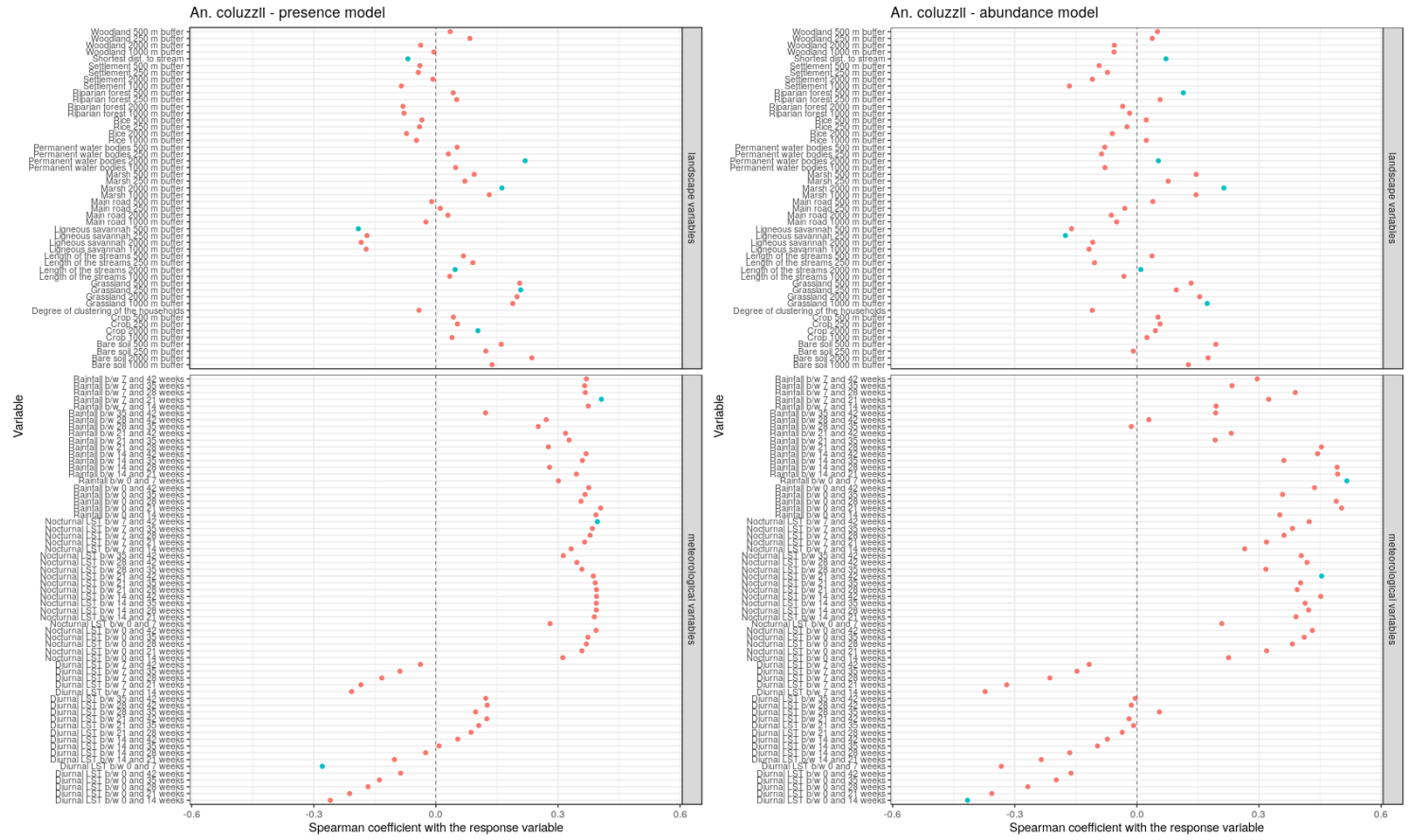
Ci-dessous : Article n°1 - Figure additionnelle n°6 : Feature selection for the multivariate models. The figure shows the Spearman correlation coefficient between the explanatory variables and each response variable (presence and abundance of An. funestus, An. gambiae s.s. and An. coluzzii). Based on these results, variables were retained for the multivariate models according to the following criteria : we first excluded variables that were poorly correlated with the response variable (i.e. correlation coefficients less than 0.1 or p-values greater than 0.2 at all time associations or buffer radii considered), except for variables related to the presence of water —i.e. possible breeding sites—that were all retained whatever their correlation. Then, for each meteorological (resp. landscape) variable, we retained the time lag interval (resp. buffer radius) showing the higher absolute correla- tion coefficient value. We finally excluded collinear variables (i.e. Pearson correlation coefficient between the variables > 0.7) based on empirical knowledge.


An. funestus - presence model
landscape variables
meteorological variables
Variable
-0.6
-0.3
0.0
0.3
0.6
Spearman correlation with the response variable


An. funestus - abundance model
landscape variables
meteorological variables
Variable
-0.6
-0.3
0.0
0.3
0.6
Spearman correlation with the response variable
multivariate_model
Variable excluded
Variable retained


An. gamblae s.s. - presence model
Variable
landscape variables
meteorological variables
Spearman correlation with the response variable
-0.6
-0.3
0.0
0.3
0.6


An. gamblae s.s. - abundance model
Variable
landscape variables
meteorological variables
Spearman correlation with the response variable
-0.6
-0.3
0.0
0.3
0.6
multivariate_model
Variable excluded
Variable retained


An. coluzzii - presence model
An. coluzzii - abundance model
Variable
landscape variables
meteorological variables
Spearman coefficient with the response variable
-0.6
-0.3
0.0
0.3
0.6
multivariate_model
Variable excluded
Variable retained

4.3 Reproduction de l'analyse dans la zone d'étude de Korhogo

Nous avons reproduit l'analyse dans la zone d'étude de Korhogo selon les méthodes décrites dans l'article. Sans entrer dans le niveau de détail de l'article, la section ci-après présente les résultats ainsi qu'une discussion, en détaillant notamment les principales similitudes et différences par rapport aux résultats obtenus dans la zone de Diébougou.

Résultats

Distribution spatio-temporelle des densités agressives des anophèles. Dans la zone de Korhogo (CI), un total de 2048 nuits-homme de capture a été réalisé (32 villages * 8 enquêtes entomologiques * 4 points de collecte * 2 lieux). 57722 anophèles ont été collectés. Les principales espèces/complexes trouvés étaient *An. gambiae s.l.* et *An. funestus* (respectivement 56267 (97% de tous les anophèles collectés) et 714 (1%) individus collectés). Parmi les 56267 *An. gambiae s.l.* collectés, 3922 (7%) ont été identifiés à l'espèce : 3726 (95% des individus identifiés à l'espèce) étaient des *An. gambiae s.s.* et 196 (5%) étaient des *An. coluzzii*. Par conséquent, dans la suite de cette étude, nous considérerons les *An. gambiae s.l.* collectés dans la zone de Korhogo comme des *An. gambiae s.s.*

An. gambiae s.s. et *An. funestus* étaient présents (cad. au moins un anophèle capturé) dans respectivement 64 % et 6 % des nuits-homme de capture. La distribution des densités agressives sur les sessions positives (cad. au moins un contact homme-vecteur) était très asymétrique, comme dans la zone de Diébougou (pour *An. gambiae s.s.* : médiane = 18, écart-type = 65, max. = 505 ; pour *An. funestus* : médiane = 2, écart-type = 12, max. = 84). La figure 4.6 présente les distributions spatio-temporelles des densités agressives des principales espèces d'anophèles dans la zone de Korhogo (pendant de la fig. 1 de l'article). La carte montre des dynamiques temporelles pour *An. gambiae s.s.* relativement similaires à celles de la zone de Diébougou : l'espèce était davantage abondante durant ou en fin de saison pluvieuse (septembre, octobre) qu'en saison sèche, où elle était malgré tout présente. Au niveau spatial, notons i) une certaine hétérogénéité de la distribution, et ii) que l'espèce était présente dans quasiment tous les villages à toutes les enquêtes entomologiques (sauf la 7ème), contrastant avec la zone de Diébougou où l'espèce était absente de plusieurs villages. La distribution

spatio-temporelle d'*An. funestus* était très déséquilibrée : l'écrasante majorité des individus (93 %) a été collectée lors de la première enquête entomologique, et presque la moitié des individus (42 %) a été collecté dans un seul village.

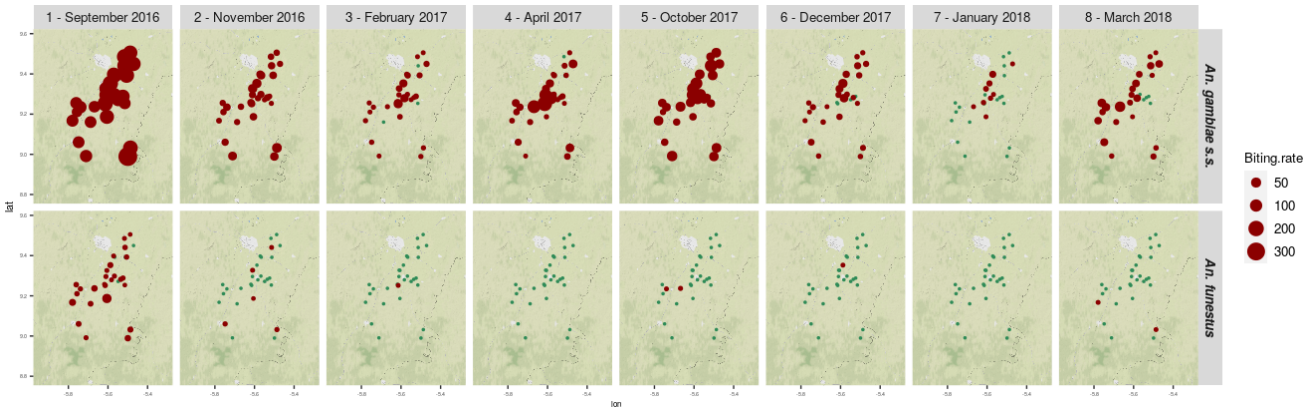


FIGURE 4.6: Distribution spatio-temporelle des densités agressives des principales espèces d'anophèles dans la zone de Korhogo (voir légende complète dans la figure 1 de l'article en section 4.2)

Modélisation bivariée. La figure 4.7 montre les variables paysagères qui étaient significativement corrélées (coefficient de corrélation de Spearman (cc) > 0.1 et p.value < 0.2) avec la présence et l'abondance des espèces d'anophèles (pendant de la fig. 3 de l'article). Comme à Diébougou, la présence et l'abondance d'*An. funestus* était corrélée à davantage de variables paysagères que celle d'*An. gambiae s.s.*, et les coefficients de corrélation les plus élevés avec les variables paysagères étaient observés pour *An. funestus*.

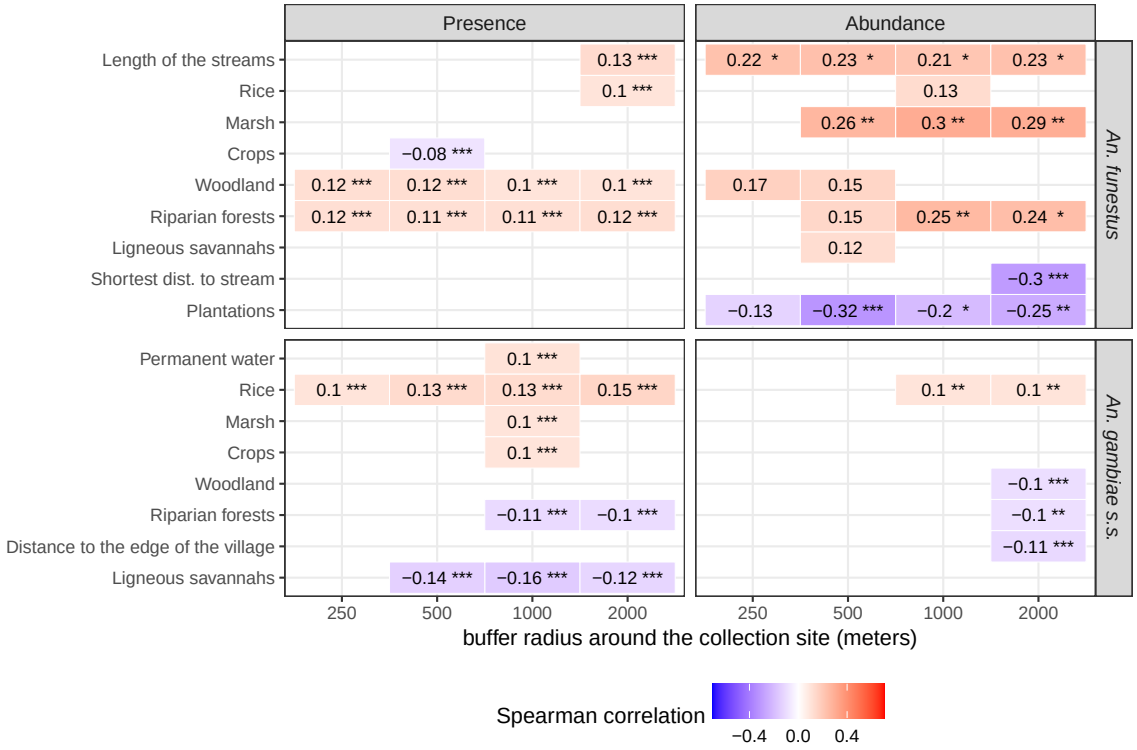


FIGURE 4.7: Coefficient de corrélation de Spearman entre entre les densités agressives des anophèles et les variables paysagères dans la zone de Korhogo (voir légende complète dans la figure 3 de l'article en section 4.2)

La présence d'*An. funestus* était positivement corrélée avec la longueur des rivières et au % de surface occupé par les zones rizicoles, dans la zone tampon de 2 km de rayon. Elle était également corrélée avec le % de surface occupé par les forêts ripicoles et les zones forestières (non ripicoles), dans toutes les zones tampon. L'abondance de l'espèce était positivement corrélée avec la longueur des rivières, les surfaces rizicoles, les surfaces marécageuses, les forêts ripicoles, et les zones forestières (non ripicoles), dans diverses tailles de zone tampon en fonction de la classe d'occupation du sol. L'abondance d'*An. funestus* était négativement corrélée avec le % de surface occupé par les plantations dans la zone tampon de 2 km de rayon, et avec à la distance à la rivière la plus proche (autrement dit, l'abondance était plus importante quand le point de capture était proche d'une rivière).

La présence d'*An. gambiae s.s.* était positivement corrélée avec le % de surface

en eaux permanentes, en zones marécageuses, et en zones agricoles dans la zone tampon d'1 km de rayon. La présence ainsi que l'abondance de l'espèce étaient également corrélée avec le % de surface occupée par les zones rizicoles, dans toutes les zones tampon pour la présence et dans les zones tampons d'1 et 2 km de rayon pour l'abondance. La présence et l'abondance d'*An. gambiae s.s.* était négativement corrélées avec le % de surface occupé par les forêts ripicoles, dans les zones tampons d'1 et 2 km de rayon pour la présence et dans la zone de 2 km pour l'abondance. L'abondance de l'espèce était négativement corrélée avec le % de surface occupée par les zones forestières dans la zone tampon de 2 km de rayon, et avec la distance à la lisière du village (autrement dit, l'abondance était plus importante dans les habitations situées proches de la lisière du village que dans celles situées proches du centre du village). La présence de l'espèce était négativement corrélée avec le % de surface occupée par les savanes ligneuses, dans les zones tampons de plus de 250 m de rayon.

Notons que les valeurs absolues des coefficients de corrélation entre la présence / abondance des espèces et les variables paysagères étaient, dans l'ensemble, moins élevées dans la zone de Korhogo que dans la zone de Diébougou.

La figure 4.8 montre les variables météorologiques qui étaient significativement corrélées (coefficient de corrélation de Spearman (cc) > 0.1 et p.value < 0.2) avec la présence et l'abondance des espèces d'anophèles (*cross-correlation maps*, ou CCM) (pendant de la fig. 4 de l'article). Comme à Diébougou, les coefficients de corrélation entre la présence/abondance des espèces et les variables météorologiques étaient plus élevés pour *An. gambiae s.s.* que pour *An. funestus*.

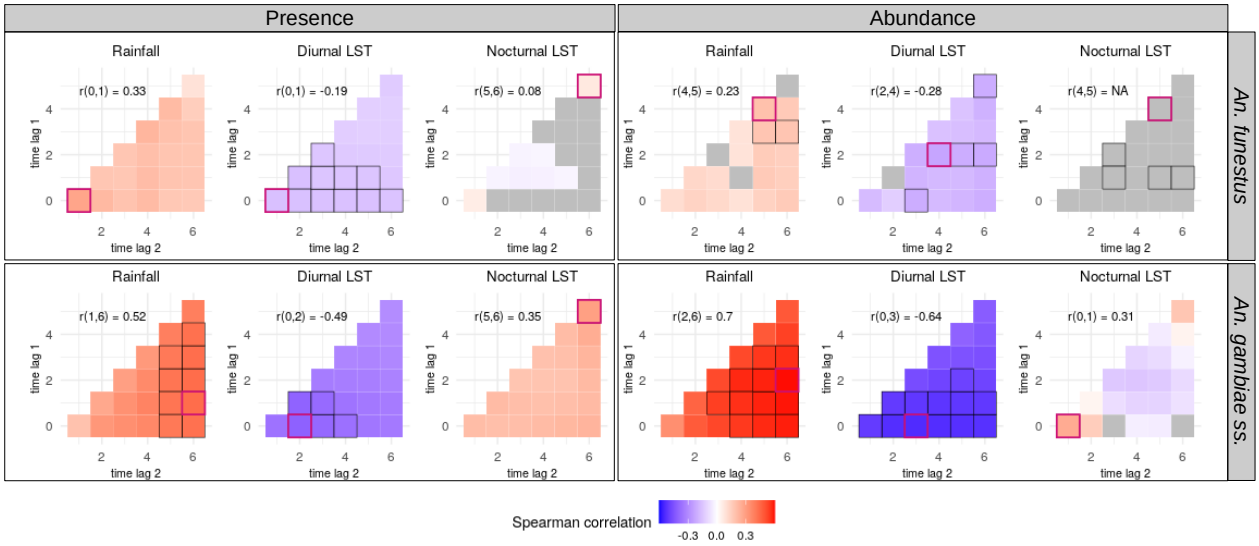


FIGURE 4.8: Coefficient de correlation de Spearman entre entre les densités agressives des anophèles et les variables météorologiques dans la zone de Korhogo (sous forme de cross-correlation maps) (voir légende complète dans la figure 4 de l'article en section 4.2)

La présence et l'abondance d'*An. funestus* étaient positivement corrélées avec le cumul des précipitations précédant la date de capture, à presque tous les 'lags' temporels - ce qui contraste avec la zone de Diébougou où les corrélations d'*An. funestus* avec les précipitations étaient négatives. La présence et l'abondance de l'espèce étaient négativement corrélées avec les températures diurnes, là aussi à presque tous les lags temporels précédant la date de capture. Les corrélations entre la présence ou l'abondance d'*An. funestus* et les températures nocturnes précédant la date de capture étaient faibles ou non-significatives.

La présence et l'abondance d'*An. gambiae s.s.* était positivement, fortement corrélées (plus encore qu'à Diébougou) avec le cumul des précipitations précédant la date de capture, à tous les lags temporels. Parmi tous les lags, c'est le cumul des précipitations enregistré entre les semaines 1 à 6 précédant la capture qui présentait le coefficient de corrélation le plus élevé avec la présence de l'espèce ; et le cumul des précipitations enregistré entre les semaines 2 à 6 précédant la capture qui présentait le coefficient de corrélation le plus élevé avec l'abondance de l'espèce. La présence d'*An. gambiae s.s.* était également positivement corrélée avec les températures nocturnes

précédant la date de capture à tous les lags temporels, et le coefficient de corrélation maximum était entre 5 et 6 semaines avant la capture. La présence et l'abondance d'*An. gambiae s.s.* était négativement corrélée avec les températures diurnes précédant la date de capture, à tous les lags temporels. Le coefficient de corrélation maximum entre les températures diurnes et la présence/abondance de l'espèce était entre 0 et 2-3 semaines avant la date de capture. Notons que les CCM (*Cross-Correlation Maps*) d'*An. gambiae s.s.* dans les zones de Korhogo et Diébougou étaient, une à une, très semblables : si les valeurs absolues des coefficients de corrélation étaient globalement légèrement supérieures dans la zone de Korhogo, les lags temporels présentant les coefficients de corrélation les plus élevés étaient quasiment identiques pour 5 des 6 CCMs.

Notons qu'à l'inverse des variables paysagères, les valeurs absolues des coefficients de corrélation entre la présence / abondance des espèces et les variables météorologiques étaient, dans l'ensemble, plus élevés dans la zone de Korhogo que dans la zone de Diébougou (en particulier pour *An. gambiae s.s.*).

Modélisation multivariée. La *Precision-Recall area under the curve* (PR-AUC) des modèles de présence était de 0.52 (baseline=0.09) et 0.91 (baseline=0.64) pour *An. funestus* et *An. gambiae s.s.* respectivement. La spécificité et la sensitivité des modèles au seuils optimaux de probabilité de présence étaient de 53 % et 98 % pour *An. funestus* et de 88 % et 61 % pour *An. gambiae s.s.* Ces résultats indiquent de bonnes puissance prédictives pour les modèles de présence, comme dans la zone de Diébougou. De même, comme dans la zone BF, les modèles d'abondance refletaient bien les tendances des abondances observées. Les figures d'évaluation des modèles multivariés de présence et d'abondance dans la zone de Korhogo (pendant des 'supplementary file' 4 et 5 de l'article) sont présentées dans les figures 4.9 et 4.10.

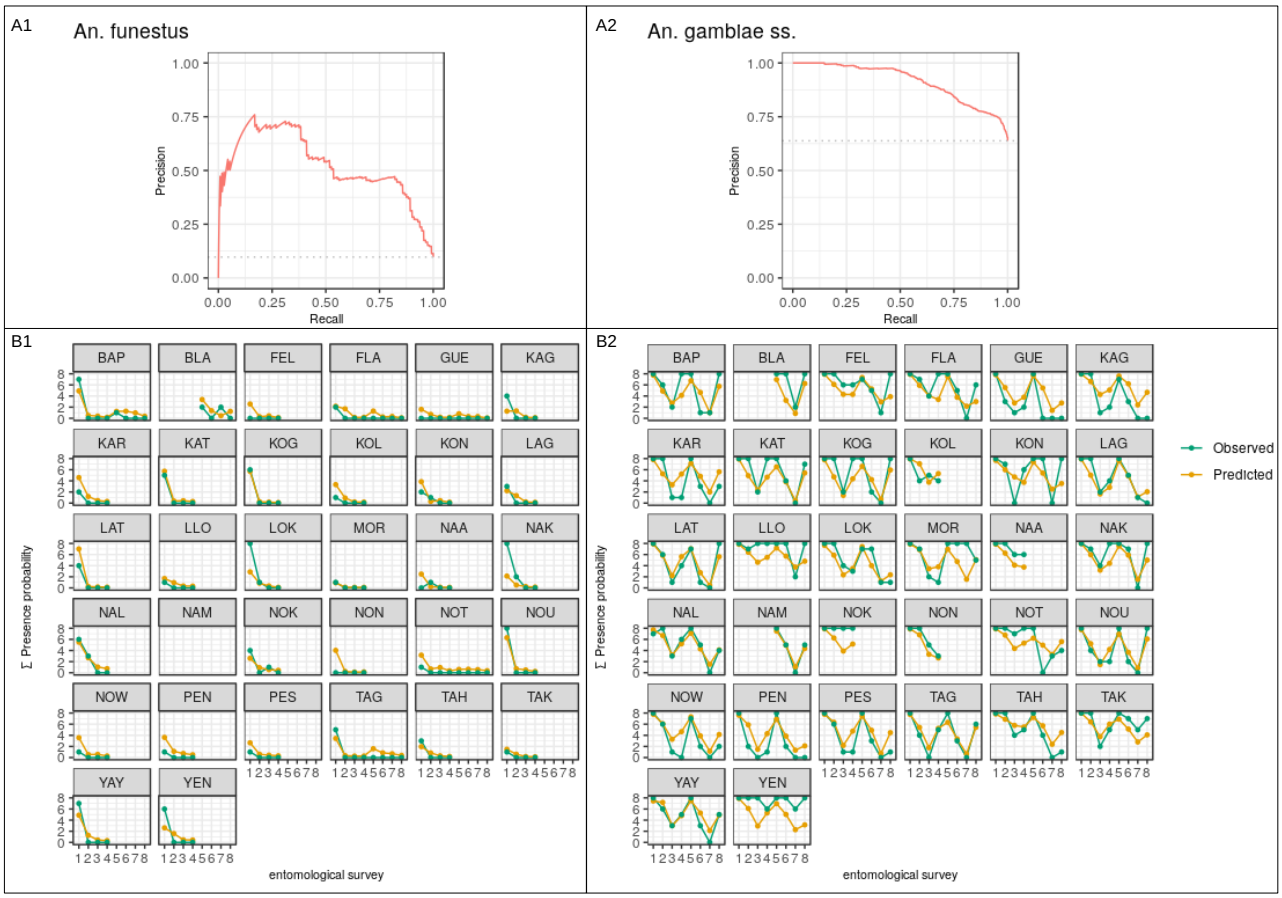


FIGURE 4.9: Evaluation de la puissance prédictive des modèles de présence des anophèles dans la zone de Korhogo (voir légende complète dans la figure additionnelle 5 (Figure S5) de l'article en section 4.2)

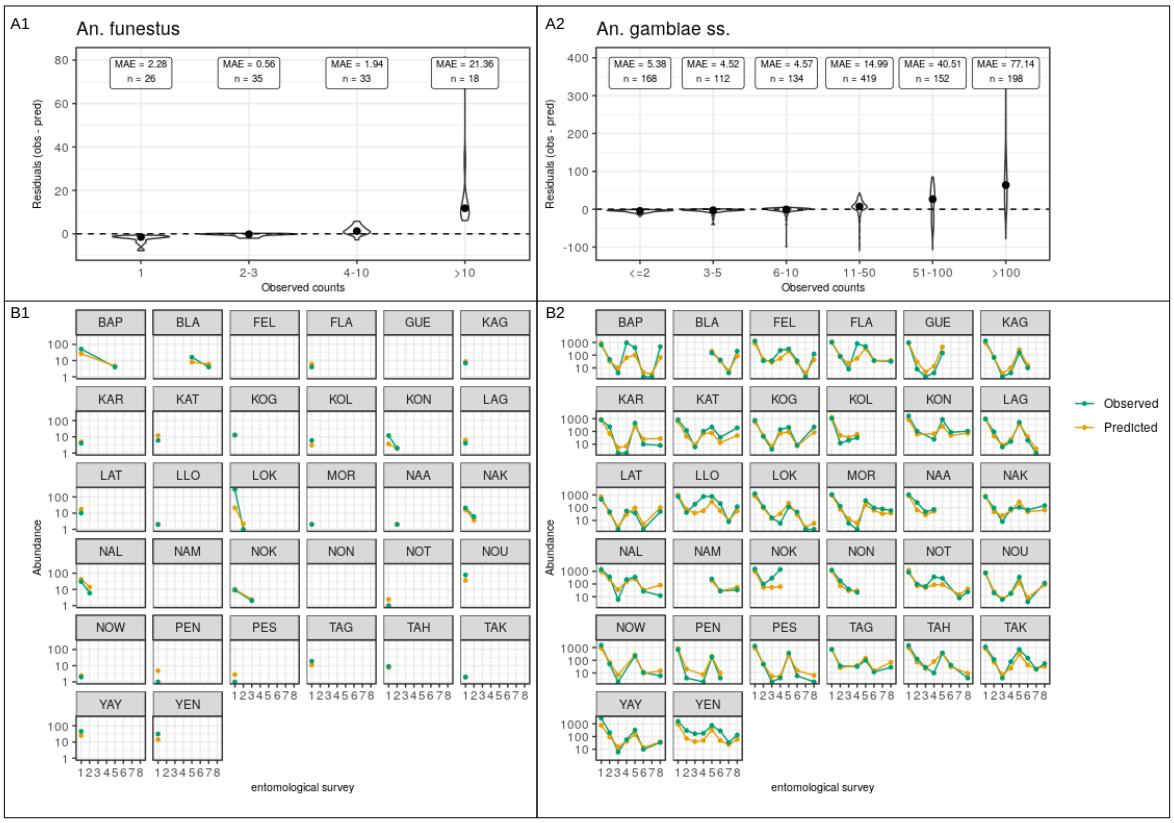


FIGURE 4.10: Evaluation de la puissance prédictive des modèles de d'abondance des anophèles dans la zone de Korhogo (voir légende complète dans la figure additionnelle 5 (Figure S5) de l'article en section 4.2)

Les figures 4.12 et 4.11 montrent les graphiques d'interprétation des modèles de forêt aléatoire (variables d'importances et les plots de dépendance partiels) pour la présence et l'abondance des deux espèces (pendant des fig. 5-6-7 de l'article).

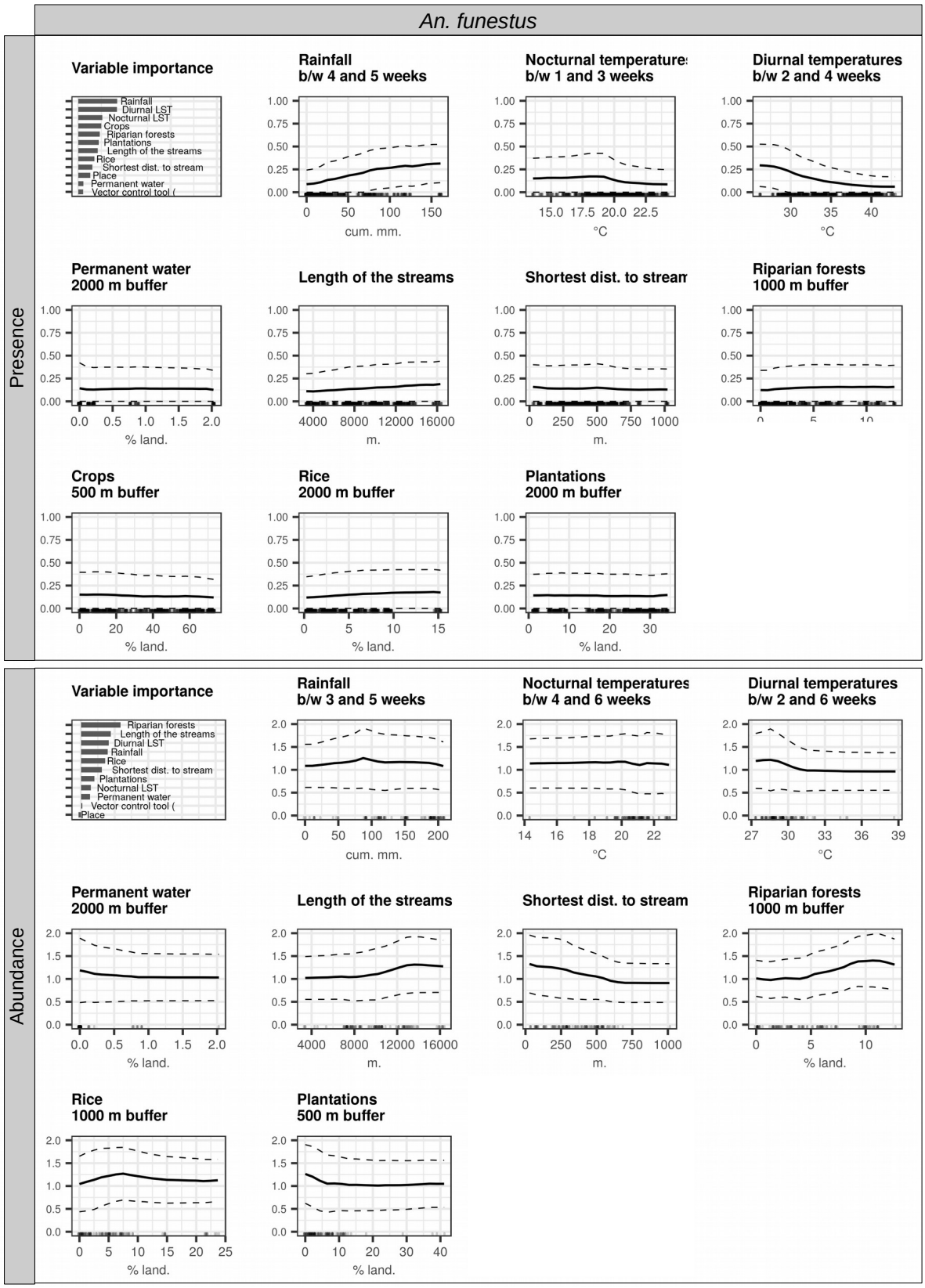


FIGURE 4.11: Graphiques d'interprétation des modèles de forêt aléatoires pour An. funestus dans la zone de Korhogo (voir légende complète dans la figure 5 de l'article en section 4.2)

Les variables les plus importantes dans le modèle de présence d'*An. funestus* étaient les trois variables descriptives de la météorologie enregistrée pendant les semaines précédant la capture : cumul des précipitations (relation linéaire positive), températures diurnes moyennes (relation négative entre 25°C et 35°C, et plafonnant entre 35° et 45°), et températures nocturnes moyennes. Les variables les plus importantes dans le modèle d'abondance de cette espèce étaient : le % de surface occupé par les forêts ripicoles (relation nulle entre 0% et 5%, positive entre 5% et 10%, et nulle entre 10% et 12%), la longueur totale des rivières dans la zone tampon de 2 km de rayon autour des points de capture (relation nulle entre 4 km et 10 km de rivières, positive entre 10 km et 13 km, et nulle entre 13 km et 16 km), et les températures diurnes moyennes (relation linéaire négative entre 27°C et 31°C et nulle entre 31°C et 39°C).

Les prédicteurs secondaires dans le modèle de présence d'*An. funestus* étaient des variables paysagères : % de surface occupé par les forêts ripicoles (relation positive linéaire), % de surface occupé par les rizicultures (relation positive linéaire), la longueur totale des rivières dans la zone tampon de 2 km de rayon autour des points de capture (relation positive linéaire). Les prédicteurs secondaires dans le modèle d'abondance étaient : le % de surface occupé par les rizicultures (relation positive), la distance à la rivière la plus proche (relation négative), et le % de surface occupé par les forêts ripicoles (relation positive).

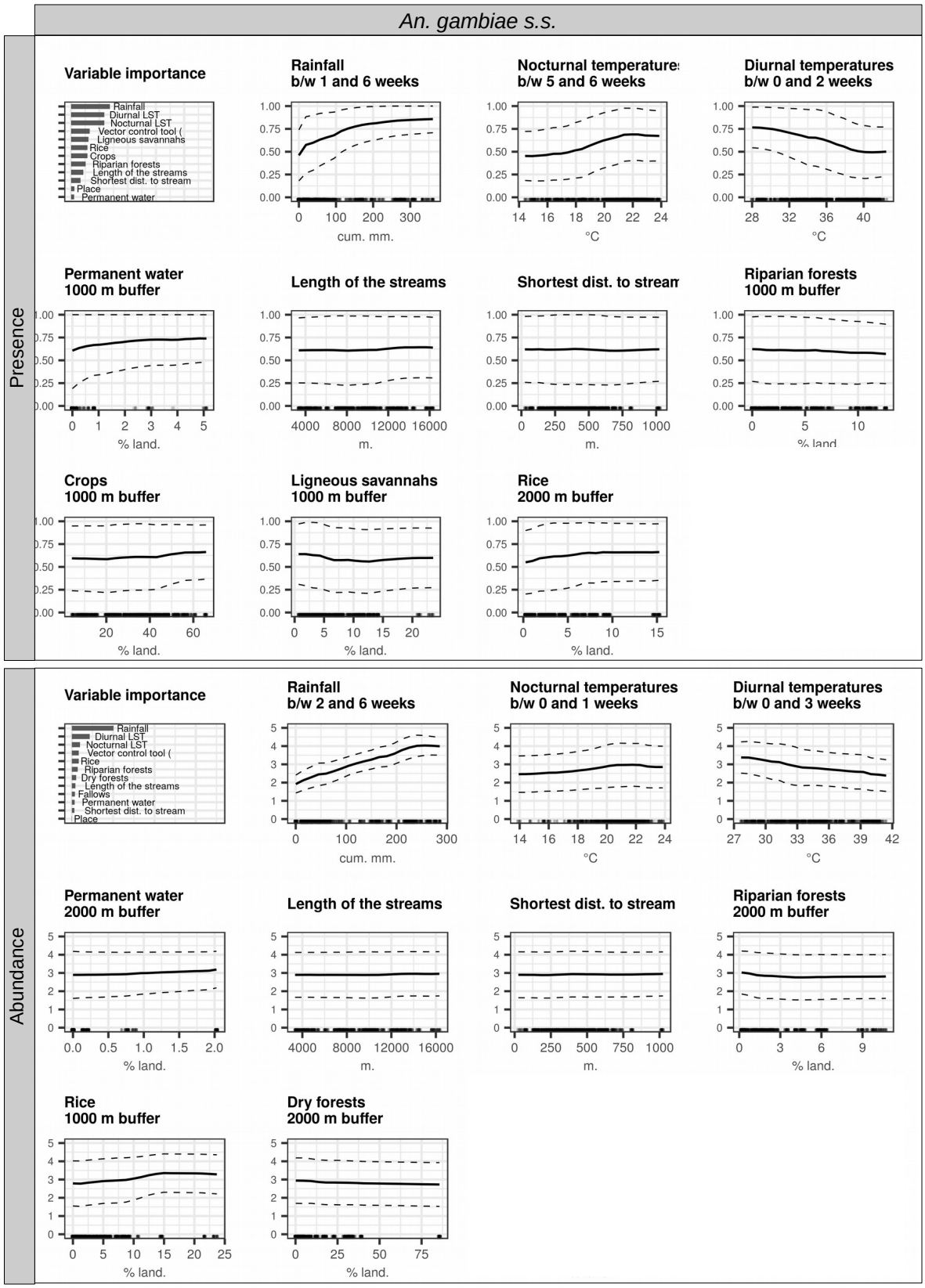


FIGURE 4.12: Graphiques d'interprétation des modèles de forêt aléatoires pour An. gambiae s.s. dans la zone de Korhogo (voir légende complète dans la figure 5 de l'article en section 4.2)

Les variables les plus importantes dans les modèles de présence et d'abondance d'*An. gambiae s.s.* étaient les trois variables descriptives de la météorologie enregistrée pendant les semaines précédant la capture : dans l'ordre, cumul des précipitations entre 1-2 et 6 semaines précédant la capture, températures diurnes moyennes entre 0 et 2-3 semaines précédant la capture, et températures nocturnes moyennes (entre 5 et 6 semaines et entre 0 et 1 semaine pour les modèles de présence et d'abondance respectivement). La probabilité de présence augmentait logarithmiquement avec les précipitations, et l'abondance augmentait linéairement avec les précipitations. La nature des relations avec les températures était relativement similaire à celle d'*An. gambiae s.s.* dans la zone de Diébougou. Notons que l'importance des précipitations était particulièrement élevée dans le modèle d'abondance d'*An. gambiae s.s.*, dominant largement l'importance de toutes les autres variables.

Les prédicteurs secondaires dans le modèle de présence d'*An. gambiae s.s.* étaient le % de surface occupé par les savanes ligneuses (relation négative entre 0% et 5% puis nulle entre 5% et 25%), le % de surface occupé par les rizicultures (relation positive entre 0% et 10% puis nulle entre 10% et 15%), et le % de surface occupé par les zones agricoles (relation nulle entre 0% et 50% puis positive entre 50% et 60%). Les prédicteurs secondaires dans le modèle d'abondance étaient le % de surface occupée par les rizières (relation positive entre 0% et 15% puis nulle entre 15% et 25%), le % de surface occupé par les forêts ripicoles (relation négative entre 0% et 3% puis nulle entre 3% et 10%), et le % de surface occupé par les forêts (relation négative linéaire).

Discussion

Dans la zone de Korhogo, le cumul des précipitations était la variable la plus importante à la fois dans les modèles de présence et d'abondance d'*An. gambiae s.s* ; et dans les deux modèles les variables de températures complétaient le trio de tête des varibles d'importance. Ces observations, similaires à celles effectuées dans la zone de Diébougou, montrent que dans la zone de Korhogo comme sur celle de Diébougou i) *An. gambiae s.s.* était attaché aux gites larvaires temporaires, remplis par les précipitations et ii) ses traits de vie (développement, survie) étaient fortement impactés par les conditions météorologiques. Le % de surface occupé par les zones rizicoles était respectivement la seconde et première variable paysagère la plus importante dans les

modèles de présence et d'abondance de l'espèce, montrant ainsi que les rizicultures abritaient très probablement une forte densité de larves d'anophèles. Cette hypothèse est confirmée par une étude de terrain menée par l'équipe du projet REACT dans la zone de Korhogo visant à caractériser les gites larvaires d'*Anopheles* spp (Zogo, Koffi, et al., 2019). La surface en savanes ligneuses, négativement corrélée avec la probabilité de présence d'*An. gambiae s.s.*, était la variable paysagère la plus importante dans le modèle d'abondance de l'espèce ; corroborant les observations effectuées dans la zone de Diébougou et les hypothèses sur l'importance de l'ouverture des milieux sur la densité agressive des vecteurs.

Les cross-correlations maps d'*An. gambiae s.s.* montrent que, comme dans la zone de Diébougou, les conditions météorologiques dans la zone de Korhogo impactaient fortement tous les stades de développement de l'espèce, et qu'elles avaient parfois eu un impact plus important encore sur les périodes précédant la durée de vie de la génération échantillonnée. Les CCMs d'*An. gambiae s.s.* dans les deux zones d'étude (Diébougou et Korhogo) se ressemblaient fortement, indiquant probablement des dynamiques de population de l'espèce très similaires sur ces deux zones - et peut-être, par extension, sur l'ensemble des zones de la sous-région présentant des conditions climatiques similaires).

Dans la zone de Korhogo, à la différence de Diébougou, les variables les plus importantes dans le modèle de présence d'*An. funestus* étaient toutes trois météorologiques. Ainsi, contrairement aux observations effectuées dans la zone de Diébougou, la distribution spatio-temporelle d'*An. funestus* dans la zone de Korhogo semblait être principalement conditionnée par les conditions météorologiques. Par contre, quand *An. funestus* était présent, son abondance semblait dépendre fortement des conditions paysagères (deux des trois variables les plus importantes du modèle d'abondance de l'espèce étaient paysagères), comme à Diébougou. En particulier, l'espèce semblait particulièrement attachée aux zones aquatiques semi-permanentes (trois des six variables les plus importantes du modèle d'abondance de l'espèce, dont les deux premières, étaient liées à des rivières inondées en saison des pluies). Les bords des rivières et autres zones aquatiques semi-permanentes semblaient donc constituer, comme à Diébougou, des gites larvaires préférentiels pour *An. funestus*.

Les modèles multivariés prédisaient correctement la présence et l'abondance des

espèces. Comme à Diébougou, les principaux déterminants de la présence et de l'abondance des principales espèces d'anophèles ont ainsi probablement été intégrés dans les modèles et identifiés.

Nous noterons pour terminer que les densités agressives moyennes ainsi que la proportion de sessions positives (avec au moins une piqûre) étaient très largement supérieures dans la zone de Korhogo que dans celle de Diébougou, bien que ces deux zones soient éloignées de 300 km seulement à vol d'oiseau. Des différences à la fois dans les régimes météorologiques et dans l'utilisation/occupation du sol de ces deux zones pourraient expliquer ces contrastes. Les précipitations plus abondantes et les températures diurnes maximum moins élevées à Korhogo qu'a Diébougou (voir section 3.2.1) peuvent impliquer, respectivement, des gites larvaires temporaires plus nombreux ou persistants et des taux de mortalité moins élevés dans la première zone que dans la seconde. Au niveau paysager, les gites larvaires permanents (zones rizicoles, barrages les irriguant) étaient plus abondants à Korhogo qu'à Diébougou, et les milieux 'fermés' (savanes ligneuses notamment) - qui réduisent *à priori* les densités agressives (voir article) - y étaient moins abondants (voir section 3.2.2). Ces différences, marquant par ailleurs un niveau d'anthropisation du territoire plus important à Korhogo qu'à Diébougou, pourraient ainsi également expliquer en partie les différences de densités agressives observées.

Conclusion

dans la zone de Korhogo, la distribution spatio-temporelle des vecteurs du paludisme, hétérogène, semblait être fortement déterminée et contrainte par les conditions météorologiques - plus encore que dans la zone de Diébougou. Les rizières, les rivières et les gites temporaires remplis par les précipitations semblaient constituer les gites larvaires des anophèles, comme cela a été confirmé par (Zogo, Koffi, et al., 2019). Les densités agressives des vecteurs étaient largement supérieures à Korhogo qu'à Diébougou. Des différences notables entre les deux territoires dans les régimes météorologiques et dans le niveau d'anthropisation pourraient expliquer ces différences. Comme dans la zone de Diébougou, la bonne prédictibilité des densités agressives des vecteurs dans la zone de Korhogo ouvre la voie au développement des outils opérationnels de gestion du risque de transmission décrits dans l'article (plans d'action de lutte antivectorielle,

cartes saisonnières de la distribution des vecteurs à l'échelle du village, systèmes d'alerte précoces). Par ailleurs, les similitudes à la fois dans les CCMs, dans l'importance relative des prédicteurs dans les modèles multivariés, et dans la nature des relations capturées par ces mêmes modèles, ouvre des perspectives intéressantes quand à la transposabilité des modèles prédictifs de présence et abondance des anophèles dans la sous-région, hors des zones d'étude du projet REACT.

Chapitre 5

Article n°2 - Modélisation des dynamiques spatio-temporelles des résistances physiologiques et comportementales des vecteurs

L'étude exposée dans le chapitre précédent nous a permis de préciser certaines caractéristiques de la niche écologique des principales espèces vectrices du paludisme dans nos deux zones d'étude. Dans cette seconde étude, nous nous intéressons aux résistances, physiologiques et comportementales, des vecteurs aux insecticides. Les objectifs, conceptuellement, sont similaires : approfondir les connaissances sur les déterminants de la prévalence des résistances physiologiques et comportementales des anophèles dans nos zones d'étude ; et évaluer la prédictibilité de la présence de ces résistances chez les vecteurs. Pour cela, au même titre que pour l'étude précédente nous faisons appel à la modélisation statistique dans une approche holistico-inductive. Nos variables explicatives sont nombreuses, variées et fines : lutte anti-vectorielle, disponibilité de l'hôte et micro-climat pendant la recherche de repas de sang, etc. L'enjeu ici est double : capturer et interpréter des associations potentiellement complexes et non-hypothétisées entre environnement et résistances des vecteurs, et quantifier précisement l'impact de certaines variables - en particulier celles liées à la lutte anti-vectorielle, principal déterminant supposé du développement des résistances. Aussi, nous faisons appel à la fois à des modèles non-paramétriques et paramétriques. Cette étude est présentée sous la forme d'un article scientifique entièrement rédigé au moment de l'écriture de ce manuscrit.

Dans ce chapitre, nous résumons puis intégrons l'article en l'état.

5.1 Résumé de l'article

Les objectifs principaux de cette étude étaient i) d'approfondir les connaissances sur les déterminants des résistances physiologiques et comportementales des anophèles dans nos zones d'étude, et ii) d'évaluer la prédictibilité de ces résistances chez les anophèles dans l'espace et dans le temps. Plus spécifiquement, les questions soulevées étaient les suivantes :

- Quelle est la contribution respective de l'agriculture et de la lutte anti-vectorielle dans le développement des résistances physiologiques sur nos territoires d'étude ?
- Quels sont les méchanismes biologiques qui sous-tendent les résistances comportementales ?
- Les comportements des vecteurs sont-ils influencés par des conditions environnementales (météorologiques, paysagères) pendant la recherche de repas de sang ?
- Les résistances physiologiques influencent-elles les résistances comportementales ?
- Quel mécanisme de résistance aux insecticides (comportemental ou physiologique) apparaît et se répand le plus rapidement dans une population de vecteurs ?
- Les résistances des vecteurs sont-elles hétérogènes dans l'espace et dans le temps à fine échelle spatiale ?
- A quel niveau peut-on expliquer et prédire la prévalence des résistances des vecteurs dans l'espace et dans le temps ?

Nous avons modélisé six indicateurs de résistance des vecteurs, dont trois de résistance comportementale et trois de résistance physiologique, pour chaque espèces d'anophèle et dans chaque zone d'étude : la probabilité pour un moustique capturé de piquer à l'extérieur (exophagie), la probabilité pour un moustique capturé de piquer précocément (avant que 50 % de la population humaine soit déclarée comme étant sous une moustiquaire le soir) (activité précoce) ou tardivement (après que 50 % de la population humaine soit déclarée comme étant hors d'une moustiquaire le matin) (activité tardive), et les probabilités pour un moustique capturé de porter un allèle résistant pour chacune des mutations kdr-w, kdr-e, et ace-1 (les modèles de résistance physiologiques ont été générés uniquement dans la zone de Diébougou, les données n'étant pas exhaustives dans la zone de Korhogo). Les variables explicatives, principalement environnementales, appartenaient à sept groupes : lutte anti-vectorielle, disponibilité de l'hôte humain au

moment de la recherche de repas de sang, conditions micro-climatiques au moment de la recherche de repas de sang, conditions météorologiques précédant la capture (mois précédant et jour de capture), conditions paysagères, résistance des vecteurs, abondance des vecteurs. Nous avons modélisé chaque indicateur de résistance à l'aide de deux modèles statistiques : un modèle paramétrique d'une part (GLMM binomial) afin de mesurer statistiquement l'impact de certaines variables explicatives d'intérêt (notamment celles liées à la LAV) ; et un modèle non-paramétrique d'autre part (forêt alétoire) pour maximiser les chances de capturer des associations entre variables potentiellement complexes. Nous avons calculé les performances explicatives et prédictives des modèles et avons interprété les modèles à l'aide des *partial dependence plots* et des informations plus classiques en sortie des GLMM binomiaux (coefficients, p-values, intervalles de confiances).

Nous avons observé que pour une espèce et un indicateur de résistance donnés, la proportion de vecteurs résistants était, dans l'ensemble, relativement stable dans l'espace (entre les villages) et dans le temps (entre les missions de captures entomologiques) ; bien que de légères variations fussent présentes. Les GLMM ont capturé de nombreuses associations statistiquement significatives entre les variables environnementales et celles représentant les résistances des vecteurs. Dans l'ensemble, les puissances explicatives et prédictives des modèles étaient cependant relativement faibles ; en particulier pour les modèles de résistances comportementales. Sur la base des associations entre variables capturées par les modèles et de leurs puissances explicatives et prédictives, nous avons émis plusieurs hypothèses sur les déterminants des résistances physiologiques et comportementales des vecteurs sur nos zones d'étude. En particulier :

— Nous avons conjecturé que le développement de la mutation kdr-e chez les anophèles dans la zone de Diébougou était davantage causé par les insecticides utilisés dans la LAV que ceux utilisés en agriculture ;
— Nous avons capturé de nombreuses associations entre les résistances physiologiques et les variables climatiques (sur le mois précédant la collecte et pendant la collecte), ce qui peut traduire un coût biologique de ces mutations génétiques pour les vecteurs, à la fois en terme de 'fitness' et d'activité ;
— Sans relever d'indices forts d'un caractère génétique et héréditaire des comportements des vecteurs (résistance constitutive), certains résultats vont malgré tout dans ce sens ;

— Nous avons noté que certaines espèces d'anophèles semblaient adapter - modérément - certains comportements de piqûre en fonction des conditions environnementales au moment de la recherche d'hôte (disponibilité de l'hôte et micro-climat) ;
— Nous avons cependant conjecturé que dans l'ensemble, les comportements de piqûre des anophèles n'étaient que marginalement déterminés par les conditions environnementales immédiates au moment de la recherche de repas de sang (cf. les faibles puissances explicatives et prédictives des modèles statistiques) ;
— Nous n'avons pas trouvé de phénotype comportemental (parmi ceux étudiés) associé à un génotype pour l'une des mutations de la cible (à savoir, pas de lien significatif entre résistances physiologiques et résistances comportementales).

Dans cette étude, nous avons donc tenté de mieux comprendre les déterminants de l'intensité et de l'hétérogénéité spatio-temporelle des résistances physiologiques et comportementales des vecteurs du paludisme, à fine échelle spatio-temporelle. Nous avons principalement (i) montré que les résistances (à la fois physiologiques et comportementales) étaient assez homogènes dans l'espace (entre les villages) et dans le temps (entre les saisons) à nos échelles d'étude, et (ii) émis l'hypothèse qu'à ces échelles spatio-temporelles, les résistances des vecteurs semblaient n'être que marginalement influencées par des facteurs environnementaux autres que ceux liés à l'utilisation d'insecticides dans la lutte antivectorielle. Afin d'éviter les rebonds de transmission, il y a donc urgence à repenser l'utilisation des insecticides dans la lutte anti-vectorielle.

5.2 Texte intégral de l'article

Insecticide resistance and biting behaviour of malaria vectors in rural West-Africa : a data mining study to adress their fine-scale spatiotemporal heterogeneity, drivers, and predictability*

PAUL TACONET, DIEUDONNÉ DILOMA SOMA, BARNABAS ZOGO, KARINE MOULINE, FRÉDÉRIC SIMARD, ALPHONSINE AMANAN KOFFI, ROCH KOUNBOBR DABIRÉ, CEDRIC PENNETIER, NICOLAS MOIROUX,

Insecticide resistance and behavioral adaptation of malaria mosquitoes impact the efficacy of long-lasting insecticide nets - currently the main malaria vector control tool. To develop and deploy complementary, efficient and cost-effective control interventions, a good understanding of the drivers of these physiological and behavioural traits is needed. In this data-mining work, we modeled a set of indicators of physiological resistances to insecticide (prevalence of three target-site mutations) and biting behaviours (early- and late-biting, exophagy) of anopheles mosquitoes in two rural areas of West-Africa, located in Burkina Faso and Cote d'Ivoire. To this aim, we used mosquito field collections along with heterogeneous, multisource and multi-scale environmental data. The objectives were i) to assess the small-scale spatial and temporal heterogeneity of the indicators, ii) to better understand their drivers, and iii) to assess their spatio-temporal predictability, at scales that are consistent with operational action. The explanatory variables covered a wide range of potential environmental determinants of vector resistance to insecticide or feeding behaviour : vector control, human availability and nocturnal behaviour, macro and micro-climatic conditions, landscape, etc. The resulting models revealed many statistically significant associations, although their predictive powers were overall weak. We interpreted and discussed these associations in light of several topics of interest, such as : respective contribution of public health and agriculture in the development of physiological resistances, biological costs associated with physiological resistances, biological mechanisms underlying biting behavior, and impact of micro-climatic conditions on the time or place of biting. To our knowledge, our work is the first studying insecticide resistance and feeding behaviour of malaria vectors at such fine spatial scale with such a large dataset of both mosquito and environmental data.

Keywords: malaria, anopheles, insectide resistance, behavioural adaptation, exophagy, early-biting, late-biting, spatiotemporal distribution, statistical modeling, data mining

Introduction

Malaria remains a major public health concern in Africa, with 241 million cases and 627 000 death over the continent in 2020 (WHO 2021). After years of steady reduction in the disease transmission mainly due to the scale-up of vector control (VC) interventions (in particular insecticide-based tools such as long lasting insecticide nets (LLIN) and indoor residual spraying (IRS)) (Bhatt et al. 2015), progress is now stalling since 2015 (WHO 2021). Involved in such worrying trends are a combination of biological, environmental and socio-economical factors. The mosquito biology, with the buildup of adaptive changes in the mosquito vectors populations enabling them to avoid or circumvent the lethal effects of insecticides, is most likely playing a very important contribution (Killeen 2014). These changes are framed as vector *resistance* to insecticides. As a consequence of the widespread

*Preprint version

use of insecticides (in agriculture and public health), vector resistance has arisen rapidly in malaria vectors in many areas of Africa and above (Durnez and Coosemans 2013; Riveron et al. 2018); and as previously indicated, is now at such level that it compromises the effectiveness of the most efficient malaria control interventions (Killeen 2014; Sokhna, Ndiath, and Rogier 2013; Hemingway et al. 2016; Gatton et al. 2013). Complementary and locally-tailored VC strategies taking into account the great diversity of vectors resistance mechanisms (see below) are therefore needed to target these vectors contributing to residual malaria transmission (Moiroux 2012; Durnez and Coosemans 2013; Sokhna, Ndiath, and Rogier 2013; Corbel and N'Guessan 2013; Hemingway et al. 2016; Riveron et al. 2018; WHO 2017).

Vector resistances to insecticide are usually split into two categories : *physiological* and *behavioural* resistance (Lockwood, Sparks, and Story 1984; Sokhna, Ndiath, and Rogier 2013). Physiological resistance refers to biochemical and morphological mechanisms (e.g. target-site modifications, metabolic resistance, cuticular thickness) that enable the mosquito to withstand the effects of insecticide despite its contact with it (Davidson 1957). Among the physiological resistances, the target-site mutations L1014F (*kdr-w*) (Martinez-Torres et al. 1998), L1014S (*kdr-e*) (Ranson et al. 2000), and G119S (*ace-1*) (Weill et al. 2004), conferring insecticide resistance to pyrethroids (*kdr-w* and *kdr-e*) and to carbamates and organophosphates (*ace-1*), have been extensively described. Behavioral resistance, on its side, refers to any modification of mosquito behavior that facilitates avoidance or circumvention of insecticides (Gatton et al. 2013; Riveron et al. 2018; Carrasco et al. 2019). Behavioral resistance of mosquitoes to insecticides can be qualitative (i.e. modifications that prevent or limit the contact with the insecticide) or quantitative (i.e. modifications that stop, limit or reduce insecticide action once contact has occurred, e.g. escaping, behavioural thermoregulation or curative self-medication) (Carrasco et al. 2019). Up-to-date, the behavioral resistance mechanisms described in the literature are mainly qualitative and consist in spatial, temporal, or trophic avoidance. In particular, in the anopheline populations, the following behavioral qualitative resistance mechanisms have been described after the scale-up of insecticide-based VC tools (Durnez and Coosemans 2013) : i) increase of exophagic or exophilic behaviours (spatial avoidance), where mosquitoes shifted from biting or resting indoor to outdoor, ii) increase of early- or late-biting behaviours (temporal avoidance), where mosquitoes shifted from biting at night to earlier in the evening or later in the morning, iii) increase of zoophagic behaviours (trophic avoidance), where mosquitoes shifted from biting on humans to biting on animals.

To help develop and deploy complementary VC strategies that are efficient and cost-effective, a better understanding of the spatiotemporal distribution and drivers of both vector physiological resistance and feeding behaviour is needed at a local scale. We raise here a set of questions that, among others, must be explored further at local scale towards this aim :

> *What is the respective contribution of public health and agriculture in the development of physiological resistances in Anopheles vectors ?* The molecular and genetic basis of physiological resistance has been widely acknowledged: under the pressure of insecticides, mutations that enable the vectors to survive are naturally selected and then spread over the generations (Martinez-Torres

et al. 1998; Labbé et al. 2017). The main force that governs the development of a physiological mechanism of resistance in a population of insects is therefore the selective pressure induced by insecticide exposure. This pressure can be induced by the vector control tools, or by the runoff of pesticides used in agriculture (in many cases, the same as those used for impregnation of bed nets) into the malaria vectors breeding sites (Chandre et al. 1999; Hien et al. 2017; Yadouleton et al. 2011; Reid and McKenzie 2016). Assessing the relative contribution of these two selection pressures on the development of resistant phenotypes is critical to further predict the relative impacts of public health and agriculture on the growth of physiological resistances and act consequently.

> *What are the biological mechanisms underlying behavioural resistances ?* Contrary to physiological resistance, the biological mechanisms underlying behavioral resistance are still poorly known (Main et al. 2016; Carrasco et al. 2019; Durnez and Coosemans 2013; Killeen 2014). In particular, a pending question, having important implications for vector control, is whether behavioural shifts reflect evolutionary adaptations in response to selection pressures from vector control tools, as for physiological resistances (*constitutive resistance*) or are manifestations of pre-existing phenotypic plasticity which is triggered when facing the insecticide or in response to environmental variation that reduces human host availability (*inducible resistance*). Inducible resistance imply that vectors rapidly revert to baseline behaviours when VC interventions are lifted, whereas constitutive resistance might progressively and durably erode the effectiveness of current VC tools. Understanding the biological mechanisms underlying behavioural resistances is therefore important to assess the long-term efficacy of insecticide-based VC interventions.

> *Are mosquito biting behaviours modulated by local-scale environmental conditions other than insecticide-related ones ?* As aforementioned, the overall rise of behavioral resistances is likely caused by the widespread of insecticide-based vector control interventions. However, local environmental conditions can modulate vector behaviours at the time of foraging activity. Local climatic conditions – e.g. wind, rain, temperature, humidity, luminosity - may for example affect the timing and location of vector biting, as it has been noted in some studies (Ngowo et al. 2017; Kreppel et al. 2020; Kirby and Lindsay 2004). Mosquitoes with natural endophagic / endophilic preferences might, for example, bite or rest outside if temperature inside is too high or humidity too low, in order to decrease their risk of desiccation-related mortality (Ngowo et al. 2017; Kreppel et al. 2020). Land cover, as well, can affect biting rhythms. It has been noted for example that distance to breeding sites may influence nocturnal host-seeking behaviour, with vectors biting on average earlier in the night in households located close to the breeding sites (Njan Nloga et al. 1993; Snow and Gilles 2002). Assessing whether and to which extent behavioural resistance traits are influenced by local environmental (climatic or landscape) settings may help design VC tools exploiting the vulnerabilities of vectors.

> *Are there associations between behavioural and physiological resistances ?* Physiological and behavioral resistances may likely coexist in mosquito populations, with the first possibly influencing the second. In fact, physiologically resistant mosquitoes may, theoretically, use the recognition of insecticide-based control tool as a proxy for host presence (framed as *behavioural exploitation* (Carrasco et al. 2019)). Several studies have actually showed that the kdr mutation can modify the

host-seeking or biting behavior of *Anopheles* in presence of insecticide-treated net (Malal M. Diop et al. 2015; Porciani et al. 2017; Malal M. Diop et al. 2021). Such behavioural exploitation could potentially lead to a better host recognition/localization and have a dramatic impact, with the control intervention having the opposite effect to the one expected. It is hence important to assess if and to which extent physiologically resistant mosquitoes exhibit different biting behaviours than their susceptible counterparts.

> *Which adaptative strategy (physiological or behavioural resistance) arises faster ?* Understanding the relative capacity of mosquitoes to develop physiological resistance and to shift their behaviour in response to vector control is necessary to highlight where and when mitigation efforts should be prioritized (Sanou et al. 2021). After introduction / re-introduction of insecticide-based tools, if vectors rapidly shift their behaviour to feed outside or at times when people are not protected by an LLIN, interventions that target such mosquitoes should be quickly deployed. In contrast, the rapid emergence of physiological resistance in vectors who continue to feed indoors and at night indicates that switching to alternative insecticide classes in indoor-based interventions may have a greater impact. Additionally, for a given environment, assessing the relative rate of development of physiological and behavioral resistances is of direct epidemiological importance : it has been showed for example that under a scenario where LLIN and IRS are both heavily used, changes in the susceptibility to insecticide is likely to have a bigger epidemiological impact than changes in biting times (Sherrard-Smith et al. 2019).

> *Are resistance rates heterogeneous at small spatiotemporal scales ?* Mosquito presence and abundance has already been found heterogeneous in space and time at fine-scale, calling for locally-tailored (species-, season-, and village-specific) control interventions (Moiroux et al. 2013, 2014; Taconet et al. 2021). However, little is known about the small-scale spatiotemporal heterogeneity of vector resistance. The potential drivers of the development or triggering of resistant phenotypes (vector control use, land cover, micro-climate, human behaviour, etc.) are likely to vary at small spatiotemporal scales, and so may, at the end of the line, vector resistance. As for abundances, assessing the level of heterogeneity of resistance rates in space and time is important to assess the spatiotemporal scale at which management of vector resistance should be considered.

> *To what extent can we explain and predict vector resistance and biting behaviour in space and time ?* Assessing the levels of explainability and predictability of vector resistance and biting behaviour is important for both scientific and operational purposes. Towards this aim, generating statistical models linking vector resistances or biting behaviours to their potential drivers and assessing their explanatory and predictive powers can help (Shmueli 2010; Shmueli and Koppius 2010). High explanatory or predictive powers in the models might suggest that the conditions driving a vector to resist are well understood, and conversely, low explanatory powers might suggest that resistances are driven by factors either yet undiscovered or not included in the models. Additionally, assessing the predictability of resistances in vector populations in space and time is an important step towards mapping vector resistance at every place (e.g. village) and time (e.g. season) in the area, with such decision-support tools important to deploy the right intervention, at the right place and time

150

(Taconet et al. 2021).

In this study, we used field mosquito collections and environmental data collected simultaneously in two rural areas of West-Africa to bring elements of answer to these questions for our areas. Guided by these questions, our overall objectives were i) to assess the fine-scale prevalence and spatiotemporal heterogeneity of physiological resistances and at-risk biting behaviours of malaria vectors in these areas and ii) to better understand their drivers. To do so, we modeled a set of indicators of physiological resistances and biting behaviours (namely exophagy, early-biting, late-biting, *kdr-w*, *kdr-e*, and *ace-1* target-site mutations) at the individual mosquito level using this fine-grained dataset and advanced statistical methods in an exploratory and data-driven approach. Patterns found in the data were interpreted and discussed in light of the topics aforementioned, of importance for the management of malaria residual transmission. We concluded with a set of recommendations to manage vector resistances in our study areas.

Material and methods

Entomological and environmental data

The data used in this work were collected in the frame of the REACT project (Soma et al. 2020; Zogo et al. 2019). In this projet, a total of fifty-five villages, distributed in two West-African rural areas (~ 2500 km² each) located in the areas of Diébougou (southwestern Burkina Faso (BF)) and Korhogo (northern Ivory Coast (IC)) were selected according to the following criteria: accessibility during the rainy season, 200–500 inhabitants per village, and distance between two villages higher than two kilometers. After an exhaustive census of the population in these villages at the beginning of the project, entomological and human behaviours surveys were regularly conducted during 15 months (1.25 year) in the Diébougou area and 18 months (1.5 year) in the Korhogo area. Vector control interventions were implemented both as part of the project and of the national malaria control programs (see below). Figure 1 shows the study areas and the corresponding timelines for data collection and vector control interventions. Entomological data were collected in the field, and environmental data were collated from specific devices (see below) or created from heterogeneous field and satellite-based sources. Below is a description of the data used in our work.

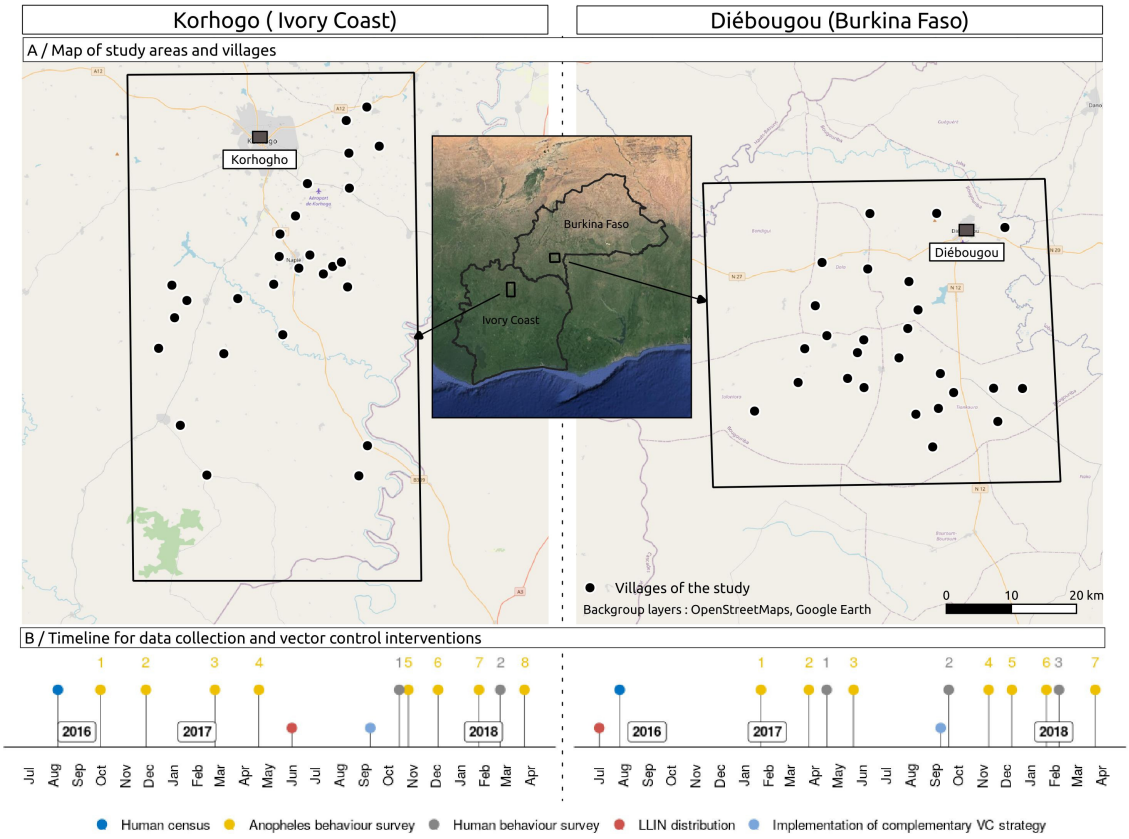


Figure 1: Villages of the study and timeline for data collection + vector control

> *Anopheles collections*

Several rounds of mosquito collections (eight in the Korhogo (IC) area, seven in the Diébougou (BF) area) were conducted in each village. The periods of the surveys span the typical climatic conditions of these tropical areas (except the high rainy season - July to September) (see Additional file 1.A for the spatiotemporal trends of the meteorological conditions). Mosquitoes were collected using the Human Landing Catch (HLC) technique from 17:00 to 09:00 both indoors and outdoors at four sites per village (i.e. eight collection points) for one night during each survey. Malaria vectors were identified using morphological keys. All individuals belonging to the *Anopheles funestus* group (in both study areas) and *Anopheles gambiae s.l.* complex (in BF) were identified to the species level using PCR. In IC, due to the very large numbers of *An. gambiae s.l.* vectors collected, a sub-sample only of these individuals (randomly selected in space and time) was identified to species. Finally, PCR assay were carried out on all the *An. gambiae s.s.* and *An. coluzzii* collected in BF to detect the L1014F (*kdr-w*), the L1014S (*kdr-e*) and the G119S (*ace-1*) target-site mutations. Detailed descriptions of the methods used to obtain these data are provided in Soma et al. (2020); Zogo et al. (2019).

> *Weather preceding mosquito collections and during mosquito collections*

Weather can impact the fitness or the activity of resistant genotypes (Kliot and Ghanim 2012),
152

as well as the biting behaviour of the mosquitoes (see Introduction). In this work, we recorded or retrieved weather conditions : (i) during mosquito collections (i.e. the HLC sessions), (ii) during the day of collection, and (iii) during the month preceding collection. Weather on the day of collection and during mosquito collection may impact the relative activity of each genotype and phenotypes associated with resistances. Weather during the month preceding the survey, on its side, can impact development and survival rates of both the current and parental generations of collected mosquitoes. Regarding our outputs (prevalence of behavioral phenotypes and target-site mutations - see next section), weather during the month preceding collection may therefore impact the fitness of the studied genotypes (for target-site mutations) or possible – and unknown - genotypes associated with studied behavioral phenotypes.

Micro-climatic conditions (temperature, relative humidity, luminosity and atmospheric pressure) were simultaneously recorded where mosquito collections were being conducted. Instruments used to record these data were : for temperature and relative humidity : Hygro Buttons 23 Data Loggers [Proges Plus DAL0084] (temporal resolution (TR): 15 minutes) ; for luminosity : HOBO Pendant® Temperature/Light 8K Data Logger (TR: 15 minutes) ; for atmospheric pressure : Extech SD700 Data Loggers (TR : 10 minutes). Hygro and Hobo loggers were positioned both inside and outside the houses where mosquito sampling was conducted, close to the sampling positions. The barometer was positioned at the center of the village. These field data were completed with satellite or modeled data available at coarser spatial resolutions : rainfall (spatial resolution (SR) : ~ 11 km, TR : 30 min, source : Global Precipitation Measurement (GPM) IMERG (At NASA GSFC 2019), wind speed (SR : ~ 28 km,TR : 1h, source : ERA5 (Hersbach et al. 2020)), apparent magnitude of the Moon (SR : 0.001°, TR : 1 day, source : Institute of celestial mechanics and ephemeris calculations).

Meteorological conditions on the day of collection and during the month preceding collection were extracted from satellite imagery. Namely, rainfall estimates were extracted from the GPM - IMERG daily Final products (Center 2019). Diurnal and nocturnal temperatures were derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) daily Land Surface Temperature (LST) Terra and Aqua products (Wan, Hook, and Hulley 2015b, 2015a). Rainfall and temperatures were averaged in a 2 km buffer zone around each HLC collection point to create meteorological variables on both the day of collection and the month preceding collection. Detailed descriptions of the methods used to collect and process these data are provided in Taconet et al. (2021).

> *Host availability and human behaviour*

The nocturnal behaviour of humans (hours inside the dwellings, hours of use of LLINs) drives host availability for the mosquitoes and can therefore impact their behaviour. For instance, high LLIN use rate can drive mosquitoes to feed outside, at times when people are not protected, or on alternative sources of blood (Durnez and Coosemans 2013). Here, human population was counted in each village, through an exhaustive census conducted at the beginning of the project. Then, several human behavioural surveys (two in IC, three in BF) were carried out in each village (see Figure 1). For each survey and village, several households (mean = 14 , SD = 2) were randomly selected, and for

each household, one to three persons in each age class (0–5 years old, 6–17 years old and ≥ 18 years old) were selected. The head of the household was then asked, for each selected person, on the night preceding the survey : i) whether he/she used an LLIN or not, ii) the time at which he/she entered and exited his own house, and iii) the time at which he/she entered and exited his LLIN-protected sleeping space (where appropriate). Households for human behavioural surveys were independently selected from households for entomological surveys. The surveys were conducted after the distribution of the LLINs (see below), and span the typical climatic conditions of the areas. Detailed descriptions of the methods used to collect these data are provided in Soma et al. (2021).

> *Landscape*

Landscape can have an impact on mosquito foraging behaviour (e.g. the distance to breeding sites can impact biting rhythms) or physiological resistance (e.g. through pesticides used in crops) (see Introduction). Digital land cover maps were produced for each study area by carrying out a Geographic Object-Based Image Analysis (Hay and Castilla 2008) using multisource very high and high resolution satellite-derived products. From these maps, several variables were derived : the percentage of landscape occupied respectively by cotton fields, by rice fields, and by the other crops (mainly leguminous crops, millet, sorghum) in a 2 km buffer size area around each collection point ; and the distance to the nearest stream (as a proxy for the distance to the breeding site). For cotton, the variable was binarized as presence / absence of cotton cultivated due to the small range of values. In addition, the geographical location of the households was recorded, and used to derive two indices : the degree of clustering of the households in each village, and the distance from each collection point to the edge of the village. Detailed descriptions of the methods used to collect, generate and process these data are provided in Taconet et al. (2021).

> *Vector control*

Repeated exposure to insecticides used in vector control interventions is undoubtedly one of the most important drivers of the development of resistance (see Introduction). In both Burkina Faso and Ivory Coast, LLINs have been universally distributed every 3-4 years since 2010 (PNLP 2014b, 2014a). In BF, a mass distribution of LLINs (PermaNet 2.0) was carried out by the National Malaria Control Program in July 2016 (i.e. 6 months before our first entomological survey). In IC, our team distributed LLINs in the villages of the project in June 2017 (i.e. height month after the first entomological survey and ten months before the last one). Complementary VC tools were implemented in some of the villages in the middle of the project - namely IRS, ivermectin to peri-domestic animals (IVM), intensive Information Education and Communication to the populations (IEC), and larval control (Larv.) (see Figure 1 and Additional file 1) - as part of a randomized controlled trial aiming at assessing the benefits of new, complementary VC strategies (Soma et al. 2020; Zogo et al. 2019).

Statistical Modeling

Dependent and independant variables

Six vector potential resistance indicators were modelled :

- three indicators of physiological resistance to insecticide : *kdr-w* mutation, *kdr-e* mutation, *ace-1* mutation,
- three indicators of biting behaviours : early biting, late biting, exophagy. **Here, it is unknown whether these behaviours are real *behavioural resistances* as defined in the introduction (i.e. induced by the use of insecticides) or are natural behaviours (i.e. behaviours that exist naturally in vector populations). However, to harmonise the vocabulary used and facilitate the reading, in the remainder of this manuscript, we will qualify these three phenotypes, associated with resistant behaviours, as 'behavioural resistance'.**

Exophagy was defined as the probability for a host-seeking mosquito to bite outdoor (as opposed to indoor). Early biting was defined as the probability for a host-seeking mosquito to bite before 50 % of the LLIN users were declared to be under their bednet in the evening, and late biting was defined as the probability for a host-seeking mosquito to bite after 50 % of the LLIN users were declared to be out of their bednet in the morning (based on the closest - in space and time - human behaviour survey). *Kdr-w*, *kdr-e* and *ace-1* mutations were defined as the probabilities for an allele of a host-seeking mosquito to be mutated (as opposed to wild type). The statistical unit was therefore the mosquito for biting behaviour models and the allele for physiological resistance models. Dependent variables were all binary (0 = absence of resistance/mutation, 1 = presence of resistance/mutation) and models outcomes were probabilities for a mosquito (resp. allele) to be resistant (resp. mutated). Each indicator was modeled separately for each main species in each study area, as determinants of resistance might be species- or site-specific (i.e. mosquitoes might respond differently to environmental variations depending on the species and study area, due to potential local chromosomal forms, adaptation, etc.) (Riveron et al. 2018; Durnez and Coosemans 2013). As three main species were found in BF and two in IC (see Results section), a total of twenty-one dependent variables were built (exophagy : 3 in BF and 2 in IC ; early biting : 3 in BF and 2 in IC ; late biting : 3 in BF and 2 in IC ; *kdr-w* : 2 in BF ; *kdr-e* : 2 in BF ; *ace-1* : 2 in BF). Based on literature (see Introduction) and available data, we then built independent variables representing potential determinants of each of these resistant phenotypes. These variables are provided in Table 1. To build these variables, the source data were possibly aggregated in space or time, at varying resolutions depending on the considered dependent variable.

Table 1 : *Dependent and independent variables for the resistance models. Each bullet point in the 'Independant variables' columns represents an independent variable built and used for the considered model. These variables belonged to several "families" of potential drivers : vector control, vector resistance, host availability, micro-climate at the time of collection, meteorological conditions preceding collection, landscape, and vector abundance. Some independent variables, mentioned as 'BF only', were available only for the BF study area.*

Resistance type	Model name	Dependent variable : Statistical unit & values	Independant variables							
			Vector control	Host availability	Vector resistance / behaviour	Micro-climatic conditions during collection	Meteorological conditions on the day of collection	Meteorological conditions on the month preceding collection	Landscape & crops	Others
Behavioural resistance	**Exophagy**	**Mosquito** **1 = exophagic** **0 = endophagic**	• Vector control intervention implemented • Time since distribution of LLIN *(BF)* / Time since first entomological survey *(IC)*	• Average LLIN use rate in the village on the season of collection • Human population in the village • % of population indoor in the village on the hour of collection • % of the population under an LLIN in the village on the hour of collection	• Kdr-w mutation alleles on the collected mosquito *(BF only, An. gambiae s.s. and An. coluzzii)* • Kdr-e mutation alleles on the collected mosquito *(BF only, An. gambiae s.s. and An. coluzzii)*	*At the hour of collection :* • Temperature indoor • Humidity indoor • Luminosity outdoor • Atmospheric pressure • Relative temperature difference b\w indoor and outdoor • Relative humidity difference b\w indoor and outdoor • Relative luminosity difference b\w indoor and outdoor • Presence / absence of rainfall • Wind speed outdoor • Apparent magnitude of the Moon	• Diurnal temperature • Cumulated rainfall	• Avg. diurnal temperature • Avg. nocturnal temperature • Cumulated rainfall	• Distance from the collection point to the edge of the village • Distance from the collection point to the nearest stream • Degree of clustering of the households in the village	
	Early biting	**Mosquito** **1 = early biter** **0 = all other**		• Average LLIN use rate in the village on the season of collection • Human population in the village	• same as above • Place of collection of the collected mosquito (indoor / outdoor)					
	Late biting	**Mosquito** **1 = late biter** **0 = all other**				*During the night of collection :* • Avg. temperature • Avg. humidity • Avg. luminosity • Avg. atmospheric pressure • Cumulated rainfall • Avg. wind speed outdoor				
Physiological resistance ***(BF area only)***	**Kdr•w mutation**	**Allele** **1 = resistant** **0 = sensible**		• Average LLIN use rate in the village on the season of collection • Human population in the village • % of population indoor in the village on the hour of collection • % of the population under an LLIN in the village on the hour of collection		*At the hour of collection :* • Temperature • Humidity • Luminosity • Atmospheric pressure		• Avg. diurnal temperature • Avg. nocturnal temperature • Cumulated rainfall	*In a 2 km buffer area around the collection points :* • Presence / absence of cotton fields • % landscape occupied by rice fields • % landscape occupied by other crops	• Number of mosquitoes collected during the night of collection
	Kdr•e mutation				• Kdr-w mutation alleles on the collected mosquito					
	Ace•1 mutation				• Kdr-w mutation alleles on the collected mosquito • Ace-1 mutation alleles on the collected mosquito					

Modeling workflow

A graphical representation of the modeling workflow (explained below) is available in Additional figure 2.

Pre-processing. Before modeling, we excluded the dependent variables that had too few 'resistant' observations, according to the following criteria : 'resistant' class ≤ 50 observations & ≤ 3% of the total observations. Next, we implemented the modeling workflow described below for each remaining dependent variable.

Bivariate modeling. We first excluded the independent variables that were poorly associated with the dependent variable (criteria for exclusion : p-value > 0.2 of a bivariate Generalized Linear binomial Mixed-effect Model (GLMM) with nested random effects at the village and collection site level). Next, we filtered-out collinear variables (Pearson correlation coefficient > 0.7) based on empirical knowledge. With the set of remaining independent variables, two distinct multivariate models were built, with complementary objectives, as explained in the Box 1 below.

Multivariate modeling part 1 : Explanatory model. A binomial GLMM was fitted to the data. Nested random effects were introduced in the model at the village and collection place level. Variables were deleted recursively using an automatic backward variable selection procedure based on the reduction of the Akaike Information Criterion (AIC). Variables belonging to the 'vector control' (for all resistance models) and 'crops' (for physiological resistance models only) groups were forced in the multivariate models (i.e. they were not filtered-out in the variable selection procedure) because there are strong *a priori* assumptions associated with these variables. Such variable selection procedure therefore retained all the 'vector control' and 'crops' variables (whether significantly associated or not with the dependent variable), and the additional variables that decreased the AIC of the multivariate model.

Multivariate modeling part 2 : Predictive model. We additionally fitted a Random Forest (RF) model (Breiman 2001a) to the data. The model hyperparameters were optimized using a random 5-combinations grid search (Chicco 2017). Whenever the dependent variable was imbalanced (more than 1/3 imbalance ratio between the positive and negative class), data were up-sampled within the model resampling procedure to cope with well-known problems of machine-learning (ML) models regarding class imbalance (Tyagi and Mittal 2020).

Assessment of effect sizes and significance of independent variables. To interpret the effect of each independent variable in the GLMM model, we plotted, for each independent variable retained in the final model, the predicted probabilities of resistance across available values of that variable (all other things being equals) (i.e. "partial dependence plot" (PDP) (Friedman and Popescu 2008)). For reporting and discussion in the manuscript, we kept only variables that had a p-value < 0.05 (results containing the 'full' models are provided in supplementary material, see Results section). To uncover the possible complex relationships that the RF model had learned, we generated smoothed versions of PDPs for each independent variable. However, we restricted the generation of PDPs to the following cases : i) the Area Under the Receiver Operating Characteristics (AUC) (see below) of the model was > 0.6 (because model interpretation tools of ML models (e.g. PDPs) should be

157

trusted only if the predictive power of the underlying model is good enough (Zhao and Hastie 2021)) and ii) the range of predicted probabilities of resistance was > 0.05 (i.e. the independent variable, over its range of available values, changed the probability of resistance by at least 5 percentage points).

Assessment of models performance. The explanatory power of the GLMM was assessed by calculating the marginal coefficient of determination (R^2) (Nakagawa and Schielzeth 2013) from the observed and in-sample predicted values. Marginal R^2 is a goodness-of-fit metric that measures the overall variance explained by the fixed effects in the GLMM. R^2 values were interpreted according to (Cohen 2013) criteria : $R^2 \in \{0; 0.02\{$: very weak ; $R^2 \in \{0.02; 0.13\{$: very weak ; $R^2 \in \{0.02; 0.13\{$: weak ; R^2 i$\in\{0.13; 0.26\{$: moderate ; $R^2 \in \{0.26; 1\{$: substantial. The predictive power of the RF model was assessed by leave village - out cross-validation (CV), with the AUC chosen as the performance metric. This CV strategy consisted in recursively leaving-out the observations belonging to one village of collection (i.e. the validation fold), training the model with the observations coming from the other villages (i.e. the training fold), and predicting on the left-out set of observations. We hence measured the ability of the model to predict resistance status ('resistant' or 'non-resistant') on individual mosquitoes caught on new - unseen villages of collection. AUC values were interpreted according to the following criteria : AUC $\in \{0.5; 0.6\{$: very weak ; AUC $\in \{0.6; 0.65\{$: weak ; AUC $\in \{0.65; 0.75\{$: moderate ; AUC $\in \{0.75; 1\{$: substantial.

Box 1 : **Why modeling with both logistic regression and random forests ?**

"Data" models (Breiman 2001b) - like linear of logistic regression - and "algorithmic" models (Breiman 2001b) – so-called *machine learning* models like random forests or support vector machines - play complementary roles in scientific theory building and testing. On the one hand, data models are useful to test existing theories and to reach to "statistical" conclusions about causal relationships that exist at the theoretical level, e.g. : vector control significantly impacts vector resistance (or not). These models are hence useful in explanatory modeling, where the main objective is to apply statistical models to data for testing causal hypotheses about theoretical constructs (Shmueli 2010). On the other hand, algorithmic models, thanks to their ability to inherently capture complex and hidden patterns contained in the data (e.g. interactions, non-linear associations), are useful for assessing the relevance of a theory and for developing new theories or improve existing ones, through predictive analytics (Shmueli and Koppius 2010). Predictive analytics is the process of building a model aimed at making empirical predictions and then assessing its predictive power. Shmueli and Koppius (2010) describe six roles for predictive analytics in theory building and testing, including generation of new theory, improvement of existing theory, assessment of relevance of a theory, and assessment of level of predictability of a phenomena. In a "big data" context such as that of this study (large datasets, containing lots of observations and variables), predictive analytics is every time more used to support scientific theory development through these roles (Karpatne et al. 2017; Breiman 2001b; Shmueli and Koppius 2010).

In our study, we used GLMMs to i) test whether vector control significantly increases vector

158

resistance, as could be expected, and ii) possibly find environmental factors that impact vector resistance in a statistically significant way. We used RFs to i) account for potential unhypothesized, complex relationships (e.g. interactions, non-linear associations) between independent and dependent variables in the statistical modeling process, ii) possibly reveal part of the complex relationships learned by the model (through PDPs), iii) assess distance between theory (potential / identified determinants of vector resistance) and practice (are they enough to predict resistance on unseen mosquitoes ? If not, why ?).

Software and libraries used

The softwares used in this work were exclusively free and open source. The R programming language (R Core Team 2018) and the R-studio environment (RStudio Team 2020) were used as the main programming tools. The QGIS software (QGIS Development Team 2021) and the 'ggplot2' R package (Wickham 2016) were used to create respectively the map of the study area and the timeline for data collection. The 'glmmTMB' (Brooks et al. 2017) package was used for the bivariate modeling. The 'buildmer' package (Voeten 2020) was used to fit the GLMM models with stepwise selection in the multivariate modeling. The 'caret' (Wing et al. 2018) and 'ranger' (Wright and Ziegler 2017) packages were used were to fit the random forest models in the multivariate modeling. The 'MLmetrics' (Yan 2016) and 'MuMIn' (Bartoń 2020) packages were used to calculate respectively the AUC of the RFs and the marginal R^2 of the GLMMs. The 'jtools' (Long 2020) and 'pdp' (Greenwell 2017) packages were used to generate the partial dependence plots of respectively the GLMMs and the RFs. The 'broom.mixed' (Bolker and Robinson 2020) package was used to extract the coefficients / odd ratios, confidence intervals and p-values of the multivariate GLMMs. The 'patchwork' (Pedersen 2019) and 'gridExtra' (Auguie 2017) packages were used to create various plot compositions. The 'tidyverse' meta-package (Wickham 2017) was used throughout the entire analysis. A few additional R packages were used to create, tidy, and transform the data used in this work (see (Taconet et al. 2021)). The LibreOffice suite was used to create the tables and other plot compositions.

Results

Spatio-temporal heterogeneity of vector abundance

In the Korhogo area (IC), a total of 2048 human-nights of HLC was conducted. A sum of 57722 vectors belonging to the *Anopheles* genus was collected. The main species/complex found were *An. gambiae s.l.* and *An. funestus* (respectively 56267 (97% of all the *Anopheles* collected) and 714 (1%) individuals collected). Among the 56267 *An. gambiae s.l.* collected, 3922 (7%) were identified to species: 3726 (95% of the individual identified to species) were *An. gambiae s.s.* and 196 (5%) were *An. coluzzii*. Hence, in the rest of this article, we will consider the *An. gambiae s.l.* collected in the Korhogo area as *An. gambiae s.s.* In the Diébougou area (BF), a total of 1512 human-nights of HLC was conducted. A sum of 3056 vectors belonging to the *Anopheles* genus was collected. The main species found were *An. coluzzii*, *An. gambiae s.s.* and *An. funestus* (respectively 1321 (43% of all the *Anopheles* collected), 616 (20%) and 708 (23%) individuals collected). As expected, mosquito abundance was heterogeneous in time and

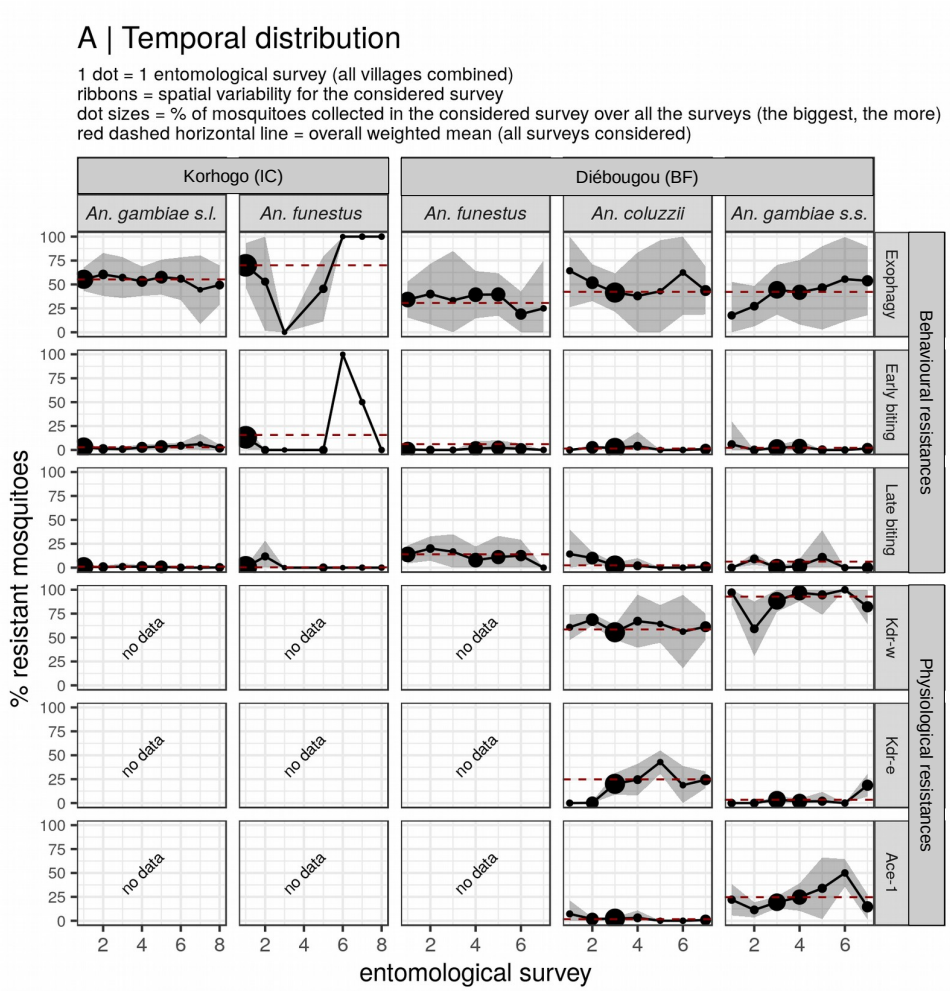
space (except for *An. funestus* in IC, for which the vast majority (93 %) of the individuals was collected in the first entomological survey, and almost half of the individuals (42 %) were collected within one single village) (see additional file 1 and additional figure 3 for maps and charts of the spatiotemporal distribution of vector abundance).

Spatio-temporal heterogeneity of vector resistance

Table 2 and Figure 2 show, respectively, global and spatiotemporal descriptive statistics on the resistances of the main vector species collected in the two areas.

Resistance type	Resistance indicator	Study area	Species	n° collected	n° resistant	% resistant	Temporal confidence interval & range	Spatial confidence interval & range
Behavioural resistance	Exophagy (rate)	Korhogo (IC)	*An. gambiae s.l.*	56267	31295	56 %	± 2 % [54 – 58] (44 – 60)	± 7 % [49 – 63] (38 – 71)
			An. funestus	714	493	69 %	± 7 % [62 – 76] (0 – 100)	± 16 % [53 – 82] (0 – 100)
		Diébougou (BF)	*An. coluzzii*	1321	577	44 %	± 5 % [39 – 49] (38 – 64)	± 9 % [35 – 53] (0 – 100)
			An. gambiae s.s.	616	268	44 %	± 7 % [37 – 51] (18 – 56)	± 12 % [32 – 56] (0 – 75)
			An. funestus	708	250	35 %	± 6 % [29 – 41] (19 – 40)	± 8 % [27 – 43] (0 – 100)
	Early biting (rate)	Korhogo (IC)	*An. gambiae s.l.*	56267	1670	3 %	± 1 % [2– 4] (1 – 6)	± 2 % [1 – 5] (0 – 10)
			An. funestus	714	92	13 %	± 6 % [7 – 19] (0 – 100)	± 12 % [1 – 25] (0 – 100)
		Diébougou (BF)	*An. coluzzii*	1321	28	2 %	± 1 % [1 – 3] (0 – 4)	± 2 % [0 – 4] (0 – 75)
			An. gambiae s.s.	616	19	3 %	± 1 % [2 – 4] (0 – 6)	± 3 % [0 – 6] (0 – 14)
			An. funestus	708	9	1 %	± 1 % [0 – 2] (0 – 2)	± 4 % [0 – 5] (0 – 100)
	Late biting (rate)	Korhogo (IC)	*An. gambiae s.l.*	56267	499	1 %	± 0 % [1 – 1] (0 – 1)	± 1 % [0 – 2] (0 – 9)
			An. funestus	714	4	1 %	± 1 % [0 – 2] (0 – 12)	± 1 % [0 – 2] (0 – 7)
		Diébougou (BF)	*An. coluzzii*	1321	46	3 %	± 3 % [0 – 6] (0 – 14)	± 3 % [0 – 6] (0 – 14)
			An. gambiae s.s.	616	8	1 %	± 3 % [0 – 4] (0 – 11)	± 5 % [0 – 6] (0 – 100)
			An. funestus	708	82	12 %	± 3 % [9 – 15] (0 – 22)	± 10 % [2 – 22] (0 – 100)
Physiological resistance *(BF only)*	Kdr-w mutation (allelic frequency)	Diébougou (BF)	*An. coluzzii*	1321	*NA*	59 %	± 5 % [55 – 64] (55 – 69)	± 8 % [51 – 67] (12 – 100)
			An. gambiae s.s.	616	*NA*	90 %	± 8 % [82 – 98] (59 – 100)	± 9 % [81 – 99] (68 – 100)
	Kdr-e mutation (allelic frequency)		*An. coluzzii*	1321	*NA*	17 %	± 8 % [9 – 25] (0 – 43)	± 10 % [7 – 27] (0 – 50)
			An. gambiae s.s.	616	*NA*	4 %	± 4 % [0 – 8] (0 – 19)	± 4 % [0 – 8] (0 – 17)
	Ace-1 mutation (allelic frequency)		*An. coluzzii*	1321	*NA*	2 %	± 1 % [1 – 3] (0 – 7)	± 1 % [1 – 3] (0 – 6)
			An. gambiae s.s.	616	*NA*	21 %	± 6 % [15 – 27] (11 – 50)	± 8 % [13 – 29] (0 – 75)

Table 2 : *Descriptive statistics for the resistances of the main vector species collected. The columns 'Temporal variability and range' and 'Spatial variability and range' provide indicators of the variability and range of resistance around the overall mean (% resistant) respectively in time (i.e. variability between the entomological surveys) and space (i.e. variability between the villages). Format of these columns: standard deviation [|mean - standard deviation| – mean + standard deviation] (minimum – maximum). Computation of standard deviation : to take into account the uneven sample size between entomological surveys (resp. villages) (i.e. to avoid excessive consideration of small / very small sample size), standard deviations for temporal (resp. spatial) variability were extracted by first calculating the resistance indicator at the entomologial survey (resp. village) level and then computing the weighted standard deviation (weights = number of mosquitoes collected in each entomologial survey (resp. village)).*


A | Temporal distribution
1 dot = 1 entomological survey (all villages combined)
ribbons = spatial variability for the considered survey
dot sizes = % of mosquitoes collected in the considered survey over all the surveys (the biggest, the more)
red dashed horizontal line = overall weighted mean (all surveys considered)
Korhogo (IC)
Diébougou (BF)
An. gambiae s.l.
An. funestus
An. funestus
An. coluzzii
An. gambiae s.s.
Exophagy
Early biting
Late biting
Kdr-w
Kdr-e
Ace-1
Behavioural resistances
Physiological resistances
no data
% resistant mosquitoes
entomological survey

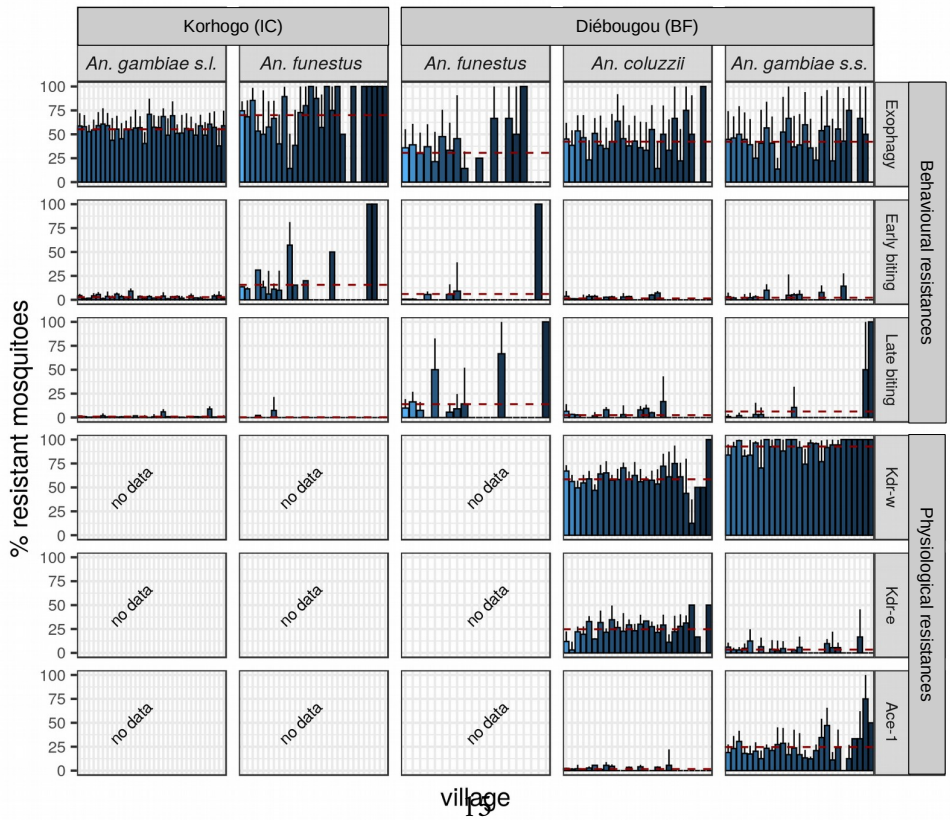

B | Spatial distribution
1 bar = 1 village (all entomological surveys combined)
error bars = temporal variability for the considered village
bar colors = % of mosquitoes collected in the considered village over all the villages (the brighter, the more)
red dashed horizontal line = overall weighted mean (all villages considered)
Korhogo (IC)
Diébougou (BF)
An. gambiae s.l.
An. funestus
An. funestus
An. coluzzii
An. gambiae s.s.
Exophagy
Early biting
Late biting
Kdr-w
Kdr-e
Ace-1
Behavioural resistances
Physiological resistances
no data
% resistant mosquitoes
village

Figure 2 : *Spatio-temporal distributions of the resistances of the main vector species collected (panel A : temporal distribution, panel B : spatial distribution). For behavioural resistances, the y-axis represents the % of resistant mosquitoes for the considered survey / village. For physiological resistances, the y-axis represents the allele frequency of the considered mutation for the considered survey / village. Confidence intervals (A : ribbons, B : lineranges) provide indicators of variability of the resistance indicator (A : mean ± standard deviation of the resistance indicator calculated at the village level for the considered entomological survey ; B : mean ± standard deviation of the resistance indicator calculated at the entomological survey level for the considered village). To avoid excessive consideration of small sample sizes, the total number of mosquito collected was represented graphically using the size of dots (A) or the color of the bars (B).*

Exophagy rates. In the Korhogo area (IC), the overall exophagy rate (% of bites received outdoor) was 56 % for *An. gambiae s.l.* and 69 % for *An. funestus*. The exophagy rate of *An. gambiae s.l.* varied little, both amongst the entomological surveys and the villages (Temporal Standard Deviation (TSD) (see legend of Table 2 for definition) = ± 2 %, Spatial Standard Deviation (SSD) (see legend of Table 2 for definition) = ± 7 %). The exophagy rate of *An. funestus* was more heterogeneously distributed in time and space (TSD = ± 7 %, SSD = ± 16 %). In the Diebougou area (BF), the overall exophagy rate was 44 % for *An. coluzzii*, 44 % for *An. gambiae s.s.* and 35 % for *An. funestus*. For the three species, the exophagy rate varied quite sensibly among the entomological surveys (TSD = ± 5%, ± 7%, ± 6% respectively) and the villages (SSD = ± 9%, ± 12%, ± 8% respectively).

Early and late biting rates. In the Korhogo area (IC), the early biting rate (i.e. % of bites received before 50% of the LLIN users were declared to be under their bednet at night) was 3% for *An. gambiae s.l.* and 13% for *An. funestus*. The early biting rate was overall stable among the surveys and villages for *An. gambiae s.l.* (TSD = ± 1%, SSD = ± 2%) and was more heterogeneously distributed for *An. funestus* (TSD = ± 6%, SSD = ± 12%). The late biting rate (i.e. % of bites received after 50% of the LLIN users were declared to be out of their bednet in the morning) was lower than the early biting rate : 1% for both *An. gambiae s.l.* and *An. funestus* (only 4 late-bites for *An. funestus*) and was overall stable among the surveys and villages (TSD = ± 0% and SSD = ± 1% for *An. gambiae s.l.*). In the Diébougou area (BF), the early biting rate was respectively 2%, 3% and 1% for *An. coluzzii*, *An. gambiae s.s.* and *An. funestus*. The early biting rate was overall stable among the surveys (TSD = ± 1% for the three species) and to some extent more heterogeneous among the villages (SSD = ± 2%, ± 3%, ± 4% respectively). The late biting rate was respectively 3%, 1% and 12% for *An. coluzzii*, *An. gambiae s.s.* and *An. funestus*. Late biting rates were more heterogeneously distributed than early biting rates, both among the surveys (TSD = ± 3% for the three species) and villages (SSD = ± 3%, ± 5%, ± 10% respectively).

Allele frequencies of kdr-e, kdr-w, ace-1 mutations. In the BF area, the allele frequency of the *kdr-w* mutation was 90% for *An. gambiae s.s.* and 59% for *An. coluzzii*. It varied to some extent among the surveys and villages (for *An. gambiae s.s.* : TSD = 8%, SSD = 9% ; for *An. coluzzii* : TSD = 5%, SSD = 8%). The allele frequency of the *kdr-e* mutation was 4% for *An. gambiae s.s.* and 17% for *An. coluzzii*. For *An. gambiae s.s.*, it remained low among the surveys and villages (TSD = SSD = 4%) and for *An. coluzzii*, it varied more sensibly (TSD = 8%, SSD = 10%). The allele frequency of the *ace-1* mutation was 21 % for *An. gambiae s.s.* and 2% for *An. coluzzii*. For *An. gambiae s.s*, it varied sensibly among the

surveys and villages (TSD = 6%, SSD = 8%), and for *An. coluzzii* it was overall stably low (TSD = SSD = 1%).

Dependent variables excluded from the modeling process

Seven of the original twenty-one dependent variables were excluded before statistical modeling due to the very small size of their 'resistant' class (see Table 2) :

- early-biting in BF for the three species,
- late-biting in BF for *An. coluzzii* and *An. gambiae s.s.*,
- late-biting in IC for *An. funestus*,
- ace-1 in BF for *An. coluzzii*.

Associations between physiological resistance and environmental variables

For the remaining five models of physiological resistance in the Diébougou area (BF), Figure 3 shows the PDPs of the independent variables retained in the modeling workflow. For the GLMMs, numerical values of odd-ratios, 95% confidence intervals, and p-values are provided in Additional file 4.

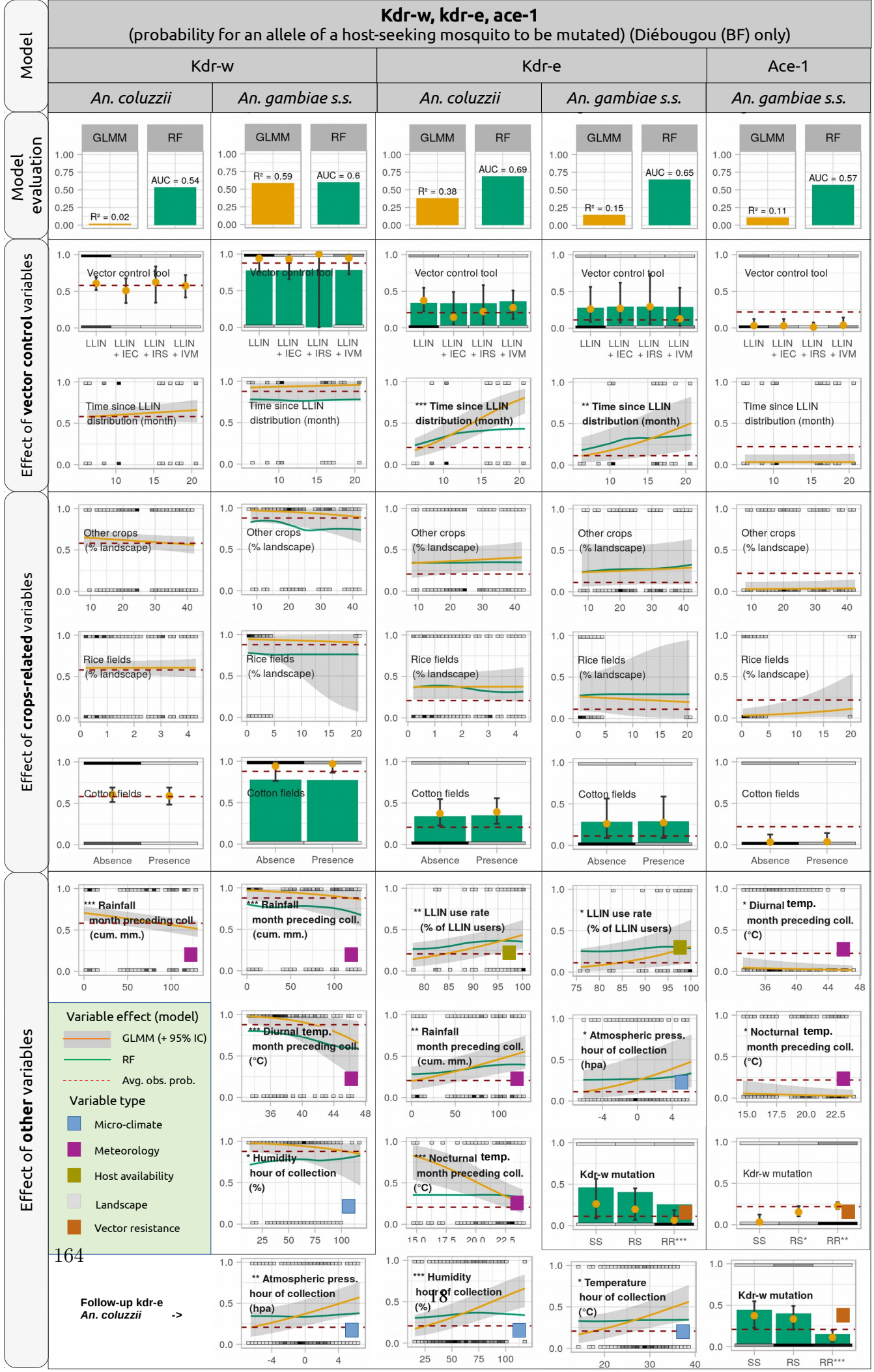

Kdr-w, kdr-e, ace-1
(probability for an allele of a host-seeking mosquito to be mutated) (Diébougou (BF) only)
Kdr-w
Kdr-e
Ace-1
An. coluzzii
An. gambiae s.s.
An. coluzzii
An. gambiae s.s.
An. gambiae s.s.
Model
Model evaluation
Effect of vector control variables
Effect of crops-related variables
Effect of other variables
GLMM
RF
R² = 0.02
AUC = 0.54
R² = 0.59
AUC = 0.6
R² = 0.38
AUC = 0.69
R² = 0.15
AUC = 0.65
R² = 0.11
AUC = 0.57
Vector control tool
LLIN
LLIN + IEC
LLIN + IRS
LLIN + IVM
Time since LLIN distribution (month)
*** Time since LLIN distribution (month)
** Time since LLIN distribution (month)
Other crops (% landscape)
Rice fields (% landscape)
Cotton fields
Absence
Presence
*** Rainfall month preceding coll. (cum. mm.)
** LLIN use rate (% of LLIN users)
* LLIN use rate (% of LLIN users)
* Diurnal temp. month preceding coll. (°C)
*** Diurnal temp. month preceding coll. (°C)
** Rainfall month preceding coll. (cum. mm.)
* Atmospheric press. hour of collection (hpa)
* Nocturnal temp. month preceding coll. (°C)
* Humidity hour of collection (%)
*** Nocturnal temp. month preceding coll. (°C)
Kdr-w mutation
SS
RS
RR***
RS*
RR**
** Atmospheric press. hour of collection (hpa)
*** Humidity hour of collection (%)
* Temperature hour of collection (°C)
Variable effect (model)
GLMM (+ 95% IC)
RF
Avg. obs. prob.
Variable type
Micro-climate
Meteorology
Host availability
Landscape
Vector resistance
Follow-up kdr-e
An. coluzzii ->
164
18

Figure 3 : *Results of the statistical models of probability of physiological resistance in the malaria vectors. For each model, the top plot shows the explanatory power (R^2) and predictive power (AUC) of respectively the GLMM and the RF model. The other plots show the predicted probabilities of collecting a resistant vector across available values of each independent variable, holding everything else in the model equal (yellow line : probability predicted by the GLMM model ; green line : probability predicted by the RF model) (short reminder : 'vector control' and 'crops' variables were forced-in, and the other variables were retained only if they improved the AIC of the model. In addition, for the GLMM models, the other variables were plotted only if their p-value was < 0.05. For the RF models, the predicted probability (i.e. green line) was plotted only if the AUC of the model was > 0.6 and the range of predicted probabilities of resistance for the considered variable was > 0.05). In these plots, the y-axis represents the probability for an allele to be resistant. The red horizontal dashed line represents the overall rate of resistance (see Table 2). The p-values of the GLMMs are indicated through the stars* (* : $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)*. The grey squares distributed along the x-axis at the top and bottom of each plot represent the measured values available in the data (the darker the square, the more the number of observations) (NB : for atmospheric pressure, the values in the x-axis are centered around the mean).*

Associations with variables encoding ***vector control interventions****.* No statistically significant association was found between the likelihood of collecting an *Anopheles* carrying any of the target-site mutations and the type of VC intervention (LLIN + complementary tool compared to LLIN only) within the time frame of the study. However, the likelihood of collecting a host-seeking *An. gambiae s.s.* or *An. coluzzii* carrying a resistant *kdr-e* allele increased with the time since LLIN distribution, and as well with the % of users of LLINs in the village population. Regarding the others target-site mutations (*kdr-w* or *ace-1*), the likelihood of collecting a host-seeking *Anopheles* carrying them did not increase with the time since LLIN distribution.

Associations with variables encoding ***crops****.* No statistically significant association was found between the likelihood of collecting a host-seeking *Anopheles* carrying any of the target-site mutations and the % of landscape occupied by crop fields (cotton, rice, or other crops) in a 2 km-wide buffer area around the collection point.

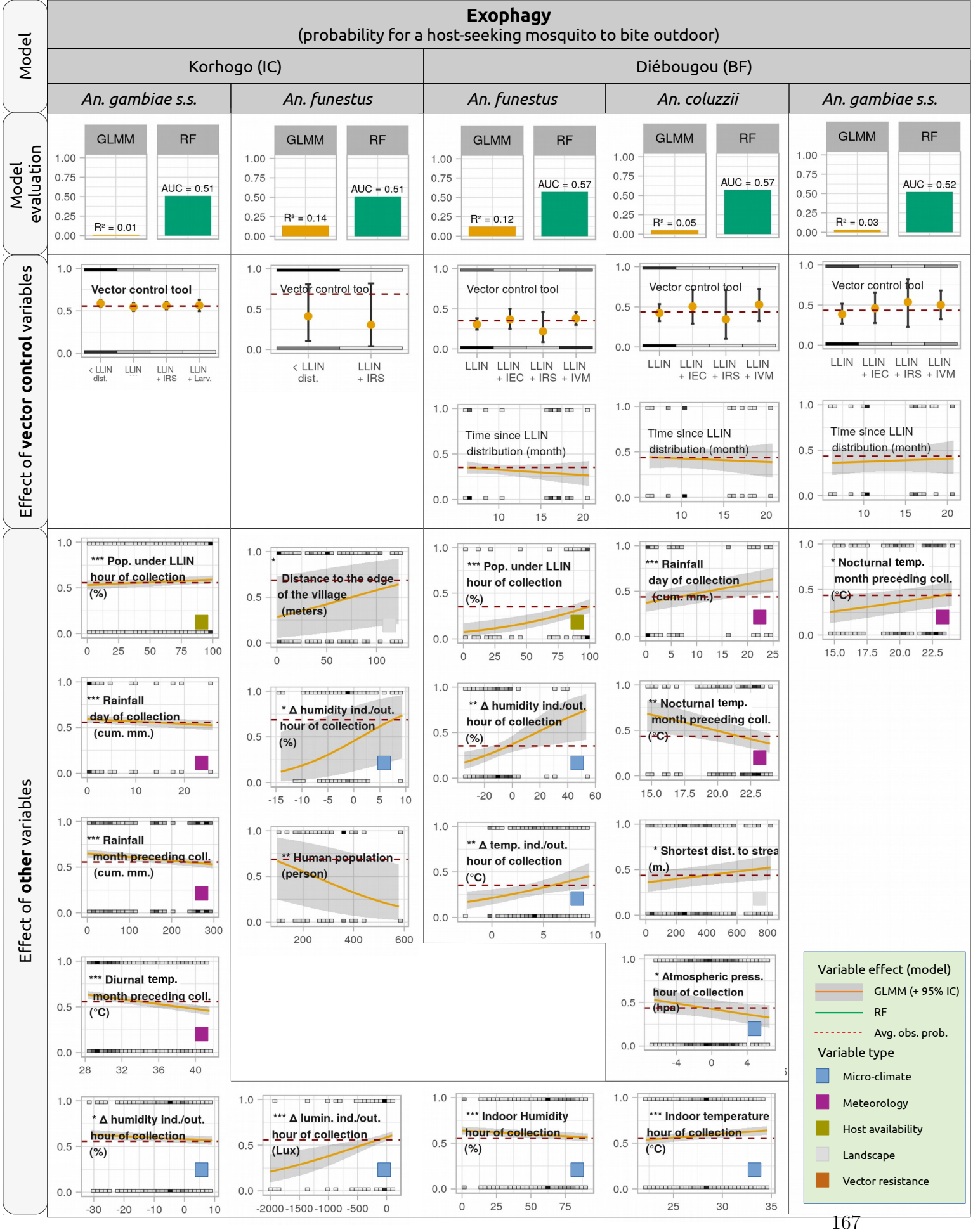
Associations with variables encoding ***micro-climate at the time (hour) of foraging activity****.* Positive associations were found between the likelihood of collecting a host-seeking *An. coluzzii* carrying the *kdr-e* mutation and atmospheric pressure, humidity and temperature at the time of collection, as well as that of collecting an *An. gambiae s.s.* carrying the *kdr-e* mutation and atmospheric pressure at the time of collection. A negative association was found between the likelihood of collecting a host-seeking *An. gambiae s.s.* carrying the *kdr-w* mutation and humidity at the time of collection.

Associations with variables encoding ***meteorological conditions during the month preceding collection****.* Negative associations were found between the likelihood of collecting a host-seeking : *An. coluzzii* carrying the *kdr-w* mutation and cumulated rainfall, *An. gambiae s.s.* carrying the *kdr-w* mutation and both cum. rainfall and mean diurnal temperatures, *An. coluzzii* carrying the *kdr-e* mutation and mean nocturnal temperatures, *An. gambiae s.s.* carrying *ace-1* mutation and

both mean diurnal and nocturnal temperatures during the month preceding collection. A positive association was found between the likelihood of collecting a host-seeking *An. coluzzii* carrying the *kdr-e* mutation and cumulated rainfall.

Associations between behavioural resistance and environmental variables

For the remaining nine models of behavioural resistance, Figure 4 shows the PDPs of the independent variables retained in the modeling workflow. For the GLMMs, numerical values of odd-ratios, 95% confidence intervals and p-values are provided in Additional file 4.


Exophagy
(probability for a host-seeking mosquito to bite outdoor)
Korhogo (IC)
Diébougou (BF)
An. gambiae s.s.
An. funestus
An. funestus
An. coluzzii
An. gambiae s.s.
Model
Model evaluation
Effect of vector control variables
Effect of other variables
GLMM
RF
AUC = 0.51
R² = 0.01
AUC = 0.51
R² = 0.14
AUC = 0.57
R² = 0.12
AUC = 0.57
R² = 0.05
AUC = 0.52
R² = 0.03
Vector control tool
< LLIN dist.
LLIN
LLIN + IRS
LLIN + Larv.
LLIN + IEC + IRS + IVM
Time since LLIN distribution (month)
*** Pop. under LLIN hour of collection (%)
Distance to the edge of the village (meters)
*** Rainfall day of collection (cum. mm.)
* Nocturnal temp. month preceding coll. (°C)
*** Rainfall month preceding coll. (cum. mm.)
* Δ humidity ind./out. hour of collection (%)
** Δ humidity ind./out. hour of collection (%)
** Nocturnal temp. month preceding coll. (°C)
** Human population (person)
** Δ temp. ind./out. hour of collection (°C)
* Shortest dist. to strea (m.)
*** Diurnal temp. month preceding coll. (°C)
* Atmospheric press. hour of collection (hpa)
* Δ humidity ind./out. hour of collection (%)
*** Δ lumin. ind./out. hour of collection (Lux)
*** Indoor Humidity hour of collection (%)
*** Indoor temperature hour of collection (°C)
Variable effect (model)
GLMM (+ 95% IC)
RF
Avg. obs. prob.
Variable type
Micro-climate
Meteorology
Host availability
Landscape
Vector resistance

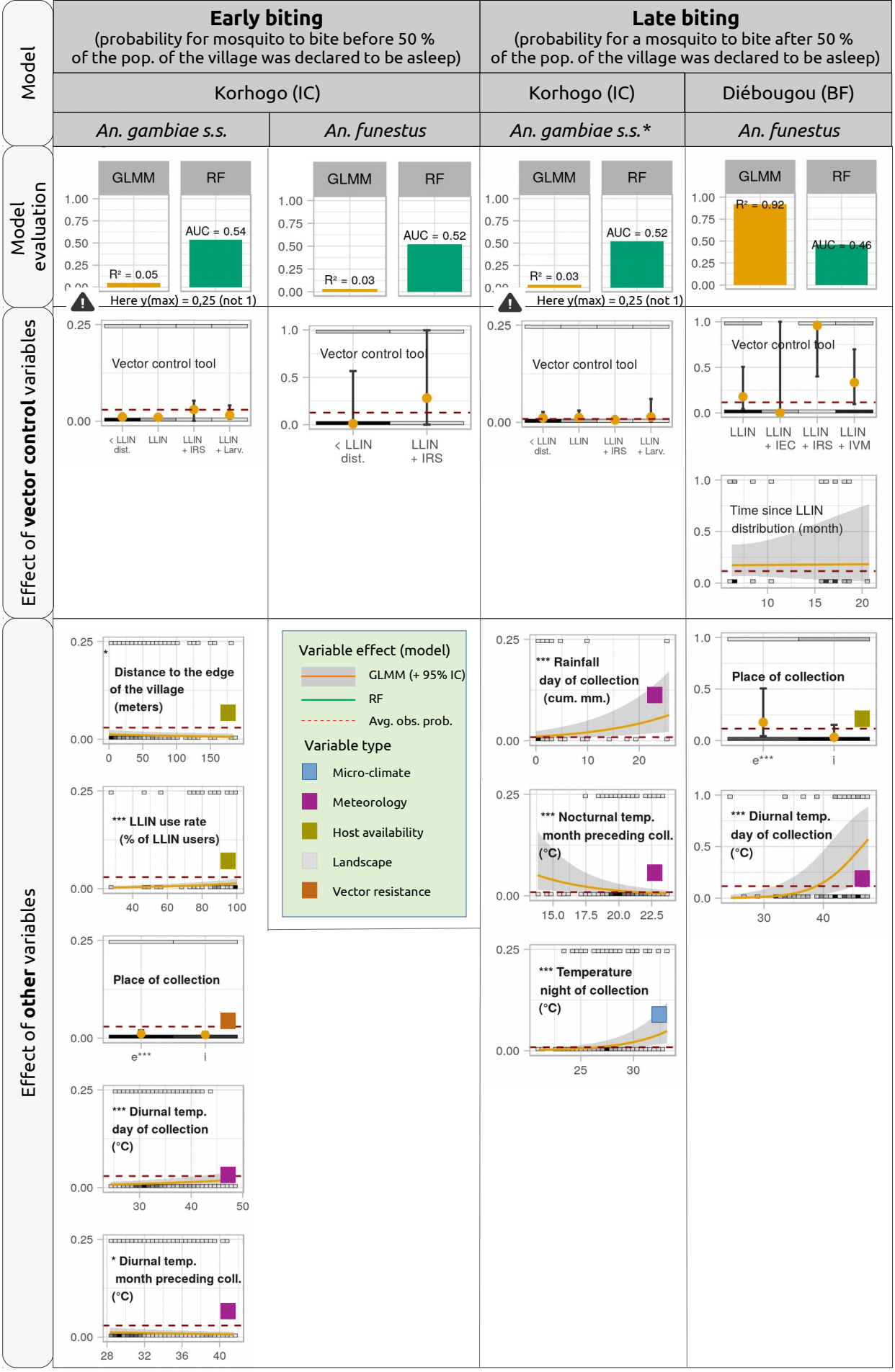

Model
Early biting
(probability for mosquito to bite before 50 % of the pop. of the village was declared to be asleep)
Late biting
(probability for a mosquito to bite after 50 % of the pop. of the village was declared to be asleep)
Korhogo (IC)
Korhogo (IC)
Diébougou (BF)
An. gambiae s.s.
An. funestus
An. gambiae s.s.*
An. funestus
Model evaluation
GLMM
RF
R² = 0.05
AUC = 0.54
R² = 0.03
AUC = 0.52
R² = 0.03
AUC = 0.52
R² = 0.92
AUC = 0.46
Here y(max) = 0,25 (not 1)
Here y(max) = 0,25 (not 1)
Effect of vector control variables
Vector control tool
< LLIN dist.
LLIN
LLIN + IRS
LLIN + Larv.
LLIN + IEC
LLIN + IVM
Time since LLIN distribution (month)
Effect of other variables
Distance to the edge of the village (meters)
*** LLIN use rate (% of LLIN users)
Place of collection
*** Diurnal temp. day of collection (°C)
* Diurnal temp. month preceding coll. (°C)
*** Rainfall day of collection (cum. mm.)
*** Nocturnal temp. month preceding coll. (°C)
*** Temperature night of collection (°C)
Variable effect (model)
GLMM (+ 95% IC)
RF
Avg. obs. prob.
Variable type
Micro-climate
Meteorology
Host availability
Landscape
Vector resistance

Figure 4 : *Results of the statistical models of probability of behavioural resistance in the malaria vectors. For each model, the top plot shows the explanatory power (R^2) and predictive power (AUC) of respectively the GLMM and the RF model. The other plots show the predicted probabilities of collecting a resistant vector across available values of each independent variable, holding everything else in the model equal (yellow line : probability predicted by the GLMM model ; green line : probability predicted by the RF model) (short reminder : 'vector control' variables were forced-in, and the other variables were retained only if they improved the AIC of the model. In addition, other variables were plotted only if their p-value was < 0.05. For the RF models, the predicted probability (i.e. green line) was plotted only if the AUC of the model was > 0.6 and the range of predicted probabilities of resistance for the considered variable was > 0.05). In these plots, the y-axis represents the probability for a mosquito to be resistant. The red horizontal dashed line represents the overall rate of resistance (see Table 2). The p-values of the GLMMs are indicated through the stars* (* : $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)*. The grey squares distributed along the x-axis at the top and bottom of each plot represent the measured values available in the data (the darker the square, the more the number of observations) (NB : for atmospheric pressure, the values in the x-axis are centered around the mean).*

Associations with variables encoding ***vector control interventions****.* No statistically significant association was found between the likelihood of collecting an exophagic, early- or late- biting *Anopheles* and neither the type of VC intervention (LLIN + complementary tool compared to LLIN only) nor the time since LLIN distribution within the time frame of the study.

Associations with variables encoding ***host availability****.* In the Korhogo area (IC), the likelihood of exophagy of *An. gambiae s.s.* slightly increased with the % of the population under an LLIN at the time of collection. The likelihood of early-biting of *An. gambiae s.s.* increased with the % of users of LLINs in the village population. In the Diébougou (BF) area, the likelihood of exophagy of *An. funestus* increased with the % of the population under an LLIN at the time of collection.

Associations with variables encoding ***landscape****.* In the Korhogo area (IC), the likelihood of exophagy of *An. funestus* increased with increasing distance to the edge of the village. The likelihood of early-biting of *An. gambiae s.s.* decreased with increasing distance to the edge of the village. In the Diébougou (BF) area, the likelihood of exophagy of *An. coluzzii* increased with increasing distance to the nearest stream.

Associations with variables encoding ***micro-climate at the time (hour) of foraging activity****.* In the Korhogo area (IC), the likelihood of exophagy of *An. gambiae s.s.* decreased when humidity indoors increased and when humidity got relatively higher indoors compared to outdoors. In addition, it increased when luminosity got relatively higher indoors compared to outdoors. In the Diébougou area (BF), the likelihood of exophagy of *An. funestus* increased when temperature or humidity got relatively higher indoors compared to outdoors.

Associations with variables encoding ***meteorological conditions on the day or night of collection****.* Positive associations were found between the likelihood of : exophagy of *An. coluzzii* and rainfall

(BF area), early-biting of *An. gambiae s.s.* and temperature (IC area), late-biting of *An. gambiae s.s.* and both rainfall and temperature (IC area), late-biting of *An. funestus* and temperature (BF area). A negative association was found between the likelihood of exophagy of *An. gambiae s.s.* and rainfall (IC area).

Associations with variables encoding ***meteorological conditions during the month preceding collection****.* Negative associations were found between the likelihood of : exophagy of *An. gambiae s.s.* and both cumulated rainfall and mean diurnal temperatures (IC area), exophagy of *An. coluzzii* and mean nocturnal temperatures (BF area), late biting of *An. gambiae s.s.* and mean nocturnal temperature (IC area). A positive association was found between the likelihood of exophagy of *An. gambiae s.s.* and mean nocturnal temperatures (BF area).

Associations with variables encoding ***physiological resistances****.* As a reminder, the genotypes for the target-site mutations of individual collected mosquitoes were introduced as independent variables in the behavioural resistance models in the Diébougou area (BF). Here, these variables were not retained in the variable selection procedure, i.e. no statistically significant association was found between any of the behavioural resistance indicator and *kdr-w*, *kdr-e*, or *ace-1* mutations.

Explanatory and predictive power of the statistical models

Additional figure 5 provides boxplots of observed resistance status vs. predicted probabilities by each model.

Exophagy. For the models of exophagy, the explanatory power of the GLMM models was : 'very weak' for *An. gambiae s.s.* in the Korhogo area (IC), 'moderate' for *An. funestus* in the Korhogo area (IC)', weak' for *An. funestus*, *An. coluzzii* and *An. gambiae s.s.* in the Diébougou area (BF). The predictive power of the RF models of exophagy was 'very weak' for all the species in the two study areas.

Early and late biting. For the models of early biting, the explanatory power of the GLMM models was 'weak' for both *An. gambiae s.s.* and *An. funestus* in the Korhogo area (IC). For the models of late biting, the explanatory power of the GLMM was 'weak' for *An. gambiae s.s.* in the Korhogo area (IC) and 'substantial' for *An. funestus* in the Diébougou area (BF). The predictive power of the RF models of early and late biting was 'very weak' for all species in the two study areas, except for the model of late biting of *An. gambiae s.s.* in the Korhogo area (IC) for which it was 'weak'.

Kdr-w, kdr-e, ace-1. For the *kdr-w* mutation in the Diébougou area (BF), the explanatory power of the GLMM models was 'weak' for *An. coluzzii* and 'substantial' for *An. gambiae s.s.* ; and the predictive power of the RF models was 'weak' for *An. coluzzii* and 'moderate' for *An. gambiae s.s.* For the *kdr-e* mutation in the Diébougou area (BF), the explanatory power of the GLMM models was 'substantial' for both *An. coluzzii* and *An. gambiae s.s.* ; and the predictive power of the RF models was 'moderate' for *An. coluzzii* and 'weak' for *An. gambiae s.s.* For the *ace-1* mutation in the Diébougou

170

area (BF), the explanatory power of the GLMM models was 'weak' for *An. gambiae s.s.* ; and the predictive power of the RF model was 'very weak'.

Discussion

In this work, we studied indicators of physiological and behavioural resistances of several malaria vectors in rural West-Africa at a fine spatial scale (approximately the extent of a health district), using longitudinal data collected in two areas belonging to two different countries, respectively 27 and 28 villages per area, and across 1.25 to 1.5 year. The objectives were to describe the spatial and temporal heterogeneity of vector resistance, and to better understand their drivers, at scales that are consistent with operational action. To our knowledge, our work is the first studying the heterogeneity of vector resistance at such fine spatial scale with such a large dataset of mosquito collection and potential drivers of resistance. In this discussion, we first use our results to provide elements of answers to the questions raised in introduction of this article. We then discuss some implications of the findings for the management of vector resistance in our areas.

Physiological resistances : potential drivers and spatiotemporal heterogeneity

The main drivers of physiological resistances are insecticides, used either in public health for vector control or in agriculture (see Introduction). In this study, we found that the probability of collecting a host-seeking *An. gambiae s.s.* or *An. coluzzii* in the Diébougou area carrying a *kdr-e* resistant allele significantly increased with both the time since LLIN distribution and the % of LLIN users in the village population. In contrast, there was no significant association between any of the target-site mutations and any of the crop-related variable. This could indicate that **within the spatiotemporal frame of our study, the development of the kdr-e mutation in the vector population was more likely due to insecticides used in public health than pesticides used in agriculture**. In Burkina Faso, pesticides are widely used for cotton and sugar cane (Ouedraogo et al. 2011), but only in lesser proportions in market gardening and cereal production (maize and rice are the only cereals that are treated to a significant extent (MERSI, CNRST, and IRSS 2016)). Here, in the 2-km wide buffer zones around our collection points crops occupied up to 40 % of the total land, but were mainly made of leguminous crops, millet, sorghum, with cotton and rice being only marginally present. Hence, pesticides are likely not much used (field surveys regarding the use of pesticides by the farmers could confirm this hypothesis). This could explain the absence of association between target-site mutations and the crops-related variables. Noteworthy, the *kdr-w* and *ace-1* mutations did not increase significantly with the time since LLIN distribution. The absence of increase of the *kdr-w* mutation may be explained by its very high baseline allelic frequencies ; while that of the *ace-1* mutation may be explained by the type of insecticide used to impregnate the LLINs - deltamethrin, which does not select the ace-1 mutation).

The statistical models captured many associations between the likelihood of collecting a physiologically resistant *Anopheles* and the variables encoding weather, both during the month preceding collection and at the hour of collection. As stated previously, weather may impact the fitness or the

activity of mosquitoes carrying resistant genotypes ; and may therefore *in fine* impact the probability of collecting a physiologically resistant mosquito. Here, the associations that were captured could hence traduce **biological costs associated with target-site mutations, both in terms of fitness and activity**, as found elsewhere for other mosquito species (Kliot and Ghanim 2012). Regarding fitness, we found that the likelihood of collecting a host-seeking mosquito (*An. gambiae s.s.* or *An. coluzzii*) carrying a mutated allele, overall, decreased (to varying extents depending on the species and mutation) in the hot seasons (i.e. when diurnal or noctural temperatures during the month preceding collection increased). Carrying a kdr mutation might be associated with a reduced ability to seek out optimal temperatures, potentially resulting in a decreased longevity, fecundity, or ovarian development rates (Foster et al. 2003). Regarding activity, we observed that the likelihood of collecting a mosquito carrying a mutated allele (for the *kdr-e* mutation) decreased when atmospheric pressure, humidity, or temperature at the hour of collection got lower ; implying that mosquitoes carrying the *kdr-e* mutation could be less active in colder or drier conditions, or when atmospheric pressure is lower.

Lastly, we observed that the allelic frequencies of the target-site mutations, within each vector species and for each mutation, were overall quite stable across the villages and seasons within the spatiotemporal frame of the study. At larger spatial and temporal scales, physiological resistances were found more heterogeneous (Moyes et al. 2020). In our study, such homogeneity might be due to a relative homogeneity in space and time of the main determinants of physiological resistance (access and use of insecticide-based vector control interventions). The relative seasonal homogeneity might traduce that fitness costs, despite their existence, might be limited within the range of meteorological conditions in our area.

Behavioural resistances : potential drivers and spatiotemporal heterogeneity

An important and pending question is the genetic (constitutive) or plastic (inducible) nature of behavioual resistances (see Introduction). In this study, we found no statistically significant association between any of the indicators of behavioural resistance and neither the time since LLIN distribution nor the VC tool implemented. There was hence no evidence of growing frequencies of behavioural resistances (exophagy, early- and late-biting) in response to vector control within the 1.25 to 1.5 years of this study, i.e. **no clear indication of constitutive resistance**. Nonetheless, some of the associations captured by the models, as well as a comparison of the measured resistance rates with that of previous studies in the same countries, suggest that there may still be a genetic component in mosquito foraging behaviour. First, we found many statistically significant associations between the likelihood of collecting a behaviourally resistant *Anopheles* and the meteorological conditions during the month preceding collection. This could mean that there might exist a fitness cost to behavioural resistance, associated with (i) maternal or paternal effects or (ii) development or survival at the larval stage (see Taconet et al. (2021)), that support (at least for the option (ii)) the hypothesis of genetic bases for some host-seeking behavioural phenotypes. Second, the exophagy rates of the main malaria vectors in both areas (on average, 41% in the Diébougou area and 56% in the Korhogo area) were substantially higher than those, overall, historically reported for these species ; in general in Africa (Huho et al. 2013; Sinka et al. 2010) and especially in Burkina Faso (Sherrard-Smith et al. 2019). Such

172

high levels of outdoor biting, in comparison with past levels, suggest that behavioural adaptations may be ongoing in the study areas, most probably in response to the widespread and prolonged use of insecticide-based vector control tools. The hypothesis of a hereditary component in the behavior of malaria vectors (at least for the biting hour) is supported by a recent study which has observed, for *Anopheles arabiensis* in Tanzania, that F2 from early-biting F0 (grandmothers) were - to some extent - more likely to bite early than F2 from mid or late-biting F0 (Govella et al. 2021). In our study, the absence of significant association between the probability of behavioural resistances and insecticide-related variables might be due to the relatively short length of the study. In fact, mosquito behaviours are likely complex multigenic traits (Main et al. 2016) and might therefore respond slowly to selection (at least, slower than target-site mutations, which are linked to single genes and may hence respond rapidly and efficiently to selection). In a recent two-years long longitudinal study in another area of Burkina Faso, the authors also observed an absence of association between time and behavioural resistance (for early biting, exophagy and exophily) (Sanou et al. 2021) ; however, such associations were actually found in a four-years long study in Tanzania (for exophagy) (Kreppel et al. 2020). Long-term monitoring of vector behaviour, particularly in areas with a long history of use of insecticides in public health, is critical to better understand the biological mechanisms underlying behavioural resistances, to potentially assess their development rate, and more broadly to assess residual malaria transmission risk (Sanou et al. 2021; Kreppel et al. 2020; Durnez and Coosemans 2013).

Weather can impact the fitness of possible genotypes associated with resistant behavioural phenotypes, but may also influence the activity of these phenotypes (see Introduction). Here, we found many associations between mosquito host-seeking behaviour and variables representing meteorological conditions on the day or at the hour of collection. For instance, the probability for an *An. gambiae s.s.* to be collected outdoor in the Korhogo area increased when the air indoor was dry, or when the air outdoor became relatively more humid than indoor. Likewise, in the Diébougou area, the probability for an *An. funestus* to be collected outdoor increased when the air outdoor became relatively cooler than indoor. These observations are consistent with the hypothesis of mosquitoes shifting from indoor to outdoor host-seeking in case of desiccation-related mortality risk indoors, as observed and discussed elsewhere (Kessler and Guerin 2008; Ngowo et al. 2017; Kreppel et al. 2020). The meteorological conditions seemed to cause not only spatial, but also temporal shifts in host-seeking activity. For instance, we found that the probability of collecting a late-biting *An. gambiae s.s.* in the Korhogo area increased when the average noctural temperature increased. Several associations also suggest that some malaria vectors may modify their behaviour in response to environmental variation that reduces host availability, as hypothesized elsewhere (Durnez and Coosemans 2013). For instance, the likelihood of collecting an *An. gambiae s.s.* (in the Korhogo area) or an *An. funestus* (in the BF area) outdoor increased at hours when people were protected by their LLINs. Likewise, the likelihood of collecting an early-biting *An. gambiae s.s.* in the Korhogo area increased when the % of LLIN users in the village increased. Altogether, all these associations suggest that in our study areas **mosquito foraging behaviour is driven – to a certain extent - by environmental conditions at the time of foraging activity**, i.e. that vectors likely modify their time or place of biting according to climatic conditions or host availability. The many associations

that were captured here in field conditions could be further tested experimentally, to quantify their effect more precisely and validate the underlying biological hypothesis.

Although many significant associations between environmental parameters and foraging behaviours have been captured by the models, their explanatory and predictive powers were overall weak. A low explanatory power can indicate either i) that variations in the dependent variable (here, individual vector resistance) are only marginally caused by the independent variables or ii) that the statistical model does not capture properly the true nature of the underlying relationships between the studied effect and its drivers (Karpatne et al. 2017) (e.g. a linear regression cannot, by definition, capture non-linear relationships that might exist in nature). Here, we minimized the risk of omitting important, complex associations by using, in addition to the binomial regression model, a machine-learning model (namely a random forest) that is inherently able to capture complex patterns contained in the data (e.g. non-linear relationships, interactions) (Breiman 2001a). Still, the models had low predictive powers. Altogether, these results indicate that very likely, despite the amount, granularity and diversity of potential factors measured and introduced in the models, most of the factors driving the individual host-seeking behaviours of the mosquitoes were not introduced in the models. Another possibility could be that some of our independent variables did not represent the actual "reality" in the field (e.g. the distance to the nearest stream is not necessary an ideal proxy for the distance to the breeding site). Nevertheless, since we used a wide range of variables encoding the environmental conditions at the time of foraging activity, we can hypothesize that within the spatiotemporal frame of the study, **mosquito foraging behaviour was only marginally driven by environmental variations**. This leaves the floor to other factors, like genetics (see above), learning, or randomness.

To test whether physiological resistance impacts the behaviour of host-seeking mosquitoes, we introduced in the behaviour resistance models of *An. coluzzii* and *An. gambiae s.s.* in the Diébougou area two variables encoding the genotypes for respectively the *kdr-w* and *kdr-e* mutations. No statistically significant association was found. In other words, **we could not find, in the field, a behavioural phenotype (among those studied, i.e. exophagy, early- and late-biting) associated with a genotype for one of the target-site mutations**. The only study, to our knowledge, having investigated the relationship between the kdr mutation and biting time or location in the field has also reported no statistically significant association between these two mechanisms of resistance to insecticide (Djènontin et al. 2021). Noteworthy, in our study, there was few variabilities in the genotypes of the collected mosquitoes (i.e. few homozygote susceptible mosquitoes captured, particularly for the *kdr-w* mutation), making it unfavorable to detect assocations between physiological and behavioural resistances. In the Korhogo area, such analysis could not be performed because physiological resistance was not available at the individual mosquito level.

Finally, we observed that the behavioural resistance rates for each vector species in each health district were, overall, relatively homogeneous across the villages and seasons within the spatiotemporal frame of the study (as for physiological resistances). This could mean that the overall dynamics of behavioural resistance occur at broader spatial and temporal scales than those of our study. At larger

scales (i.e. among countries and across years in Africa), exophagy rates of *Anopheles* mosquitoes seem, actually, to be more variable (Sherrard-Smith et al. 2019).

Implications of the findings for the management of vector resistance in the study areas

Long-lasting insecticidal nets have undoubtedly played a major role in reducing malaria cases throughout Africa, thanks both to their barrier and killing effects. More locally, we highlighted the efficacy of their barrier role in the Diébougou area by showing that, for their users, they prevented more than 80% of *Anopheles* bite exposure in the area (Soma et al. 2021). Despite these successes, our study adds to the growing body of evidence that the insecticides they are impregnated with are responsible for the rise of physiological resistances in the malaria vectors populations. We also highlighted that in response to an LLIN distribution, physiological resistance seems to grow quite rapidly in a susceptible population. Besides the development of physiological resistance, comparison with historical data suggests that the vectors may also be progressively changing their feeding behaviour to avoid the effects of the insecticides - although there was no clear evidence of this in the strict context of this study. Such trends in vector resistance may have an important epidemiological impact (Sherrard-Smith et al. 2019). Altogether, these results show, if still necessary, that we urgently need to think more strategically about our use of insecticides in public health tools in our areas. Switching to alternative insecticides, rotating or mixing insecticides, using current or novel insecticides in vector control tools others than long-lasting nets, entirely removing the insecticides from the vector control toolbox, or fostering the use of insecticidal-free tools, are all actions that could be envisaged (Paaijmans and Huijben 2020). Burkina Faso has, actually, distributed LLINs that mixes pyrethroid with Piperonyl butoxide (PBO) in its last universal LLIN distribution, in 2019.

Here, we observed that both behavioural and physiological resistances of mosquitoes were quite stable across the villages and seasons within the spatiotemporal frame of the study. This contrasts with their biting rates, which was found, in another study (Taconet et al. 2021), highly variable across the villages, seasons, and amongst the species. At small spatiotemporal scales, this calls for different strategies for respectively vector control (interventions aiming at reducing the human-vector contact) and resistance management (interventions aiming at reducing the development of physiological or behavioural resistance) at such spatiotemporal scales. While vector control plans should be very locally-tailored (species-, season-, and village-specific) (Taconet et al. 2021), resistance management strategies would probably not gain much in being adapted to the season or village within our areas. In other words, while resistance management plans are undoubtedly urgently needed, there is no compelling evidence – in the current state of the knowledge - that they should be tailored at very fine scales (village, season). Noteworthy, no entomological survey was performed during the high rainy season (July to September), at the known mosquito abundance and malaria transmission peaks. It would be worth collecting mosquitoes at this season to confirm the observed resistance rates.

Conclusion

Less than a decade after the first universal LLIN distribution, malaria vectors in two areas of rural West-Africa seem to be growingly adapting to avoid or circumvent the lethal effects of insecticides used in control interventions. After an LLIN distribution, rapid widespread of physiological resistance occurring in tandem with probable lower acting behavioral adaptations, are very likely contributing to the erosion of the insecticides impact. In an attempt to better understand the drivers of the intensity and spatio-temporal heterogeneity of physiological and behavioural resistance in malaria vectors, at the scale of a rural health district over a period of 1.5 years, we have mainly (i) shown that resistance (both physiological and behavioural) was quite homogeneous across the villages and seasons at theses scales, and (ii) hypothesized that at these spatiotemporal scales, vector resistance seemed to be only marginally driven by environmental factors other than those linked to insecticide use in current vector control. We believe that without waiting to understand precisely the underlying drivers, mechanisms, and rates of development of resistances, the malaria control community needs to think very strategically about the use and usefulness of current and novel insecticide-based control interventions.

References

At NASA GSFC, Precipitation Processing System (PPS). 2019. "GPM IMERG Final Precipitation L3 Half Hourly 0.1 Degree x 0.1 Degree V06." NASA Goddard Earth Sciences Data; Information Services Center. https://doi.org/10.5067/GPM/IMERG/3B-HH/06.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Bartoń, Kamil. 2020. *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn.

Bhatt, S., D. J. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. E. Battle, et al. 2015. "The Effect of Malaria Control on Plasmodium Falciparum in Africa Between 2000 and 2015." *Nature* 526 (7572): 207–11. https://doi.org/10.1038/nature15535.

Bolker, Ben, and David Robinson. 2020. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Breiman, Leo. 2001a. "Random Forests." *Machine Learning* 45 (1): 5–32.

———. 2001b. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–215. http://www.jstor.org/stable/2676681.

Brooks, Mollie E., Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. 2017. "glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling." *The R Journal* 9 (2): 378–400. https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

Carrasco, David, Thierry Lefèvre, Nicolas Moiroux, Cédric Pennetier, Fabrice Chandre, and Anna Cohuet. 2019. "Behavioural Adaptations of Mosquito Vectors to Insecticide Control." *Current Opinion in Insect Science* 34 (August): 48–54. https://doi.org/10.1016/j.cois.2019.03.005.

Center, NASA Goddard Earth Sciences Data And Information Services. 2019. "GPM IMERG Final Precipitation L3 1 Day 0.1 Degree x 0.1 Degree V06." NASA Goddard Earth Sciences Data; Information Services Center. https://doi.org/10.5067/GPM/IMERGDF/DAY/06.

Chandre, Fabrice, Frédéric Darriet, Sylvie Manguin, Cécile Brengues, Pierre Carnevale, and Pierre Guillet. 1999. "Pyrethroid Cross Resistance Spectrum Among Populations of Anopheles Gambiae s.s. From Côte d'Ivoire." *Journal of the American Mosquito Control*

Association 15: 53–59. https://www.documentation.ird.fr/hor/fdi:010018255.

Chicco, Davide. 2017. "Ten Quick Tips for Machine Learning in Computational Biology." *BioData Mining* 10 (1). https://doi.org/10.1186/s13040-017-0155-3.

Cohen, Jacob. 2013. *Statistical Power Analysis for the Behavioral Sciences*. 0th ed. Routledge. https://doi.org/10.4324/9780203771587.

Corbel, Vincent, and Raphael N'Guessan. 2013. "Distribution, Mechanisms, Impact and Management of Insecticide Resistance in Malaria Vectors: A Pragmatic Review." *Anopheles Mosquitoes - New Insights into Malaria Vectors*, July. https://doi.org/10.5772/56117.

Davidson, G. 1957. "Insecticide Resistance in Anopheles Sundaicus." *Nature* 180 (4598): 1333–35. https://doi.org/10.1038/1801333a0.

Diop, Malal M, Fabrice Chandre, Marie Rossignol, Angélique Porciani, Mathieu Chateau, Nicolas Moiroux, and Cédric Pennetier. 2021. "Sub-Lethal Insecticide Exposure Affects Host Biting Efficiency of Kdr-Resistant Anopheles Gambiae." *Peer Community Journal* 1 (November): e28. https://doi.org/10.24072/pcjournal.15.

Diop, Malal M., Nicolas Moiroux, Fabrice Chandre, Hadrien Martin-Herrou, Pascal Milesi, Olayidé Boussari, Angélique Porciani, Stéphane Duchon, Pierrick Labbé, and Cédric Pennetier. 2015. "Behavioral Cost & Overdominance in Anopheles Gambiae." Edited by Claudio R. Lazzari. *PLOS ONE* 10 (4): e0121755. https://doi.org/10.1371/journal.pone.0121755.

Djènontin, Armel, Aziz Bouraima, Christophe Soares, Seun Egbinola, and Gilles Cottrell. 2021. "Human Biting Rhythm of Anopheles Gambiae Giles, 1902 (Diptera: Culicidae) and Sleeping Behaviour of Pregnant Women in a Lagoon Area in Southern Benin." *BMC Research Notes* 14 (1): 200. https://doi.org/10.1186/s13104-021-05615-7.

Durnez, Lies, and Marc Coosemans. 2013. "Residual Transmission of Malaria: An Old Issue for New Approaches." In *Anopheles Mosquitoes - New Insights into Malaria Vectors*, edited by Sylvie Manguin. InTech. https://doi.org/10.5772/55925.

Foster, S P, S Young, M S Williamson, I Duce, I Denholm, and G J Devine. 2003. "Analogous Pleiotropic Effects of Insecticide Resistance Genotypes in Peach–Potato Aphids and Houseflies." *Heredity* 91 (2): 98–106. https://doi.org/10.1038/sj.hdy.6800285.

Friedman, Jerome H., and Bogdan E. Popescu. 2008. "Predictive Learning via Rule Ensembles." *The Annals of Applied Statistics* 2 (3): 916–54. https://doi.org/10.1214/07-AOAS148.

178

Gatton, Michelle L., Nakul Chitnis, Thomas Churcher, Martin J. Donnelly, Azra C. Ghani, H. Charles J. Godfray, Fred Gould, et al. 2013. "THE IMPORTANCE OF MOSQUITO BEHAVIOURAL ADAPTATIONS TO MALARIA CONTROL IN AFRICA." *Evolution* 67 (4): 1218–30. https://doi.org/10.1111/evo.12063.

Govella, Nicodem J., Paul C. D. Johnson, Gerry F. Killeen, and Heather M. Ferguson. 2021. "Heritability and Phenotypic Plasticity of Biting Time Behaviors in the Major African Malaria Vector *Anopheles Arabiensis*." Preprint. Evolutionary Biology. https://doi.org/10.1101/2021.05.17.444456.

Greenwell, Brandon M. 2017. "Pdp: An R Package for Constructing Partial Dependence Plots." *The R Journal* 9 (1): 421–36. https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

Hay, G. J., and G. Castilla. 2008. "Geographic Object-Based Image Analysis (GEOBIA): A New Name for a New Discipline." In *Object-Based Image Analysis*, edited by Thomas Blaschke, Stefan Lang, and Geoffrey J. Hay, 75–89. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77058-9_4.

Hemingway, Janet, Hilary Ranson, Alan Magill, Jan Kolaczinski, Christen Fornadel, John Gimnig, Maureen Coetzee, et al. 2016. "Averting a Malaria Disaster: Will Insecticide Resistance Derail Malaria Control?" *The Lancet* 387 (10029): 1785–88. https://doi.org/10.1016/S0140-6736(15)00417-1.

Hersbach, Hans, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, et al. 2020. "The Era5 Global Reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146 (730): 1999–2049. https://doi.org/10.1002/qj.3803.

Hien, Aristide Sawdetuo, Dieudonné Diloma Soma, Omer Hema, Bazoma Bayili, Moussa Namountougou, Olivier Gnankiné, Thierry Baldet, Abdoulaye Diabaté, and Kounbobr Roch Dabiré. 2017. "Evidence That Agricultural Use of Pesticides Selects Pyrethroid Resistance Within Anopheles Gambiae s.l. Populations from Cotton Growing Areas in Burkina Faso, West Africa." Edited by Luzia Helena Carvalho. *PLOS ONE* 12 (3): e0173098. https://doi.org/10.1371/journal.pone.0173098.

Huho, Bernadette, Olivier Briët, Aklilu Seyoum, Chadwick Sikaala, Nabie Bayoh, John Gimnig, Fredros Okumu, et al. 2013. "Consistently High Estimates for the Proportion of Human Exposure to Malaria Vector Populations Occurring Indoors in Rural Africa." *International Journal of Epidemiology* 42 (1): 235–47. https://doi.org/10.1093/ije/dys214.

Karpatne, Anuj, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. "Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data." *IEEE Trans-*

179

actions on Knowledge and Data Engineering 29 (10): 2318–31. https://doi.org/10.1109/TKDE.2017.2720168.

Kessler, Sébastien, and Patrick M. Guerin. 2008. "Responses of Anopheles Gambiae, Anopheles Stephensi, Aedes Aegypti, and Culex Pipiens Mosquitoes (Diptera: Culicidae) to Cool and Humid Refugium Conditions." *Journal of Vector Ecology* 33 (1): 145–49. https://doi.org/10.3376/1081-1710(2008)33%5B145:ROAGAS%5D2.0.CO;2.

Killeen, Gerry F. 2014. "Characterizing, Controlling and Eliminating Residual Malaria Transmission." *Malaria Journal* 13 (1): 330. https://doi.org/10.1186/1475-2875-13-330.

Kirby, M. J., and S. W. Lindsay. 2004. "Responses of Adult Mosquitoes of Two Sibling Species, *Anopheles Arabiensis* and *A. Gambiae* s.s. (Diptera: Culicidae), to High Temperatures." *Bulletin of Entomological Research* 94 (5): 441–48. https://doi.org/10.1079/BER2004316.

Kliot, Adi, and Murad Ghanim. 2012. "Fitness Costs Associated with Insecticide Resistance: Fitness Cost and Insecticide Resistance." *Pest Management Science* 68 (11): 1431–37. https://doi.org/10.1002/ps.3395.

Kreppel, K. S., M. Viana, B. J. Main, P. C. D. Johnson, N. J. Govella, Y. Lee, D. Maliti, F. C. Meza, G. C. Lanzaro, and H. M. Ferguson. 2020. "Emergence of Behavioural Avoidance Strategies of Malaria Vectors in Areas of High LLIN Coverage in Tanzania." *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-71187-4.

Labbé, P., J.-P. David, H. Alout, P. Milesi, L. Djogbénou, N. Pasteur, and M. Weill. 2017. "Evolution of Resistance to Insecticide in Disease Vectors." In *Genetics and Evolution of Infectious Diseases*, 313–39. Elsevier. https://doi.org/10.1016/B978-0-12-799942-5.00014-7.

Lockwood, J. A., T. C. Sparks, and R. N. Story. 1984. "Evolution of Insect Resistance to Insecticides: A Reevaluation of the Roles of Physiology and Behavior." *Bulletin of the Entomological Society of America* 30 (4): 41–51. https://doi.org/10.1093/besa/30.4.41.

Long, Jacob A. 2020. *Jtools: Analysis and Presentation of Social Scientific Data*. https://cran.r-project.org/package=jtools.

Main, Bradley J, Yoosook Lee, Heather M. Ferguson, Katharina S. Kreppel, Anicet Kihonda, Nicodem J. Govella, Travis C. Collier, et al. 2016. "The Genetic Basis of Host Preference and Resting Behavior in the Major African Malaria Vector, Anopheles Arabiensis." Edited by Laurence J. Zwiebel. *PLOS Genetics* 12 (9): e1006303. https://doi.org/10.1371/journal.pgen.1006303.

Martinez-Torres, D., F. Chandre, M. S. Williamson, F. Darriet, J. B. Berge, A. L. Devonshire, P. Guillet, N. Pasteur, and D. Pauron. 1998. "Molecular Characterization of Pyrethroid

Knockdown Resistance (Kdr) in the Major Malaria Vector Anopheles Gambiae s.s." *Insect Molecular Biology* 7 (2): 179–84. https://doi.org/10.1046/j.1365-2583.1998.72062.x.

MERSI, CNRST, and IRSS. 2016. "Utilisation Des Pesticides Agricoles Dans Trois Régions à l'ouest Du Burkina Faso Et Évaluation de Leur Impact Sur La Santé Et l'environnement: Cas Des Régions de La Boucle Du Mouhoun, Des Cascades Et Des Hauts-Bassins."

Moiroux, Nicolas. 2012. "Modélisation Du Risque d'exposition Aux Moustiques Vecteurs de Plasmodium Spp. Dans Un Contexte de Lutte Anti-Vectorielle." PhD thesis. http://www.theses.fr/2012MON20177/document.

Moiroux, Nicolas, Abdul S. Bio-Bangana, Armel Djènontin, Fabrice Chandre, Vincent Corbel, and Hélène Guis. 2013. "Modelling the Risk of Being Bitten by Malaria Vectors in a Vector Control Area in Southern Benin, West Africa." *Parasites & Vectors* 6 (1): 71. https://doi.org/10.1186/1756-3305-6-71.

Moiroux, Nicolas, Armel Djènontin, Abdul S Bio-Bangana, Fabrice Chandre, Vincent Corbel, and Hélène Guis. 2014. "Spatio-Temporal Analysis of Abundances of Three Malaria Vector Species in Southern Benin Using Zero-Truncated Models." *Parasites & Vectors* 7 (1): 103. https://doi.org/10.1186/1756-3305-7-103.

Moyes, Catherine L., Duncan K. Athinya, Tara Seethaler, Katherine E. Battle, Marianne Sinka, Melinda P. Hadi, Janet Hemingway, Michael Coleman, and Penelope A. Hancock. 2020. "Evaluating Insecticide Resistance Across African Districts to Aid Malaria Control Decisions." *Proceedings of the National Academy of Sciences* 117 (36): 22042–50. https://doi.org/10.1073/pnas.2006781117.

Nakagawa, Shinichi, and Holger Schielzeth. 2013. "A General and Simple Method for Obtaining r^2 from Generalized Linear Mixed-Effects Models." Edited by Robert B. O'Hara. *Methods in Ecology and Evolution* 4 (2): 133–42. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Ngowo, Hs, Ew Kaindoa, J Matthiopoulos, Hm Ferguson, and Fo Okumu. 2017. "Variations in Household Microclimate Affect Outdoor-Biting Behaviour of Malaria Vectors." *Wellcome Open Research* 2 (October): 102–2. https://doi.org/10.12688/wellcomeopenres.12928.1.

Njan Nloga, A., Vincent Robert, J. C. Toto, and Pierre Carnevale. 1993. "La Durée Du Cycle Gonotrophique d'Anopheles Moucheti Varie de Trois à Quatre Jours En Fonction de La Proximité Par Rapport Aux Gites de Ponte." *Bulletin de Liaison Et de Documentation - OCEAC* 26: 69–72. https://www.documentation.ird.fr/hor/fdi:39013.

Ouedraogo, Moustapha, Adama M., Theodore Z., and Pierre I. 2011. "Pesticides in Burkina Faso: Overview of the Situation in a Sahelian African Country." In *Pesticides in the Modern World - Pesticides Use and Management*, edited by Margarita Stoytcheva. InTech. https://doi.org/
181

10.5772/16507.

Paaijmans, Krijn P., and Silvie Huijben. 2020. "Taking the 'I' Out of LLINs: Using Insecticides in Vector Control Tools Other Than Long-Lasting Nets to Fight Malaria." *Malaria Journal* 19 (1). https://doi.org/10.1186/s12936-020-3151-x.

Pedersen, Thomas Lin. 2019. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

PNLP. 2014a. "Directives Nationales Pour La Prise En Charge Du Paludisme Dans Les Formations Sanitaires Du Burkina Faso. Ministère de La Santé/Burkina Faso."

———. 2014b. "Programme National de Lutte Contre Le Paludisme En Côte d'ivoire. 2014. Plan Stratégique National de Lutte Contre Le Paludisme 2012–2015 (Période Replanifiée 2014–2017). Approche Stratifiée de Mise à l'échelle Des Interventions de Lutte Contre Le Paludisme En Côte d'ivoire Et Consolidation Des Acquis. Abidjan: Ministère de La Santé Et l'hygiène Publique. 149 p."

Porciani, Angélique, Malal Diop, Nicolas Moiroux, Tatiana Kadoke-Lambi, Anna Cohuet, Fabrice Chandre, Laurent Dormont, and Cédric Pennetier. 2017. "Influence of Pyrethroïd-Treated Bed Net on Host Seeking Behavior of Anopheles Gambiae s.s. Carrying the Kdr Allele." Edited by Guido Favia. *PLOS ONE* 12 (7): e0164518. https://doi.org/10.1371/journal.pone.0164518.

QGIS Development Team. 2021. *QGIS Geographic Information System*. QGIS Association. https://www.qgis.org.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ranson, H., B. Jensen, J. M. Vulule, X. Wang, J. Hemingway, and F. H. Collins. 2000. "Identification of a Point Mutation in the Voltage-Gated Sodium Channel Gene of Kenyan Anopheles Gambiae Associated with Resistance to DDT and Pyrethroids." *Insect Molecular Biology* 9 (5): 491–97. https://doi.org/10.1046/j.1365-2583.2000.00209.x.

Reid, Molly C., and F. Ellis McKenzie. 2016. "The Contribution of Agricultural Insecticide Use to Increasing Insecticide Resistance in African Malaria Vectors." *Malaria Journal* 15 (1). https://doi.org/10.1186/s12936-016-1162-4.

Riveron, Jacob M., Magellan Tchouakui, Leon Mugenzi, Benjamin D. Menze, Mu-Chun Chiang, and Charles S. Wondji. 2018. "Insecticide Resistance in Malaria Vectors: An Update at a Global Scale." In *Towards Malaria Elimination - A Leap Forward*, edited by Sylvie Manguin and Vas Dev. InTech. https://doi.org/10.5772/intechopen.78375.

182

RStudio Team. 2020. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Sanou, Antoine, Luca Nelli, W. Moussa Guelbéogo, Fatoumata Cissé, Madou Tapsoba, Pierre Ouédraogo, N'falé Sagnon, Hilary Ranson, Jason Matthiopoulos, and Heather M. Ferguson. 2021. "Insecticide Resistance and Behavioural Adaptation as a Response to Long-Lasting Insecticidal Net Deployment in Malaria Vectors in the Cascades Region of Burkina Faso." *Scientific Reports* 11 (1): 17569. https://doi.org/10.1038/s41598-021-96759-w.

Sherrard-Smith, Ellie, Janetta E. Skarp, Andrew D. Beale, Christen Fornadel, Laura C. Norris, Sarah J. Moore, Selam Mihreteab, et al. 2019. "Mosquito Feeding Behavior and How It Influences Residual Malaria Transmission Across Africa." *Proceedings of the National Academy of Sciences* 116 (30): 15086–95. https://doi.org/10.1073/pnas.1820646116.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310. https://doi.org/10.1214/10-STS330.

Shmueli, Galit, and O. Koppius. 2010. "Predictive Analytics in Information Systems Research." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1606674.

Sinka, Marianne E, Michael J Bangs, Sylvie Manguin, Maureen Coetzee, Charles M Mbogo, Janet Hemingway, Anand P Patil, et al. 2010. "The Dominant Anopheles Vectors of Human Malaria in Africa, Europe and the Middle East: Occurrence Data, Distribution Maps and Bionomic Précis." *Parasites & Vectors* 3 (1). https://doi.org/10.1186/1756-3305-3-117.

Snow, Robert W., and Herbert M. Gilles. 2002. "The Epidemiology of Malaria." *Essential Malariology* 4.

Sokhna, C., M. O. Ndiath, and C. Rogier. 2013. "The Changes in Mosquito Vector Behaviour and the Emerging Resistance to Insecticides Will Challenge the Decline of Malaria." *Clinical Microbiology and Infection* 19 (10): 902–7. https://doi.org/10.1111/1469-0691.12314.

Soma, Barnabas Mahugnon Zogo, Anthony Somé, Bertin N'Cho Tchiekoi, Domonbabele François de Sales Hien, Hermann Sié Pooda, Sanata Coulibaly, et al. 2020. "Anopheles Bionomics, Insecticide Resistance and Malaria Transmission in Southwest Burkina Faso: A Pre-Intervention Study." *PLOS ONE* 15 (8): e0236920. https://doi.org/10.1371/journal.pone.0236920.

Soma, B. Zogo, P. Taconet, A. Somé, S. Coulibaly, L. Baba-Moussa, G. A. Ouédraogo, et al. 2021. "Quantifying and Characterizing Hourly Human Exposure to Malaria Vectors Bites to Address Residual Malaria Transmission During Dry and Rainy Seasons in Rural Southwest Burkina Faso." *BMC Public Health* 21 (1). https://doi.org/10.1186/s12889-021-10304-y.

Taconet, Paul, Angélique Porciani, Dieudonné Diloma Soma, Karine Mouline, Frédéric Simard, Alphonsine Amanan Koffi, Cedric Pennetier, Roch Kounbobr Dabiré, Morgan Mangeas, and Nicolas Moiroux. 2021. "Data-Driven and Interpretable Machine-Learning Modeling to Explore the Fine-Scale Environmental Determinants of Malaria Vectors Biting Rates in Rural Burkina Faso." *Parasites & Vectors* 14 (1). https://doi.org/10.1186/s13071-021-04851-x.

Tyagi, Shivani, and Sangeeta Mittal. 2020. "Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning." In *Proceedings of ICRIC 2019*, edited by Pradeep Kumar Singh, Arpan Kumar Kar, Yashwant Singh, Maheshkumar H. Kolekar, and Sudeep Tanwar, 597:209–21. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-29407-6_17.

Voeten, Cesko C. 2020. *Buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*. https://CRAN.R-project.org/package=buildmer.

Wan, Zhengming, Simon Hook, and Glynn Hulley. 2015a. "Mod11a1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006." NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD11A1.006.

———. 2015b. "Myd11a1 MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006." NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MYD11A1.006.

Weill, M., C. Malcolm, F. Chandre, K. Mogensen, A. Berthomieu, M. Marquine, and M. Raymond. 2004. "The Unique Mutation in Ace-1 Giving High Insecticide Resistance Is Easily Detectable in Mosquito Vectors." *Insect Molecular Biology* 13 (1): 1–7. https://doi.org/10.1111/j.1365-2583.2004.00452.x.

WHO. 2017. "WHO | Global Vector Control Response 2017–2030." *WHO*. http://www.who.int/vector-control/publications/global-control-response/en/.

———. 2021. "World Malaria Report 2021." Licence: CC BY-NC-SA 3.0 IGO. https://www.who.int/publications-detail-redirect/9789240040496.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wing, Max Kuhn Contributions from Jed, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, et al. 2018. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

184

Wright, Marvin N., and Andreas Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77 (1): 1–17. https://doi.org/10.18637/jss.v077.i01.

Yadouleton, Anges, Thibaud Martin, Gil Padonou, Fabrice Chandre, Alex Asidi, Luc Djogbenou, Roch Dabiré, et al. 2011. "Cotton Pest Management Practices and the Selection of Pyrethroid Resistance in Anopheles Gambiae Population in Northern Benin." *Parasites & Vectors* 4 (1). https://doi.org/10.1186/1756-3305-4-60.

Yan, Yachen. 2016. *MLmetrics: Machine Learning Evaluation Metrics*. https://CRAN.R-project.org/package=MLmetrics.

Zhao, Qingyuan, and Trevor Hastie. 2021. "Causal Interpretations of Black-Box Models." *Journal of Business & Economic Statistics* 39 (1): 272–81. https://doi.org/10.1080/07350015.2019.1624293.

Zogo, Barnabas, Dieudonné Diloma Soma, Bertin N'Cho Tchiekoi, Anthony Somé, Ludovic P. Ahoua Alou, Alphonsine A. Koffi, Florence Fournet, et al. 2019. "*Anopheles* Bionomics, Insecticide Resistance Mechanisms, and Malaria Transmission in the Korhogo Area, Northern Côte d'Ivoire: A Pre-Intervention Study." *Parasite* 26: 40. https://doi.org/10.1051/parasite/2019040.

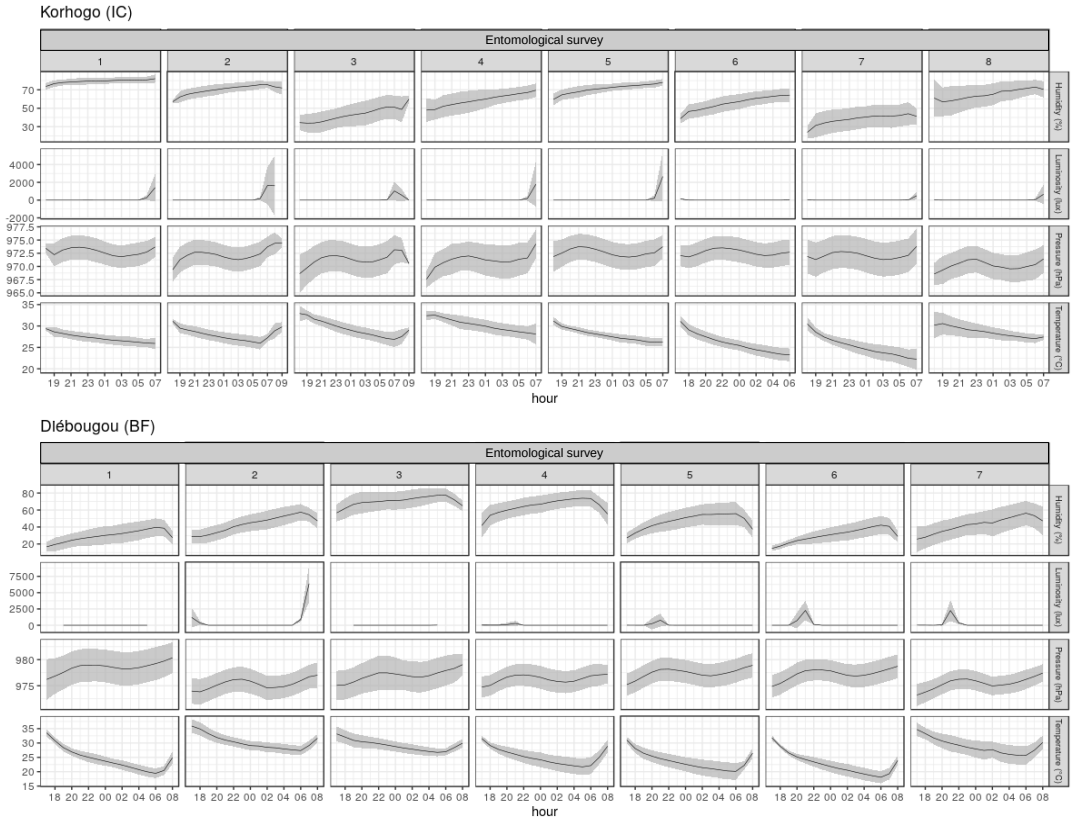
Supplementary material

Supplementary file 1 : Spatio-temporal distribution of the independent variables used in the models

These plots show the spatiotemporal distribution of the various independent variables used in the resistance models. Namely, the following plots are presented : Biting rates of the main malaria vectors in each village and entomological survey, Vector control tool is use in each village for each entomological survey, % of population that use an LLIN for each human behaviour survey, Hourly trends of % of population indoor and under an LLIN for each human behaviour survey, Hourly trends of micro-climatic conditions for each entomological survey, Weekly meteorological conditions, Lanscape variables.

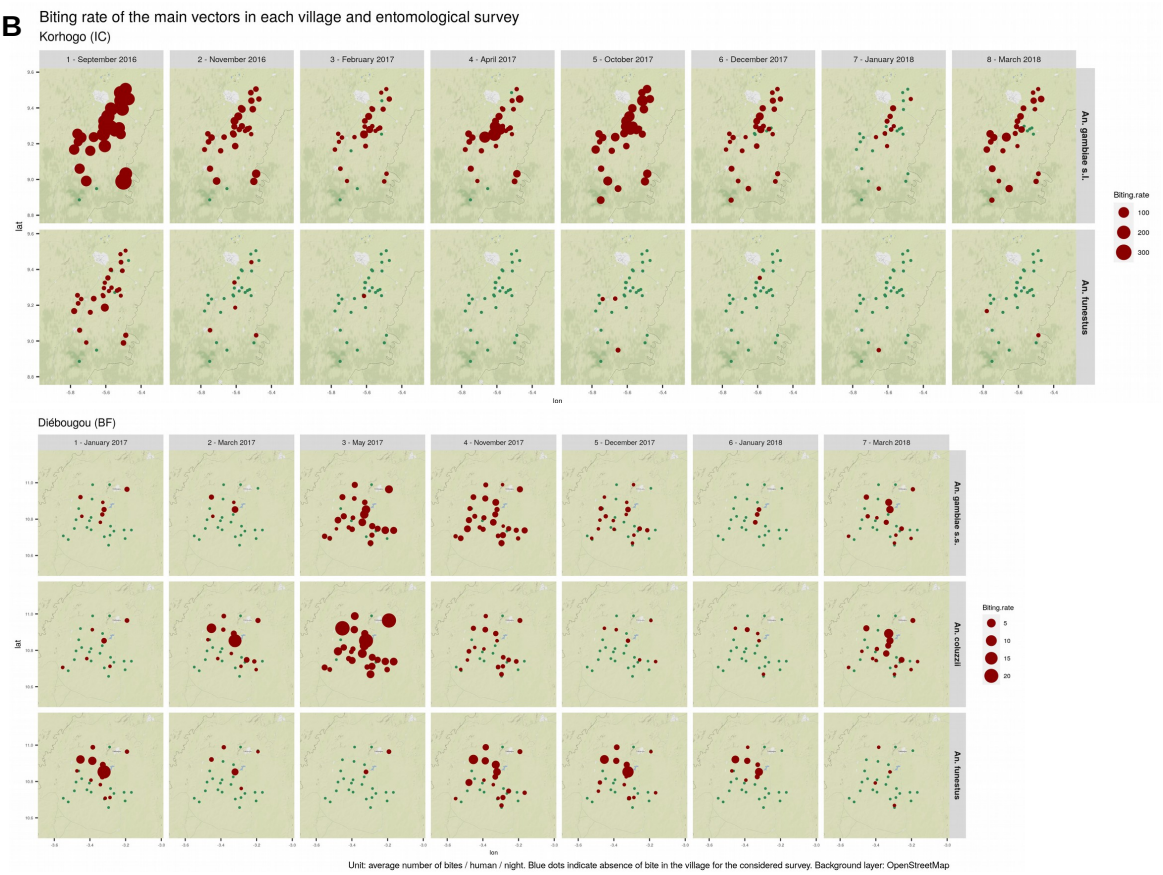
1.A Micro-climate

Micro-climatic conditions during each entomological survey



Solid bars indicate the mean for all the collection points in the considered entomological survey.
Grey ribbons indicates the mean ± the standard deviation (i.e. spatial variability).

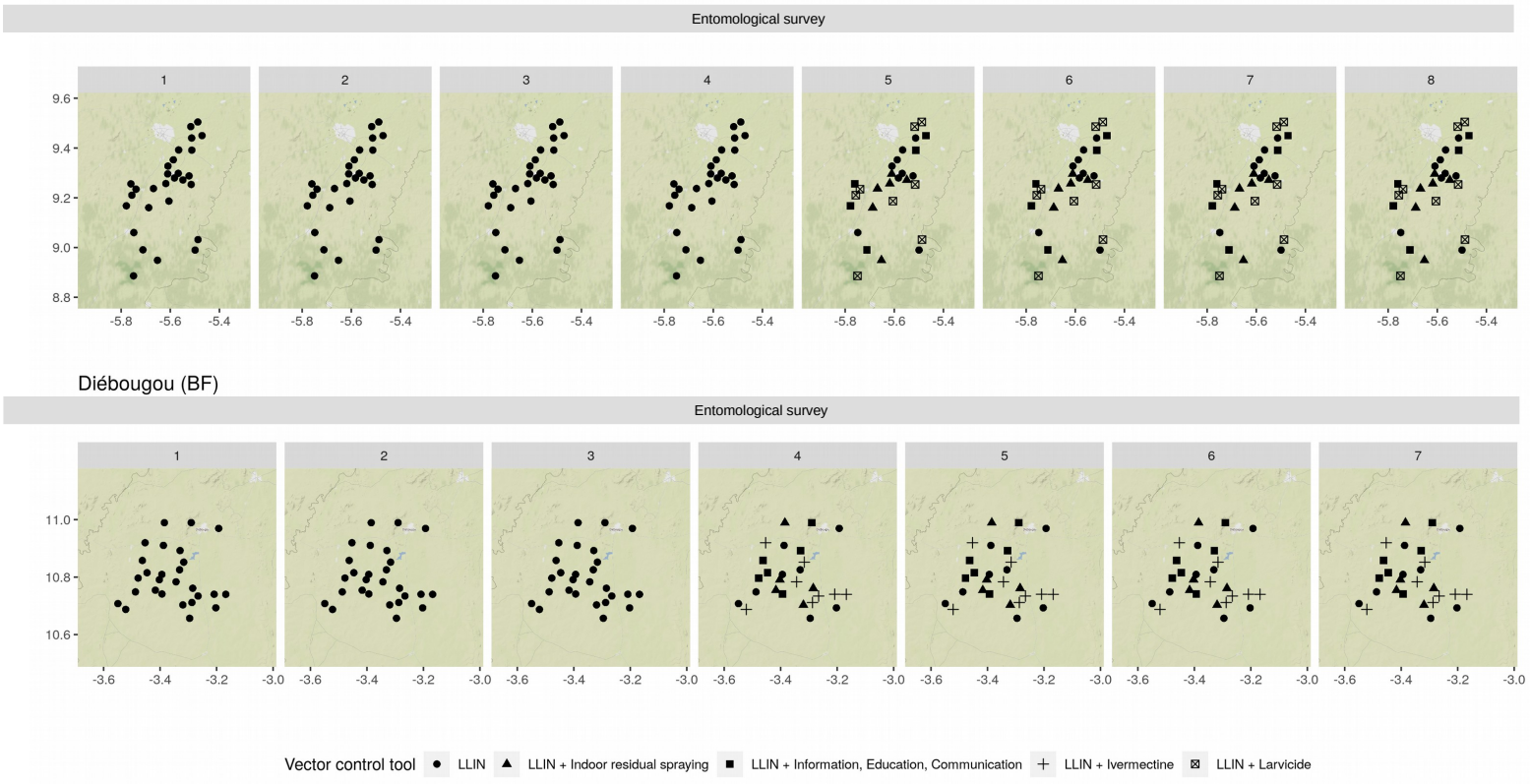
1.B Biting rate of the main vectors in each village and entomological survey



Unit: average number of bites / human / night. Blue dots indicate absence of bite in the village for the considered survey. Background layer: OpenStreetMap

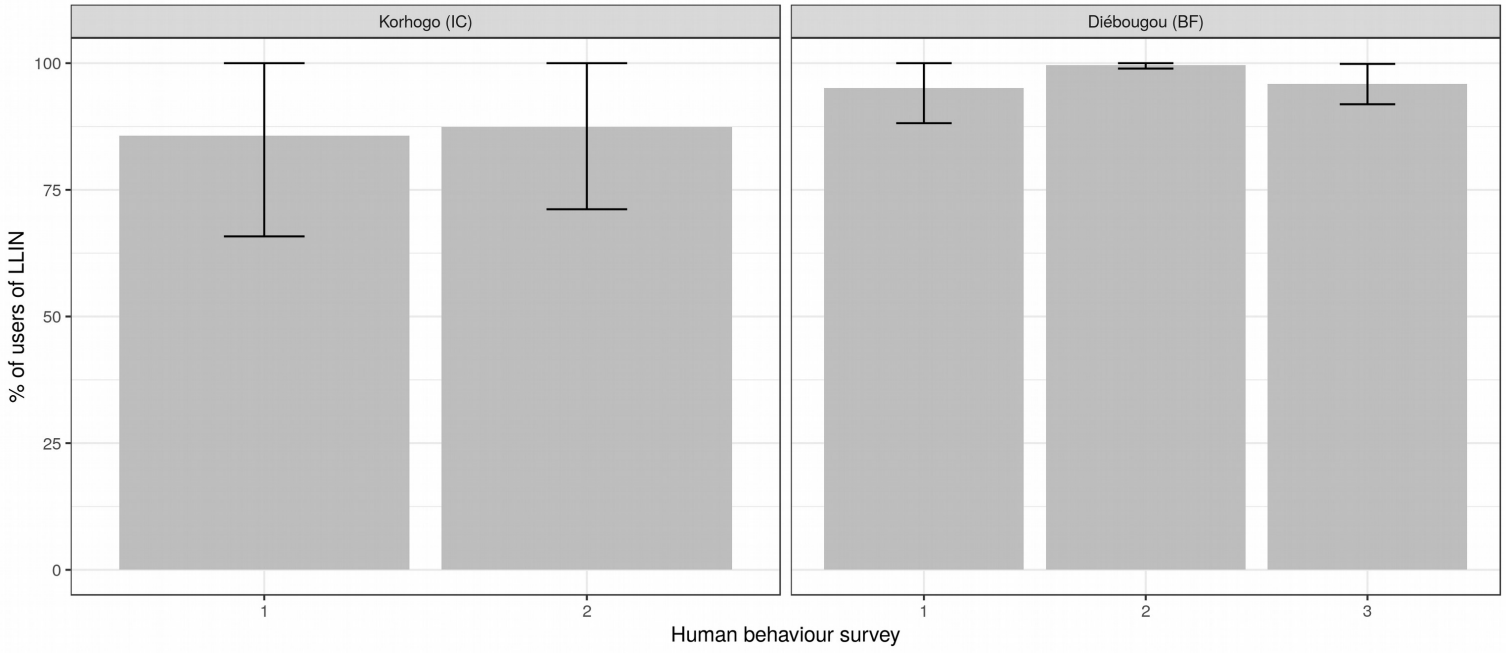
1.C Vector control

Vector control tool in use for each village and entomological survey



1.D Human behaviour

% of population that use LLIN for each study area and human behaviour survey

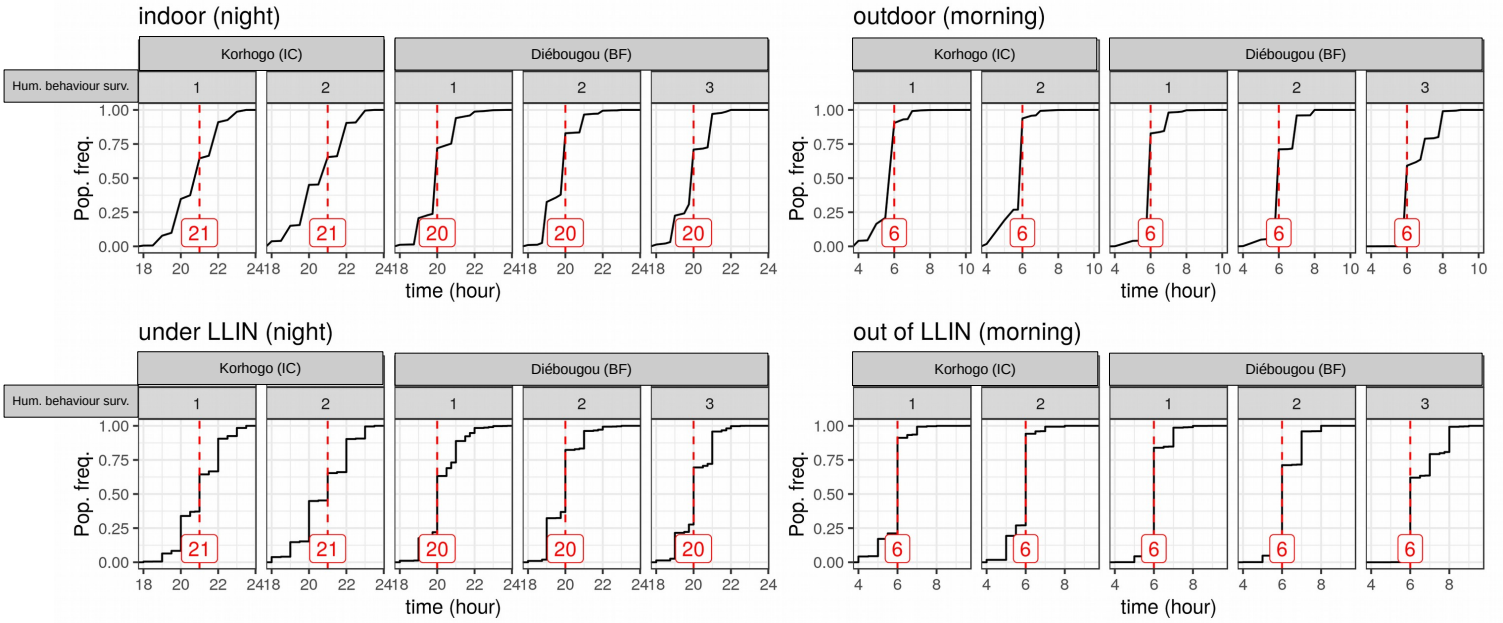


Values were averaged over all the villages.
Error bars indicate the mean +- sd for all the study villages for the considered survey

190

1.E Human behaviour

% of population indoor, outdoor, under an LLIN and out of LLIN for each study area and human behaviour survey

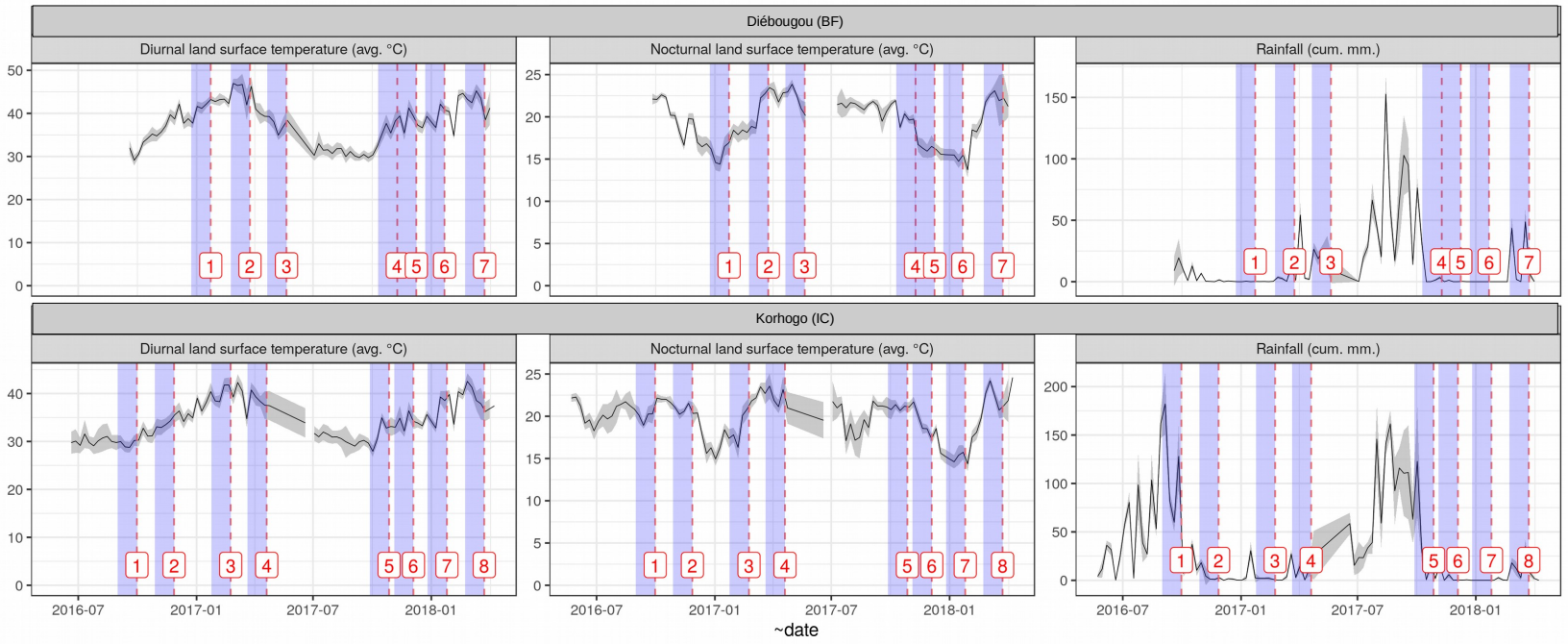


The red line indicates the median time.

1.F

Past weather

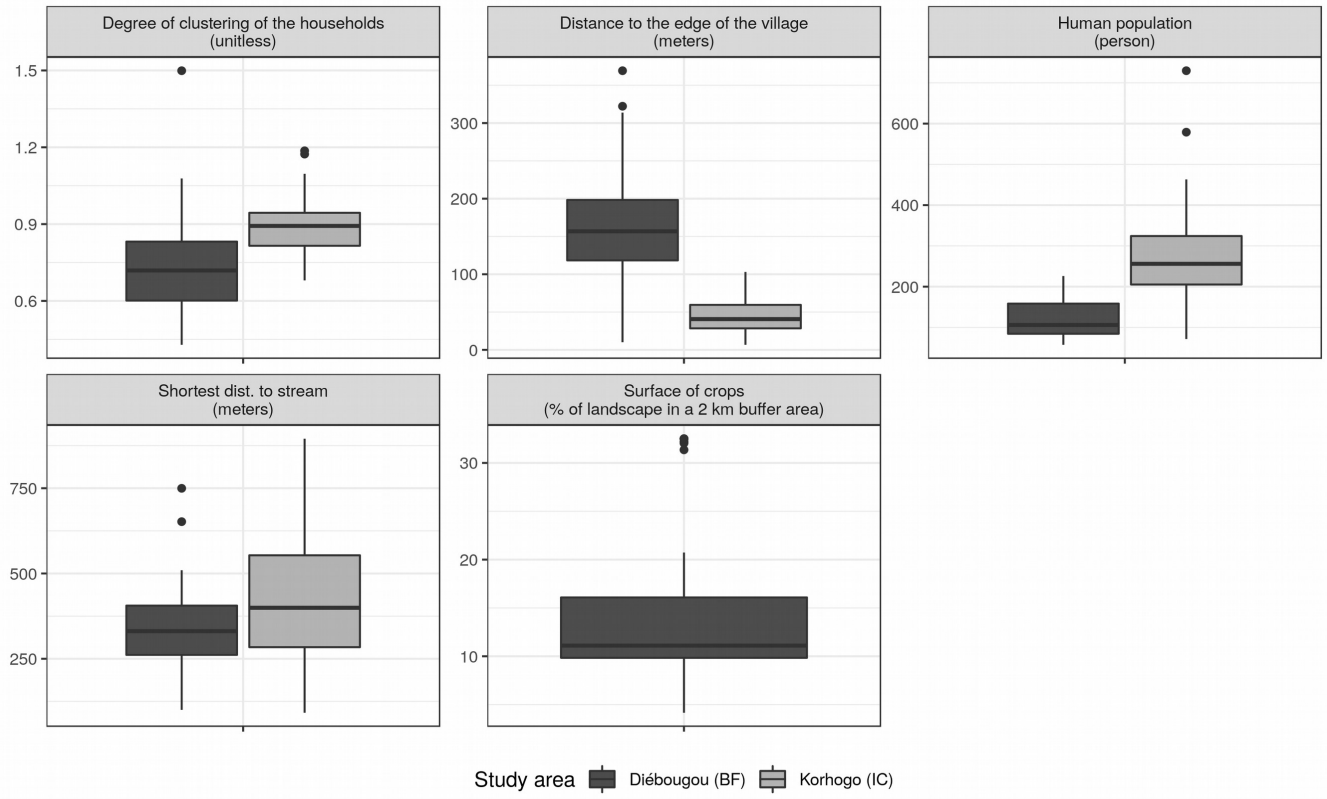
Weekly mean (for land surface temperature) and sum (for rainfall) in a 2 km buffer around each collection point.



Solid bars indicate the mean for all the collection points in the considered week.
Grey ribbons indicates the mean ± the standard deviation (i.e. spatial variability).
Red vertical lines indicate the dates of the entomological surveys.
Blue ribbons indicate a time frame of one month before each entomological survey.
data sources : GPM (rainfall), MODIS (temperature)

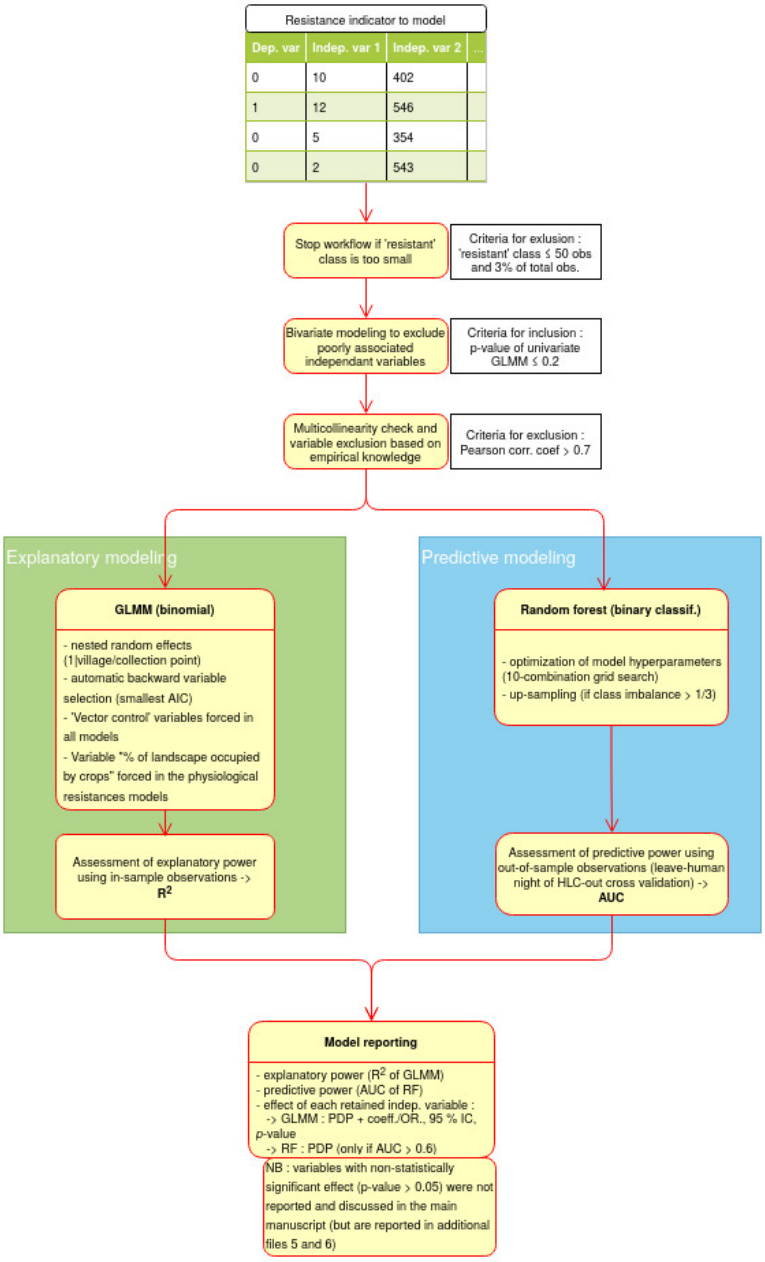
1.G Landscape

Landscape variables for each study area



Boxplots summarize the distribution over all the collection points for the considered study area.

Supplementary figure 2 : Graphical representation of the modeling workflow



Supplementary figure 3 : Charts of spatio-temporal distribution of vector abundance, by species and study area

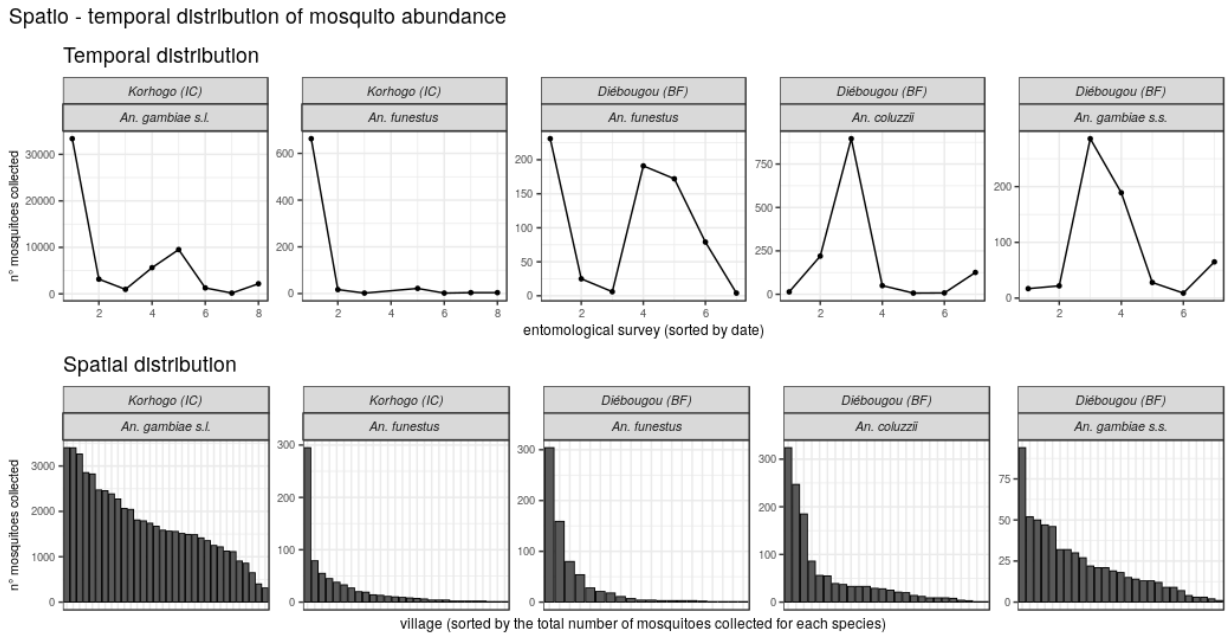


Figure 3: Charts of spatio-temporal distribution of vector abundance, by species and study area

Additional file 4 : Table of results of the GLMM (numerical values).

These tables were used to create Figure 3 of the main article. They provide the numerical values of the GLMM models (coefficients / odd ratios, 95 % IC, p-values).

Variable	*unit*	*estimate*	*Conf. Low (95 %)*	*Conf. High (95 %)*	*p.value*
exophagy - Korhogo (IC) - An. gambiae s.s.					
VC : LLIN only	comp. to before introduction of LLIN	0.84045	0.78977	0.89439	0.06
VC : LLIN + IRS		0.89969	0.79929	1.0127	0.08
VC : LLIN : Larvicide		0.91187	0.73841	1.12606	0.391
Pop. under LLIN (hour of collection) * **	**%**	**1.00243**	**1.0014**	**1.00345**	**< 0.001**
Δ luminosity ind./out. (hour of collection) * **	**Lux**	**1.00085**	**1.0004**	**1.00131**	**< 0.001**
Indoor Humidity (hour of collection) * **	**%**	**0.99675**	**0.99544**	**0.99806**	**< 0.001**
Indoor Temperature (hour of collection) * **		**0.99658**	**0.99527**	**0.99789**	**< 0.001**
**Δ humidity ind./out. (hour of collection) * **	**%**	**0.9929**	**0.98716**	**0.99867**	**0.016**
Rainfall (day of collection) * **	**cum. mm.**	**0.9867**	**0.98135**	**0.99208**	**< 0.001**
Diurnal temp. (month preceding coll.) * **	**°C**	**0.94927**	**0.93905**	**0.95959**	**< 0.001**
Rainfall (month preceding coll.) * **	**cum. mm.**	**0.99838**	**0.99799**	**0.99878**	**< 0.001**
exophagy - Korhogo (IC) - An. funestus					
VC : LLIN + IRS	comp. to before introduction of LLIN	0.63	0.11	3.64	0.604
Indoor temperature (hour of collection)	°C	1.31737	0.9723	1.7849	0.075
**Δ humidity ind./out. (hour of collection) * **	**%**	**1.13832**	**1.03053**	**1.25739**	**0.011**
**Human population ** **	**person**	**0.99526**	**0.99183**	**0.99869**	**0.007**
**Distance to the edge of the village * **	**meters**	**1.01241**	**1.00071**	**1.02424**	**0.038**
Atmospheric pressure (hour of collection)	hpa	1.13222	0.95531	1.34189	0.152
Nocturnal temp. (month preceding coll.)	°C	0.72019	0.46172	1.12333	0.148
exophagy - Diébougou (BF) - An. funestus					
VC : LLIN + IEC	comp. to LLIN only	1.31	0.66	2.59	0.441
VC : LLIN + IRS		0.63	0.19	2.08	0.452
VC : LLIN + ivermectine		1.37	0.77	2.46	0.284
Time since distrib. of LLIN	month	0.97	0.92	1.03	0.344
Pop. under LLIN (hour of collection) * **	**%**	**1.0194**	**1.01014**	**1.02874**	**< 0.001**
**Δ humidity ind./out. (hour of collection) ** **	**%**	**1.03053**	**1.00814**	**1.05341**	**0.007**
**Δ temperature ind./out. (hour of collection) ** **	**°C**	**1.12609**	**1.03029**	**1.2308**	**0.009**
Nocturnal temp. (month preceding coll.)	°C	1.07719	0.9881	1.17432	0.091
exophagy - Diébougou (BF) - An. coluzzii					
VC : LLIN + IEC	comp. to LLIN only	1.38	0.58	3.28	0.463
VC : LLIN + IRS		0.72	0.16	3.21	0.663
VC : LLIN + ivermectine		1.52	0.68	3.4	0.306
Time since distrib. of LLIN	month	0.99	0.92	1.05	0.661
**Atmospheric pressure (hour of collection) * **	**hpa**	**0.93766**	**0.88005**	**0.99904**	**0.047**
**Shortest dist. to stream * **	**m.**	**1.00083**	**1.00017**	**1.00149**	**0.013**
Rainfall (day of collection) * **	**cum. mm.**	**1.04334**	**1.02337**	**1.06369**	**< 0.001**
**Nocturnal temp. (month preceding coll.) ** **	**°C**	**0.86299**	**0.78727**	**0.94601**	**0.002**
exophagy - Diébougou (BF) - An. gambiae ss.					
VC : LLIN + IEC	comp. to LLIN only	1.36	0.65	2.86	0.419
VC : LLIN + IRS		1.85	0.5	6.91	0.357
VC : LLIN + ivermectine		1.61	0.82	3.15	0.164
Time since distrib. of LLIN	month	1.01355	0.94546	1.08654	0.704
Δ luminosity ind./out. (hour of collection)	Lux	1.00054	0.99978	1.00129	0.163
Δ temperature ind./out. (hour of collection)	°C	0.93502	0.85516	1.02235	0.14
**Nocturnal temp. (month preceding coll.) * **	**°C**	**1.09949**	**1.00554**	**1.20223**	**0.037**

Variable	*unit*	*estimate*	*Conf. Low (95 %)*	*Conf. High (95 %)*	*p.value*
late_biting - Korhogo (IC) - An. gambiae s.s.					
VC : LLIN only	comp. to before introduction of LLIN	1.2	0.82	1.76	0.345
VC : LLIN + IRS		0.54	0.294	1.006	0.052
VC : LLIN : Larvicide		0.86	0.24	3.01	0.81
Temperature (night of collection) ***	**°C**	**1.24835**	**1.1334**	**1.37496**	**< 0.001**
Rainfall (day of collection) ***	**cum. mm.**	**1.0704**	**1.03891**	**1.10284**	**< 0.001**
Nocturnal temp. (month preceding coll.) **	**°C**	**0.81169**	**0.70589**	**0.93334**	**0.001**
late_biting - Diébougou (BF) - An. funestus					
VC : LLIN + IEC	comp. to LLIN only	0	0	Inf	1
VC : LLIN + IRS		108	5.5	1380	0.061
VC : LLIN + ivermectine		2.31	0.34	15.67	0.390
Time since distrib. of LLIN	month	1.004	0.84	1.19	0.960
Place (exterior) ***	**comp. to interior**	**0.15**	**0.06**	**0.35**	**< 0.001**
Degree of clustering of the households	NA	0.06	0.0025	1.68	0.10
Diurnal temp. (day of collection) ***	**°C**	**1.30**	**1.16**	**1.47**	**< 0.001**
early_biting - Korhogo (IC) - An. gambiae s.s.					
VC : LLIN only	comp. to before introduction of LLIN	0.85	0.7	1.03	0.067
VC : LLIN + IRS		2.88	1.01	4.13	0.053
VC : LLIN : Larvicide		1.18	0.55	2.54	0.672
LLIN use rate ***	**% of LLIN users**	**1.022**	**1.011**	**1.031**	**< 0.001**
Place (exterior) ***	**comp. to interior**	**0.77**	**0.7**	**0.86**	**< 0.001**
Distance to the edge of the village *	**meters**	**0.99564**	**0.99191**	**0.99938**	**0.031**
Diurnal temp. (month preceding coll.) **	**°C**	**0.96181**	**0.93173**	**0.98287**	**0.019**
Diurnal temp. (day of collection) ***	**°C**	**1.03622**	**1.01426**	**1.06865**	**< 0.001**
Rainfall (month preceding coll.)	cum. mm.	1.0011	0.9999	1.002	0.067
early_biting - Korhogo (IC) - An. funestus					
VC : LLIN + IRS	comp. to before introduction of LLIN	37.69	0.23	6286.6	0.164
Rainfall (month preceding coll.)	cum. mm.	1.00988	0.99737	1.02254	0.122

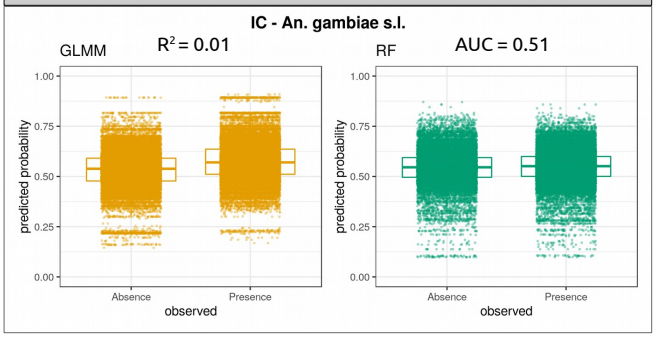
Variable	*unit*	*estimate*	*Conf. Low (95 %)*	*Conf. High (95 %)*	*p.value*
physiological_resistance_kdrw - Diébougou (BF) - An. coluzzii					
VC : LLIN + IEC		0.69876	0.36345	1.3434	0.282
VC : LLIN + IRS	comp. to LLIN only	1.12889	0.36818	3.46133	0.832
VC : LLIN + ivermectine		0.90256	0.49908	1.6322	0.734
Presence of cotton field	comp. to absence	0.936	0.684	0.681	1.29
% landscape used by rice fields	% land.	1.0014	0.89547	1.11986	0.98
% landscape used by other crops	% land.	0.99023	0.97474	1.00597	0.223
Time since distrib. of LLIN	month	1.02239	0.97445	1.07268	0.366
Rainfall (month preceding coll.) * **	**cum. mm.**	**0.99374**	**0.99119**	**0.99629**	**< 0.001**
physiological_resistance_kdrw - Diébougou (BF) - An. gambiae ss.					
VC : LLIN + IEC		0.74948	0.2099	2.67615	0.657
VC : LLIN + IRS	comp. to LLIN only	20030806	0	Inf	0.997
VC : LLIN + ivermectine		1.00618	0.32877	3.07937	0.991
Presence of cotton field	comp. to absence	1.94	0.627	6.00	0.25
% landscape used by rice fields	% land.	0.97285	0.76284	1.24069	0.824
% landscape used by other crops	% land.	0.97004	0.92564	1.01657	0.203
Time since distrib. of LLIN	month	1.04652	0.94669	1.15688	0.374
Humidity (hour of collection) *	**%**	**0.98154**	**0.9653**	**0.99805**	**0.029**
Diurnal temp. (month preceding coll.) * **	**°C**	**0.79397**	**0.71185**	**0.88556**	**< 0.001**
Rainfall (month preceding coll.) * **	**cum. mm.**	**0.98743**	**0.98051**	**0.9944**	**< 0.001**
physiological_resistance_kdre - Diébougou (BF) - An. coluzzii					
VC : LLIN + IEC		0.27634	0.10426	0.73248	0.51
VC : LLIN + IRS	comp. to LLIN only	0.47876	0.11241	2.03905	0.319
VC : LLIN + ivermectine		0.64367	0.30667	1.35099	0.244
Presence of cotton field	comp. to absence	1.09	0.788	0.595	1.98
% landscape used by rice fields	% land.	1.00576	0.80677	1.25382	0.959
% landscape used by other crops	% land.	1.01184	0.98093	1.04372	0.457
Time since distrib. of LLIN * **	**month**	**1.22976**	**1.1494**	**1.31574**	**< 0.001**
**LLIN use rate ** **	**% of LLIN users**	**1.07058**	**1.02013**	**1.12352**	**0.006**
Humidity (hour of collection) * **	**%**	**1.02221**	**1.01089**	**1.03365**	**< 0.001**
**Atmospheric pressure (hour of collection) ** **	**hpa**	**1.13422**	**1.04231**	**1.23423**	**0.003**
**Temperature (hour of collection) * **	**°C**	**1.07941**	**1.01805**	**1.14446**	**0.01**
**Rainfall (month preceding coll.) ** **	**cum. mm.**	**1.01243**	**1.0048**	**1.02011**	**0.001**
Nocturnal temp. (month preceding coll.) * **	**°C**	**0.75079**	**0.6361**	**0.88615**	**0.001**
physiological_resistance_kdre - Diébougou (BF) - An. gambiae ss.					
VC : LLIN + IEC		0.98472	0.37974	2.55351	0.975
VC : LLIN + IRS	comp. to LLIN only	1.15975	0.27173	4.94979	0.841
VC : LLIN + ivermectine		0.41481	0.17695	0.97241	0.053
Presence of cotton field	comp. to absence	1.06	0.897	0.429	2.63
% landscape used by rice fields	% land.	0.98149	0.79241	1.21569	0.864
% landscape used by other crops	% land.	1.01141	0.97087	1.05364	0.587
**Time since distrib. of LLIN ** **	**month**	**1.15876**	**1.06122**	**1.26526**	**0.001**
**LLIN use rate * **	**% of LLIN users**	**1.08402**	**1.01426**	**1.15858**	**0.017**
**Atmospheric pressure (hour of collection) ** **	**hpa**	**1.17771**	**1.04061**	**1.33287**	**0.01**
physiological_resistance_ace1 - Diébougou (BF) - An. gambiae ss.					
VC : LLIN + IEC		1.0264	0.59355	1.77491	0.926
VC : LLIN + IRS	comp. to LLIN only	0.37011	0.10394	1.31793	0.125
VC : LLIN + ivermectine		1.13731	0.70414	1.83695	0.599
Presence of cotton field	comp. to absence	1.13	0.509	0.788	1.62
% landscape used by rice fields	% land.	1.07333	0.98084	1.17453	0.124

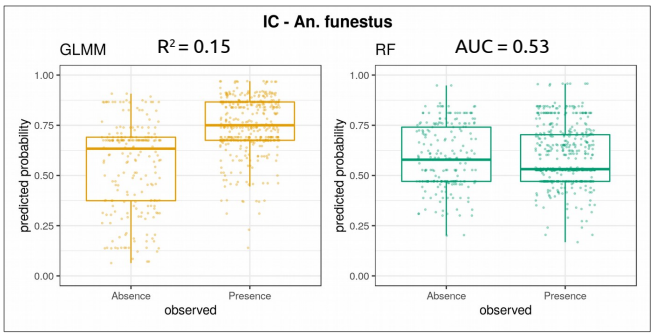
% landscape used by other crops	% land.	1.00899	0.99045	1.02787	0.344
Time since distrib. of LLIN	month	1.00036	0.94788	1.05573	0.99
Diurnal temp. (month preceding coll.) **	**°C**	**0.92129**	**0.86766**	**0.97824**	**0.007**
Nocturnal temp. (month preceding coll.) *	**°C**	**0.9144**	**0.85191**	**0.98148**	**0.013**

Additional figure 5 : Boxplots of observed resistance status vs. predicted probabilities for all the models.

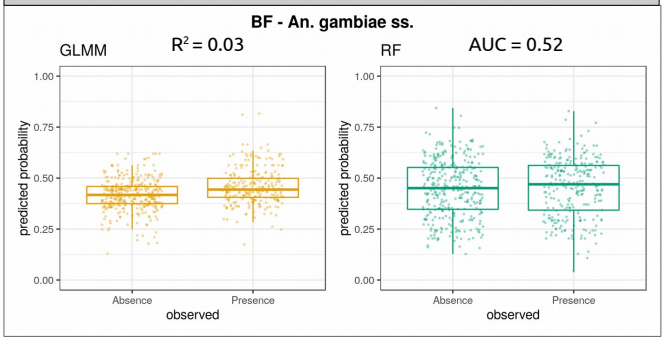
Exophagy

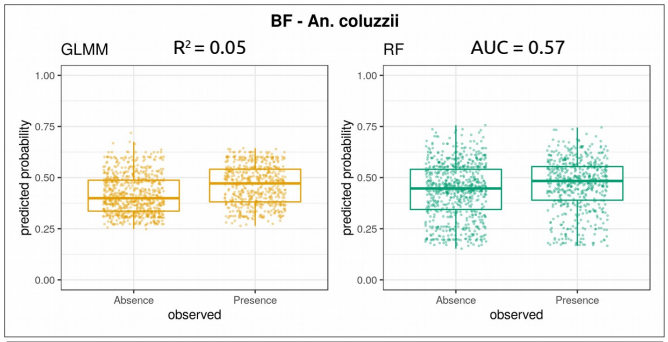
Korhogo (IC)

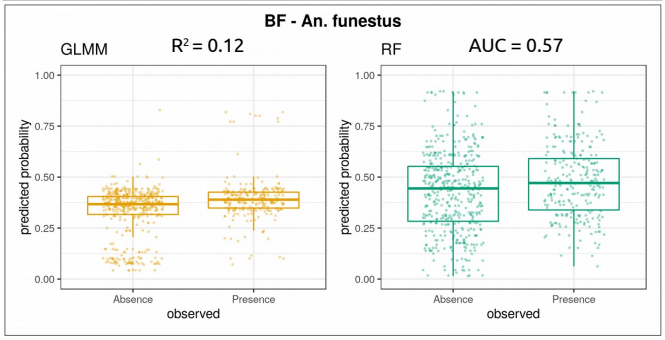




Diébougou (BF)



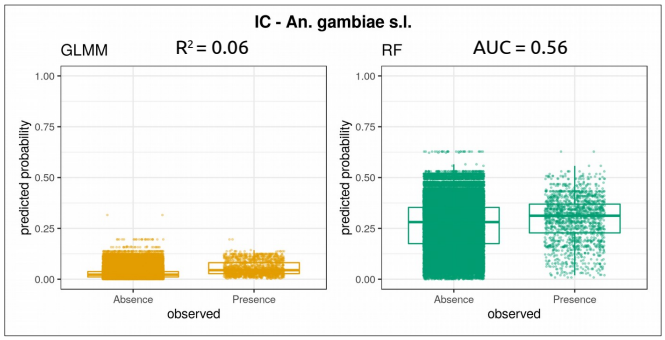


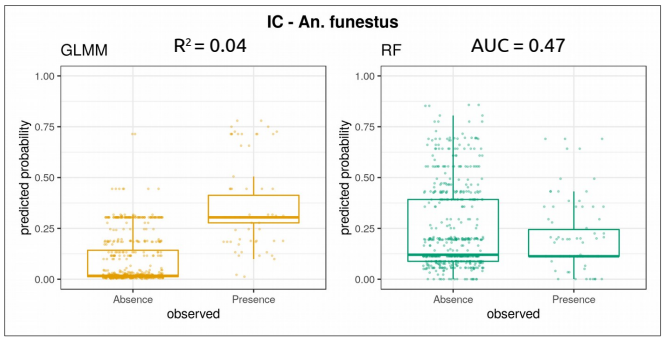


202

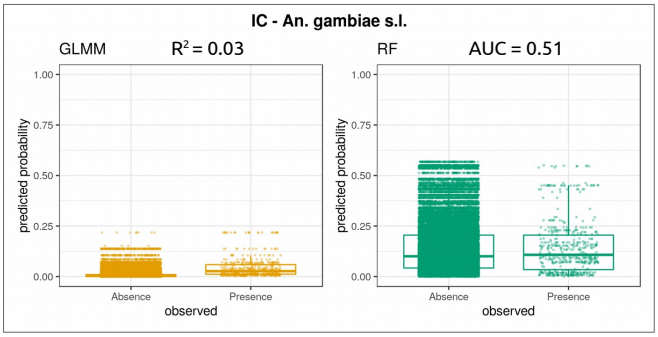
Early & late biting

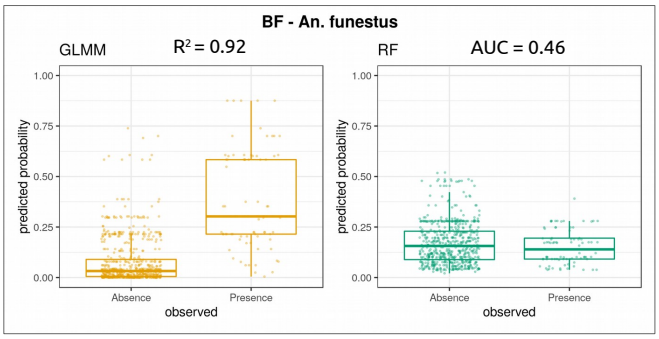
Early biting





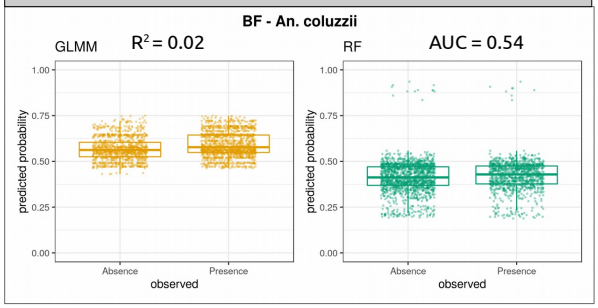
Late biting

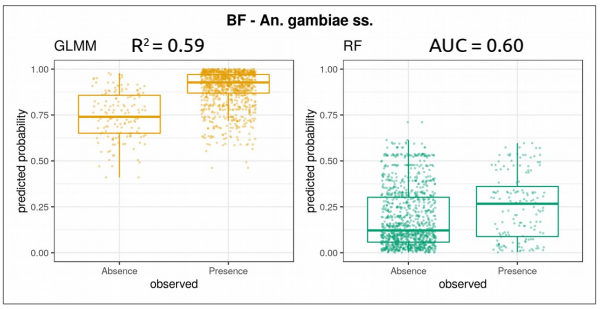




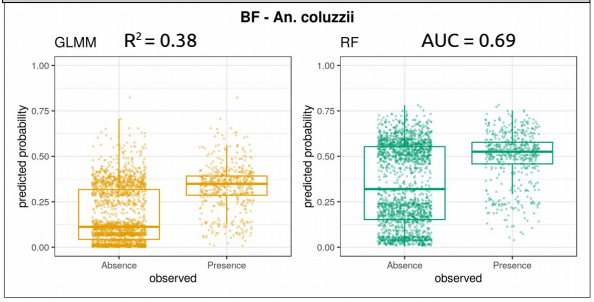
Kdr-w, Kdr-e, Ace-1 (Diébougou (BF) only)

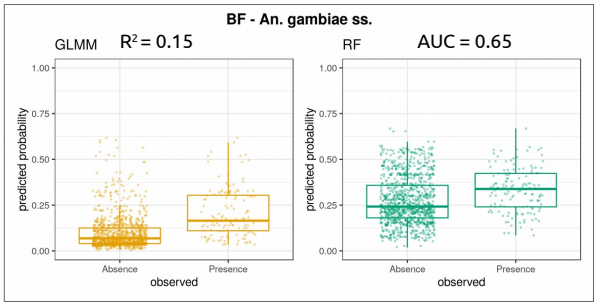
Kdr-w



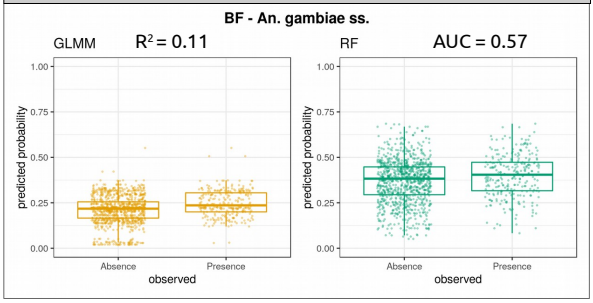


Kdr-e





Ace-1



Chapitre 6

Articles n°3 et 4 - Etudes complémentaires : contributions à des travaux de modélisation liés à la transmission du paludisme

Ce chapitre présente deux travaux complémentaires, auxquels nous avons contribué dans le cadre de la thèse, de modélisation du risque de transmission (article n°3) ou de la transmission à proprement parler (article n°4) du paludisme dans la zone d'étude de Diébougou. Ces deux études ont fait l'objet de publications scientifiques que nous résumons ci-après.

6.1 Article n°3 - Modélisation de l'exposition humaine à la piqûre d'anophèle

Introduction à l'article

Le risque de transmission résiduelle du paludisme, à savoir la probabilité de contact homme-vecteur, dépend du comportement d'une part anophélien - ses horaires et sites d'activités de recherche de repas de sang - et d'autre part humain - ses horaires d'utilisation des moustiquaires et habitudes nocturnes. Le contact se produit quand les humains ne sont pas protégés par les moustiquaires et que simultanément,

les anophèles sont à la recherche d'un repas de sang. Aussi, avant de concevoir et déployer des mesures de LAV complémentaires aux MIILDA, il est important de mesurer le niveau de protection conféré par les moustiquaires (quantifier la transmission résiduelle) et de comprendre où (intérieur ou extérieur des habitations) et quand (à quels horaires de la nuit) les populations sont exposées à la piqûre (caractériser la transmission résiduelle). Cette connaissance permet d'élaborer des outils complémentaires de LAV efficaces, qui ciblent la part résiduelle de la transmission. L'étude présentée dans ce chapitre avait ainsi pour objectif de quantifier et caractériser le risque de transmission résiduelle du paludisme dans la zone d'étude de Diébougou, à l'aide des données de comportements humains et anophéliens et de modèles mathématiques éprouvés d'exposition à la piqûre. Cette étude a fait l'objet d'une publication scientifique, que nous résumons ci-dessous. Le texte intégral est disponible en annexe E et à l'adresse suivante : https://doi.org/10.1186/s12889-021-10304-y.

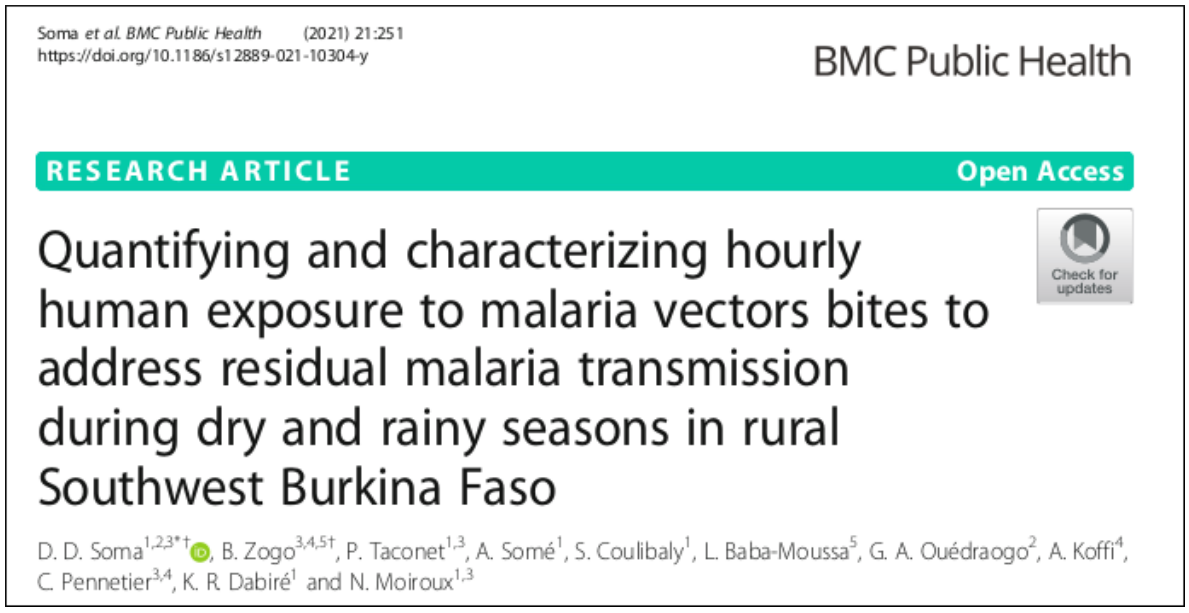


FIGURE 6.1: Couverture de la publication n°3

6.1.1 Résumé de l'article

Cette étude avait pour objectif de caractériser et quantifier la transmission résiduelle dans la zone d'étude de Diébougou (avant la mise en place de l'intervention dans le cadre de l'ERC) : mesurer le taux de protection conféré par les MIILDA, caractériser les

sites et horaires où la population humaine est exposée à la piqûre d'anophèle, mesurer l'hétérogénéité spatio-temporelle de l'exposition au sein de la zone.

Nous avons utilisé une méthode permettant l'étude des intéractions comportementales entre les moustiques et les humains, décrite par Gerry F. Killeen et al. (2006) puis améliorée par Geissbühler et al. (2007). Cette approche mathématique consiste à croiser des données de comportements nocturnes horaires des anophèles (densités agressives horaires à l'intérieur et à l'extérieur des habitations) et des personnes (horaires d'entrée et sortie des habitations, horaires de sommeil et utilisation ou non de moustiquaire). En sortie, les modèles procurent des informations sur l'exposition des humains aux piqûres d'anophèles, à chaque heure de la nuit. Nous avons ainsi utilisé les données horaires de comportement humain [1] et d'agressivité des anophèles collectées dans la zone de Diébougou en entrée de ce modèle mathématique. Nous avons stratifié l'étude par village, saison et classe d'âge de la population, afin d'affiner la connaissance sur les populations les plus à risque et les éventuelles hétérogénéités spatio-temporelles de l'exposition à la piqûre. Nous avons finalement judicieusement agrégé les sorties des modèles pour calculer les indicateurs de transmission résiduelle suivants :

— efficacité moyenne réelle de la protection personnelle offerte par l'utilisation d'une MIILDA (proportion de l'exposition aux piqûres qui est évitée par l'utilisation d'une MIILDA),
— proportion de l'exposition se produisant à l'intérieur des habitations,
— proportion de l'exposition se produisant avant 20 h ou après 5 h, correspondant aux horaires respectivement précédant et succédant les périodes ou la majorité (> 50 %) des utilisateurs de MIILDA sont protégés.

Le taux moyen déclaré d'utilisation des MIILDA était très élevé, quelle que soit la saison ou la tranche d'âge (minimum : 92.45% pour les + de 18 ans en saison sèche-chaude ; maximum : 100% pour les 0 à 5 ans en saison pluvieuse). Nous avons noté de légères variations dans les taux d'utilisation des MIILDA selon les saisons (taux légèrement inférieurs en saison sèche par rapport à la saison humide), les tranches d'âge (taux légèrement inférieurs chez les adultes par rapport aux enfants), et les villages. Les populations humaines étaient exposées quasiment exclusivement à l'intérieur de leurs

1. issues des enquêtes de comportement humain relatif à l'utilisation des moustiquaires et aux habitudes horaires nocturnes, voir annexe A

habitations (94 % de l'exposition se déroulait en intérieur) ; cependant, les MIILDA protégaient très efficacement de cette exposition (efficacité moyenne réelle comprise entre 80% et 85% selon les saisons). Le pic d'exposition résiduelle avait lieu à l'intérieur entre 5h et 6h du matin (33% à 57% de l'exposition résiduelle pour les utilisateurs de MIILDA), entre la sortie de l'espace de sommeil protégé par la moustiquaire et la sortie de l'habitation. Les piqûres précoces (avant 20h) représentaient moins de 12% de l'exposition résiduelle.

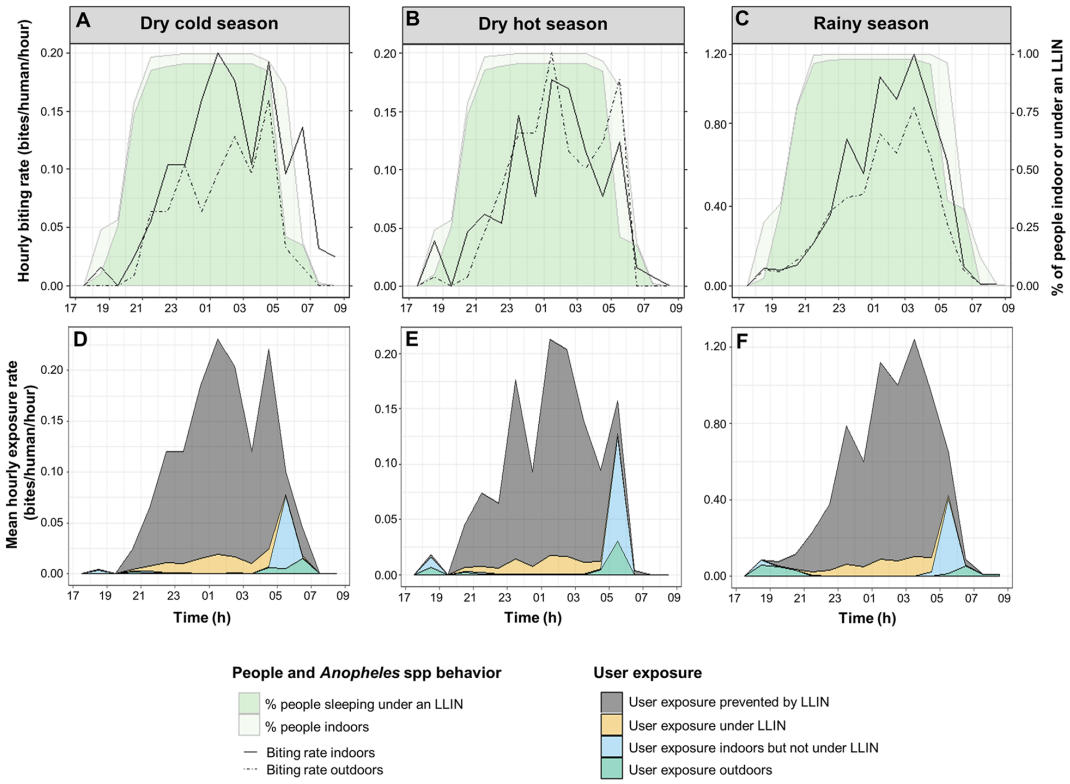


FIGURE 6.2: Comportement horaire humain et anophélien (A, B, C) et exposition humaine horaire aux piqûres des utilisateurs de MIILDA (D, E, F)

Cette étude a mis en évidence l'importance des MIILDA dans la lutte contre la transmission du paludisme dans la zone de Diébougou. Elle a notamment montré que les populations les plus à risque (enfants de moins de 5 ans) étaient aussi les plus protégées. Les taux déclarés d'utilisation des MIILDA étaient supérieurs à ceux rapportés par l'OMS, ce qui pourrait être dû au fait que les enquêtes de comportement humain ont été effectuées peu de temps après la dernière distribution universelle de

moustiquaires. L'exposition résiduelle à la piqûre d'anophèle s'effectuant en majorité à l'intérieur des habitations, les outils ou mesures de LAV complémentaires devraient cibler prioritairement les vecteurs endophages (mesures telles que peintures insecticides, fermeture des avant-toits, fermeture des plafonds, ou encore grillage/moustiquaires aux fenêtres).

Dans la zone de Diébougou, les MIILDA protègent à priori considérablement des piqûres des vecteurs du paludisme. Cependant, les utilisateurs de moustiquaires restent exposés principalement à l'intérieur de leurs habitations, le matin. Aussi, la combinaison MIILDA + outil complémentaire de LAV visant les vecteurs endophages est probablement la plus efficace pour lutter contre la transmission résiduelle du paludisme dans cette zone.

6.2 Article n°4 - Modélisation des dynamiques spatio-temporelles des cas de paludisme

Introduction à l'article

Les travaux des chapitres 4 et 5 ont montré comment les images satellitaires et les modèles statistiques peuvent aider à comprendre et prédire les dynamiques entomologiques spatio-temporelles sur nos territoires d'étude ; ces éléments permettant d'optimiser la conception et le déploiement des outils de LAV. Mais ces mêmes outils peuvent également être utilisés en épidémiologie, pour expliquer ou prédire la distribution spatio-temporelle d'indicateurs épidémiologiques tels que la prévalence du paludisme. Une telle prédiction permet alors d'optimiser le déploiement de mesures de prévention, diagnostic et de traitement de la maladie. Dans cette dernière étude, nous avons étudié les dynamiques spatio-temporelles des cas de paludisme dans la zone de Diébougou. Nous montrons i) que la distribution spatiale des cas de paludisme est hétérogènes au sein du district sanitaire ; et ii) que nous pouvons y anticiper les pics épidémiques plusieurs semaines à l'avance grâce aux données satellitaires et aux modèles statistiques. Cette étude a fait l'objet d'une publication scientifique, que nous résumons ci-dessous. Le texte intégral est disponible en annexe F et à l'adresse suivante : https://doi.org/10.1038/s41598-021-99457-9.

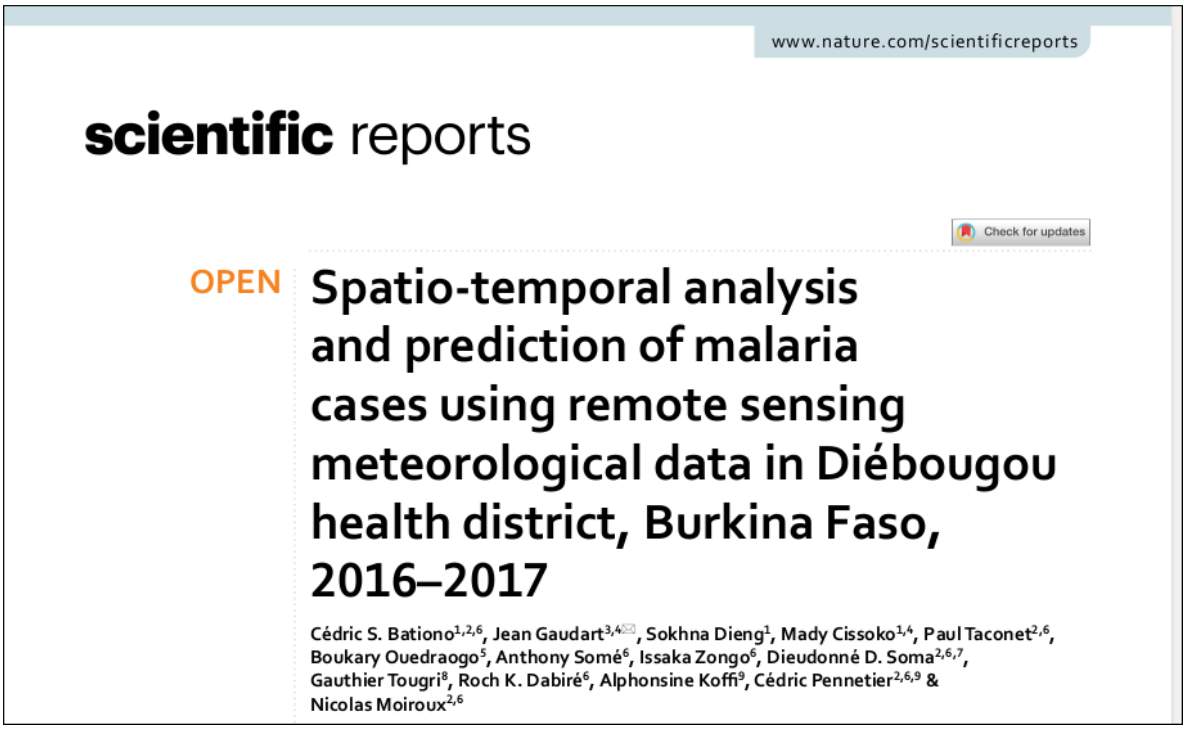


FIGURE 6.3: Couverture de la publication n°4

6.2.1 Résumé de l'article

Les objectifs de cette étude étaient i) d'étudier la dynamique spatiale de l'épidémiologie du paludisme dans la zone de Diébougou, en détectant d'éventuels 'points chauds' spatiaux de cas de paludisme et ii) de modéliser les dynamiques temporelles des cas de paludisme dans la zone d'étude en utilisant des données météorologiques satellitaires, afin d'appréhender la capacité à anticiper les épidémies de la maladie dans le district sanitaire.

Les données hebdomadaires de cas de paludisme en 2016 et 2017 ont été collectées dans 13 centres de santé de la zone. Les points chauds spatiaux de cas de paludisme ont été détectés en utilisant des méthodes d'analyse statistique spatiale. Les cas reportés de paludisme ont été comparés à la distance euclidienne du village au centre de santé et à la densité agressive des vecteurs afin de tenter d'expliquer les différences observées entre les villages. Pour l'analyse temporelle, les données météorologiques ont été extraites de produits satellitaires ou issues de modèles météorologiques, disponibles à l'échelle mondiale. Ces données sont produites en routine par le Centre Européen pour les Prévisions Météorologiques à Moyen Terme. Deux variables statistiques synthétisant les

données météorologiques à l'échelle hebdomadaire (*Synthetic Meteorological Indicator* ou SMI) ont été construites par analyse en composantes principales. Dans une première analyse bivariée, des modèles additifs généralisés (*Generalized Additive Model* ou GAM) ont été entrainés pour modéliser le nombre de cas de paludisme à l'échelle de la zone d'étude entière et au pas de temps hebdomadaire en fonction de chacune des deux variables météorologiques, utilisées tour à tour sur chacune des 30 semaines précédant les cas de paludisme à expliquer/prédire. Les variables dont les 'lags' temporels conduisaient à la plus petite erreur dans ces analyses bivariées ont ensuite été utilisées en tant que variables indépendantes dans un GAM mixte multivarié, entrainé sur les données d'une année entière. La capacité prédictive de ce modèle multivarié a finalement été évaluée en prédisant le nombre de cas de paludisme sur une période de 17 semaines non utilisées pour entrainer le modèle (correspondant au pic épidémique suivant, en 2017) et en évaluant graphiquement la superposition des courbes épidémiologiques observées et prédites par le modèle statistique.

Nous avons observé que la distribution des cas de paludisme était hétérogène à l'échelle du district sanitaire, à la fois spatialement et temporellement. Quatre points chauds spatiaux ont été identifiés, regroupant chacun entre un et trois villages (figure 6.4). La distribution spatiale des cas de paludisme n'était corrélée ni à la distance euclidienne au centre de santé, ni aux densités agressives des anophèles. Aussi, l'hétérogénéité spatiale de la densité des vecteurs n'explique à priori pas, dans cette zone, celle de la prévalence des cas. Pour l'accès aux centres de santé, des variables plus réalistes pourraient être testées (par exemple, distance réelle au centre, ou encore accessibilité pendant la saison pluvieuse). D'autres facteurs, restants à explorer, pourraient expliquer l'hétérogénéité spatiale des cas de paludisme dans notre zone (différences dans les niveaux d'éducation, revenus, activités profesionnelles, possession et utilisation des MIILDA, etc.).

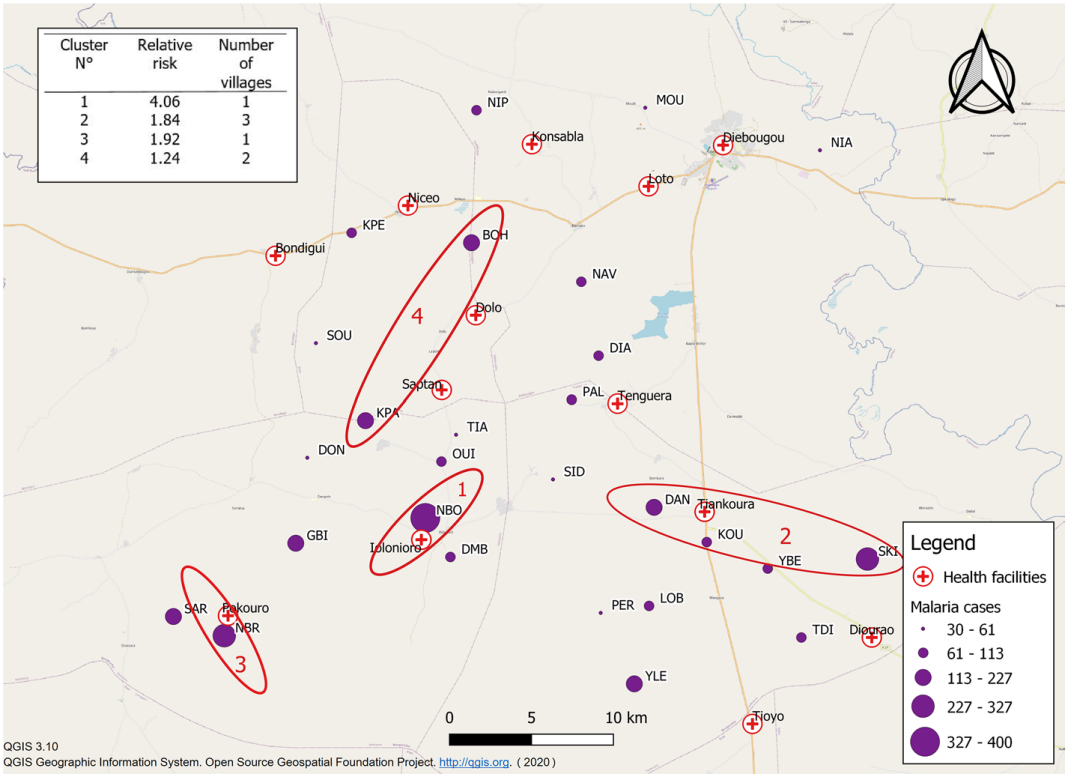


FIGURE 6.4: Distribution spatiale des cas de paludisme reportés dans les 27 villages de la zone de Diébougou pour l'année épidémique 2016-2017. Les cercles rouges représentent les point chauds

Au niveau temporel, un pic épidémique a été relevé entre les mois d'août et novembre 2016. Les SMI présentant les meilleurs capacités prédictives (c.a.d. les plus faibles erreurs) étaient situés respectivement 9 et 16 semaines avant les cas prédits, pour chacun des deux SMI. Le modèle multivarié généré avec ces deux SMI prédisait correctement le départ de la prochaine épidémie, 9 semaines à l'avance, avec cependant un décalage (retard) de 3 semaines environ sur la courbe épidémiologique prédite (figure 6.5).

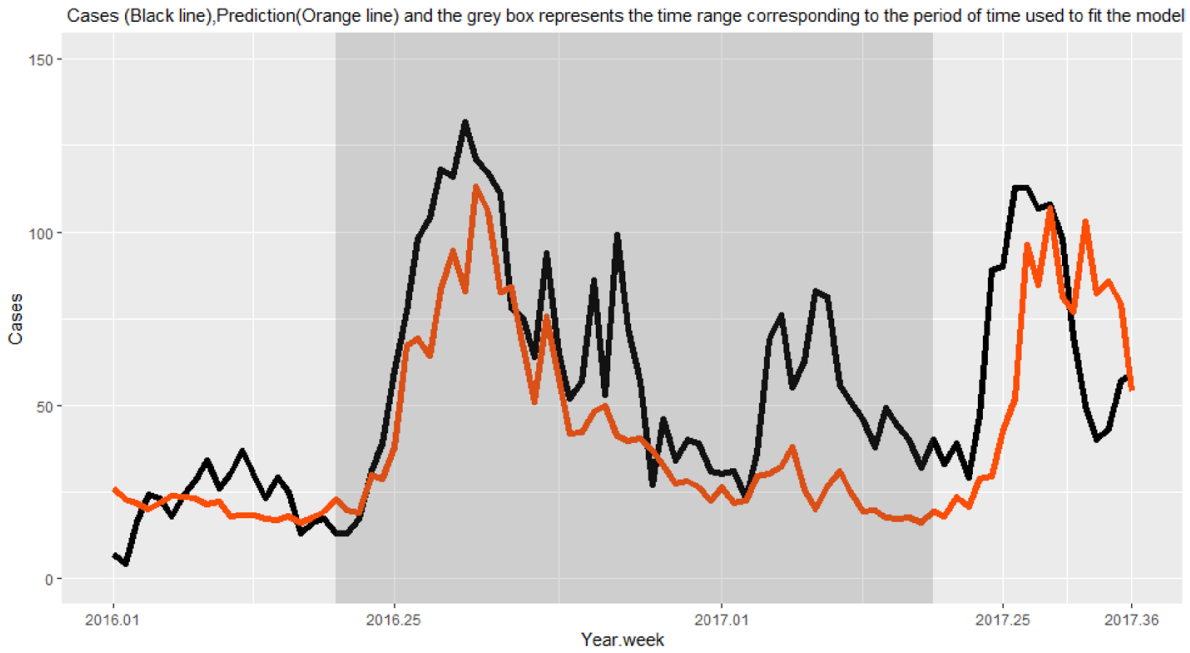
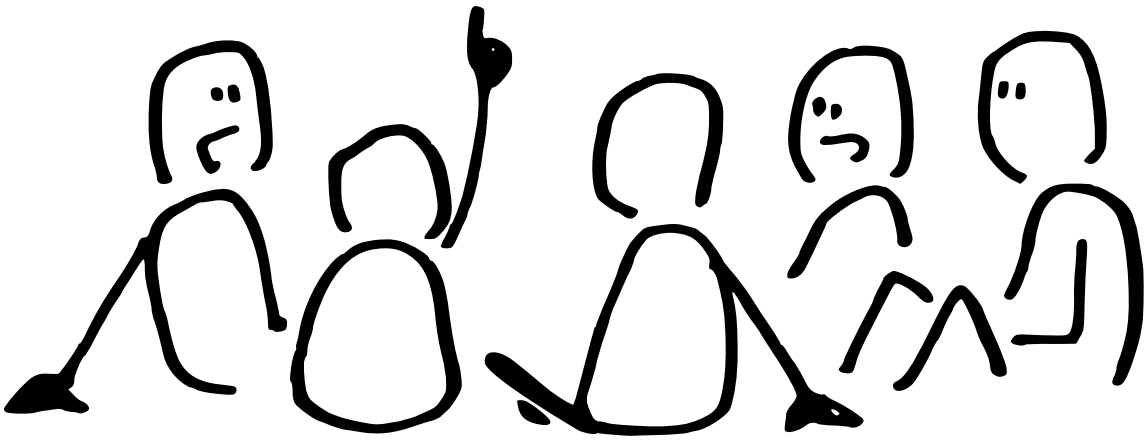


FIGURE 6.5: Nombre cumulé de cas reportés (courbe noire) et prédits par le modèle statistique basé sur des données météorologiques (courbe orange) dans les 27 villages de la zone de Diébougou

Cette étude a ainsi montré que la distribution spatio-temporelle des cas de paludisme était hétérogène à l'échelle du district sanitaire, justifiant ainsi de cibler les interventions de prévention, diagnostic, et traitement. Nous avons également montré qu'il est possible d'anticiper la prévalence du paludisme plusieurs semaines à l'avance dans la zone de Diébougou grâce aux données météorologiques issues des images satellitaires et aux modèles statistiques. Ces données et méthodes pourraient être utilisées pour mettre en place un système d'alerte précoce qui identifierait les zones (villages ou points chauds) et périodes prioritaires pour les campagnes de lutte contre le paludisme. Un tel système d'alerte précoce gagnerait à être alimenté en continu par les données épidémiologiques des centres de santé, afin d'améliorer les prédictions.

Partie 3

Discussion générale



Chapitre 7

Discussion générale

Après d'importants progrès, la morbidité et la mortalité liées au paludisme dans le monde ne recule plus – voire ré-augmente. Redynamiser le progrès nécessite de repenser certaines approches, entre autres : élargir la palette d'outils de lutte contre la transmission, déployer des interventions adaptées au contexte local, prioriser leur déploiement. Dans cette optique, la recherche scientifique se doit de proposer, concevoir, et tester de nouvelles stratégies et outils pour la surveillance et le contrôle du paludisme. Ces propositions doivent être construites sur la base d'une connaissance fine et localisée du risque de transmission du paludisme.

Dans ce cadre, cette thèse avait pour objectif principal d'approfondir les connaissances sur le risque de transmission résiduelle du paludisme – à savoir, la probabilité de contact homme-vecteur - dans deux zones d'étude d'environ 50 km x 50 km, en milieu rural ouest-africain, en implémentant, en partie, trois approches : mesurer et caractériser le risque, comprendre le risque, prédire le risque. Pour ce faire, nous avons étudié certains indicateurs entomologiques du risque de transmission en utilisant des méthodes de modélisation statistique descriptives et prédictives. Nous nous sommes intéressés en particulier à la bio-écologie des vecteurs du paludisme : déterminants et prédictibilité spatio-temporelle de leur probabilité de présence, de leur abondance, de leurs résistances physiologiques, et de leurs résistances comportementales.

Basé sur les résultats et limites des travaux de thèse, ce dernier chapitre est une discussion principalement constituée de propositions, axes de recherche ou opérationnels, pour la recherche et le contrôle du paludisme dans nos zones d'étude et au-delà. Cette

discussion est composée de deux grandes parties. La première partie est une discussion traitant d'entomologie médicale pour la prévention du paludisme. En nous appuyant sur les résultats de la thèse, nous définissons tout d'abord un ensemble de caractéristiques générales pour améliorer la LAV dans nos zones d'étude, décrivons ou mentionnons brièvement quelques outils de LAV existants répondant à ces caractéristiques, et présentons quelques limites de la thèse et perspectives de recherche qui permettraient d'améliorer la définition de ces caractéristiques. La deuxième partie de la discussion est davantage méthodologique et orientée autour de la science des données pour l'entomologie médicale et la géo-épidémiologie. En nous appuyant sur les travaux de la thèse et certaines réflexions ou observations personnelles, nous proposons dans un premier temps quelques pistes pour mieux exploiter le potentiel offert par la science et l'ingénierie des données dans la recherche en entomologie médicale et en géo-épidémiologie du paludisme. Dans un deuxième temps et pour terminer la discussion, nous définissons les contours d'un outil de surveillance et prévention du paludisme qui pourrait être développé au regard des résultats de la thèse pour nos zones d'étude et au-delà, pour le milieu rural ouest-africain.

7.1 Propositions pour des stratégies de réduction du risque de transmission résiduelle sur nos zones d'étude

7.1.1 Définition des caractéristiques de la LAV et stratégies potentielles concrètes

La portée opérationnelle de ces recherches est la conception et le déploiement optimisé de mesures de LAV durables et adaptées au faciès local. Sur la base des résultats de nos travaux de recherche, nous proposons ici quelques pistes pour une LAV plus efficace sur nos deux zones d'étude.

> ***Dans les deux zones, conserver la moustiquaire comme outil de base et principal de la lutte anti-vectorielle.*** Notre travail supporte, s'il en est encore besoin, la moustiquaire, distribuée universellement, comme outil premier et fondamental de la LAV : dans la zone de Diébougou, nous avons mesuré que les moustiquaires protégeaient leurs utilisateurs, grâce à leur effet barrière, de plus de 80% des piqûres

potentielles d'anophèles. Bien que la méthode utilisée (enquêtes déclaratives) puisse avoir tendance à sur-évaluer les taux d'utilisation réels, ces ordres de grandeur confirment l'importance de cet outil dans la lutte contre la transmission du paludisme. Il serait opportun de réaliser une étude similaire dans la zone de Korhogo avec les données du projet REACT.

Concrètement : Il apparaît essentiel de maintenir l'effort de distribution universelle des moustiquaires, tels que le font le Burkina Faso et la Côte d'Ivoire depuis 12 années maintenant. La question de l'imprégnation des moustiquaires par les insecticides, davantage discutable, fait l'objet d'un paragraphe à suivre. Par ailleurs, la fréquence optimale de distribution des moustiquaires, aujourd'hui de 3 à 4 ans dans les deux pays, reste à évaluer. L'OMS estime à trois ans la durée de conservation des moustiquaires (WHO, 2009), mais certaines études montrent qu'elle pourrait être bien inférieure (Bertozzi-Villa et al., 2021).

> ***Dans les deux zones, déployer urgemment des stratégies de LAV complémentaires à la MIILDA.*** Nous avons observé que les densités agressives des vecteurs sur nos zones d'étude étaient élevées (en particulier dans la zone de Korhogo), avec en moyenne, sur la durée du projet, 2,0 piqûres/homme/nuit à Diébougou et 28,2 piqûres/homme/nuit à Korhogo . Les données recueillies dans le cadre du projet REACT ont montré un taux d'inoculation entomologique (nombre de piqûres infectieuses/homme/an) de 898 dans la zone de Korhogo (Zogo, Soma, et al., 2019) et d'environ 100 dans la zone de Diébougou (Soma et al., 2020), plaçant ces zones au dessus de la moyenne africaine en ce qui concerne le risque de transmission du paludisme (Gething et al., 2011 ; G. F. Killeen, Githure, & Beier, 1999). Ces indicateurs particulièrement élevés montrent que la moustiquaire, bien qu'essentielle, doit très probablement être complémentée d'autres stratégies de LAV pour prévenir la transmission du paludisme.

Concrètement : La recherche, le développement et l'évaluation de nouveaux outils de LAV sont des secteurs actifs (Barreaux et al., 2017, ; Gerry F. Killeen, 2014 ; Seynabou Sougoufara, Ottih, & Tripet, 2020 ; WHO, 2017b ; Wilson et al., 2020). Les outils actuellement développés ou testés sont très variés, ciblant une vaste gamme de stades de vie des vecteurs et de leurs comportements ; et vont du conceptuellement très

simple (amélioration des habitations) au technologiquement complexe (modifications génétiques). Pour ne mentionner que quelques-unes de ces stratégies, citons : la lutte anti-larvaire (à base de bio-larvicides ou d'espèces prédatrices), la lutte génétique (techniques de l'insecte stérile et de l'insecte incompatible, forçage génétique), la protection personnelle (répulsifs cutanés, serpentins fumigènes, etc.), l'aménagement du territoire (drainage des eaux de surface, etc.), l'administration d'endectocides, les pulvérisations spatiales d'insecticide à l'extérieur des habitations, etc. Il s'agit d'autant de stratégies qui pourraient être envisagées pour réduire le risque de transmission sur nos zones d'étude.

> ***À Diébougou, déployer des stratégies complémentaires à la MIILDA ciblant spécifiquement les vecteurs endophages et endophiles.*** En caractérisant le risque de contact homme-vecteur dans la zone de Diébougou, nous avons mis en évidence que les populations restaient exposées à la piqûre d'anophèle, principalement le matin à l'intérieur des habitations. Aussi, dans cette zone au moins, il apparaît nécessaire de déployer également des outils complémentaires ciblant les vecteurs endophages pour réduire le risque de transmission.

Concrètement : Afin de cibler spécifiquement les vecteurs endophages ou endophiles, en sus des stratégies citées précédemment et des PID, des mesures relativement simples ayant pour objectif d'améliorer les habitations pour réduire l'entrée des vecteurs par des barrières physiques pourraient être déployées : la fermeture des avants-toits, les grillages sur les fenêtres et les portes, les tubes d'avant-toits, etc. (Animut, Balkew, & Lindtjørn, 2013 ; Atieli, Menya, Githeko, & Scott, 2009 ; Bradley et al., 2013, ; Sternberg et al., 2016 ; Yé et al., 2006). Des mesures telles que l'optimisation de l'utilisation des MIILDA peuvent également avoir un impact positif. Par exemple, dans le cadre du projet REACT, la stratégie d'IEC déployée en complément des MIILDA (voir section 3.1) a significativement réduit la morbidité de la maladie sur les deux zones d'étude par rapport aux MIILDA seules (Moiroux 2021, soutenance d'Habilitation à Diriger des Recherches).

> ***Dans les deux zones, déployer des stratégies complémentaires à la MIILDA ciblant spécifiquement les vecteurs exophages.*** L'étude des comportements trophiques des anophèles a mis en évidence un niveau d'exophagie

moyen important dans les deux zones (41 % à Diébougou, 56 % à Korhogo). Ces vecteurs ne sont pas ciblés par les outils actuellement utilisés (MIILDA principalement). Aussi, il semble nécessaire, pour réduire le risque de transmission, de déployer des outils ciblant ces vecteurs qui piquent à l'extérieur. Nous avons également observé que les proportions de vecteurs exophages étaient relativement homogènes spatio-temporellement au sein de chaque district sanitaire, et avons émis l'hypothèse que les conditions environnementales (autres que celles liées à l'utilisation des insecticides) n'impactaient que marginalement le comportement trophique des vecteurs. Il n'y aurait ainsi probablement pas de bénéfice particulier à adapter la lutte spécifique contre les vecteurs exophages au village ou à la saison. Notons, cependant, que cela n'implique pas qu'il n'y ait pas de bénéfice à prioriser les villages et/ou saisons dans lesquels ces outils seront déployés, si les ressources sont limitées.

Concrètement : En plus de toutes les stratégies précédemment citées (qui visent indistinctement les vecteurs exophages / exophiles et endophages / endophiles), une des stratégies existantes et déployables à l'extérieur (mais aussi à l'intérieur) des habitations consiste à piéger les vecteurs à l'aide d'appâts sucrés (*Attractive Toxic Sugar Baits*). Ces pièges relativement simples contiennent du sucre (alimentation des moustiques mâles et femelles) et des substances toxiques qui sont sélectives et qui ont des effets minimes sur les espèces non ciblées et sur l'environnement (Beier, Müller, Gu, Arheart, & Schlein, 2012 ; Müller et al., 2010 ; Stewart et al., 2013).

> ***Dans les deux zones, modifier l'usage des insecticides dans la LAV.***
Dans les deux zones d'étude, nous avons observé que les vecteurs étaient en mesure de résister efficacement aux effets létaux des insecticides : en évitant ou en contournant leurs effets létaux. Nous avons émis l'hypothèse que ces résistances, qu'elles soient physiologiques ou comportementales, continuaient à se développer dans les zones d'étude et qu'elles étaient principalement causées par les insecticides utilisés dans la LAV. Ces constats appellent à repenser l'utilisation des insecticides dans la LAV sur nos zones d'étude.

Concrètement : Différentes pistes sont envisageables, citons par exemple (Paaijmans & Huijben, 2020 ; WHO, 2017a) : la modification de la nature des insecticides, les mosaïques ou la rotation d'insecticides, l'utilisation des insecticides dans des outils

alternatifs aux MIILDA, ou encore le retrait pur et simple des insecticides de la boite à outils des mesures de LAV. Depuis la tenue du projet REACT, le Burkina Faso et la Côte d'Ivoire ont implémenté de nouvelles stratégies d'utilisation des insecticides. En effet, lors de la distribution universelle de 2019, le Burkina Faso a distribué des moustiquaires imprégnées d'un mélange pyréthrinoïdes - piperonyl butoxide (PBO), dont l'efficacité épidémiologique et entomologique par rapport aux moustiquaires imprégnées de pyréthrinoïdes uniquement a été démontrée en zone de forte résistance des vecteurs aux insecticides – et notamment au Burkina Faso (Gleave, Lissenden, Chaplin, Choi, & Ranson, 2021). La Côte d'Ivoire a, de son côté, stratifié la distribution des moustiquaires en 2021 selon quatre zones en fonction des informations disponibles sur les résistances physiologiques des vecteurs dans le pays. Un suivi des indicateurs entomologiques et épidémiologiques sur le long terme semble essentiel pour évaluer l'efficacité de ces nouvelles stratégies.

> ***Dans les deux zones, cibler et prioriser le déploiement des stratégies complémentaires à la MIILDA.*** Nous avons observé dans les deux zones que la distribution des populations de vecteurs était fortement hétérogène temporellement (entre les saisons) – comme l'on pouvait s'y attendre – mais aussi spatialement (entre les villages). Aussi, il y aurait très probablement un bénéfice à cibler les mesures de LAV complémentaires à l'échelle de la saison et du village, ainsi que prioriser leur déploiement si les ressources sont limitées. Par ailleurs, nous avons identifié nombre de déterminants de la probabilité de présence et de l'abondance des vecteurs - autant d'informations permettant d'identifier les leviers d'action les plus pertinents pour lutter localement contre les vecteurs.

Concrètement : Des outils de surveillance et de prédiction spatio-temporelle du risque de transmission, tels que des cartes ou des systèmes d'alerte précoces, pourraient être développés (ce point spécifique fait l'objet de la section 7.2.2 en fin de discussion). De tels outils de surveillance permettent d'envisager une LAV ciblée et priorisée dans l'espace et dans le temps. En effet, les cartes saisonnières de la distribution spatiale des densités agressives des vecteurs pourraient permettre de cibler ou prioriser le déploiement de mesures ponctuelles (amélioration des habitations, aménagement de l'environnement, etc.) ou nécessitant une certaine récurrence (lutte anti-larvaire, IEC, etc.). Les systèmes d'alerte précoce, de leur côté, pourraient permettre de déployer des mesures préventives

exceptionnelles (campagnes de larvicides, alertes à la population, etc.) dans les villages et moments pour lesquels un seuil de risque (à définir) est franchi. L'ensemble des procédures liées à la LAV (nature des interventions, sites et fréquences de de déploiement, cartes, seuils d'alerte de risque et interventions à déployer, etc.) gagneraient à être décrites dans des plans de gestion de LAV ad hoc.

7.1.2 Principales limites et perspectives de recherche

Bien que la thèse ait permis de définir certaines caractéristiques pour la LAV sur nos zones d'étude, de nombreuses questions – dont les réponses permettraient d'améliorer la définition de ces caractéristiques - restent en suspens. Nous exposons ci-dessous quelques-unes des limites de la thèse et perspectives de recherche associées.

> ***Tester, de manière expérimentale, les hypothèses soulevées dans les travaux de modélisation descriptive holistico-inductive.*** Les travaux de la thèse ont soulevé un certain nombre d'hypothèses ou de questions de recherche. Par exemple : À quoi sont dues les corrélations fortes entre les densités agressives des vecteurs et les variables météorologiques précédant la durée de vie des moustiques capturés (> 3 semaines avant) : dynamiques de population , effets paternels / maternels , ou préparation de conditions biotiques ou abiotiques favorables ? Quelle est l'explication biologique de l'association positive entre le niveau d'ouverture du paysage et les densités agressives ? Pour ces mêmes associations, les seuils révélés par les modèles sont-ils liés à un biais d'échantillonnage ou représentent-ils une réalité biologique ? Dans le même ordre d'idée, les associations entre la prévalence des résistances comportementales des vecteurs et la météorologie au cours du mois précédant les captures entomologiques sont-elles une réalité biologique (coût de mutations liées à un caractère éventuellement héréditaire des résistances comportementales) ou bien un biais d'échantillonnage ? Puisque les variables introduites dans les modèles de résistance n'expliquaient que peu la probabilité de résistance comportementale des vecteurs, quels sont les autres déterminants des résistances comportementales des vecteurs (génétique, hasard) ? Puisqu'à l'échelle du district sanitaire, la prévalence des résistances est globalement stable dans l'espace et dans le temps, à quelle(s) échelle(s) fluctuent, de manière significative, les dynamiques spatiales et temporelles des résistances des vecteurs ? Ces diverses questions et hypothèses pourraient être étudiées

et testées expérimentalement, dans des approches hypothético-déductives réductionnistes.

> ***Expliquer et prédire les composantes du risque de transmission résiduelle non étudiées dans la thèse (possession et utilisation des moustiquaires, niche d'activité des vecteurs).*** Dans cette thèse, nous n'avons étudié qu'une partie des composantes du risque de transmission résiduelle (en l'occurrence : présence, abondances, et résistances des vecteurs) identifiés dans le chapitre 1 (voir section 1.3.1). Pour compléter la définition des caractéristiques des outils de LAV complémentaires et cibler ou prioriser leur déploiement, il serait important d'expliquer et de prédire d'autres composantes du risque de transmission résiduelle, liées à l'homme ou au vecteur. En particulier, il serait intéressant d'étudier les déterminants de la possession, utilisation, et horaires d'utilisation des moustiquaires ; ainsi que ceux des horaires d'activité de recherche de repas du sang (« niche d'activité ») des vecteurs. Les conditions et horaires d'utilisation (ou absence d'utilisation) des moustiquaires peuvent être modulées par de nombreux facteurs socio-culturels, environnementaux, climatiques, entomologiques (Monroe, Moore, Koenker, Lynch, & Ricotta, 2019, ; Koenker et al., 2019). La niche d'activité du vecteur, quand à elle, reste peu étudiée, et pourrait être contrainte - entre autres – par les conditions micro-climatiques (Yin et al., 2019). Les données du projet REACT (de terrain et satellitaires) et les approches de modélisation statistique utilisées dans cette thèse pourraient être utilisées pour étudier ces composantes du risque.

> ***Tester l'efficacité des moustiquaires sans insecticides.*** Parmi les stratégies de gestion des résistances des vecteurs aux insecticides, la moustiquaire sans insecticides est une approche qu'il serait intéressant de tester (Paaijmans & Huijben, 2020). La question qui se pose ici est celle de l'origine de la protection communautaire conférée par les moustiquaires. En effet, l'insecticide dont la moustiquaire est imprégnée a pour objectif principal de réduire la longévité des vecteurs, et donc la population globale de vecteurs, conférant finalement la protection communautaire (voir section 1.1.5). Cependant, la barrière physique que constitue la moustiquaire empêche le vecteur de se gorger de sang, bloquant ainsi son cycle biologique, et *in fine* est également susceptible de réduire la densité globale des vecteurs. La mesure dans laquelle l'intensité de la transmission locale est réduite par la barrière chimique ou physique de la moustiquaire n'a pas été étudiée (Paaijmans & Huijben, 2020). Dans un contexte de fortes résistances aux insecticides – où l'effet de la barrière chimique pourrait être perdu,

voire pourrait avoir l'effet inverse à celui attendu (*behavioural exploitation*) - il pourrait être intéressant de comparer expérimentalement l'efficacité de moustiquaires imprégnées et de moustiquaires non imprégnées d'insecticides. De telles moustiquaires auraient par ailleurs certains avantages non-négligeables par rapport aux moustiquaires imprégnées : coût, durabilité, réduction des risques sur la santé humaine et sur l'environnement (Paaijmans & Huijben, 2020).

> ***Quantifier l'impact épidémiologique des résistances des vecteurs aux insecticides.*** Enfin, la question de l'impact des résistances des vecteurs sur la transmission effective du paludisme, et plus globalement, de l'association entre entomologie et épidémiologie, reste à développer. De nombreuses questions restent ainsi en suspens, par exemple : quel est l'impact des résistances des vecteurs (physiologiques, comportementales, ou leur interaction) sur l'épidémiologie du paludisme dans nos zones d'étude ? Les données entomologiques et épidémiologiques du projet REACT pourraient être croisées pour répondre à ces questions. Rappelons qu'à l'échelle de l'Afrique, le modèle mathématique utilisé par Sherrard-Smith et al. (2019) prédisait que les résistances des vecteurs étaient susceptibles d'avoir un impact significatif sur la morbidité de la maladie (voir section 1.2.2).

7.2 Propositions pour une meilleure exploitation de la science et de l'ingénierie des données pour la recherche et le contrôle du paludisme

7.2.1 Connaître et exploiter le potentiel de la science des données en entomologie médicale et géo-épidémiologie

Les travaux de cette thèse ont largement fait appel à des méthodes et outils de la science des données. Dans un monde toujours plus numérisé, dans lequel les données sont toujours plus abondantes, fines, accessibles, abordables, la recherche et la gestion du paludisme ne doit pas manquer les opportunités offertes par la science des données. Celles-ci sont nombreuses, et concernent des aspects à la fois de recherche et opérationnels. Cette thèse a montré comment l'utilisation judicieuse de données hétérogènes (spatio-temporelles, multi-source, multi-échelle) permet à la fois

d'approfondir les connaissances et peut avoir une portée très opérationnelle, via le développement d'outils concrets de surveillance du risque de transmission. Dans cette section, nous discutons d'aspects méthodologiques – là aussi sous forme de propositions – liés à l'exploitation de la science et de l'ingénierie des données pour la recherche ou le contrôle du paludisme.

> ***Connaître et exploiter pleinement le potentiel offert par la modélisation statistique.*** Comme nous l'avons expliqué dans le chapitre 2 et montré à travers les travaux de la thèse, aujourd'hui la modélisation statistique n'est plus seulement un outil de vérification d'hypothèse scientifique : c'est également un véritable outil à part entière de création de connaissances, d'hypothèses scientifiques. En effet, un modèle statistique peut capturer des associations inattendues pouvant soulever de nouvelles hypothèses qui peuvent ensuite être testées expérimentalement. Il semble aujourd'hui que la recherche dans des domaines tels que la géo-épidémiologie n'exploite pas encore pleinement son potentiel. Par exemple, une revue de littérature récente a montré que les modèles statistiques utilisés en géo-épidémiologie pour prédire la distribution spatio-temporelle de certaines grandes maladies vectorielles (paludisme, dengue, virus du Nil occidental) sont principalement des modèles paramétriques type GLMM (Parselia et al., 2019). À l'heure des modèles d'apprentissage automatique en capacité de capturer des associations complexes, des outils permettant d'interpréter ces modèles, et des données volumineuses, il semble nécessaire que les domaines de l'entomologie médicale et de la géo-épidémiologie prennent davantage conscience du potentiel offert par la modélisation statistique, à la fois pour la recherche afin de construire des hypothèses scientifiques à partir des données, et pour la gestion afin, par exemple, de prédire des indicateurs de transmission.

> ***Connaître et exploiter les données libres et ouvertes disponibles.*** En parallèle du développement des modèles statistiques, le volume et la diversité des données disponibles sont toujours plus importants, et leur granularité est toujours plus fine. Ces données fines et hétérogènes ont un fort potentiel pour la recherche et le contrôle du paludisme. A titre d'exemple, dans cette thèse, nous avons montré de quelle manière il était possible d'exploiter la diversité et la granularité spatiale et temporelle des images satellitaires pour mieux comprendre la bio-écologie des vecteurs (article n°1) et prédire l'abondance des vecteurs à fine échelle spatiale. De même,

nous avons montré comment des données hétérogènes, multi-échelles, collectées avec des instruments différents, et issues de sources diverses, pouvaient être utilisées pour étudier des systèmes biologiques complexes (article n°2). De nombreuses données, prêtes à l'emploi, sont produites à des échelles spatio-temporelles fines, tout en couvrant l'intégralité du continent africain, voire du globe. Ces données constituent des ressources de grande valeur pour constituer des variables indépendantes dans des travaux de modélisation explicatives, descriptives ou prédictives d'indicateurs de la transmission du paludisme ou de son risque. Parmi les données disponibles à l'échelle mondiale, à fine granularité, et pouvant être d'intérêt pour l'étude du paludisme, mentionnons, pêle-mêle (en sus des données de précipitations, température et altitude utilisées dans la thèse) : le produit WorldCover sur l'occupation du sol à 10 mètres de résolution spatiale couvrant l'ensemble du globe (Zanaga et al., 2021), les données SMAP sur l'humidité du sol (ONeill et al., 2021), les données Global Surface Water sur la présence et la persistance d'eaux de surface (Pekel, Cottam, Gorelick, & Belward, 2016), la collection VNP46A1 sur les lumières nocturnes (indicateur du niveau d'électrification) (Román et al., 2018), les données WorldPop sur la démographie (incluant la structure par âge) (Bondarenko, Kerr, Sorichetta, & Tatem, 2020), les données cartographiques du Malaria Atlas Project sur des indicateurs liés au paludisme (distribution des vecteurs, possession et utilisation des moustiquaires, etc.) (S. I. Hay & Snow, 2006), etc.

> ***Connaître et valoriser des formes d'inférence logique alternatives au raisonnement hypothético-déductif dans la recherche en biologie, entomologie médicale, épidémiologie.*** Dans le chapitre 2, nous avons présenté les différentes formes d'inférence logique ainsi que, en lien direct, les différentes approches de l'étude des systèmes biologiques complexes. Nous avons ainsi expliqué qu'il existe des formes d'inférence logique alternatives au raisonnement hypothético-déductif très largement utilisé en sciences biologiques, et avons montré avec les travaux de thèse en quoi elles pouvaient consister. L'approche holistico-inductive utilisée dans la thèse est de prime abord inconfortable et déroutante, car il n'y a pas d'hypothèse spécifique à tester. Mais elle laisse peut-être davantage de place à la découverte de nouvelles hypothèses, de nouvelles questions de recherche. Elle s'apparente à ce que François Jacob, chercheur en biologie du milieu du siècle dernier, appelait la « science de nuit » (Yanai & Lercher, 2019b) (figure 7.1) : cette recherche moins formelle, moins structurée que celle qui est généralement valorisée à travers les publications scientifiques ; mais qui

exploite pleinement la créativité du chercheur, sans l'enfermer dans des approches et une structure de pensée qui pourraient, parfois, être trop "restrictives". Les données et les modèles statistiques non-paramétriques sont des outils qui, sans en être les seuls, permettent d'implémenter ces approches. Il semble important de réhabiliter, accepter et promouvoir l'approche holistico-inductive dans la recherche dans des disciplines telles que la biologie, l'entomologie médicale, l'épidémiologie. N'oublions pas que certaines découvertes majeures ont été faites par raisonnement inductif. Ainsi en est-il de la théorie de l'évolution de Darwin : l'objectif et le travail initial de ce dernier était "simplement" de recenser et établir une forme de classification des espèces vivantes sur la planète. C'est en organisant ses observations, autrement dit données, que l'hypothèse de la sélection naturelle a émergé chez Darwin, dans une approche purement holistico-inductive (Kell & Oliver, 2004) (bien que les modèles statistiques n'aient, selon toute vraisemblance, pas été utilisés dans ce cas).

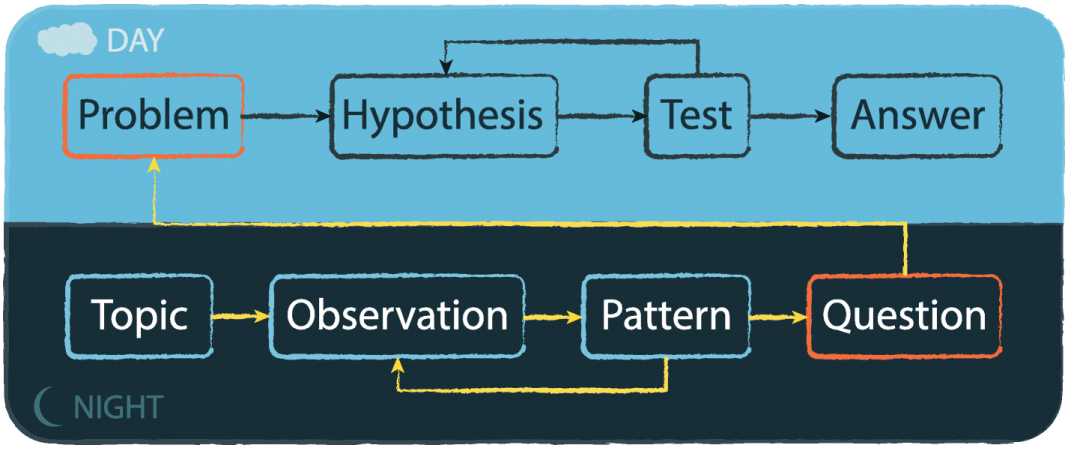


FIGURE 7.1: Concepts de 'Science de nuit' et 'Science de jour' (développés par le chercheur François Jacob). Le raisonnement inductif est associé à la 'science de nuit', celle qui explore le domaine non structuré des hypothèses possibles, des idées qui n'ont pas encore été pleinement concrétisées ; et le raisonnement déductif à la 'science de jour', qui teste formellement des hypothèses pré-établies, possiblement grâce à des approches de science de nuit (Yanai & Lercher, 2019b)

> ***Décrire, harmoniser et ouvrir les données de recherche en entomologie / épidémiologie du paludisme.*** Les modèles statistiques et les données environnementales resteront peu utiles pour la recherche ou la gestion du paludisme sans

des données entomologiques ou épidémiologiques trouvables, accessibles, interopérables, et réutilisables (FAIR [1] data). Par exemple, dans cette thèse, les analyses ont pu être reproduites à moindre coût sur les deux zones d'étude grâce au caractère standardisé des données environnementales mais aussi entomologiques. Les laboratoires de recherche côtoyés dans la thèse (MIVEGEC (France), IRSS (Burkina Faso) et IPR (Côte d'Ivoire)) fournissent un effort considérable pour recueillir de nombreuses données entomologiques ou épidémiologiques sur le terrain, à travers de très nombreux projets de recherche, mais ces données ne semblent ensuite être que trop rarement ouvertes et valorisées autrement que par les publications scientifiques écrites par ces laboratoires les recueillant. Ces données représentent pourtant une mine d'or, pour la recherche mais aussi pour la gestion directe du paludisme (par exemple, via le développement et le maintien d'outils de surveillance et prévention du paludisme, voir section 7.2.2). Les ressources et efforts humains, matériels ou financiers à fournir pour décrire et publier ces données de recherche sont relativement minimes par rapport à ceux déployés pour les collecter sur le terrain ; et le gain potentiel à court, moyen et long terme, à la fois pour la recherche et la gestion, est important. Par ailleurs, les outils informatiques permettant de rendre les données FAIR sont aujourd'hui performants, efficaces et relativement accessibles à tout-un-chacun. Citons par exemple la libraire R `geoflow` (Blondel, Barde, Heintz, & Bennici, 2020), qui offre un cadre simple en R pour exécuter et orchestrer des tâches de gestion et de publication de données et métadonnées géospatiales de manière automatisée. Le travail de sensibilisation à l'importance de la description et l'ouverture des données de la recherche en entomologie et épidémiologie, déjà entamé, est donc à poursuivre et intensifier. En parallèle de ce travail à effectuer au niveau de chaque laboratoire de recherche, il pourrait être opportun de travailler à l'élaboration de standards pour les données entomologiques (en particulier, standards pour les référentiels métiers, type méthode de capture), co-construits au sein de groupes de travail internationaux, s'appuyant sur des standards ouverts tels que ceux développés par l'Open Geospatial Consortium (OGC [2]) ; dans l'objectif final de faciliter l'utilisation, réutilisation, et interopérabilité des données entomologiques. La communauté de l'entomologie médicale pourrait s'inspirer d'initiatives de ce genre initiées dans d'autres disciplines – pour ne citer qu'un exemple, les pêcheries mondiales (FAO, 1995).

1. Findable, Accessible, Interoperable, Reusable
2. L'OGC est un consortium international pour développer et promouvoir des standards ouverts, les spécifications OpenGIS, afin de garantir l'interopérabilité des contenus, des services et des échanges dans les domaines de la géomatique et de l'information géographique (définition Wikipédia)

> ***Scripter les analyses de données et ouvrir les codes.*** Dans cette thèse, un effort important a été fourni pour rendre les méthodes de recueil, production, analyses de données autant que possible transparentes, génériques et reproductibles ; par le développement de codes en langage R, leur description et leur ouverture. Si ce travail requiert un effort initial important, le gain est rapidement tangible. Ainsi, dans le cadre strict de la thèse, ce travail a permis de réaliser à moindre effort des analyses complexes sur deux zones d'étude distinctes. Au delà du strict cadre de la thèse, il est tout à fait envisageable de réutiliser à moindre coût les méthodes (codes) développées dans la thèse dans d'autres cadres (par exemple, urbains), à d'autres échelles spatiales (par exemple, quartiers dans les villes), ou encore sur d'autres zones géographiques dans le monde. De tels exemples de réutilisation du travail existent d'ailleurs déjà : les codes R de cartographie de l'occupation du sol développés dans le cadre de cette thèse ont été réutilisés pour cartographier l'occupation du sol dans plusieurs quartiers de la ville de Bouaké (Côte d'Ivoire) avec des images collectées par drone dans le cadre d'une autre thèse (*travail en cours de publication*) ; et la librairie R `opendapr` a été réutilisée dans une autre étude nécessitant des données de précipitations (Sondo et al., 2020). Bien que la description des codes développés dans la thèse ne soit pas encore totalement aboutie, ces exemples montrent l'intérêt immédiat de scripter les analyses de données, de les rendre génériques, de décrire ces codes, et de les rendre accessibles à tous.

> ***Promouvoir l'utilisation des logiciels libres et ouverts.*** Les travaux de la thèse nécessitant l'outil informatique - de la production des données à la rédaction de ce manuscrit en passant par la modélisation statistique - ont été réalisés exclusivement avec des logiciels et librairies informatiques gratuits et à code source libre et ouvert (*Free and Open Source Software* (FOSS)). Cela montre la nature et l'étendue des possibilités offertes par les logiciels libres, en entomologie, géo-épidémiologie, et au-delà. Un temps réservés aux informaticiens ou adeptes de la programmation informatique, les logiciels libres semblent aujourd'hui de plus en plus accessibles à tout-un-chacun (interfaces graphiques améliorées, « bugs » informatiques moins fréquents, etc.) et constituent ainsi de réelles alternatives, performantes, aux logiciels propriétaires. Il est essentiel de les promouvoir dans la recherche publique. Plus globalement, il semble important que la recherche publique privilégie et soutienne l'utilisation des outils informatiques (logiciels, standards liés aux données géospatiales, etc.) dont le code source est ouvert, libre

d'accès, et développés collaborativement par de multiples institutions ; par opposition aux services proposés par les grands groupes privés [3] - souvent gratuits et performants mais potentiellement délétères pour la recherche sur le long terme (surveillance de la recherche, dépendance à ces outils, absence de garantie quand à la continuité de la gratuité de ces services ou leur interopérabilité avec d'autres systèmes informatiques, etc.). D'une manière générale, les alternatives FOSS aux logiciels propriétaires existent presque toujours, et nous estimons qu'elles devraient être énergétiquement défendues dans des milieux tels que celui de la recherche publique – même s'il existe parfois un coût à leur prise en main.

La thèse et ces paragraphes ont présenté quelques usages et intérêts de la science et ingénierie des données pour la recherche et le contrôle du paludisme. Afin d'appréhender l'utilité mais aussi les limites de ces outils, il est essentiel d'en connaître l'existence et de les maîtriser. A cet égard, il est important de développer des collaborations entre chercheurs et ingénieurs issus des différentes disciplines scientifiques thématiques (entomologie médicale, épidémiologie) et méthodologiques (science des (géo)données, ingénierie des données).

Afin d'illustrer nos propos, nous terminons cette discussion en proposant le développement d'un outil de surveillance et prévention du paludisme dans les zones du projet REACT - et éventuellement au-delà - basé sur les données (entomologiques, épidémiologiques, satellitaires) et les modèles statistiques prédictifs.

7.2.2 Vers la création d'outils de surveillance et prévention du paludisme, dans les zones du projet REACT et au-delà

Les stratégies de gestion de la LAV proposées dans la première partie de cette discussion reposent largement sur le ciblage et la priorisation des interventions, éléments par ailleurs largement préconisés par la communauté des acteurs de la lutte contre le paludisme (WHO, 2021). Planifier, cibler et prioriser les interventions requièrent l'accès à une information spatialisée et temporalisée sur ce risque. Ici, nous avons préconisé que l'échelle spatiale du **village** semblait pertinente pour cibler les

3. Google Scholar pour la recherche bibliographique, Google Earth Engine pour la télédétection, Google Drive pour le travail collaboratif, etc.

interventions de LAV, requérant ainsi de générer et mettre à disposition de la communauté (chercheurs, décideurs, grand public) une information à une telle échelle spatiale.

À notre connaissance, l'information spatiale la plus fine existante sur les indicateurs du paludisme à ce jour sur nos zones d'étude est celle produite par le Malaria Atlas Project [4] (MAP). Le MAP est un projet de recherche international qui produit des données spatio-temporelles sur le paludisme de 5 km de résolution spatiale, de couverture mondiale ou à minima africaine, et mises à jour en général annuellement. Différents indicateurs sont générés, à la fois entomologiques (probabilité de présence de chaque espèce majeure d'anophèles (Wiebe et al., 2017)), épidémiologiques (prévalence, incidence, mortalité liée à *Plasmodium falciparum* (Weiss et al., 2019)), ou encore socio-économiques (accès et utilisation des MIILDA (Bertozzi-Villa et al., 2021)). Ces produits sont générés par modélisation statistique prédictive. Ils sont mis à disposition de tous (données ouvertes) et visualisables via une plateforme cartographique interactive [5]. Si ces produits et outils sont sans aucun doute utiles pour la recherche et la gestion du paludisme à large échelle (suivi des tendances spatiales et temporelles de la maladie, planification et priorisation de certaines interventions à large échelle), il semblent moins adaptés pour cibler concrètement les interventions et aider à la décision aux échelles spatio-temporelles compatibles avec l'action immédiate de terrain (Nosten & Phyo, 2019) : hautes incertitudes de prédiction à fine échelle, paucité des données d'entraînement des modèles (relativement à l'étendue spatio-temporelle des prédictions), etc. Ainsi, une information plus fine (à la fois spatialement et temporellement) et des outils de surveillance plus proches du besoin des gestionnaires sur le terrain (PNLPs, centres de santé, etc.) sont nécessaires. Dans notre thèse, nous avons montré que nous étions en capacité de prédire et anticiper correctement la présence et l'abondance des vecteurs à l'échelle du village dans les zones du projet REACT, grâce à des produits satellitaires disponibles en tout point de l'espace et du temps. Ce résultat ouvre des perspectives intéressantes quand à la création de tels outils de surveillance, d'intérêt direct pour la prévention du paludisme dans nos zones d'étude.

À partir de nos résultats, nous pourrions dans un premier temps créer un ensemble de cartes saisonnières de la présence et abondance des vecteurs dans chacune des zones,

4. https://malariaatlas.org/
5. https://malariaatlas.org/explorer/#/

à l'échelle du village et pour chaque espèce (à l'image des travaux de Moiroux et al. (2013) ; Moiroux et al. (2014)). Ces cartes pourraient servir à cibler, dans l'espace et dans le temps, certaines interventions de LAV « récurrentes ».

Par ailleurs, nous avons observé que les variables temporelles (climatiques) montrant la plus forte corrélation avec les abondances observées se situaient plusieurs semaines en amont de la capture, impliquant qu'il est à priori possible non seulement de prédire les densités agressives à l'échelle du village, mais également de les anticiper plusieurs semaines à l'avance. Cette observation ouvre la voie au développement d'outils de surveillance et prévention des épidémies (WHO, 2018), tels qu'un système d'alerte précoce du risque de piqûre par un vecteur du paludisme. Un tel système d'alerte précoce pourrait prédire, en routine à une fréquence à définir (par exemple, hebdomadaire), certains indicateurs entomologiques (ou épidémiologiques, voir ci-après) de la transmission du paludisme, dans chaque village des zones d'étude, à courte ou moyenne échéance (échéance t+1 semaine, t+2 semaines, t+3 semaines, etc.) (figure 7.2).

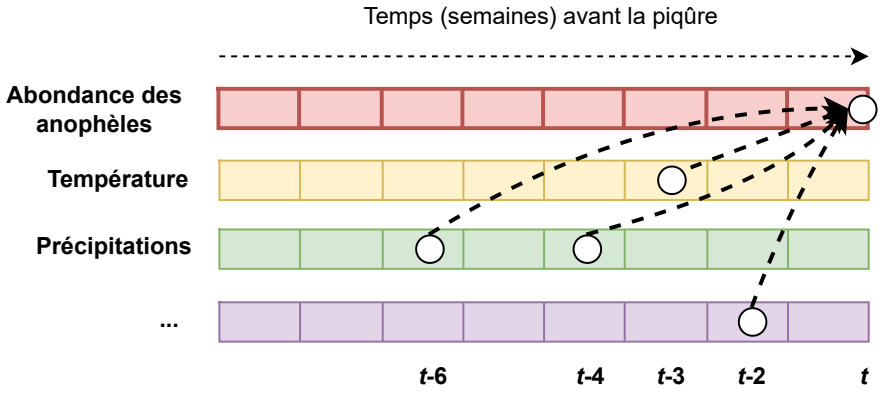


FIGURE 7.2: Concept de prédiction précoce d'un indicateur de risque de transmission (ici : abondance des anophèles)

Un système d'alerte précoce est un système d'information permettant de détecter un risque donné suffisamment précocement pour permettre d'agir dans l'optique de réduire les dommages potentiels causés par ce risque (WHO, 2001, 2018). Si le concept et l'histoire des systèmes d'alerte précoces du paludisme sont loin d'être récents (WHO, 2001), la diversité et la granularité spatio-temporelle des données aujourd'hui disponibles (notamment satellitaires) et leur rapidité de mise à disposition suivant l'aquisition,

combinés avec les performances prédictives des modèles non-paramétriques, permet d'envisager l'existence de systèmes dynamiques prédisant et anticipant les indicateurs à des résolutions spatio-temporelles fines, comme nous l'avons montré dans la thèse, et adaptées à la prise de décision rapide. Certaines initiatives récentes montrent l'intérêt croissant des pouvoirs publics pour ce genre d'outils. Notons, par exemple, le Prix pour l'Alerte Précoce des Epidémies (*Prize for Early Warning for Epidemic*) lancé en 2018 par le Conseil Européen de l'Innovation (Commission, Research, & Innovation, 2022). Ce concours, dont le prix était de 5 millions d'euros, consistait à développer un système d'alerte précoce des maladies transmises par les moustiques, basé sur des produits satellitaires d'observation de la Terre. Le vainqueur du prix est le projet *Early Warning System for Mosquito Borne Diseases*[6], un système d'alerte précoce qui se base sur un outil informatique que ses développeurs ont nommé MAMOTH (Tsantalidou et al., 2021). MAMOTH est un système d'information générique de prédiction de l'abondance des moustiques vecteurs de maladies infectieuses dans l'espace et dans le temps, qui a montré de bonnes capacités prédictives pour plusieurs espèces de moustiques en Europe. Dans le même ordre d'idée, notons le projet *MosquitoAlertBCN*[7], qui s'appuie sur un réseau d'observation des moustiques tigres (*Aedes albopictus*) et des modèles statistiques prédictifs pour calculer un indice de risque de contact homme-vecteur à t+7 jours à l'échelle spatiale du quartier dans la ville de Barcelone (Espagne).

Ce genre de systèmes d'alerte précoce, sur les territoires concernés par notre thèse, pourrait prédire/anticiper non seulement des indicateurs entomologiques mais aussi des indicateurs épidémiologiques. En effet, dans nos travaux nous avons montré que nous étions en mesure de prédire effectivement la présence et l'abondance des vecteurs à l'échelle du village (article n°1), mais également le nombre de cas de paludisme à l'échelle du district sanitaire (article n°4). Nous pourrions donc envisager un système d'alerte précoce multi-indicateurs, cet ensemble d'informations spatio-temporalisées permettant alors de servir différents aspects de la lutte (prévention, diagnostic, traitement) (J. M. Cohen et al., 2017).

Ces outils d'aide à la décision (cartes, systèmes d'alerte précoce multi-indicateurs) pourraient dans un premier temps être développés et testés dans les zones d'étude du

6. http://beyond-eocenter.eu/index.php/web-services/eywa
7. https://mosquito-alert.github.io/MosquitoAlertBCN/

projet REACT, en se basant sur les données entomologiques et épidémiologiques du projet et les données environnementales issues d'images satellitaires. S'ils s'avèrent utiles et efficaces, ils pourraient être étendus à des zones géographiques extérieures à celles des districts sanitaires de Diébougou et Korhogo. Afin de tester la capacité des modèles entraînés dans les zones de REACT à prédire en dehors (spatialement et temporellement) de ces zones, nous pourrions évaluer, avec les données de REACT, la performance prédictive des modèles entraînés dans la zone de Korhogo, dans la zone de Diébougou ; et inversement. Si les performances prédictives restent correctes, nous pourrions alors envisager d'étendre les prédictions à l'extérieur des zones REACT ; par exemple dans l'ensemble des zones rurales de la sous-région bioclimatique soudanienne. Notons que dans tous les cas, pour tout indicateur prédit (en ou hors zone d'entraînement du modèle), il sera important de fournir des informations (cartes) sur l'erreur de prédiction spatio-temporelle, afin de ne pas sur-interpréter l'information délivrée par les produits (Meyer & Pebesma, 2021 ; Wardrop, Geary, Osborne, & Atkinson, 2014). Dans le domaine des sciences des géo-données, certaines méthodes innovantes commencent à être développés à cette fin (Meyer & Pebesma, 2021).

Enfin, la durabilité d'un tel système et son extensibilité à d'autres régions reposerait sur la capacité à recalibrer et améliorer continuellement les modèles prédictifs sur lequel il se base ; elle-même reposant sur le recueil régulier de données entomologiques ou épidémiologiques provenant du terrain et sur la capacité à y accéder rapidement et simplement. Les données collectées en routine dans les centres de santé et dans le cadre des différents projets de recherche menés sur le terrain pourraient constituer une première source précieuse à cet égard. Il est important pour cela d'améliorer la chaîne de recueil et stockage de ces données (description, harmonisation, anonymisation, ouverture), comme présenté dans la section 7.2.1. Ces données pourraient être complétées par des observations régulières, long-terme, issues de réels réseaux d'observations entomologiques, qui restent à développer en Afrique de l'Ouest.

Conclusion

Deux décennies après le lancement de programmes révolutionnaires dans la lutte contre le paludisme, celle-ci est à nouveau à un tournant de son histoire. Il y a vingt ans, l'approche préventive défendue par la communauté des acteurs de la lutte contre le paludisme était celle de la couverture universelle en un outil qui avait été éprouvé expérimentalement : la moustiquaire imprégnée d'insecticide. Si cette approche a largement fait ses preuves sur le terrain également, force est de constater qu'aujourd'hui, les limites en sont tangibles. Le paradigme d'une gestion « universelle » est donc peu à peu abandonné, et remplacé par celui d'une gestion localisée. Ces approches sont notamment soutenues par l'OMS à travers l'initiative *High burden to high impact* (« D'une charge élevée à un fort impact »).

À nouvelles approches de gestion, nouvelles perspectives de recherche. Au niveau de la prévention, l'enjeu pour les chercheurs est de mieux comprendre ce qui amène, en un lieu et un moment donnés, le moustique et l'homme à entrer mutuellement en contact. La recherche doit donc se concentrer sur les composantes et comportements respectifs de ces deux agents favorisant la probabilité de cette rencontre. C'est sur la base des connaissances dégagées sur ces éléments que pourront être prises des mesures adéquates, adaptées aux contextes et problématiques locaux.

Dans cette thèse, nous avons tenté de comprendre une partie de ces composantes du risque, du côté du vecteur, en étudiant sa bio-écologie. Nous avons pour cela travaillé à fine échelle spatiale, et nous sommes largement appuyés sur des données de tous types et des modèles statistiques variés. La science et l'ingénierie des données, disciplines grandissantes et incontournables aujourd'hui, offrent des opportunités à saisir pleinement pour la communauté des acteurs oeuvrant contre le fardeau du paludisme, scientifiques comme gestionnaires. Les données et modèles renferment un potentiel

d'innovation important ; notamment, celui d'approcher la maladie plus finement – spatialement, temporellement, dimensionnellement. De part leurs potentiels à mieux décrire, expliquer et prédire les systèmes complexes, la science et l'ingénierie des données sont en capacité de servir la lutte contre la maladie sur tous les fronts : celui de la recherche mais aussi, très directement, de la gestion au jour-le-jour par tous les acteurs.

Redynamiser le progrès passera par un savant mélange de capacité à innover mais également à consolider l'existant, de projets de recherche mais également d'ingénierie, de maîtrise des nouvelles technologies sans oublier l'efficacité des plus anciennes ou des « low tech ». Les innovations sont essentielles mais quelles qu'elles soient, elles n'auront de sens que si elles sont accompagnées à minima du maintien, et idéalement du renforcement, des efforts qui ont permis la réduction spectaculaire du fléau de la maladie au début du XXI^{ème} siècle. Pour n'en citer qu'un, et peut-être le principal : la volonté politique de diminuer le fardeau du paludisme.

Bibliographie

Cette bibliographie inclut uniquement les références citées dans le corps du manuscrit de thèse (les références mentionnées dans les publications sont intégrées à la fin de chaque article)

Allaire, J., Horner, J., Xie, Y., Marti, V., & Porte, N. (2019). *Markdown : Render markdown with the c library 'sundown'*. Retrieved from https://CRAN.R-project.org/package=markdown

Alout, H., Krajacich, B. J., Meyers, J. I., Grubaugh, N. D., Brackney, D. E., Kobylinski, K. C., ... Foy, B. D. (2014). Evaluation of ivermectin mass drug administration for malaria transmission control across different West African environments. *Malaria Journal*, *13*(1), 417. http ://doi.org/10.1186/1475-2875-13-417

Amboise, G., & Audet, J. (1996). *Le projet de recherche en administration : Un guide général à sa préparation (chapitre 4)*. Faculté des sciences de l'administration, Université Laval. Retrieved from https://books.google.fr/books?id=U2BcnQEACAAJ

Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). *A land use and land cover classification system for use with remote sensor data* (Report No. 964). Retrieved from http://pubs.er.usgs.gov/publication/pp964

Animut, A., Balkew, M., & Lindtjørn, B. (2013). Impact of housing condition on indoor-biting and indoor-resting Anopheles arabiensis density in a highland area, central Ethiopia. *Malaria Journal*, *12*(1), 393. http ://doi.org/10.1186/1475-

2875-12-393

Arnaud, A.-J. (1986). Karl R. Popper, Conjectures et réfutations. La croissance du savoir scientifique, trad. M.I. Et M.B. De Launay, 1985. *Droit Et Société*, *4*(1), 464–465. Retrieved from https://www.persee.fr/doc/dreso_0769-3362_1986_num_4_1_1528_t1_0464_0000_2

Atieli, H., Menya, D., Githeko, A., & Scott, T. (2009). House design modifications reduce indoor resting malaria vector densities in rice irrigation scheme area in western Kenya. *Malaria Journal*, *8*(1), 108. http ://doi.org/10.1186/1475-2875-8-108

Aubréville, A. (1957). Accord à Yangambi sur la nomenclature des types africains de végétation. *Bois Et Forêts Des Tropiques*, (51), 23–27.

Baatz, M., & Schape, A. (2000). Multiresolution segmentation : An optimization approach for high quality multi-scale image segmentation. *In : Strobl, J., Blaschke, T. And Griesbner, G., Eds., Angewandte Geographische Informations-Verarbeitung, XII, Wichmann Verlag, Karlsruhe, Germany,* 12–23.

Barreaux, P., Barreaux, A. M. G., Sternberg, E. D., Suh, E., Waite, J. L., Whitehead, S. A., & Thomas, M. B. (2017). Priorities for Broadening the Malaria Vector Control Tool Kit. *Trends in Parasitology*, *33*(10), 763–774. http ://doi.org/10.1016/j.pt.2017.06.003

Bar-Yam, Y. (2002). General Features of Complex Systems. *Encyclopedia Of Life Support Systems*, 10.

Baudon, D., Molez, J.-F., & Guiguemde, T. R. (1984). Aspects classiques et modernes des cycles de développement des plasmodiums humains. *Etudes Médicales*, 61–78. Retrieved from https://www.documentation.ird.fr/hor/fdi:15172

Beier, J. C., Müller, G. C., Gu, W., Arheart, K. L., & Schlein, Y. (2012). Attractive toxic sugar bait (ATSB) methods decimate populations of Anopheles malaria

vectors in arid environments regardless of the local availability of favoured sugar-source blossoms. *Malaria Journal*, *11*(1), 31. http ://doi.org/10.1186/1475-2875-11-31

Bertozzi-Villa, A., Bever, C. A., Koenker, H., Weiss, D. J., Vargas-Ruiz, C., Nandi, A. K., ... Bhatt, S. (2021). Maps and metrics of insecticide-treated net access, use, and nets-per-capita in Africa from 2000-2020. *Nature Communications*, *12*(1), 3589. http ://doi.org/10.1038/s41467-021-23707-7

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., ... Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, *526*(7572), 207–211. http ://doi.org/10.1038/nature15535

Bivand, R. (2018). *rgrass7 : Interface Between GRASS 7 Geographical Information System and R*. Retrieved from https://CRAN.R-project.org/package=rgrass7

Bivand, R., Keitt, T., & Rowlingson, B. (2019). *Rgdal : Bindings for the 'Geospatial' Data Abstraction Library*. Retrieved from https://CRAN.R-project.org/package=rgdal

Blondel, E., Barde, J., Heintz, W., & Bennici, A. (2020). geoflow : R engine to orchestrate and run geospatial (meta)data workflows (Version 0.0.20201116). Zenodo. http ://doi.org/10.5281/zenodo.4275926

Bondarenko, M., Kerr, D., Sorichetta, A., & Tatem, A. (2020). Census/projection-disaggregated gridded population datasets for 189 countries in 2020 using Built-Settlement Growth Model (BSGM) outputs. University of Southampton. http ://doi.org/10.5258/SOTON/WP00684

Bradley, J., Rehman, A. M., Schwabe, C., Vargas, D., Monti, F., Ela, C., ... Kleinschmidt, I. (2013). Reduced Prevalence of Malaria Infection in Children Living in Houses with Window Screening or Closed Eaves on Bioko Island, Equatorial Guinea. *PLoS ONE*, *8*(11), e80626. http ://doi.org/10.1371/journal.pone.0080626

Breiman, L. (1996). OUT-OF-BAG ESTIMATION.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical Modeling : The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. http ://doi.org/10.1214/ss/1009213726

Brenning, A., Bangs, D., & Becker, M. (2018). *RSAGA : SAGA Geoprocessing and Terrain Analysis*. Retrieved from https://CRAN.R-project.org/package=RSAGA

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. http ://doi.org/10.1038/nmeth.4642

Carnevale, P., Robert, V., Manguin, S., Corbel, V., Fontenille, D., Garros, C., ... Roux, J. (2009). *Les anophèles : Biologie, transmission du Plasmodium et lutte antivectorielle*. IRD. Retrieved from http://www.documentation.ird.fr/hor/fdi:010047862

Carrasco, D., Lefèvre, T., Moiroux, N., Pennetier, C., Chandre, F., & Cohuet, A. (2019). Behavioural adaptations of mosquito vectors to insecticide control. *Current Opinion in Insect Science*, *34*, 48–54. http ://doi.org/10.1016/j.cois.2019.03.005

CILSS, 2016. (2016). Landscapes of West Africa—A window on a changing world : Ouagadougou, Burkina Faso, CILSS, 219 p. (Comité Permanent Inter-états de Lutte contre la Sécheresse dans le Sahel) [Also available at https ://eros.usgs.gov/westafrica]. http ://doi.org/http://dx.doi.org/10.5066/F7N014QZ

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, J. M., Le Menach, A., Pothin, E., Eisele, T. P., Gething, P. W., Eckhoff, P.

A., ... Smith, D. L. (2017). Mapping multiple components of malaria risk for improved targeting of elimination interventions. *Malaria Journal*, *16*(1), 459. http ://doi.org/10.1186/s12936-017-2106-3

Cohuet, A., Simard, F., Berthomieu, A., Raymond, M., Fontenille, D., & Weill, M. (2002). Isolation and characterization of microsatellite DNA markers in the malaria vector Anopheles funestus. *Molecular Ecology Notes*, *2*(4), 498–500.

Commission, E., Research, D.-G. for, & Innovation. (2022). *Early warning for epidemics : EIC horizon prize*. http ://doi.org/doi/10.2777/64345

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., ... Böhner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, *8*(7), 1991–2007. http ://doi.org/10.5194/gmd-8-1991-2015

Corbel, V., & N'Guessan, R. (2013). Distribution, Mechanisms, Impact and Management of Insecticide Resistance in Malaria Vectors : A Pragmatic Review. *Anopheles Mosquitoes - New Insights into Malaria Vectors*. http ://doi.org/10.5772/56117

Cox, F. E. (2010). History of the discovery of the malaria parasites and their vectors. *Parasites & Vectors*, *3*(1), 5. http ://doi.org/10.1186/1756-3305-3-5

Davidson, G. (1957). Insecticide Resistance in Anopheles Sundaicus. *Nature*, *180*(4598), 1333–1335. http ://doi.org/10.1038/1801333a0

Davies, T. G. E., Field, L. M., Usherwood, P. N. R., & Williamson, M. S. (2007). DDT, pyrethrins, pyrethroids and insect sodium channels. *IUBMB Life*, *59*(3), 151–162. http ://doi.org/10.1080/15216540701352042

Derua, Y. A., Alifrangis, M., Hosea, K. M., Meyrowitsch, D. W., Magesa, S. M., Pedersen, E. M., & Simonsen, P. E. (2012). Change in composition of the Anopheles gambiae complex and its possible implications for the transmission of malaria and lymphatic filariasis in north-eastern Tanzania. *Malaria Journal*,

11(1), 188. http ://doi.org/10.1186/1475-2875-11-188

Djènontin, A., Pennetier, C., Zogo, B., Soukou, K. B., Ole-Sangba, M., Akogbéto, M., ... Corbel, V. (2014). Field Efficacy of Vectobac GR as a Mosquito Larvicide for the Control of Anopheline and Culicine Mosquitoes in Natural Habitats in Benin, West Africa. *PLoS ONE*, *9*(2), e87934. http ://doi.org/10.1371/journal.pone.0087934

Durnez, L., & Coosemans, M. (2013). Residual Transmission of Malaria : An Old Issue for New Approaches. In S. Manguin (Ed.), *Anopheles mosquitoes - New insights into malaria vectors*. InTech. http ://doi.org/10.5772/55925

Ebhuoma, O., & Gebreslasie, M. (2016). Remote Sensing-Driven Climatic/Environmental Variables for Modelling Malaria Transmission in Sub-Saharan Africa. *International Journal of Environmental Research and Public Health*, *13*(6). http ://doi.org/10.3390/ijerph13060584

FAO. (1995). *The coordinating working party on fishery statistics : Its origin, role and structure*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, 18.

Fontenille, D. (2009). Qu'est ce qu'un insecte vecteur de maladies ? Quel risque entomologique dans la région ? *Actes Du Colloque Maladies Vectorielles Et Moustiques Vecteurs : Actualités Et Prévention Sur Le Littoral Méditerranéen. Journée d'information, CRDP, Allée de La Citadelle, Montpellier Centre.*

Friedman, J. H. (2001). Greedy function approximation : A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. http ://doi.org/10.1214/aos/1013203451

Garrett-Jones, C., & Organization, W. H. (1964). A method for estimating the man-biting rate / by c. Garrett-jones. World Health Organization.

Gatton, M. L., Chitnis, N., Churcher, T., Donnelly, M. J., Ghani, A. C., Godfray,

H. C. J., ... Lindsay, S. W. (2013). THE IMPORTANCE OF MOSQUITO BEHAVIOURAL ADAPTATIONS TO MALARIA CONTROL IN AFRICA. *Evolution*, *67*(4), 1218–1230. http ://doi.org/10.1111/evo.12063

Geissbühler, Y., Chaki, P., Emidi, B., Govella, N. J., Shirima, R., Mayagaya, V., ... Killeen, G. F. (2007). Interdependence of domestic malaria prevention measures and mosquito-human interactions in urban Dar es Salaam, Tanzania. *Malaria Journal*, *6*(1), 126. http ://doi.org/10.1186/1475-2875-6-126

Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I. R., Johnston, G. L., ... Hay, S. I. (2011). A new world malaria map : Plasmodium falciparum endemicity in 2010. *Malaria Journal*, *10*(1), 378. http ://doi.org/10.1186/1475-2875-10-378

Gillies, M. T. (1953). The duration of the gonotrophic cycle in Anopheles gambiae and Anopheles funestus, with a note on the efficiency of hand catching. *East African Medical Journal*, *30*(4), 129–135.

Gillies, M. T., & B. De Meillon. (1968). The Anophelinae of Africa south of the Sahara (Ethiopian Zoogeographical Region). *Publications of the South African Institute for Medical Research*, *54*.

Gillies, M. T., & Coetzee, M. (1987). A supplement to the Anophelinae of Africa South of the Sahara. *Publ S Afr Inst Med Res*, *55*, 1–143.

Gleave, K., Lissenden, N., Chaplin, M., Choi, L., & Ranson, H. (2021). Piperonyl butoxide (PBO) combined with pyrethroids in insecticide-treated nets to prevent malaria in Africa. *Cochrane Database of Systematic Reviews*, *2021*(6). http ://doi.org/10.1002/14651858.CD012776.pub3

Govella, N. J., Johnson, P. C. D., Killeen, G. F., & Ferguson, H. M. (2021). *Heritability and phenotypic plasticity of biting time behaviors in the major African malaria vector* Anopheles arabiensis (preprint). Evolutionary Biology. Retrieved from http://biorxiv.org/lookup/doi/10.1101/2021.05.17.444456

GRASS Development Team. (2017). *Geographic resources analysis support system (GRASS GIS) software, version 7.2*. Open Source Geospatial Foundation. Retrieved from http://grass.osgeo.org

Grizonnet, M., Michel, J., Poughon, V., Inglada, J., Savinaud, M., & Cresson, R. (2017). Orfeo ToolBox : Open source processing of remote sensing images. *Open Geospatial Data, Software and Standards*, *2*(1), 15.

Hawley, W. A., Phillips-Howard, P. A., Kuile, F. O. ter, Terlouw, D. J., Vulule, J. M., Ombok, M., ... Hightower, A. W. (2003). Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. *The American Journal of Tropical Medicine and Hygiene*, *68*(4 Suppl), 121–127.

Hay, G. J., & Castilla, G. (2008). Geographic Object-Based Image Analysis (GEOBIA) : A new name for a new discipline. In T. Blaschke, S. Lang, & G. J. Hay (Eds.), *Object-Based Image Analysis* (pp. 75–89). Berlin, Heidelberg : Springer Berlin Heidelberg. http ://doi.org/10.1007/978-3-540-77058-9_4

Hay, S. I., & Snow, R. W. (2006). The Malaria Atlas Project : Developing Global Maps of Malaria Risk. *PLoS Medicine*, *3*(12), e473. http ://doi.org/10.1371/journal.pmed.0030473

Hijmans, R. J. (2020). *Raster : Geographic Data Analysis and Modeling*. Retrieved from https://CRAN.R-project.org/package=raster

Holstein, M. (1952). *Biologie d'Anopheles gambiae : Recherches en Afrique-Occidentale Française*. Genève : OMS. Retrieved from http://www.documentation.ird.fr/hor/fdi:42581

Horning, N., Fleishman, E., Ersts, P. J., Fogarty, F. A., & Wohlfeil Zillig, M. (2020). Mapping of land cover with open-source software and ultra-high-resolution imagery acquired with unmanned aerial vehicles. *Remote Sensing in Ecology and Conservation*, *6*(4), 487–497. http ://doi.org/10.1002/rse2.144

INSD. (2015). Tableau de bord économique et social 2014 de la région du sud ouest.

INSD. (2017). Enquête nationale sur le secteur de l'orpaillage (ENSO).

Ismay, C., & Solomon, N. (n.d.). *Thesisdown : An updated r markdown thesis template using the bookdown package*.

Jenson, S. K., & Domingue, J. O. (1988). Extracting topographic structure from digital elevation data for geographic information-system analysis. *Photogrammetric Engineering and Remote Sensing*, *54*(11), 1593–1600. Retrieved from http://pubs.er.usgs.gov/publication/70142175

Johnson-Laird, P. N. (2013). *Human and Machine Thinking*. Taylor & Francis. Retrieved from https://books.google.fr/books?id=gUb7AQAAQBAJ

JPL, N. (2013). NASA Shuttle Radar Topography Mission Global 1 arc second. NASA EOSDIS Land Processes DAAC. http ://doi.org/10.5067/MEASURES/SRTM/SRTMGL1

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ... Kumar, V. (2017). Theory-Guided Data Science : A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. http ://doi.org/10.1109/TKDE.2017.2720168

Kell, D. B., & Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis ? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, *26*(1), 99–105. http ://doi.org/10.1002/bies.10385

Killeen, Gerry F. (2014). Characterizing, controlling and eliminating residual malaria transmission. *Malaria Journal*, *13*(1), 330. http ://doi.org/10.1186/1475-2875-13-330

Killeen, G. F., Githure, J. I., & Beier, J. C. (1999). Short report : Entomologic inoculation rates and Plasmodium falciparum malaria prevalence in Africa. *The American Journal of Tropical Medicine and Hygiene*, *61*(1), 109–113.

http ://doi.org/10.4269/ajtmh.1999.61.109

Killeen, Gerry F., Kihonda, J., Lyimo, E., Oketch, F. R., Kotas, M. E., Mathenge, E., ... Drakeley, C. J. (2006). Quantifying behavioural interactions between humans and mosquitoes : Evaluating the protective efficacy of insecticidal nets against malaria transmission in rural Tanzania. *BMC Infectious Diseases*, *6*(1), 161. http ://doi.org/10.1186/1471-2334-6-161

Killeen, Gerry F., & Smith, T. A. (2007). Exploring the contributions of bed nets, cattle, insecticides and excitorepellency to malaria control : A deterministic model of mosquito host-seeking behaviour and mortality. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *101*(9), 867–880. http ://doi.org/10.1016/j.trstmh.2007.04.022

Killeen, Gerry F., Smith, T. A., Ferguson, H. M., Mshinda, H., Abdulla, S., Lengeler, C., & Kachur, S. P. (2007). Preventing Childhood Malaria in Africa by Protecting Adults from Mosquitoes with Insecticide-Treated Nets. *PLoS Medicine*, *4*(7), e229. http ://doi.org/10.1371/journal.pmed.0040229

Koekemoer, L. L., Kamau, L., Hunt, R. H., & Coetzee, M. (2002). A cocktail polymerase chain reaction assay to identify members of the Anopheles funestus (Diptera : Culicidae) group. *The American Journal of Tropical Medicine and Hygiene*, *66*(6), 804–811.

Koenker, H., Taylor, C., Burgert-Brucker, C. R., Thwing, J., Fish, T., & Kilian, A. (2019). Quantifying seasonal variation in insecticide-treated net use among those with access. *The American Journal of Tropical Medicine and Hygiene*, *101*(2), 371–382. http ://doi.org/10.4269/ajtmh.19-0249

Labbé, P., David, J.-P., Alout, H., Milesi, P., Djogbénou, L., Pasteur, N., & Weill, M. (2017). Evolution of Resistance to Insecticide in Disease Vectors. In *Genetics and Evolution of Infectious Diseases* (pp. 313–339). Elsevier. http ://doi.org/10.1016/B978-0-12-799942-5.00014-7

Le Goff, G., Carneval, P., & Robert, V. (1997). Low dispersion of anopheline

malaria vectors in the African equatorial forest. *Parasite*, *4*(2), 187–189. http ://doi.org/10.1051/parasite/1997042187

Le Monde. (2020, December). Le Covid-19 est-il la plus grande épidémie actuelle ? (VIH, paludisme, tuberculose...). Retrieved from https://www.youtube.com/watch?v=ZizfWA7fyto

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Lockwood, J. A., Sparks, T. C., & Story, R. N. (1984). Evolution of Insect Resistance to Insecticides : A Reevaluation of the Roles of Physiology and Behavior. *Bulletin of the Entomological Society of America*, *30*(4), 41–51. http ://doi.org/10.1093/besa/30.4.41

Main, B. J., Lee, Y., Ferguson, H. M., Kreppel, K. S., Kihonda, A., Govella, N. J., ... Lanzaro, G. C. (2016). The Genetic Basis of Host Preference and Resting Behavior in the Major African Malaria Vector, Anopheles arabiensis. *PLOS Genetics*, *12*(9), e1006303. http ://doi.org/10.1371/journal.pgen.1006303

Martinez-Torres, D., Chandre, F., Williamson, M. S., Darriet, F., Berge, J. B., Devonshire, A. L., ... Pauron, D. (1998). Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector Anopheles gambiae s.s. *Insect Molecular Biology*, *7*(2), 179–184. http ://doi.org/10.1046/j.1365-2583.1998.72062.x

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space ? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*. http ://doi.org/10.1111/2041-210X.13650

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, *101*, 1–9. http ://doi.org/10.1016/j.envsoft.2017.12.001

Moiroux, N. (2012). *Modélisation du risque d'exposition aux moustiques vecteurs de plasmodium spp. Dans un contexte de lutte anti-vectorielle* (PhD thesis). Retrieved from http://www.theses.fr/2012MON20177/document

Moiroux, N., Bio-Bangana, A. S., Djènontin, A., Chandre, F., Corbel, V., & Guis, H. (2013). Modelling the risk of being bitten by malaria vectors in a vector control area in southern Benin, west Africa. *Parasites & Vectors*, *6*(1), 71. http ://doi.org/10.1186/1756-3305-6-71

Moiroux, N., Djènontin, A., Bio-Bangana, A. S., Chandre, F., Corbel, V., & Guis, H. (2014). Spatio-temporal analysis of abundances of three malaria vector species in southern Benin using zero-truncated models. *Parasites & Vectors*, *7*(1), 103. http ://doi.org/10.1186/1756-3305-7-103

Molnar, C. (2019). *Interpretable Machine Learning : A Guide for Making Black Box Models Explainable*.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., ... Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI* (Vol. 13200, pp. 39–68). Cham : Springer International Publishing. http ://doi.org/10.1007/978-3-031-04083-2_4

Monroe, A., Moore, S., Koenker, H., Lynch, M., & Ricotta, E. (2019). Measuring and characterizing night time human behaviour as it relates to residual malaria transmission in sub-Saharan Africa : A review of the published literature. *Malaria Journal*, *18*(1), 6. http ://doi.org/10.1186/s12936-019-2638-9

Moyes, C. L., Athinya, D. K., Seethaler, T., Battle, K. E., Sinka, M., Hadi, M. P., ... Hancock, P. A. (2020). Evaluating insecticide resistance across African districts to aid malaria control decisions. *Proceedings of the National Academy of Sciences*, *117*(36), 22042–22050. http ://doi.org/10.1073/pnas.2006781117

Müller, G. C., Beier, J. C., Traore, S. F., Toure, M. B., Traore, M. M., Bah, S., ...

Schlein, Y. (2010). Successful field trial of attractive toxic sugar bait (ATSB) plant-spraying methods against malaria vectors in the Anopheles gambiae complex in Mali, West Africa. *Malaria Journal*, *9*(1). http ://doi.org/10.1186/1475-2875-9-210

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. http ://doi.org/10.1073/pnas.1900654116

Mwangangi, J. M., Mbogo, C. M., Orindi, B. O., Muturi, E. J., Midega, J. T., Nzovu, J., ... Beier, J. C. (2013). Shifts in malaria vector species composition and transmission dynamics along the Kenyan coast over the past 20 years. *Malaria Journal*, *12*(1), 13. http ://doi.org/10.1186/1475-2875-12-13

NASA. (2019). GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V06. NASA Goddard Earth Sciences Data ; Information Services Center. http ://doi.org/10.5067/GPM/IMERGDF/DAY/06

NASA GSFC, P. P. S. (PPS). (2019). GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06. NASA Goddard Earth Sciences Data ; Information Services Center. http ://doi.org/10.5067/GPM/IMERG/3B-HH/06

Nosten, F. H., & Phyo, A. P. (2019). New malaria maps. *The Lancet*, *394*(10195), 278–279. http ://doi.org/10.1016/S0140-6736(19)31273-5

O'Reilly, A. O., Khambay, B. P. S., Williamson, M. S., Field, L. M., WAllace, B. A., & Davies, T. G. E. (2006). Modelling insecticide-binding sites in the voltage-gated sodium channel. *Biochemical Journal*, *396*(2), 255–263. http ://doi.org/10.1042/BJ20051925

ONeill, P. E., Chan, S., Njoku, E. G., Jackson, T., Bindlish, R., Chaubell, M. J., & Colliander, A. (2021). SMAP Enhanced L3 Radiometer Global and Polar Grid Daily 9 km EASE-Grid Soil Moisture, Version 5. NASA National Snow ;

Ice Data Center DAAC. http ://doi.org/10.5067/4DQ54OUIJ9DL

OSS. (2015). Burkina faso : Atlas des cartes d'occupation du sol» - projet amélioration de la résilience des populations sahéliennes aux mutations environnementales - REPSAHEL.

Ouedraogo, A. L., Bastiaens, G. J. H., Tiono, A. B., Guelbeogo, W. M., Kobylinski, K. C., Ouedraogo, A., ... Bousema, T. (2015). Efficacy and Safety of the Mosquitocidal Drug Ivermectin to Prevent Malaria Transmission After Treatment : A Double-Blind, Randomized, Clinical Trial. *Clinical Infectious Diseases*, *60*(3), 357–365. http ://doi.org/10.1093/cid/ciu797

Paaijmans, K. P., & Huijben, S. (2020). Taking the "I" out of LLINs : Using insecticides in vector control tools other than long-lasting nets to fight malaria. *Malaria Journal*, *19*(1). http ://doi.org/10.1186/s12936-020-3151-x

Parselia, E., Kontoes, C., Tsouni, A., Hadjichristodoulou, C., Kioutsioukis, I., Magiorkinis, G., & Stilianakis, N. I. (2019). Satellite Earth Observation Data in Epidemiological Modeling of Malaria, Dengue and West Nile Virus : A Scoping Review. *Remote Sensing*, *11*(16), 1862. http ://doi.org/10.3390/rs11161862

Pebesma, E. (2018). Simple Features for R : Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. http ://doi.org/10.32614/RJ-2018-009

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*(7633), 418–422. http ://doi.org/10.1038/nature20584

PNLP. (2014a). Directives nationales pour la prise en charge du paludisme dans les formations sanitaires du burkina faso. Ministère de la santé/burkina faso.

PNLP. (2014b). Programme national de lutte contre le paludisme en côte d'ivoire. 2014. Plan stratégique national de lutte contre le paludisme 2012–2015 (période replanifiée 2014–2017). Approche stratifiée de mise à l'échelle des interventions

de lutte contre le paludisme en côte d'ivoire et consolidation des acquis. Abidjan : Ministère de la santé et l'hygiène publique. 149 p.

QGIS Development Team. (2021). *QGIS Geographic Information System*. QGIS Association. Retrieved from https://www.qgis.org

R Core Team. (2018). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ranson, H., Jensen, B., Vulule, J. M., Wang, X., Hemingway, J., & Collins, F. H. (2000). Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan Anopheles gambiae associated with resistance to DDT and pyrethroids. *Insect Molecular Biology*, *9*(5), 491–497. http ://doi.org/10.1046/j.1365-2583.2000.00209.x

Reisen, W. K. (2010). Landscape Epidemiology of Vector-Borne Diseases. *Annual Review of Entomology*, *55*(1), 461–483. http ://doi.org/10.1146/annurev-ento-112408-085419

Riveron, J. M., Tchouakui, M., Mugenzi, L., Menze, B. D., Chiang, M.-C., & Wondji, C. S. (2018). Insecticide Resistance in Malaria Vectors : An Update at a Global Scale. In S. Manguin & V. Dev (Eds.), *Towards Malaria Elimination - A Leap Forward*. InTech. http ://doi.org/10.5772/intechopen.78375

Rodhain, F. (2015). Le microbe, l'insecte, l'homme et les autres... : Le monde des maladies à vecteurs. *Bulletin de l'Académie Vétérinaire de France*, *168*(1), 5–11. http ://doi.org/10.4267/2042/56539

Román, M. O., Wang, Z., Sun, Q., Kalb, V., Miller, S. D., Molthan, A., ... Masuoka, E. J. (2018). NASA's Black Marble nighttime lights product suite. *Remote Sensing of Environment*, *210*, 113–143. http ://doi.org/10.1016/j.rse.2018.03.017

RStudio Team. (2020). *RStudio : Integrated Development Environment for R*. Boston, MA : RStudio, PBC. Retrieved from http://www.rstudio.com/

Russell, T. L., Govella, N. J., Azizi, S., Drakeley, C. J., Kachur, S. P., & Killeen, G. F. (2011). Increased proportions of outdoor feeding among residual malaria vector populations following increased use of insecticide-treated nets in rural Tanzania. *Malaria Journal*, *10*(1). http ://doi.org/10.1186/1475-2875-10-80

Service, M. W. (2008). Sampling adults by animal bait catches and by animal-baited traps. In *Mosquito ecology : Field sampling methods* (pp. 493–675). Dordrecht : Springer Netherlands. http ://doi.org/10.1007/978-1-4020-6666-5_6

Shapiro, L. L. M., Whitehead, S. A., & Thomas, M. B. (2017). Quantifying the effects of temperature on mosquito and parasite traits that determine the transmission potential of human malaria. *PLoS Biology*, *15*(10). http ://doi.org/10.1371/journal.pbio.2003489

Sherrard-Smith, E., Skarp, J. E., Beale, A. D., Fornadel, C., Norris, L. C., Moore, S. J., ... Churcher, T. S. (2019). Mosquito feeding behavior and how it influences residual malaria transmission across Africa. *Proceedings of the National Academy of Sciences*, *116*(30), 15086–15095. http ://doi.org/10.1073/pnas.1820646116

Shmueli, G. (2010). To Explain or to Predict ? *Statistical Science*, *25*(3), 289–310. http ://doi.org/10.1214/10-STS330

Shmueli, G., & Koppius, O. (2010). Predictive Analytics in Information Systems Research. *SSRN Electronic Journal*. http ://doi.org/10.2139/ssrn.1606674

Sinka, M. E., Bangs, M. J., Manguin, S., Coetzee, M., Mbogo, C. M., Hemingway, J., ... Hay, S. I. (2010). The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East : Occurrence data, distribution maps and bionomic précis. *Parasites & Vectors*, *3*(1). http ://doi.org/10.1186/1756-3305-3-117

Sinka, M. E., Bangs, M. J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., ... Hay, S. I. (2012). A global map of dominant malaria vectors. *Parasites & Vectors*, *5*(1). http ://doi.org/10.1186/1756-3305-5-69

Sokhna, C., Ndiath, M. O., & Rogier, C. (2013). The changes in mosquito vector behaviour and the emerging resistance to insecticides will challenge the decline of malaria. *Clinical Microbiology and Infection*, *19*(10), 902–907. http ://doi.org/10.1111/1469-0691.12314

Soma, D. D., Zogo, B. M., Somé, A., Tchiekoi, B. N., Hien, D. F. de S., Pooda, H. S., … Dabiré, R. K. (2020). Anopheles bionomics, insecticide resistance and malaria transmission in southwest Burkina Faso : A pre-intervention study. *PLOS ONE*, *15*(8), e0236920. http ://doi.org/10.1371/journal.pone.0236920

Sondo, P., Derra, K., Rouamba, T., Nakanabo Diallo, S., Taconet, P., Kazienga, A., … Tinto, H. (2020). Determinants of Plasmodium falciparum multiplicity of infection and genetic diversity in Burkina Faso. *Parasites & Vectors*, *13*(1). http ://doi.org/10.1186/s13071-020-04302-z

Sougoufara, S., Harry, M., Doucouré, S., Sembène, P. M., & Sokhna, C. (2016). Shift in species composition in the *Anopheles gambiae* complex after implementation of long-lasting insecticidal nets in Dielmo, Senegal. *Medical and Veterinary Entomology*, *30*(3), 365–368. http ://doi.org/10.1111/mve.12171

Sougoufara, Seynabou, Ottih, E. C., & Tripet, F. (2020). The need for new vector control approaches targeting outdoor biting anopheline malaria vector communities. *Parasites & Vectors*, *13*(1), 295. http ://doi.org/10.1186/s13071-020-04170-7

Sternberg, E. D., Ng'habi, K. R., Lyimo, I. N., Kessy, S. T., Farenhorst, M., Thomas, M. B., … Mnyone, L. L. (2016). Eave tubes for malaria control in Africa : Initial development and semi-field evaluations in Tanzania. *Malaria Journal*, *15*(1), 447. http ://doi.org/10.1186/s12936-016-1499-8

Stewart, Z. P., Oxborough, R. M., Tungu, P. K., Kirby, M. J., Rowland, M. W., & Irish, S. R. (2013). Indoor Application of Attractive Toxic Sugar Bait (ATSB) in Combination with Mosquito Nets for Control of Pyrethroid-Resistant Mosquitoes. *PLoS ONE*, *8*(12), e84168. http ://doi.org/10.1371/journal.pone.0084168

Stresman, G. H. (2010). Beyond temperature and precipitation : Ecological risk factors that modify malaria transmission. *Acta Tropica*, *116*(3), 167–172. http ://doi.org/10.1016/j.actatropica.2010.08.005

Tchuinkam, T., Simard, F., Lélé-Defo, E., Téné-Fossog, B., Tateng-Ngouateu, A., Antonio-Nkondjio, C., ... Awono-Ambéné, H.-P. (2010). Bionomics of Anopheline species and malaria transmission dynamics along an altitudinal transect in Western Cameroon. *BMC Infectious Diseases*, *10*(1). http ://doi.org/10.1186/1471-2334-10-119

Tsantalidou, A., Parselia, E., Arvanitakis, G., Kyratzi, K., Gewehr, S., Vakali, A., & Kontoes, C. (2021). MAMOTH : An Earth Observational Data-Driven Model for Mosquitoes Abundance Prediction. *Remote Sensing*, *13*(13), 2557. http ://doi.org/10.3390/rs13132557

Wan, Z., Hook, S., & Hulley, G. (2015a). MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. http ://doi.org/10.5067/MODIS/MOD11A1.006

Wan, Z., Hook, S., & Hulley, G. (2015b). MYD11A1 MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. http ://doi.org/10.5067/MODIS/MYD11A1.006

Wardrop, N. A., Geary, M., Osborne, P. E., & Atkinson, P. M. (2014). Interpreting predictive maps of disease : Highlighting the pitfalls of distribution models in epidemiology. *Geospatial Health*, 237–246. http ://doi.org/10.4081/gh.2014.397

Weill, M., Malcolm, C., Chandre, F., Mogensen, K., Berthomieu, A., Marquine, M., & Raymond, M. (2004). The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors. *Insect Molecular Biology*, *13*(1), 1–7. http ://doi.org/10.1111/j.1365-2583.2004.00452.x

Weiss, D. J., Lucas, T. C. D., Nguyen, M., Nandi, A. K., Bisanzio, D., Battle, K. E., ... Gething, P. W. (2019). Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17 : A spatial and temporal model-

ling study. *The Lancet*, *394*(10195), 322–331. http ://doi.org/10.1016/S0140-6736(19)31097-9

WHO. (2001). A framework for field research in Africa : Malaria early warning systems : Concepts, indicators and partners. World Health Organization.

WHO. (2009). Report of the twelfth WHOPES working group meeting, WHO/HQ, geneva, 8-11 december 2008 : Review of bioflash GR, permanet 2.0, permanet 3.0, permanet 2.5, lambda-cyhalothrin LN. World Health Organization.

WHO. (2017a). *Framework for a national plan for monitoring and management of insecticide resistance in malaria vectors*. World Health Organization.

WHO. (2017b). *Global vector control response 2017-2030* (pp. 51 p.). World Health Organization.

WHO. (2018). *Malaria surveillance, monitoring and evaluation : A reference manual* (pp. ix, 196 p.). World Health Organization.

WHO. (2020). World malaria report 2020 : 20 years of global progress and challenges. Licence : CC BY-NC-SA 3.0 IGO.

WHO. (2021). World malaria report 2021. Licence : CC BY-NC-SA 3.0 IGO.

Wiebe, A., Longbottom, J., Gleave, K., Shearer, F. M., Sinka, M. E., Massey, N. C., ... Moyes, C. L. (2017). Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance. *Malaria Journal*, *16*(1), 85. http ://doi.org/10.1186/s12936-017-1734-y

Wilson, A. L., Courtenay, O., Kelly-Hope, L. A., Scott, T. W., Takken, W., Torr, S. J., & Lindsay, S. W. (2020). The importance of vector control for the control and elimination of vector-borne diseases. *PLOS Neglected Tropical Diseases*, *14*(1), e0007831. http ://doi.org/10.1371/journal.pntd.0007831

Xie, Y. (2019). *Bookdown : Authoring books and technical documents with r markdown*. Retrieved from https://github.com/rstudio/bookdown

Xie, Y. (2020). *Knitr : A general-purpose package for dynamic report generation in r*. Retrieved from https://yihui.org/knitr/

Yanai, I., & Lercher, M. (2019a). Night science. *Genome Biology*, *20*(1). http ://doi.org/10.1186/s13059-019-1800-6

Yanai, I., & Lercher, M. (2019b). What is the question ? *Genome Biology*, *20*(1), 289, s13059-019-1902-1. http ://doi.org/10.1186/s13059-019-1902-1

Yé, Y., Hoshen, M., Louis, V., Séraphin, S., Traoré, I., & Sauerborn, R. (2006). Housing conditions and Plasmodium falciparum infection : Protective effect of iron-sheet roofed houses. *Malaria Journal*, *5*(1), 8. http ://doi.org/10.1186/1475-2875-5-8

Yin, Q., Li, L., Guo, X., Wu, R., Shi, B., Wang, Y., ... Zhou, X. (2019). A field-based modeling study on ecological characterization of hourly host-seeking behavior and its associated climatic variables in Aedes albopictus. *Parasites & Vectors*, *12*(1), 474. http ://doi.org/10.1186/s13071-019-3715-1

Yu, Q., Ji, W., Prihodko, L., Ross, C. W., Anchang, J. Y., & Hanan, N. P. (2021). Study becomes insight : Ecological learning from machine learning. *Methods in Ecology and Evolution*. http ://doi.org/10.1111/2041-210X.13686

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., ... Arino, O. (2021, October). ESA WorldCover 10 m 2020 v100. Zenodo. http ://doi.org/10.5281/ZENODO.5571936

Zhao, Q., & Hastie, T. (2021). Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, *39*(1), 272–281. http ://doi.org/10.1080/07350015.2019.1624293

Zhu, H. (2019). *kableExtra : Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

Zogo, B., Koffi, A. A., Alou, L. P. A., Fournet, F., Dahounto, A., Dabiré, R. K.,

... Pennetier, C. (2019). Identification and characterization of Anopheles spp. Breeding habitats in the Korhogo area in northern Côte d'Ivoire : A study prior to a Bti-based larviciding intervention. *Parasites & Vectors*, *12*(1), 146. http ://doi.org/10.1186/s13071-019-3404-0

Zogo, B., Soma, D. D., Tchiekoi, B. N., Somé, A., Ahoua Alou, L. P., Koffi, A. A., ... Pennetier, C. (2019). *Anopheles* bionomics, insecticide resistance mechanisms, and malaria transmission in the Korhogo area, northern Côte d'Ivoire : A pre-intervention study. *Parasite*, *26*, 40. http ://doi.org/10.1051/parasite/2019040

Annexe A

Description des données recueillies sur le terrain au cours du projet REACT

Cette annexe décrit les protocoles de recueil des données de terrain collectées dans le cadre du projet REACT, utilisées dans les travaux de thèse

Données entomologiques

Dans le cadre du projet REACT, plusieurs enquêtes entomologiques (huit en Côte d'Ivoire, sept au Burkina Faso) ont été effectuées dans chaque village au cours des 2 années du projet. Les périodes des enquêtes ont couvert les conditions climatiques typiques de ces régions tropicales (à l'exception de la haute saison des pluies - juillet à septembre). Les moustiques ont été collectés en utilisant la technique de la capture sur sujet humain (Service, 2008), de 17h00 à 09h00, à l'intérieur et à l'extérieur des habitations, à raison de quatre sites (habitations) par village. Ainsi, dans la zone de Korhogo (CI), un total de 2048 nuits-homme de capture a été réalisé (32 villages * 8 enquêtes entomologiques * 4 points de collecte * 2 lieux) ; tandis que dans la zone de Diébougou (BF), un total de 1512 nuits-homme de capture a été réalisé (27 villages * 7 enquêtes entomologiques * 2 points de collecte * 2 lieux). Au total, cela représente environ 52000 heures de collecte effectuées dans le cadre du projet REACT. Les anophèles ont été identifiés à l'aide de clés morphologiques (Gillies & B. De Meillon, 1968 ; Gillies & Coetzee, 1987). Tous les individus appartenant au groupe *Anopheles funestus* (dans les deux zones d'étude) et au complexe *Anopheles gambiae* (sur la zone burkinabé uniquement) ont été identifiés à l'espèce par PCR (Cohuet

et al., 2002 ; Koekemoer, Kamau, Hunt, & Coetzee, 2002). Sur la zone ivoirienne, en raison du très grand nombre d'*An. gambiae s.l.* collectés, un sous-échantillon de ces individus (sélectionnés aléatoirement dans l'espace et le temps) a été identifié à l'espèce. Enfin, les mutations de la cible L1014F (*kdr-w*), L1014S (*kdr-e*) et G119S (*ace-1*) ont été détectées par PCR sur tous les *An. gambiae s.l.* et *An. coluzzii* collectés sur la zone burkinabé. Des descriptions détaillées des méthodes utilisées pour collecter ces données sont fournies dans (Soma et al., 2020 ; Zogo, Soma, et al., 2019).

Le tableau A.1 résume la composition spécifique des anophèles capturés sur chaque zone d'étude.

TABLE A.1: Composition spécifique des anophèles capturés au cours des missions entomologiques du projet REACT

	Zone CI	Zone BF
Nombre total capturés	57722	3056
Densité agressive moyenne*	28.18	2.02
% An. gambiae s.s.	97 %	20 %
% An. coluzzii	< 1 %	43 %
% An. funestus	1%	23 %
% autres espèces	< 1 %	14%

* nombre moyen de piqûres / homme / nuit de capture

Données de comportement humain relatif à l'utilisation de moustiquaires et aux habitudes horaires nocturnes

Au total, cinq enquêtes de comportement humain (trois sur la zone de Diébougou, deux sur la zone de Korhogo) ont été menées. Les enquêtes ont été effectuées après la distribution des moustiquaires (voir figure 3.1) et couvrent au mieux les conditions climatiques typiques des zones d'étude. Pour chaque enquête, une quinzaine de ménages a été séléctionnée aléatoirement, et pour chacun de ces ménages trois personnes (maximum) appartenant à chacun des trois groupes d'âge suivants ont été aléatoirement séléctionnées : 0–5 ans, 6–17 ans, et supérieur ou égal à 18 ans. Le chef du ménage a ensuite été questionné sur l'heure à laquelle chaque personne sélectionnée, la nuit précédant l'enquête, est

entrée (le soir) et sortie (le matin) de i) l'habitation, et ii) son espace de sommeil éventuellement protégé par une MIILDA. Les ménages pour ces enquêtes ont été sélectionnés indépendamment de ceux des enquêtes entomologiques. Le protocole de collecte de ces données est plus largement détaillé dans le chapitre 4.

Données de micro-climat au cours des collectes entomologiques

Les paramètres climatiques et environnementaux mesurés simultanément à chaque collecte entomologique étaient les suivants : la température, l'humidité relative, la luminosité et la pression atmosphérique. Les instruments utilisés pour mesurer ces données étaient les suivants : pour la température et l'humidité relative : capteur Hygro Buttons 23 [Proges Plus DAL0084] (résolution temporelle (RT) : 15 minutes) ; pour la luminosité : capteur HOBO Pendant® Temperature/Light 8K (RT : 15 minutes) ; pour la pression atmosphérique : capteur Extech SD700 (RT : 10 minutes). Les capteurs Hygro et Hobo étaient positionnés à l'intérieur des maisons où les captures étaient effectuées (au milieu de la pièce) et à l'extérieur (près du point captureur). Le baromètre était positionné au centre du village.

Nous avons complété ces paramètres mesurés sur le terrain avec des données issues de produits satellitaires ou de modèles météorologiques disponibles à des résolutions spatiales bien plus large. En particulier, nous avons extrait i) les précipitations semi-horaires des produits GPM (NASA GSFC, 2019) (voir section suivante) (résolution spatiale : 10 km, résolution temorelle : 30 minutes), et ii) la vitesse du vent des produits ERA-5 (résolution spatiale : 28 km, résolution temorelle : 1 heure).

La figure A.1 montre les séries temporelles horaires de température, humidité, luminosité et pression atmosphérique pour chaque enquête entomologique sur chaque zone d'étude.

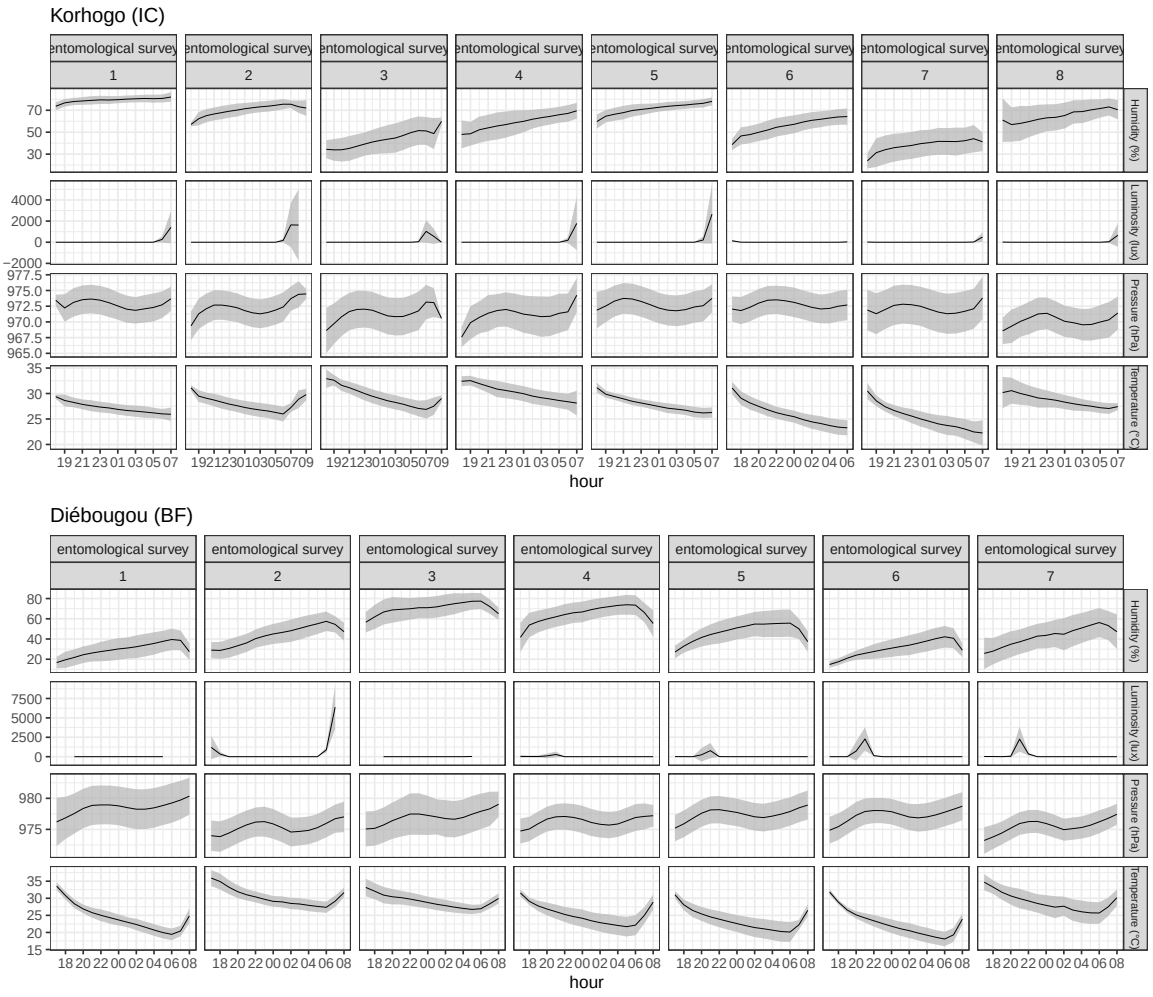


FIGURE A.1: Conditions micro-climatiques horaires au cours des enquêtes entomologiques

Données de population et localisation des habitations

Afin de décrire l'attractivité et la pénétrabilité des des ménages pour les vecteurs du paludisme, nous avons recensé la population et enregistré l'emplacement géographique des habitations dans les villages, au cours du recensement de la population en début de projet, à l'aide d'un récépteur Global Positioning System (GPS).

Annexe B

Détails sur les travaux de cartographie de l'occupation du sol

Cette annexe donne des informations supplémentaires sur les travaux de cartographie de l'occupation du sol dans les deux zones d'étude du projet REACT.

Résumé textuel et graphique des traitements

Méthode. [*NB : la méthode est détaillée dans la section 3.2.2. Ci-après, nous la résumons en quelques lignes.*] Nous avons généré des produits d'occupation du sol dans nos deux zones d'étude selon la méthode décrite ci-dessus. Les produits satellitaires utilisés étaient les suivants : images SPOT 6/7 (Satellite Pour l'Observation de la Terre), images Sentinel-2, et Modèle Numérique de Terrain (MNT) Shuttle Radar Topography Mission (SRTM) (JPL, 2013). Nous avons mené des campagnes de terrain sur les deux zones, en novembre et décembre 2018, afin de constituer les jeux de données d'apprentissage et de validation. Nous avons établi les classes d'occupation du sol dans chacune des zones sur la base de recherches bibliographiques sur les types de paysages potentiellement rencontrés dans nos zones (Aubréville, 1957 ; CILSS, 2016 ; OSS, 2015) et de nos observations du paysage sur le terrain. Nous avons collecté un minimum de 20 parcelles par classe, en tentant de les répartir au mieux sur l'étendue de chacune des zones. L'algorithme de segmentation utilisé est celui proposé par (Baatz & Schape, 2000). Nous avons calculé au total une centaine de variables prédictives basés sur les images SPOT, Sentinel-2 et le MNT, ainsi que la forme des objets. Nous avons ensuite entrainé un modèle de forêts aléatoires (Breiman, 2001a) sur le jeu

de données d'entrainement. Nous avons généré la matrice de confusion en utilisant la procédure de validation interne aux forêts aléatoires (basée sur les observations 'out-of-bag' (Breiman, 1996)). En se basant sur cette matrice, nous avons ensuite regroupé, dans le jeu de données d'entrainement, les classes d'occupation du sol dont la confusion était importante (par exemple, zones de culture de mil et de sorgho) ; en prenant cependant soin de conserver la distinction entre les différentes classes à priori favorables à la présence de gîtes larvaires (par exemple, zones marécageuses et rizicoles) ou aux résistances physiologiques ou comportementales des vecteurs. Nous avons entrainé un modèle de forêt aléatoires sur cette nouvelle version du jeu de données d'entraînement puis l'avons utilisé pour prédire la classe d'occupation du sol sur chaque objet issu de la segmentation. Comme précédemment, nous avons généré la matrice de confusion puis en avons extrait un indice de qualité de la classification (*accuracy* (J. Cohen, 1960)) mesurant la proportion d'objets correctement classés.

Les différentes étapes de la classification sont résumées graphiquement dans la figure suivante (B.1).

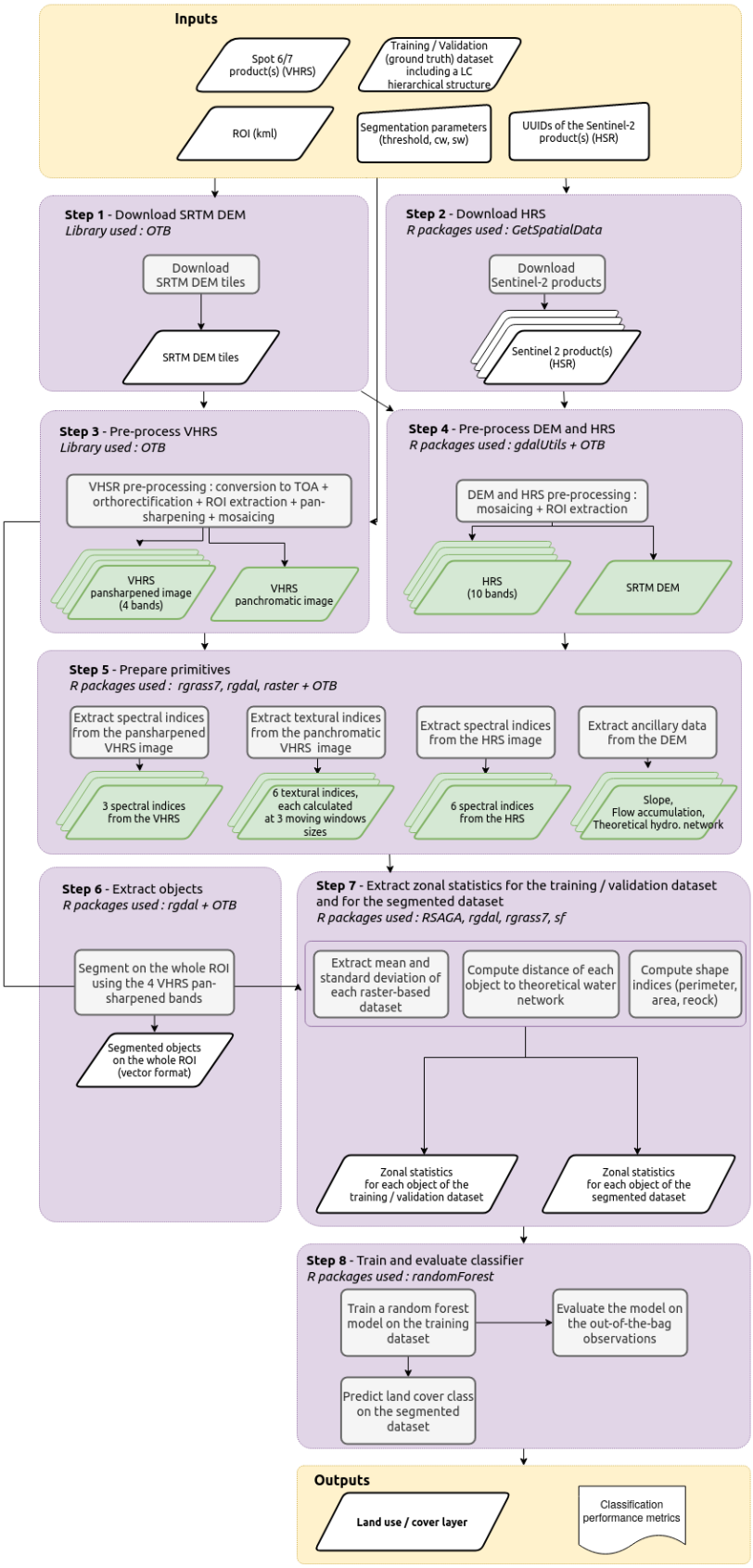


FIGURE B.1: Etapes du processus de cartographie de l'occupation du sol par classification supervisée orientée objet d'images satellitaires

Définitions des classes d'occupation du sol

Classe	Définition
Eau dormante	Zone d'eaux profondes, généralement conséquences de la mise en place de barrages ou autres infrastructures de retenues d'eaux
Eau courante	Cours d'eau (rivière ou fleuve en eau tout au long de l'année)
Culture à maïs	Zones de culture à maïs
Culture à poids de terre	Zones de culture à poids de terre
Culture à arachide	Zones de culture à arachide
Culture à mil	Zones de culture à mil
Culture à sorgho	Zones de culture à sorgho
Culture à haricot	Zones de culture à haricot
Culture à sésame	Zones de culture à sésame
Jachère	Terrain laissé en jachère et recouvert d'herbes envahissantes denses peu pénétrables
Culture cotonière	Zones de culture à coton
Rizière	Zones de culture à riz
Plantation à mangue	Zones de plantations à manguiers
Plantation à anacarde	Zones de plantations à anacardiers
Savane arbustive	Formation herbeuse comportant un tapis de grandes herbes graminéennes mesurant, en fin de saison de végétation, au moins 80 cm de hauteur, avec des feuilles planes disposées à la base ou sur les chaumes, des herbes et plantes herbacées de moindre taille - où seuls les arbustes sont présents
Savane arborée	Formation herbeuse comportant un tapis de grandes herbes graminéennes mesurant, en fin de saison de végétation, au moins 80 cm de hauteur, avec des feuilles planes disposées à la base ou sur les chaumes, des herbes et plantes herbacées de moindre taille - où arbres et arbustes sont disséminés
Savane boisée	Formation herbeuse comportant un tapis de grandes herbes graminéennes mesurant, en fin de saison de végétation, au moins 80 cm de hauteur, avec des feuilles planes disposées à la base ou sur les chaumes, des herbes et plantes herbacées de moindre taille - où arbres et arbustes forment un couvert clair et continu
Savane dégradée	Savane anciennement arborée ou boisée, dont la dégradation a affecté le peuplement et le couvert (moins important que ceux des savanes non dégradées). La formation herbacée y est souvent absente. Ces zones pourraient aussi être des savanes déboisées actuellement en cours de recolonisation. En terme de recouvrement ligneux ce milieu ressemble à la savane boisée.
Prairie	Formation herbeuse ouverte constituée de graminées (et parfois quelques plantes ligneuses). Si présente, la strate ligneuse est de faible densité et de hauteur inférieure à 10 mètres. ; les graminées sont vivaces et ne dépassent pas généralement 80 cm de haut
Forêt dense	Forêt non inféodées aux zones humides (non inondables), dont le peuplement est fermé, sans ouverture majeure du couvert (supérieur à 80 %), les arbres ont une hauteur de l'ordre d'une vingtaine de mètres
Forêt claire	Forêt non inféodées aux zones humides (non inondables), dont le peuplement est ouvert, avec des arbres de petite et moyenne taille (10 à 20 mètres de hauteur), et les cimes sont plus ou moins jointives, l'ensemble du couvert laissant largement filtrer la lumière (40 à 60 % de recouvrement)
Forêt ripicole	Formation forestière des bords de cours d'eau, temporairement inondées
Prairie marécageuse	Formation herbeuse comportant un tapis de grandes herbes graminéennes (minimum 1 mètre), fréquemment inondée en saison humide, présente dans les zones de bas fond ou en bordure des marais ou des eaux dormantes
Bâti	Milieux construits (villes ou villages)
Sol nu	Sols nu dégradés ou sols nu agricoles
Routes principales	Routes ou sentiers de largeur supérieure à 5 mètres

Photographies représentatives des principales classes d'occupation du sol rencontrées sur les zones d'étude

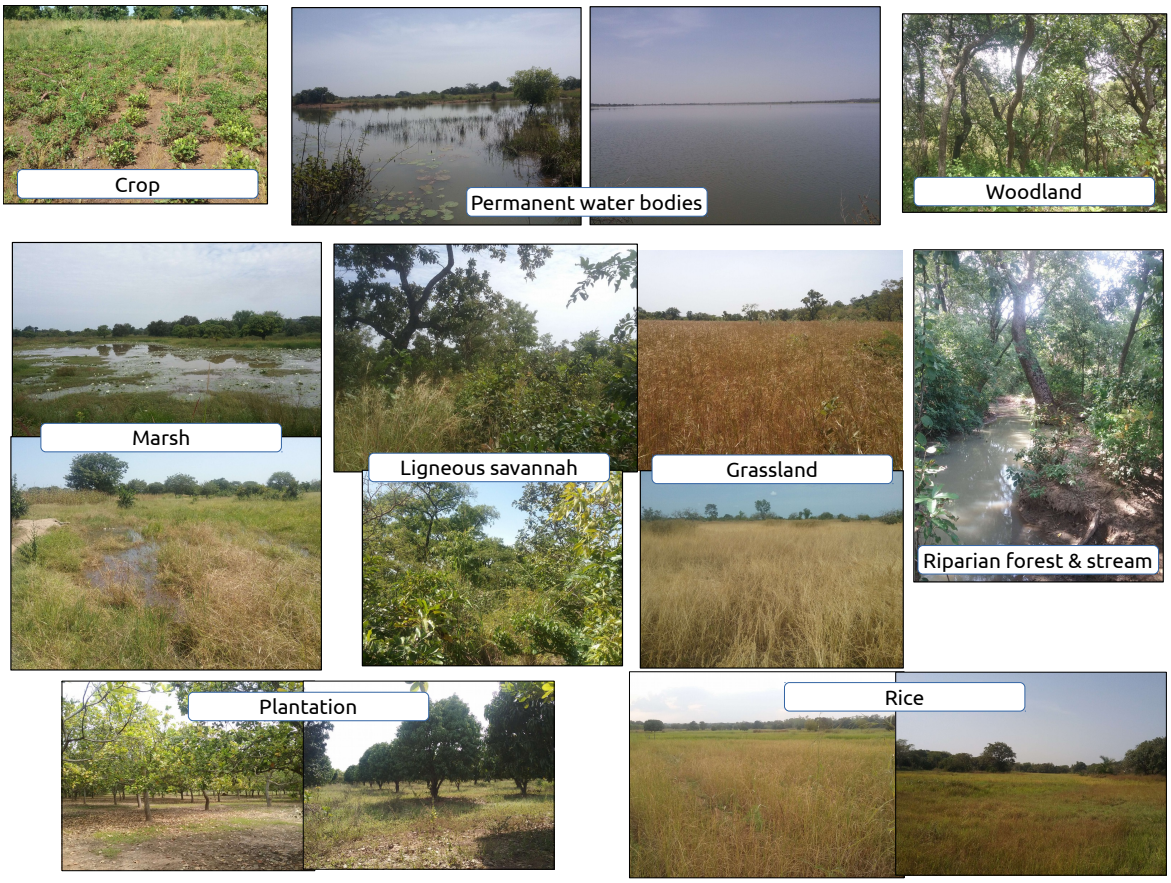


FIGURE B.2: Photographies représentatives des principales classes d'occupation du sol rencontrées sur les zones d'étude

Annexe C

Présentation de la librarie R opendapr

Cette annexe présente succintement la librairie R opendapr, développée dans le cadre de la thèse afin d'extraire les séries temporelles satellitaires météorologiques (MODIS et GPM) sur les zones d'étude. La libraire est hebergée à l'adresse suivante : `https://github.com/ptaconet/opendapr`. *Le texte de cette annexe est extrait de la description de la librairie, disponible à cette adresse.*

opendapr is an R package that provides functions to **harmonize** and **speed-up** the **download** of some well-known and widely-used **spatiotemporal Earth science datacubes** (e.g. MODIS, VIIRS, GPM or SMAP) using the OPeNDAP framework (*Open-source Project for a Network Data Access Protocol*)

Harmonize ? opendapr proposes a single function to query the various data servers, and another single function to download the data.

Speed-up ? opendapr uses the abilities offered by the OPeNDAP to download a subset of data cube, along spatial, temporal or any other data dimension (depth, ...). This way, it reduces downloading time and disk usage to their minimum : no more 1° x 1° MODIS tiles when your region of interest is only 100 km x 100 km wide ! Moreover, opendapr supports parallelized downloads.

Below is a comparison of opendapr with other packages available for downloading chunks of remote sensing data :

Package	Data	Spatial subsetting*	Dimensional subsetting*
opendapr	MODIS, VIIRS, SMAP, GPM	Yes	Yes
MODIS	MODIS	No	No
MODIStsp	MODIS	No	Yes
MODISTools	MODIS	Yes	Yes
smapr	SMAP	No	No

* at the downloading phase

By enabling to download subsets of data cubes, opendapr facilites the access to Earth science data for R users in places where internet connection is slow or expensive and promotes digital sobriety for our research work.

The OPeNDAP, over which the package builds, is a project developed by the non-profit OPeNDAP, Inc. and advanced openly and collaboratively. By using this data access protocol, opendapr support the open-source-software movement.

Installation

The package can be installed with :

```
# install.packages("devtools")
devtools::install_github("ptaconet/opendapr", build_vignettes = T, build_manual = T)
```

Work is ongoing to publish the package on the CRAN.

Collections available in opendapr

Currently **opendapr** supports download of 77 data collections, extracted from the following meta-collections :

— MODIS land products made available by the NASA / USGS LP DAAC (source OPeNDAP server) ;

— VIIRS land products made available by the NASA / USGS LP DAAC (source OPeNDAP server) ;
— VIIRS land products made available by the NASA LAADS DAAC (source OPeNDAP server) ;
— Global Precipitation Measurement (GPM) made available by the NASA / JAXA GES DISC (source OPeNDAP server) ;
— Soil Moisture Active-Passive (SMAP) made available by the NASA NSIDC DAAC (source OPeNDAP server)

Details of each product available for download are provided through the function `odr_list_collections()`.

Get Started

Downloading the data with **opendapr** is a simple two-steps workflow :

— With the function `odr_get_url()`, get the URL(s) of the data for :
  — a collection : see previous section,
  — variables,
  — region of interest,
  — time range,
  — output data format (netcdf, ascii, json)
— Next, with the function `odr_download_data()` : download the data to your computer.

Additional functions include : list collection available for download (`odr_list_collections()`), list variables available for each collection (`odr_list_variables()`), login to EOSDIS Earthdata before querying the servers and downloading the data (`odr_login()`).

Have a look at the `vignette("opendapr1")` to get started with a simple example, and for a more advanced workflow see the `vignette("opendapr2")` !

Next steps

Next developments may involve :

— Short term : including more SMAP collections (at now only SPL3SMP_3.003 collection is available)

— Longer term : including access to more collections and OPeNDAP servers

Any contribution is welcome !

Acknowledgments

We thank NASA and its partners for making all their Earth science data freely available, and implementing open data access protocols such as OPeNDAP. opendapr heavily builds on top of the OPeNDAP, so we thank the non-profit OPeNDAP, Inc. for developing the eponym tool in an open and collaborative way.

We also thank the contributors that have tested the package, reviewed the documentation and brought valuable feedbacks to improve the package : Florian de Boissieu, Julien Taconet, Nicolas Moiroux

The initial development and first release of this package were financed by the MIVEGEC unit of the French Research Institute for Sustainable Development, as part of the REACT project.

Annexe D

Tutoriel d'initiation à la cartographie de l'occupation du sol par télédétéction spatiale sur logiciel libre

Cette annexe présente un tutoriel pour réaliser des produits d'occupation du sol par classification supervisée orientée objet d'images satellitaires en se basant exclusivement sur des logiciels libres et à code source ouvert.

Initiation à la télédétection spatiale sur logiciel libre

Paul Taconet
Institut de Recherche pour le Développement (IRD), UMR MIVEGEC

30/10/2019

Contents

1 Introduction

Ce document est un tutoriel d'initiation à la manipulation d'images satellites d'observation de la Terre. L'objectif est de fournir quelques clés pour appréhender ce type de données sur des logiciels informatiques dédiés libres et à code source ouverts (*free and open source software*), via un cas d'utilisation classique en télédétéction spatiale : la cartographie l'occupation/utilisation du sol.

Nous proposons d'utiliser le logiciel QGIS comme interface principale de visualisation des données et de paramétrisation des algorithmes, tout en "augmentant" ses fonctionnalités de base grâce à sa capacité à intégrer de nombreuses librairies spatiales externes.

Plus spécifiquement, nous utilisons les logiciels et librairies suivantes :

- *QGIS v3.10*
- *SAGA v7.4.0*
- La chaine de traitements *Sen2cor v2.5.5* et l'extension QGIS *Sen2Cor Adapter*

Les images satellites manipulées sont des produits Sentinel-2 au niveau 1C ainsi que le modèle numérique de terrain Shuttle Radar Topography Mission (SRTM). Ces produits sont en libre accès et disponibles sur l'ensemble de la surface terrestre.

Prérequis pour aborder sereinement le document: concepts de bases en SIG et télédétéction, utilisation basique de QGIS (ouvrir et manipuler des données vectorielles et rasterisées sur QGIS)

Mots clés : télédétéction spatiale, Sentinel 2, occupation du sol, classification supervisée orientée pixel, classification supervisée orientée objet,

Note importante n°1 : Avant de commencer le tutoriel, veillez à installer et configurer l'ensemble des logiciels utiles ! Rendez-vous à l'annexe n°1 pour un guide d'installation et de configuration des logiciels.

Note importante n°2 : Les jeux de données mentionnées dans ce tutoriel sont fournies aux personnes formées. Afin de contrôler si les données générées aux différentes étapes du tutoriel sont conformes à ce qui est attendu, l'ensemble des données intermédiaires est également fourni.

2 Présentation du cas d'études

Dans cette étude fictive, nous souhaitons améliorer les connaissances sur les habitats favorables aux moustiques vecteurs du paludisme dans le Sud-Ouest du Burkina Faso. Pour cela, nous avons collecté des moustiques par capture sur sujet humain dans plusieurs villages de la région de Diebougou. Nous allons à présent cartographier l'occupation du sol sur la zone d'étude. En croisant finalement les deux jeux de données (captures de moustiques et occupation du sol), nous pourrons répondre à notre problématique initiale.

Les villages dans lequels nous avons effectué les captures sont fournis en tant que fichier vectoriel sous `vecteur/villages.gpkg`. Nous souhaitons définir la région d'intérêt (en anglais 'Region Of Interest') de notre étude, c'est-à-dire, la surface géographique pour laquelle nous allons faire une carte d'occupation du sol.

Objectifs :

- Définir l'emprise de la région d'intérêt de notre étude (avec `QGIS > Créer une couche`).

Etapes :

- Ouvrez la couche `vecteur/villages.gpkg` sur QGIS.
- Ajoutez une carte en arrière plan pour contextualiser la zone (par exemple si vous disposez d'une connexion internet : une image Google Satellite en utilisant l'extension Quick Map Services de QGIS).
- Créez une nouvelle couche vectorielle polygonale qui englobe l'ensemble des villages. Enregistrez la sous `vecteur/roi.gpkg`. Veillez à bien lui attribuer une projection en UTM (EPSG : 32630).

Avez-vous réussi ?

- Quelle surface (en km^2) la zone d'étude couvre-t-elle ?

3 Recueillir les données

Dans le cadre d'une classification supervisée à des fins de cartographie d'occupation du sol, nous avons besoin de deux types de données :

- *Les vérités terrains.* C'est un échantillon reprénsentatif des classes d'occupation du sol présentes sur notre zone d'étude. Ce sont les données terrain qui vont nous permettre d'entraîner et de valider notre modèle de classification (la classe d'occupation du sol constitue la **variable à expliquer** du modèle de cartographie de l'occupation du sol que l'on va constituer) ;
- *Les produits satellitaires.* Ce sont les images satellites qui nous permettre de cartographier l'occupation du sol à grande échelle. Ces données vont représenter ce qu'on appelle les **variables explicatives** du modèle.

3.1 Recueillir les données terrain

Une classification supervisée requiert, par définition, un jeu de données d'entraînement et de validation du modèle. Dans le cas d'une classification de l'occupation du sol, il s'agit de collecter un échantillon représentatif des parcelles d'occupation du sol de notre zone d'étude.

L'objectif de cette étape est donc d'acquérir une couche SIG vectorielle de polygones représentant des parcelles dont on est certain de la classe d'occupation du sol, en procédant en deux temps :

1. Définir les classes d'occupation du sol présentes sur notre territoire d'étude, en consultant la littérature par exemple. Les classes définies doivent être exhaustives (c'est-à-dire couvrir l'ensemble des classes présentes sur notre zone d'étude), pertinentes et adaptées à la problématique (par exemple, dans le cas d'une étude sur la modélisation de la présence de moustiques, établir une classe 'zones humides' sera important).
2. Collecter les vérités terrain, sur le terrain directement et/ou via d'autres méthodes pertinentes (par exemple photo-interprétation). Il faut avoir un nombre de parcelles minimum pour chacune des classes définies. La classification sera d'autant plus performante que le nombre de parcelles du jeu de données de vérité terrain est important. Le nombre minimum de parcelles à acquérir pour chacune des classes fait l'objet de débats dans la littérature, mais retenez qu'un minimum de 20 parcelles pour chacune des classes d'occupation du sol définie est nécéssaire.

Dans ce tutoriel, nous utiliserons un jeu de données terrain qui a été aquis en novembre 2018 dans notre région d'intérêt. Le jeu de données est stocké sous `vecteur/ground_truth.shp`. Nous avons établi 5 classes d'occupation du sol : eau permanente, milieux dégradés, zones humides, cultures et jachères, milieux naturels.

Etapes :

- Ouvrez le jeu de données `vecteur/ground_truth.shp` et explorez son contenu (colonnes, nombre de parcelles pour chacune des classes, etc.).
- Ajoutez une symbologie catégorielle au jeu de données, afin de colorier les parcelles en fonction de la classe d'occupation du sol qu'ils représentent.

Avez-vous réussi ?

- Combien y a-t-il d'éléments (c'est-à-dire de polygones) dans la classe "eau permanente" ?

3.2 Identifier les produits satellitaires

Les produits satellitaires vont nous permettre de cartographier l'occupation du sol sur l'ensemble de notre zone d'étude. Dans notre cas, nous allons utiliser deux types de produits : des images satellites optiques et un modèle numérique de terrain (MNT).

Il existe de nombreuses sources d'images satellites optiques. Les images se distinguent par leurs résolutions spatiale (de quelques centimetres à plusieurs kilomètres) et spectrale, leur étendue (quelques dizaines de kilomètre carrés à plusieurs dizaines de milliers de km^2), leur date d'aquisition, leur coût (gratuit à plusieurs milliers d'euros), etc. C'est la nature de notre projet et nos contraintes qui guident le choix des images à utiliser. L'Annexe 2 présente une liste (non-exhautive) de site internet pour pré-visualiser et récupérer des produits satellitaires.

Dans notre cas, nous utiliserons des données issues du satellite Sentinel-2. Sentinel-2 fait partie de la constellation de satellites du programme Sentinel de l'Agence Spatiale Européenne. Le capteur dont il est équipé (MultiSpectral imager) capture des images dans 13 bandes spectrales dans les domaines du visible et de l'infrarouge, allant de 10 à 60 mètres de résolution spatiale. La période de revisite du satellite est de 5 jours. Les données sont libres d'accès, comme toutes les données du programme Sentinel. Il existe de multiples manières de récupérer des images Sentinel 2. Une des plus simples est de passer par le **Copernicus Open Access Hub**, un portail dedié avec une interface utilisateur permettant de rechercher et télécharger des images du programme Sentinel.

276

Le modèle numérique de terrain, de son côté, nous servira à affiner la classification. Le MNT founit l'altitude en tout point. Grâce à cette information, nous pouvons extraire nombre d'informations supplémentaires (pente, orientiation, etc.). Il existe deux MNT globaux gratuits : le Shuttle Radar Topography Mission (SRTM) et le Global Digital Elevation Model (GDEM). Ces deux MNT ont une résolution spatiale de 30 mètres, et se valent en terme de qualité et précision. Dans notre cas nous utiliserons le SRTM.

Objectifs :

- Rechercher, identifier et télécharger des images satellites Sentinel-2 (avec le site `Copernicus Open Access Hub`)
- Identifier et télécharger un modèle numérique de terrain (avec le site `30-Meter SRTM Tile Downloader`)

Etapes :

1. Télécharger les images Sentinel 2

- Créez un compte auparavent sur le site des produits d'Observation de la Terre de l'ESA.
- Allez sur le Copernicus Open Access Hub: https://scihub.copernicus.eu/dhus/ puis identifiez vous via le bouton *login*.
- Affinez la recherche avec les éléments suivants:
  - **Sensing period** doit correspondre à une fenêtre de dates autour de notre date de campagne de relevé des vérités terrain (mettre tout le mois de novembre 2018) ;
  - **Mission** est Sentinel-2 ;
  - Tracez une région autour de notre zone d'étude ;
  - Cliquez finalement sur la loupe pour rechercher.

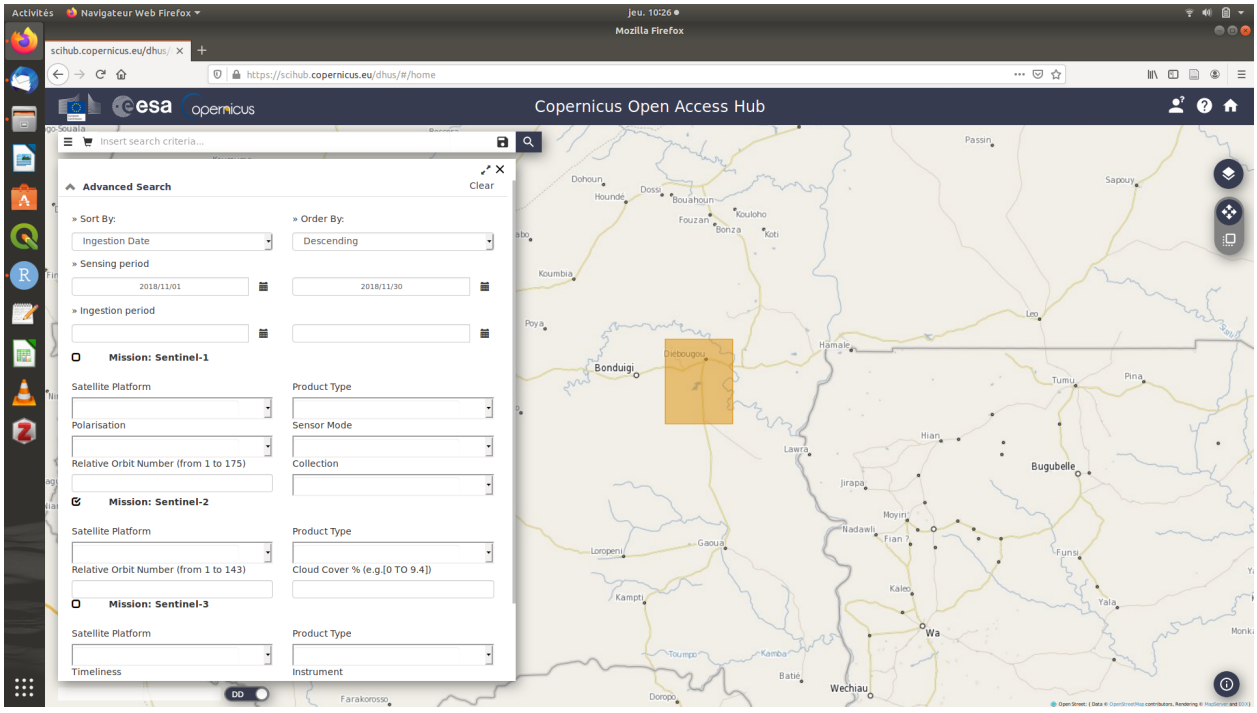


Figure 1: Copernicus Open Access Hub - recherche de produits satellites

- Les produits répondant à la recherche apparaissent alors. Dans le panneau de gauche, l'ensemble des images correspondant à notre recherche sont affichées. Dans notre cas, nous observons que notre zone d'études se trouve à la frontière entre deux images, il va donc falloir télécharger deux images pour couvrir entièrement notre zone. Observez le panneau de gauche avec les images disponibles. Pour

277

chacune, l'information 'Sensing date' donne la date à laquelle elle a été photographiée. Nous avons 6 couples d'images sur l'ensemble du mois de novembre. Le choix de l'image que l'on utilisera se fera généralement selon les critères suivants (par ordre d'importance):

- Couverture nuageuse : sachant que dans le domaine de l'imagerie optique on perd l'information de tout pixel sous un nuage, on cherchera à conserver l'image qui a la couverture nuageuse la moins importante. Ainsi par exemple, on excluera l'image du 2018-11-21 qui est quasiment entièrement sous les nuages.
- Niveau de traitement de l'image : les images Sentinel 2 sont distribuées principalement selon deux niveaux de traitements : niveau 1C ou niveau 2A. Au niveau 2A, les corrections géométriques et atmosphériques sont effectuées. Il est donc préférable d'utiliser une image de niveau 2A si disponible. Dans notre cas, il nous est proposé uniquement des images de niveau 1C, nous allons donc effectuer les corrections nous mêmes (voir section Préparer les images satellites optiques)
- Proximité temporelle à notre campagne de relevé des vérités terrain.

Téléchargez les images *S2B_MSIL1C_20181116T103309_N0207_R108_T30PVT_20181116T160025* et *S2B_MSIL1C_20181116T103309_N0207_R108_T30PVS_20181116T160025* puis décompressez les dans un dossier de votre choix. Note : ne modifiez ni le nom ni le contenu des dossier décompréssés !

2. Télécharger le MNT SRTM

- Créez un compte sur le site des produits d'Observation de la Terre de la NASA.
- Allez sur le site *30-Meter SRTM Tile Downloader* : https://dwtkns.com/srtm30m/ .
- Télécharger le MNT en cliquant sur la tuile recouvrant notre zone d'étude puis décompressez le dossier le fichier dans un dossier de votre choix.

Avez-vous réussi ? :

- Vous devriez avoir les dossiers dans votre ordinateur tels qu'indiqués dans l'image ci-dessous :

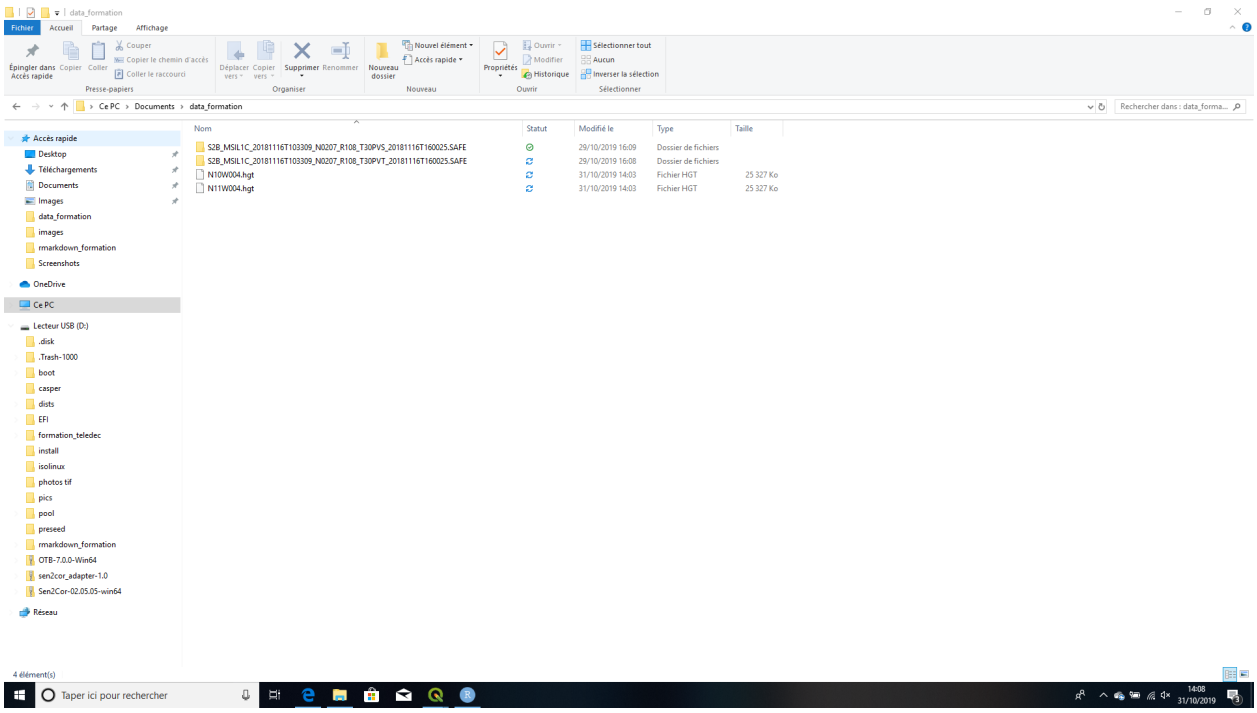


Figure 2: Dossiers contenant les produits satellites téléchargés

- Ouvrez le dossier `S2B_MSIL1C_20181116T103309_N0207_R108_T30PVT_20181116T160025.SAFE/GRANULE/L1C_T30PVT_`

278

Ce dossier contient les 13 bandes spectrales de l'image (B01, B02, etc.) + la bande TCI qui donne l'image en composition de couleurs réelle. Ouvrez les 14 bandes sur QGIS.
- Naviguez à travers les images ouvertes sur QGIS. En particulier, observez l'image TCI.
- Chargez le MNT sur QGIS, c'est-à-dire les fichiers `N10W004.hgt` et `N11W004.hgt` du dossier dans lequel vous avez stocké le MNT.
- Quelle est le système de projection de l'image satellite Sentinel-2? et celui du MNT ?
- Pouvez-vous dire à quelle altitude se trouve la ville de Diébougou ? Le barage au sud de Diébougou ?

4 Préparer les données pour la classification

Toutes les données qui serviront à établir notre carte d'occupation du sol sont maintenant dans nos mains (ou plutôt notre ordinateur) ! Mais avant de se lancer dans la classification, il faut préparer les données afin de s'assurer que notre classification sera pertinente et performante. C'est l'objectif de cette section.

4.1 Préparer les images satellites optiques

La préparation des images satellites comprend les points suivants :

- Les corrections atmosphériques et géométriques : un ensemble de traitements qui garantissent que l'image est correctement géoréférencée et que les valeurs des pixels sont des valeurs de réfléctance en Bottom of Atmosphere ;
- Le mosaïquage des images, pour générer une image unique dans le cas où notre zone d'études est couverte par plusieurs images ;
- Le découpage des images sur l'étendue de notre zone d'étude

Dans notre cas, nous avons récupéré des images avec les propriétés suivantes (cf. section Recueillir les produits satellitaires) :

- de niveau 1C, donc, pour lesquelles les corrections l'ensemble des corrections atmosphériques et géométriques n'a pas été effectuées (rappel : ce sont les images de niveau 2A qui sont corrigées au niveau Bottom of Atmosphere) ;
- nous avons récupéré deux images pour recouvrir l'ensemble de notre zone d'étude ;
- la zone totale recouverte par l'ensemble des deux images est plus vaste que notre zone d'études

Nous allons donc devoir effectuer les trois étapes de pré-traitements sus-mentionnées.

Objectifs :

- Executer la chaine de traitement Sen2Cor pour effectuer les corrections géométriques et atmosphériques des images Sentinel-2 (passer du niveau 1C au niveau 2A) (avec l'outil `QGIS > Sen2Cor Adapter`)
- Mosaïquer des images satellites (avec l'outil `QGIS > Raster > Fusionner`)
- Découper des images satellites selon une emprise (avec l'outil `QGIS > Raster > Découper un raster selon une couche de masque`)

Etapes :

1. Corrections atmosphériques et géométriques

Les corrections atmosphériques et géométriques sont effectuées par la chaîne de traitement Sen2Cor. Cette chaîne permet de passer du niveau 1C au niveau 2A pour les images Sentinel-2. Elle effectue automatiquement les corrections atmosphériques et géométriques des produits Sentinel 2 de niveau 1C. Rendez-vous sur le site de l'ESA pour plus d'informations sur Sen2Cor : http://step.esa.int/main/third-party-plugins-2/sen2cor/

- Dans QGIS, cliquez sur `Raster > Sen2Cor Adapter` et ouvrer Sen2Cor Adapter ;
- Remplissez les cases suivantes :
  - `SEN2COR tool path` : lien vers le dossier dans votre ordinateur contenant l'application Sen2Cor (voir la section Télécharger et installer les logiciels) ;

  - `Input (.SAFE folder)` : lien vers le dossier `S2B_MSIL1C_20181116T103309_N0207_R108_T30PVS_20181116T1600` comprenant l'image Sentinel-2 au niveau 1C ;
  - `Output (optional)` : lien vers le dossier dans lequel seront stockées les images au niveau 2A qui vont être générées par Sen2Cor ;
  - `Resolution` : séléctionnez ALL ;
- Laissez les autres paramètres avec leurs valeurs par défaut et cliquez sur `RUN`.

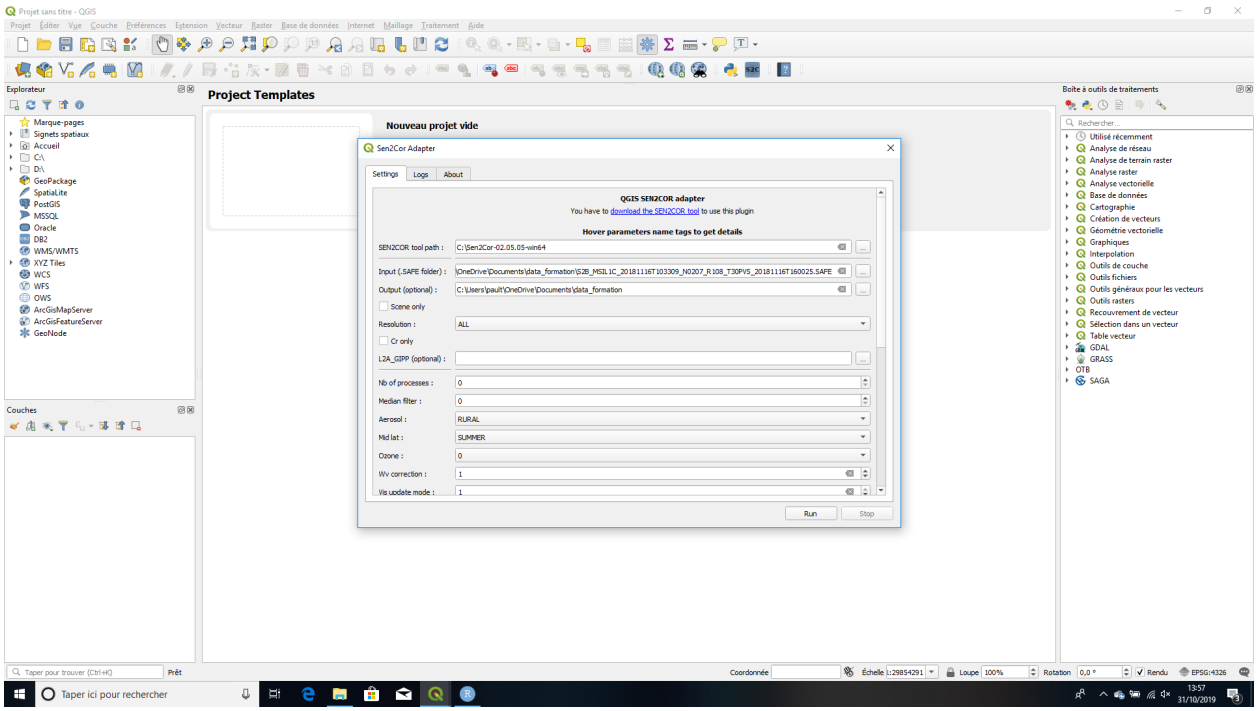


Figure 3: Sen2Cor sous QGIS

La chaine de traitements Sen2Cor est alors lancée. Cela dure une quinzaine de minutes environ, en fonction de la puissance de calcul de votre ordinateur. En fin de traitement, les images corrigées au niveau 2A seront stockées dans le dossier `Output (optional)` que vous avez défini.

- Répétez l'opération avec l'autre image (`S2B_MSIL1C_20181116T103309_N0207_R108_T30PVT_20181116T160025.SAFE`)

Note : Nous avons présenté ici les étapes de corrections atmosphériques et radiométriques pour les images Sentinel-2. Notez que, si conceptuellement les corrections sont identiques pour toutes les images satellites, les outils que l'on utilisera dépendront la source, du capteur, etc. Ainsi dans le cas de Sentinel-2, le travail est facilité par la chaine de traitement Sen2Cor. Pour d'autres images, il faudra procéder différemment. Par exemple, si l'on désire corriger des images SPOT6/7, Orfeo Toolbox peut faire le travail avec les algorithmes `OpticalCalibration` et `OrthoRectification`.

2. Mosaïquage des images

Vous avez à présent à votre disposition des images satellites de niveau 2A, c'est-à-dire, la valeur des pixels est une réflectance réelle (Bottom-of-Atmosphere) et l'image est orthorectifiée. Cependant, comme indiqué à la section Recueillir les produits satellitaires, notre zone d'étude est couverte par 2 images satellites, à savoir, `S2B_MSIL1C_20181116T103309_N0207_R108_T30PVS_20181116T160025` et `S2B_MSIL1C_20181116T103309_N0207_R108_T30PVT_20181116T160025`. Il s'agit donc de mosaiquer ces images, c'est-à-dire, de les assembler pour n'en faire qu'une.

- Dans QGIS, ouvrez les bandes spectrales B02, B03, B04, B08 à 10 mètres de résolution et les bandes B05,

280

B06, B07, B08A, B11 et B12 à 20m de résolution de l'image n°1 (`S2B_MSIL2A_20181116T103309_N0207_R108_T30PVS_20`
et de l'image n°2 (`S2B_MSIL2A_20181116T103309_N0207_R108_T30PVT_20181116T160025.SAFE\GRANULE\L2A_T30PVS`
;

- Ouvrez le menu QGIS **`Raster > Divers > Fusionner`**. Ce menu permet de fusionner (autrement dit mosaiquer) deux ou plusieurs images raster.
- Dans la case **`Couches en entrée`**, cliquez sur les 3 petits points. Une fenêtre s'ouvre alors avec les 20 couches raster que vous avez ouvertes dans QGIS. Vous désirez mosaiquer les bandes 2 à 2 : la bande B02 de l'image n°1 avec la bande B02 de l'image n°2, etc. Ainsi, cochez les cases `T30PVS_20181116T103309_B02_10m` et `T30PVT_20181116T103309_B02_10m` (comme indiqué sur l'image ci-dessous) puis cliquez sur **`OK`**

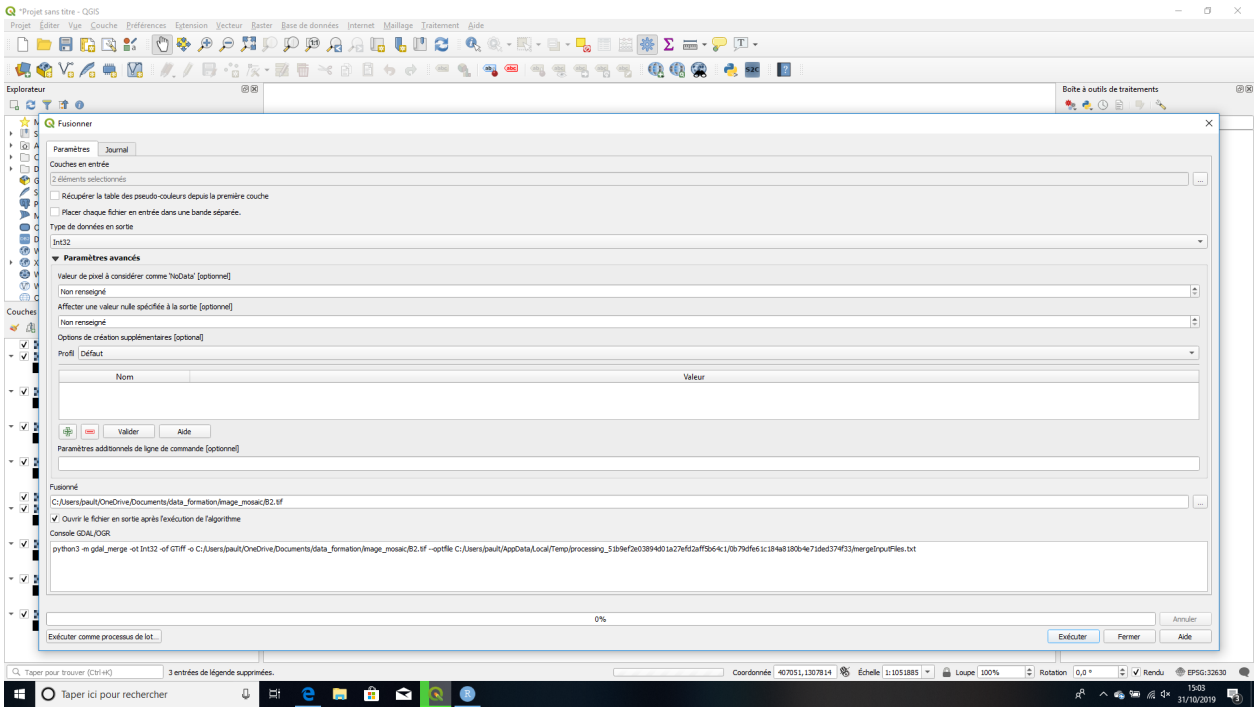


Figure 4: QGIS - menu Raster > Fusionner

- Dans la case **`Type de données en sortie`**, séléctionnez **`Int32`** (l'image en sortie sera ainsi moins volumineuse)
- Dans la case **`Fusionner`**, cliquez sur **`enregistrer vers un fichier`** puis séléctionnez un dossier dans lequel sera généré votre image mosaiquée (par exemple, **`image_mosaic`**) et appelez l'image en sortie **`B02.TIF`**
- Cliquez sur **`Executer`** en bas à droite de la fenêtre pour lancer le mosaiquage.
- Vous avez ainsi fusionné les bandes B02 des deux images satellites. Répétez alors l'opération avec les bandes B03, B04, B05, B06, B07, B08, B08A, B11, B12.

3. Découpage selon l'emprise de la zone d'étude

Les images à présent fusionnées couvrent une surface plus large que notre zone d'études. Or plus la surface couverte par les images est importante, plus les calculs à venir sur ces images seront longs. Nous allons donc découper notre image selon l'emprise de la zone d'étude, afin de conserver uniquement la zone qui nous intéresse.

- Dans QGIS, ouvez les 10 bandes B02, B03, B04, etc. générées à l'étape précédente (mosaiquage)
- Ouvez également la couche **`roi.gpkg`** générée à l'étape de définition de l'emprise de la région d'intérêt

281

de notre étude.

- Vérifiez que les couches se superposent bien et qu'elles sont dans le même système de projection (EPSG: 32630 dans notre cas), comme indiqué dans l'image ci-dessous.

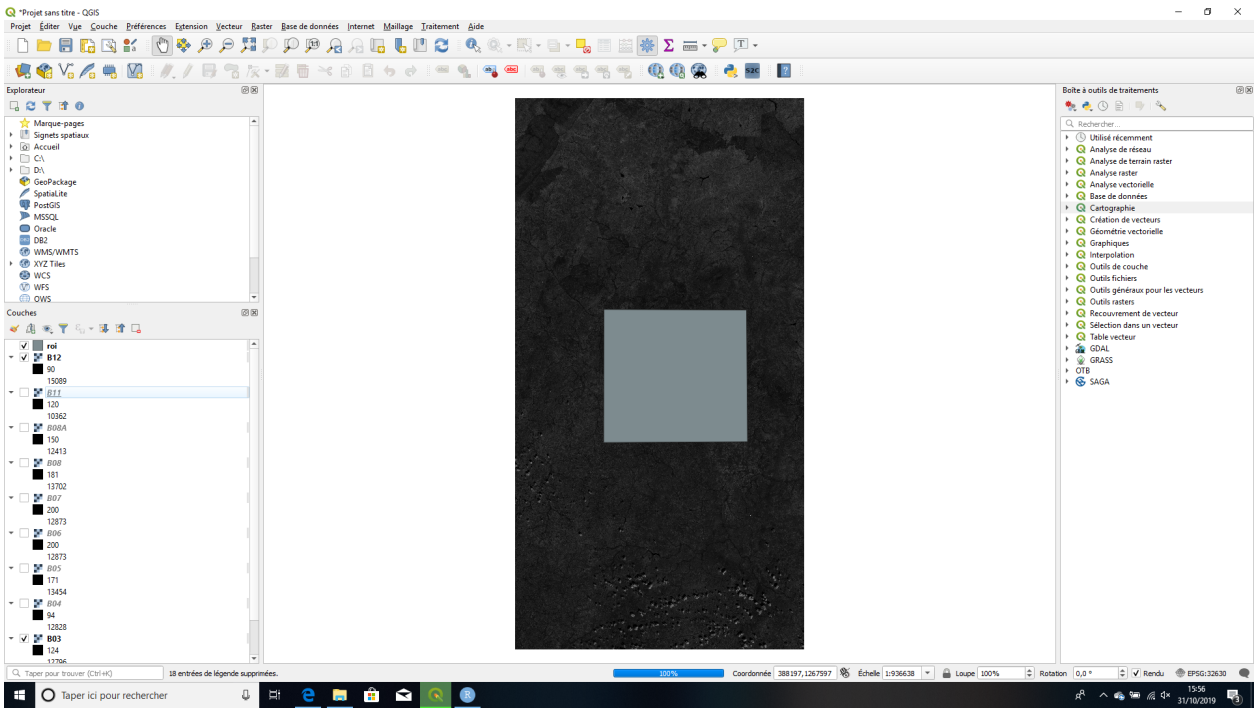


Figure 5: Images satellites et région d'intérêt

- Aller dans le menu `Raster > Extension > Découper un raster selon une couche de masque`
- Dans le menu déroulant `Couche source`, cliquez la bande `B02`
- Dans le menu déroulant `Couche de masquage`, cliquez la couche `roi` (couche vectorielle de l'emprise de la zone d'étude)
- Tout en bas, dans le menu `Découpé (masque)`, choisissez `Enregistrer vers un fichier` puis sélectionnez un dossier dans lequel sera généré votre image mosaiquée (par exemple, `image_mosaic_roi`) et appelez l'image en sortie `B02.TIF`
- Cliquez sur `Executer` en bas à droite de la fenêtre pour lancer le découpage.
- Vous avez ainsi découpé la bande B02. Répétez alors l'opération avec les bandes B03, B04, B05, B06, B07, B08, B08A, B11, B12.

Avez-vous réussi ? :

- Dans QGIS, chargez les 10 bandes mosaiquées et découpées selon l'emprise de la zone d'étude, la couche de l'emprise de la zone d'étude (`vecteur/roi.gpkg`) et la couche de données de vérités terrain (`vecteur/ground_truth.shp`)
- La couche de données des vérités terrain couvre-t-elle à peu près homogènement toute la zone d'étude ? Si ce n'est pas le cas, quelles peuvent-en être les raisons à votre avis ?

4.2 Préparer le modèle numérique de terrain

Les images satellite Sentinel-2 sont maintenant prêtes. Il s'agit à présent de faire les mêmes pré-traitements pour le MNT : mosaiquage et découpage selon l'emprise de la zone d'étude. Cependant, notez que les images Sentinel-2 et le MNT ne sont pas projetés dans le même système de projection : les images sont en

WGS84/UTM zone 30N (EPSG 32630) alors que le MNT est en WGS84 non projeté (EPSG 4326). Nous allons donc, avant de mosaiquer et découper le MNT, le reprojeter en WGS84/UTM zone 30N.

Objectifs :

- Reprojeter le MNT (avec l'outil `QGIS > Raster > Projection`)
- Mosaïquer le MNT (avec l'outil `QGIS > Raster > Fusionner`)
- Découper le MNT selon une emprise (avec l'outil `QGIS > Raster > Découper un raster selon une couche de masque`)

Etapes :

1. Reprojeter le MNT

- Charger dans QGIS les 2 fichiers de MNT téléchargés à l'étape de Téléchargement des produits satellites(`N10W004.hgt` et `N11W004.hgt`)
- Dans le menu, aller dans `Raster > Projection > Projection (warp)`
- Dans la liste déroulante `Couche en entrée`, choisissez la couche `N10W004`
- Dans `SCR cible`, choisissez le SCR des images satellites, c'est-à-dire EPSG 32630 : WGS84/UTM zone 30N
- Tout en bas, dans le menu `Reprojeté`, choisissez `Enregistrer vers un fichier` puis séléctionnez un dossier dans lequel sera généré votre MNT reprojeté (par exemple, `mnt`) et appelez l'image en sortie `N10W004.TIF`
- Cliquez sur `Executer` en bas à droite de la fenêtre pour lancer la reprojection
- Vous avez ainsi reprojeté la tuile N10W004 du MNT. Répétez alors l'opération avec la tuile N11W004.

2. Mosaïquer le MNT

Suivre la suite d'opération expliquée dans la partie précédente, section Mosaiquage

3. Découper le MNT

Suivre la suite d'opération expliquée dans la partie précédente, section Découpage selon l'emprise de la zone d'étude

Avez-vous réussi ? :

- Vérifiez que votre MNT final est bien projeté en EPSG 32630 : WGS84/UTM zone 30N et qu'il recouvre la zone d'étude.

5 Calculer des indices spectraux et topographiques

Nos images sont maintenant prêtes, et l'on pourrait à présent se lancer dans la classification. Cependant, nous allons augmenter nos chances d'obtenir une bonne classification en générant un certain nombre de couches dérivées de l'image Sentinel-2 et du MNT, qui seront ensuite intégrées dans la classification. Nous allons calculer des indices spectraux à partir de l'image satellite et des indices topographiques à partir du MNT.

5.1 Calculer des indices spectraux avec les images satellite

Extrait de : https://e-cours.univ-paris1.fr/modules/uved/envcal/html/vegetation/indices/index.html et http://agritrop.cirad.fr/585651/

En télédétection, les indices font parties des méthodes de traitement que l'on appelle les transformations multispectrales. Ils consistent à convertir les luminances mesurées au niveau du capteur satellitaire en grandeurs ayant une signification dans le domaine de l'environnement.

Basés sur le caractère multispectral des données satellitaires, ils permettent de décrire l'état d'un phénomène. Un indice de végétation par exemple, peut rendre compte du stade de croissance végétale à un moment donné.

Tous les indices, que ce soient les indices de végétation, les indices des sols, les indices relatifs à la colonne d'eau, etc., reposent sur une approche empirique basée sur des données expérimentales. Les indices de végétation sont très utilisés d'une part, pour identifier et suivre la dynamique de la végétation, mais aussi pour estimer certains paramètres biophysiques caractéristiques des couverts végétaux, comme la biomasse, l'indice de surface foliaire, la fraction de rayonnement photosynthétique actif, etc.

Les indices spectraux sont obtenus à partir d'équations appliquées à la valeur des pixels dans bandes différentes, dans le but de tirer profit des particularités du comportement radiométrique de différents types d'objets. Par exemple le NDVI (indice normalisé de végétation) utilise la haute réflectance de la végétation dans le proche infrarouge et sa basse réflectance dans le rouge ; plus dense et vigoureuse est la végétation, plus cette tendance s'accentue.

Il existe de très nombreux indices spectraux, qui ont été développés au cours du temps par les scientifiques et utilisateurs des images satellites optiques. Vous pouvez en trouver une liste assez complète sur ce site.

Le choix des indices à générer dans le cadre d'une classification d'occupation du sol dépend des classes que l'on souhaite discriminer. Ainsi, on peut s'attendre par exemple à ce que la classe 'eau' soit particulièrement bien discriminée par un indice de présence d'eau. Dans notre cas, nous allons intégrer 6 indices, dont les noms et equations sont donnés dans le tableau suivant :

Indice	Type	Equation
NDVI (Normalized Difference Vegetation Index)	Vegetation	$\frac{NIR-R}{NIR+R}$
BRI (Brillance du sol)	Sol	$\sqrt{\frac{R^2}{2\times G^2}}$
NDWI (Normalized Difference Water Indice)	Eau	$\frac{G-NIR}{G+NIR}$
MNDWI (Modified Normalized Difference Water Indice)	Eau	$\frac{NIR-SWIR}{NIR+SWIR}$
NDBI (Normalized Difference Built-up Index)	Bâti	$\frac{SWIR-NIR}{SWIR+NIR}$

Pour rappel, les bandes spectrales du satellite Sentinel-2 sont les suivantes :

Sentinel-2 Bands	Central Wavelength (µm)	Resolution (m)
Band 1 - Coastal aerosol	0.443	60
Band 2 - Blue	0.490	10
Band 3 - Green	0.560	10
Band 4 - Red	0.665	10
Band 5 - Vegetation Red Edge	0.705	20
Band 6 - Vegetation Red Edge	0.740	20
Band 7 - Vegetation Red Edge	0.783	20
Band 8 - NIR	0.842	10
Band 8A - Vegetation Red Edge	0.865	20
Band 9 - Water vapour	0.945	60
Band 10 - SWIR - Cirrus	1.375	60
Band 11 - SWIR	1.610	20
Band 12 - SWIR	2.190	20

Figure 6: Caractéristiques des bandes spectrales de Sentinel-2

Objectifs :

- Calculer des indices spectraux (avec l'outil `QGIS > Raster > Calculatrice Raster`)

Etapes :

- Chargez dans QGIS les 10 bandes spectrales générées à l'étape précédente (images mosaiquées et découpées selon l'emprise de la zone d'étude) ;
- Dans le menu, allez dans `Raster > Calculatrice Raster`
- Utilisez les tableaux ci-dessus (formules des indices spectraux et tableau des bandes spectrale de Sentinel-2) pour calculer chaque indice spectral du tableau à l'aide de la calculatrice raster, comme indiqué sur l'image ci-dessous
- Enregistrez chaque indice dans un dossier de votre choix (par exemple, `indices_spectraux`) et donnez au fichier raster généré le nom de l'indice en question

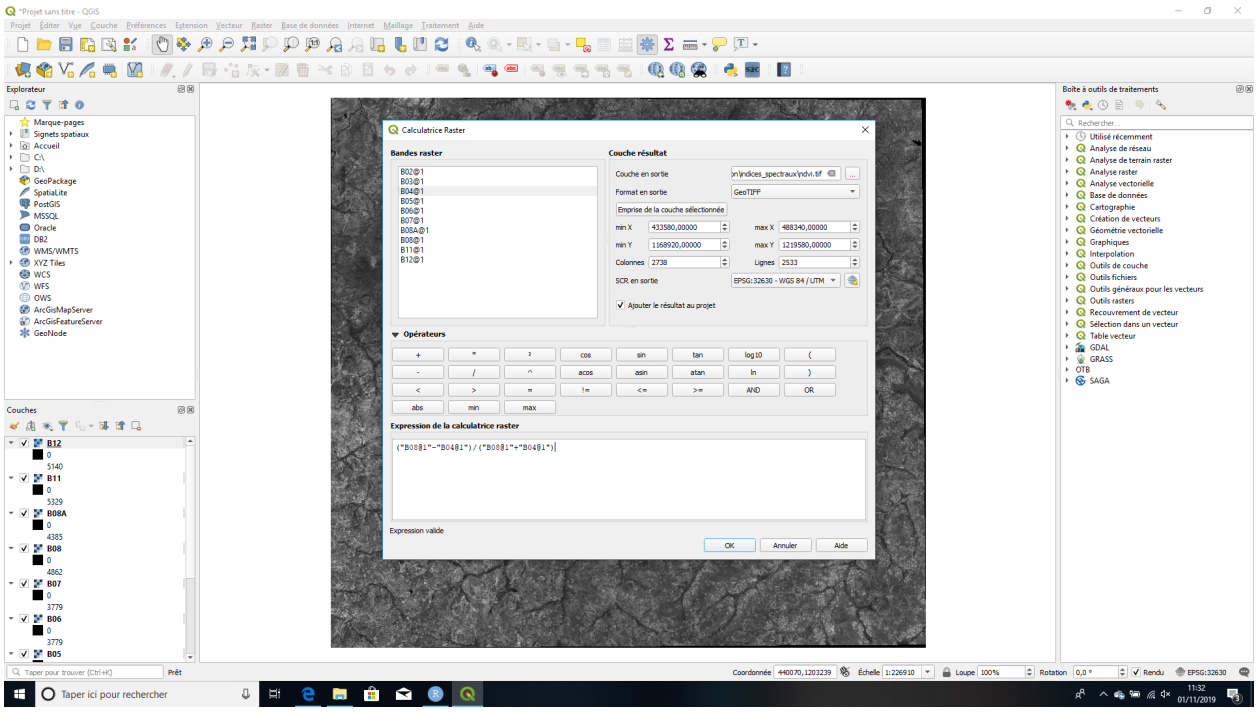


Figure 7: Calculatrice raster - calcul d'indices spectraux

www.gisresources.com/ndvi-ndbi-ndwi-ranges-1-1/

Avez-vous réussi ? :

- Quelle est la valeur du NDVI au niveau des bas-fonds ? du barrage ?
- De même pour le NDWI.

5.2 Calculer des indices topographiques avec le MNT

Nous allons utiliser le MNT pour dériver des indices topographiques. Pour rappel, le MNT donne la valeur de l'altitude en tout point de l'espace (pour le MNT SRTM que l'on utilise : sur une grille de 30m x 30m de résolution). A partir de la valeur de l'altitude en un pixel donné et des valeurs dans les pixels adjacents, nous pouvons calculer de nombreux indices de terrain : pente, orientiation, accumulation de flux, etc. Cela sera utile pour discriminer certaines classes : par exemple, dans notre région les zones humides se trouvent dans les zones où l'altitude et la pente sont faibles, et où l'accumulation de flux est importante.

Objectifs :

- Calculer des indices topographiques de pente et d'aspect (avec l'outil `Grass > r.slope.aspect`)
- Calculer des indices topographiques d'accumulation de flux (avec l'outil `Grass > r.terraflow`)

Note importante :

Veillez à utiliser la version de QGIS `QGIS Desktop 3.10.0 with GRASS 7.6.1` (et non `QGIS Desktop 3.10.0`) pour cette étape.

Etapes :

1. Calcul de la pente et de l'orientation

- Charger dans QGIS la couche `mnt_mosaic_roi.tif` générée dans la section Préparer le Modèle numérique de terrain
- Dans la boite à outils de traitements, recherchez l'algorithme `r.slope.aspect` de GRASS et l'ouvrir
- Dans la case `Elevation`, séléctionnez le MNT
- Dans les cases `Pente` et `Exposition`, saisissez respectivement les chemins de sortie vers les fichiers de pente, orientation et courbature qui vont être générés par l'algorithme, comme indiqué sur la figure ci-dessous.
- Pour tous les autres paramètres (`Profile curvature`, `Courbe tangentielle`, etc.), séléctionnez `Ignorer la sortie`
- Cliquez sur `Executer` pour lancer l'algorithme

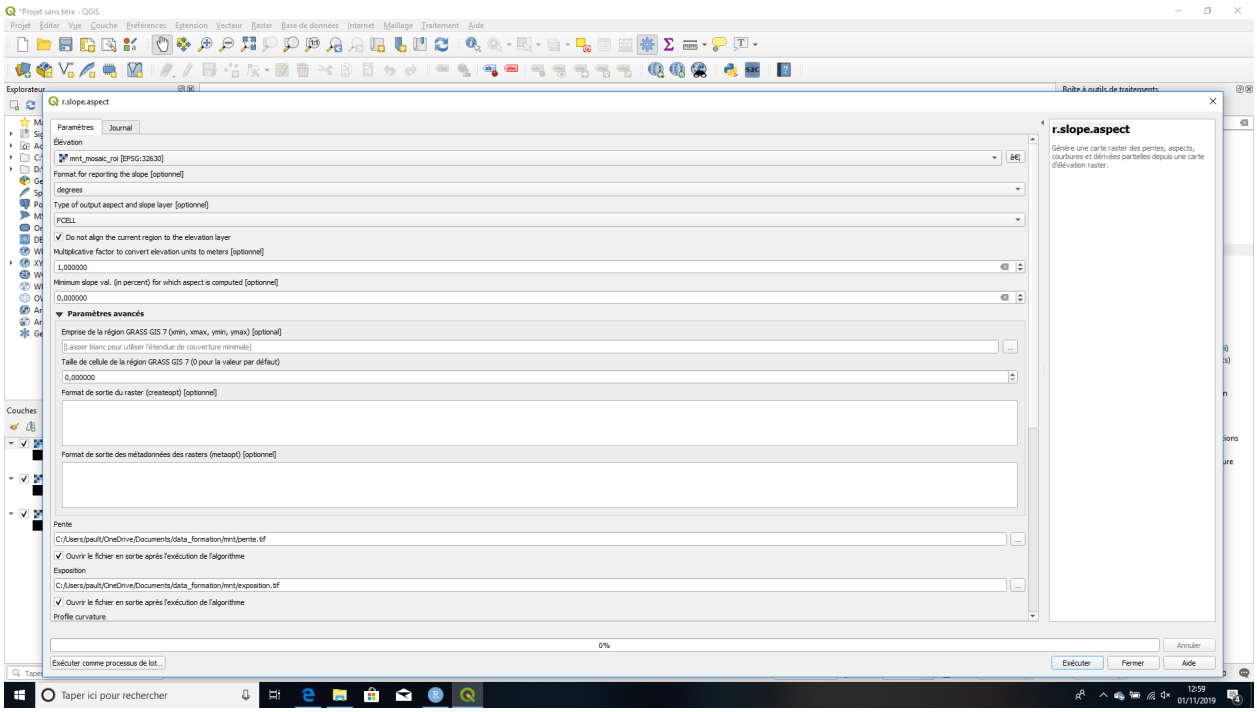


Figure 8: Algorithme r.slope.aspect de GRASS

2. Calcul de l'accumulation de flux

- Dans la boite à outils de traitements, recherchez l'algorithme `r.terraflow` de GRASS et l'ouvrir
- Dans la case `Nom de la carte d'élévation raster`, séléctionnez le MNT
- Dans la case `Flow accumulation`, saisissez le chemins de sortie vers le fichier d'accumulation, comme indiqué sur la figure ci-dessous
- Cliquez sur `Executer` pour lancer l'algorithme

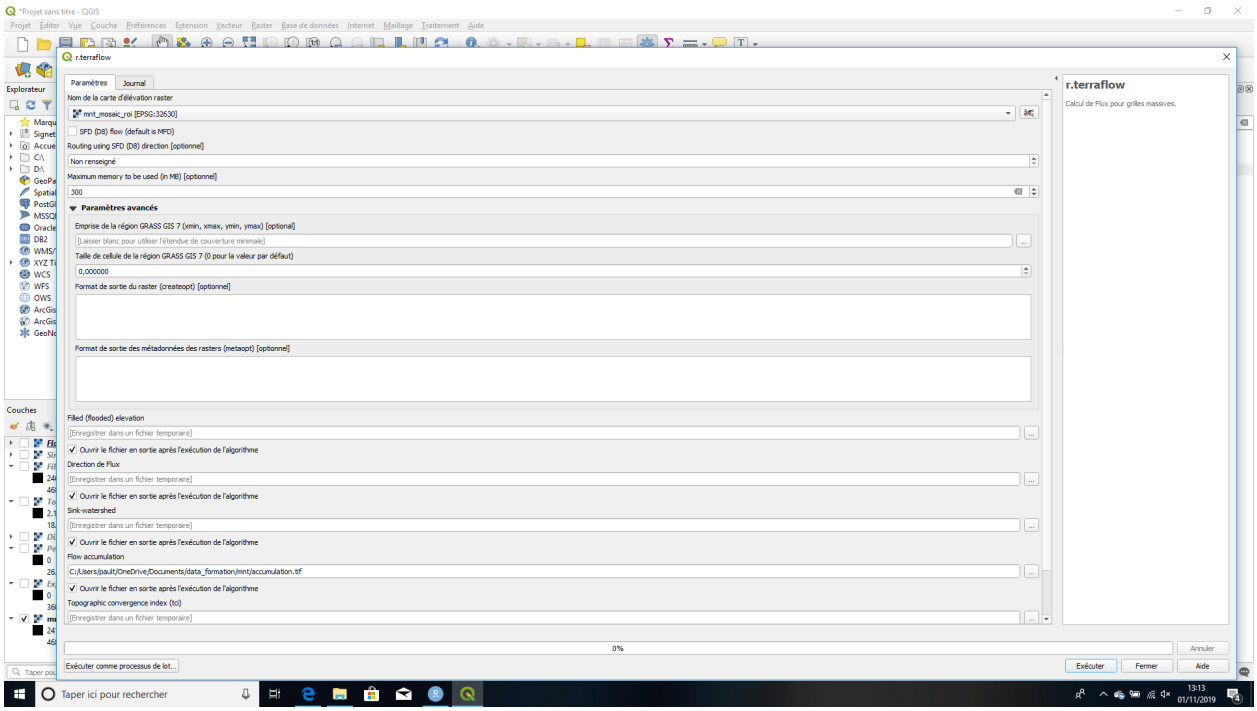


Figure 9: Calculatrice raster - calcul d'indices spectraux

Avez-vous réussi ? :

- Quelle est la valeur de la pente des bas-fonds ? du barrage ?
- Même question pour l'accumulation de flux

6 Classer les images pour cartographier l'occupation du sol

Nos données sont maintenant prêtes. Nous allons utiliser un algorithme de classification nommé "Random Forest" pour générer notre carte d'occupation du sol en réalisant une classification supervisée par pixel.

Le court texte qui suit est extrait et adapté de http://perso.ens-lyon.fr/lise.vaudor/classification-par-forets-aleatoires/ et explique succintement le fonctionnement des Random Forest :

Random forest est un algorithme d'apprentissage supervisé basé sur la génération d'arbres décisionnels. Random forest génère un nombre n d'arbres décisionnels à partir des données d'apprentissage. Ces arbres se distinguent les uns des autres par le sous-échantillon de données sur lequel ils sont entraînés. Ces sous-échantillons sont tirés au hasard dans le jeu de données initial.

Chaque arbre de la forêt est construit sur une fraction aléatoire ("in bag") des données (c'est la fraction qui sert à l'entraînement de l'algorithme), [et utilise une fraction aléatoire des descripteurs pour réaliser la segmentation des arbres]. *Pour chacun des individus de la fraction de données restante ("out of bag") l'arbre peut prédire une classe.*

Nous utiliserons l'algorithme *Random Forest Classification* de la librairie SAGA pour réaliser et valider notre classification. Dans un premier temps, nous allons séparer notre jeu de données de vérités terrain (généré dans la section Recueillir les données terrain) en un jeu de données pour entraîner le modèle (*données d'entraînement*) et un jeu de données pour le valider (*données de validation*). Nous allons utiliser 80% des parcelles de chaque classe d'occupation du sol pour consituer le jeu de données d'entraînement et le reste

288

(20%) pour le jeu de données de validation. Puis nous allons générer notre classification avec Random Forest en utilisant les 18 couches raster et le jeu de données d'entrainement. Enfin nous allons évaluer la qualité de notre classification avec le jeu de données de validation.

Objectifs :

- Créer un jeu de données d'entraînement et un jeu de données de validation à partir des données terrains (avec l'outil `Saga > Split Shapes Layer randomly`)
- Réaliser une classification supervisée en utilisant l'algorithme Random Forest (avec l'outil `Saga > Random Forest Classification (OpenCV)`)
- Valider une classification avec la matrice de confusion (avec l'outil `Saga > Confusion Matrix (Polyon / Grid)`)

Note importante :

Nous allons utiliser des algorithmes de la librairie SAGA pour réaliser la classification. En principe, la librairie SAGA est disponible sur QGIS, via la boite à outils de traitements. Cependant, pour une raison inconnue, il s'est avéré impossible lors de la préparation de ce tutoriel d'executer les traitements SAGA via QGIS sous le système d'exploitation Windows. Nous allons donc réaliser la classification à travers l'interface utilisateur de SAGA.

Etapes

1. Aligner les résolutions et étendues des différentes couches qui serviront à la classification

L'algorithme `Random Forest Classification (OpenCV)` de SAGA requiert que les couches raster qui serviront à la classification soient de résolution et étendue identiques. Dans notre cas, nous avons des couches de résolutions spatiales différentes (10 m et 20 m pour les images satellites et les indices spectraux, 30 m pour le MNT et les indices topographiques). Nous allons donc aligner les résolutions avant de lancer la classification. Notez que lorsque le classification est lancée via la librairie SAGA dans QGIS directement, cette étape est automatiquement réalisée par l'algorithme avant la classification, il n'est donc pas nécéssaire de la réaliser à la main comme nous allons le faire ici.

- Chargez dans QGIS l'ensemble des couches qui seront utilisées pour la classification (les 10 bandes spectrales, les 5 indices spectraux, le MNT, les 2 indices topographiques)
- Dans le menu, allez dans `Raster > Aligner les rasters`
- Cliquez sur le `+`
- Dans `Couche raster d'entrée`, choisissez `B02`
- Dans `Nom de fichier du raster en sortie`, saisissez le chemin de sortie de la couche qui sera redimensionné, comme indiqué sur l'image ci-dessous

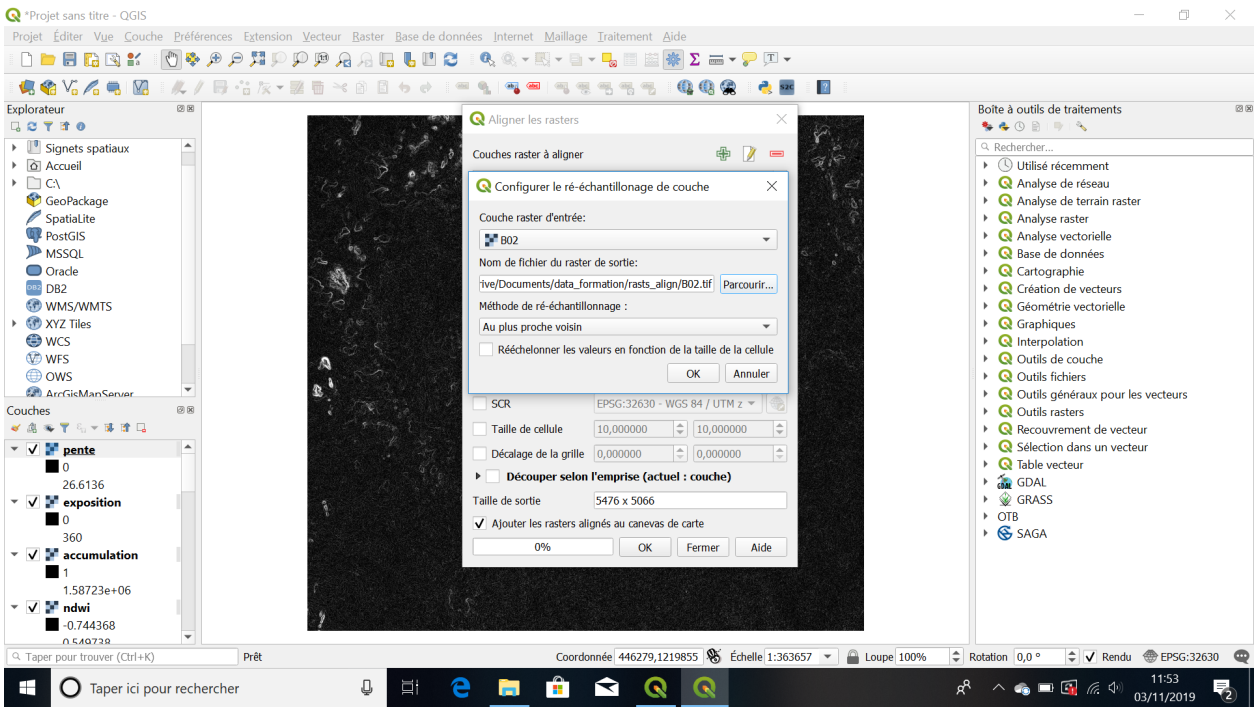


Figure 10: QGIS - menu Aligner rasters

- Répétez l'opération avec l'ensemble des couches
- Dans le menu `Couche de référence`, séléctionnez `B02 (meilleures référence)`
- Cliquez sur OK pour lancer le traitement. Les couches sont alors redimensionnées et stockées dans l'ordinateur.

2. Créer un jeu de données d'entraînement et de validation

- Ouvrir le logiciel SAGA.
- Chargez les vérités terrain : cliquez sur `File > Open`, séléctionnez "all files" dans la liste déroulante en bas de la fenêtre, et séléctionnez le fichier `ground_truth.shp`
- Dans le menu, allez dans `Geoprocessing > Shapes > Construction > Split Shapes Layer Randomly`
- Remplissez les menus comme dans l'image ci-dessous :

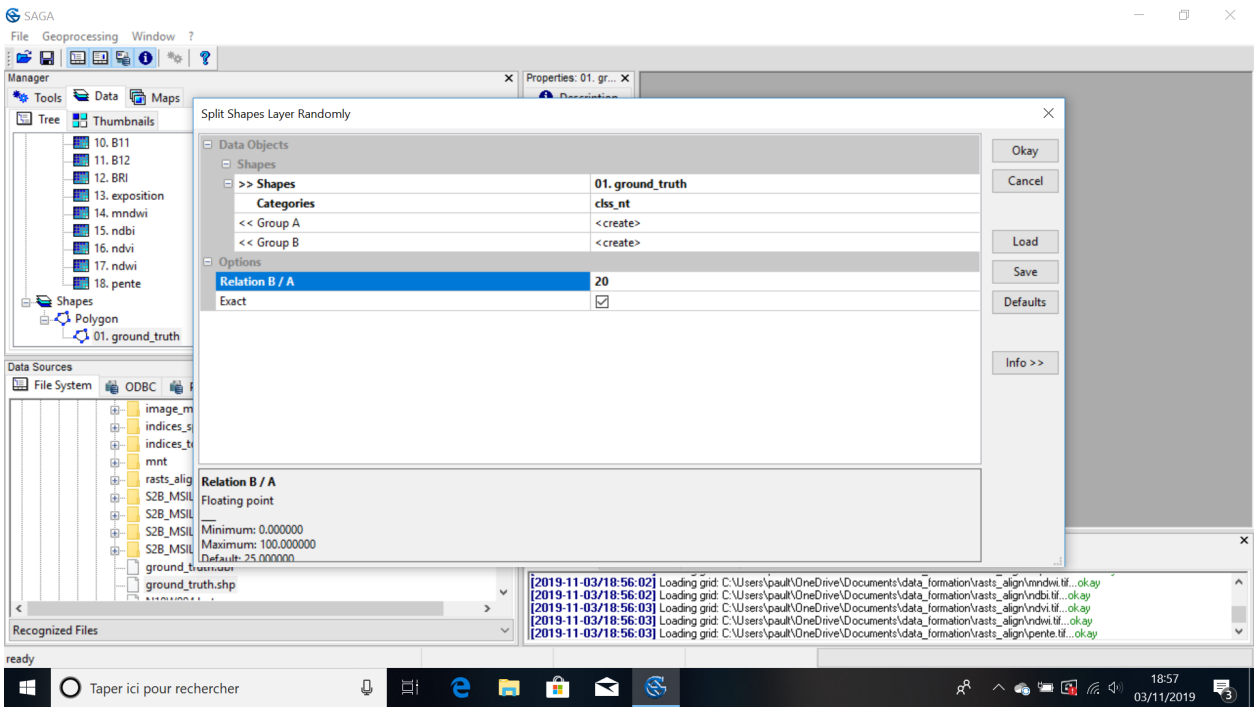


Figure 11: SAGA - Split shapes

- SAGA sépare alors le jeu de deonnées `ground_truth.shp` en deux jeux distincts :
  - `ground_truth [80%]` qui représentera le jeu de données d'entrainement de notre modèle de classification
  - `ground_truth [20%]` qui représentera le jeu de données de validation de notre modèle de classification

3. Réaliser la classification supervisée

- Dans Saga, chargez les 18 couches raster qui serviront à la classification : cliquez sur `File > Open`, séléctionnez "all files" dans la liste déroulante en bas de la fenêtre, et allez dans le dossier dans lequel sont stockées vos images (redimensionnées à l'étape 1)
- Dans le menu, allez dans `Geoprocessing > Imagery > Classification > Machine Learning > Random forest classification (OpenCV)`
- Remplissez les menus comme dans l'image ci-dessous :

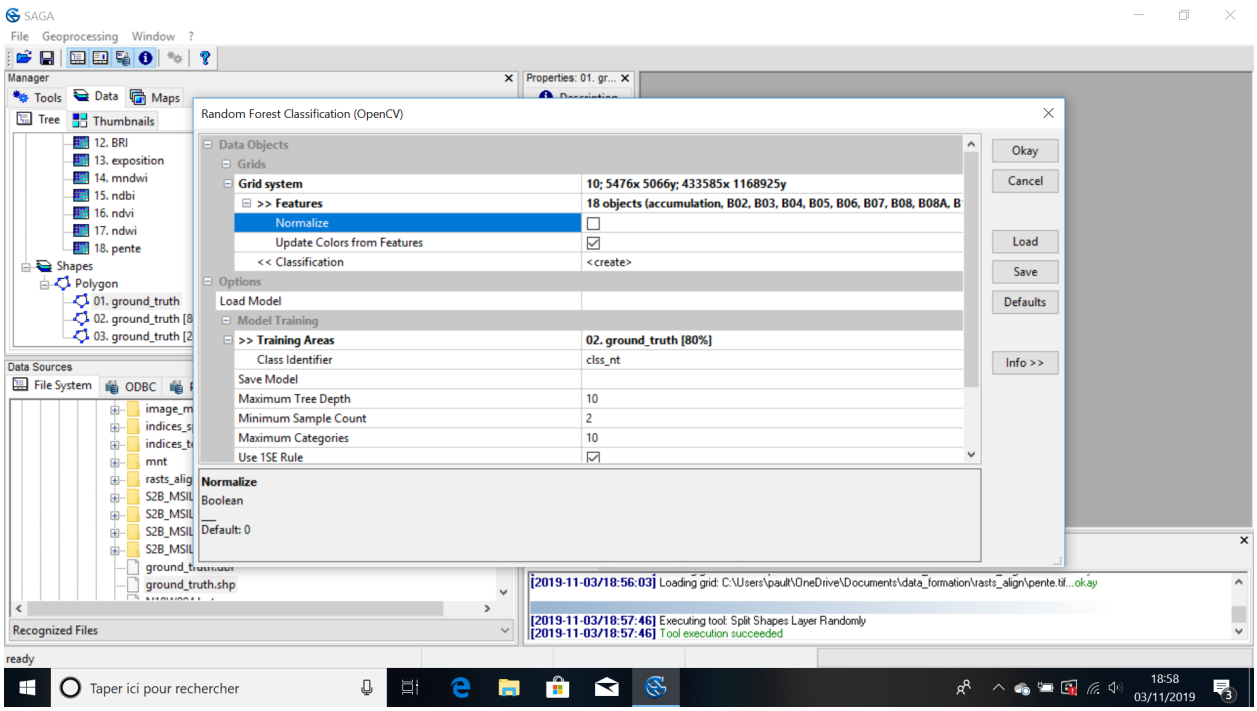


Figure 12: SAGA - Classer avec Random forest

- Cliquez sur `Okay` pour lancer la classification. A la fin du processus, la couche `Random forest classification (OpenCV)` est créée.

4. Valider la classification

L'étape de validation a pour objectif d'évaluer la qualité de notre classification. Nous allons utiliser notre jeu de données de validation pour compter le nombre de pixels correctement classés dans notre classification finale.

- Dans le menu de SAGA, allez dans `Geoprocessing > Imagery > Classification > Confusion Matrix (Polygons / Grid)`
- Remplissez les menus comme dans l'image ci-dessous :

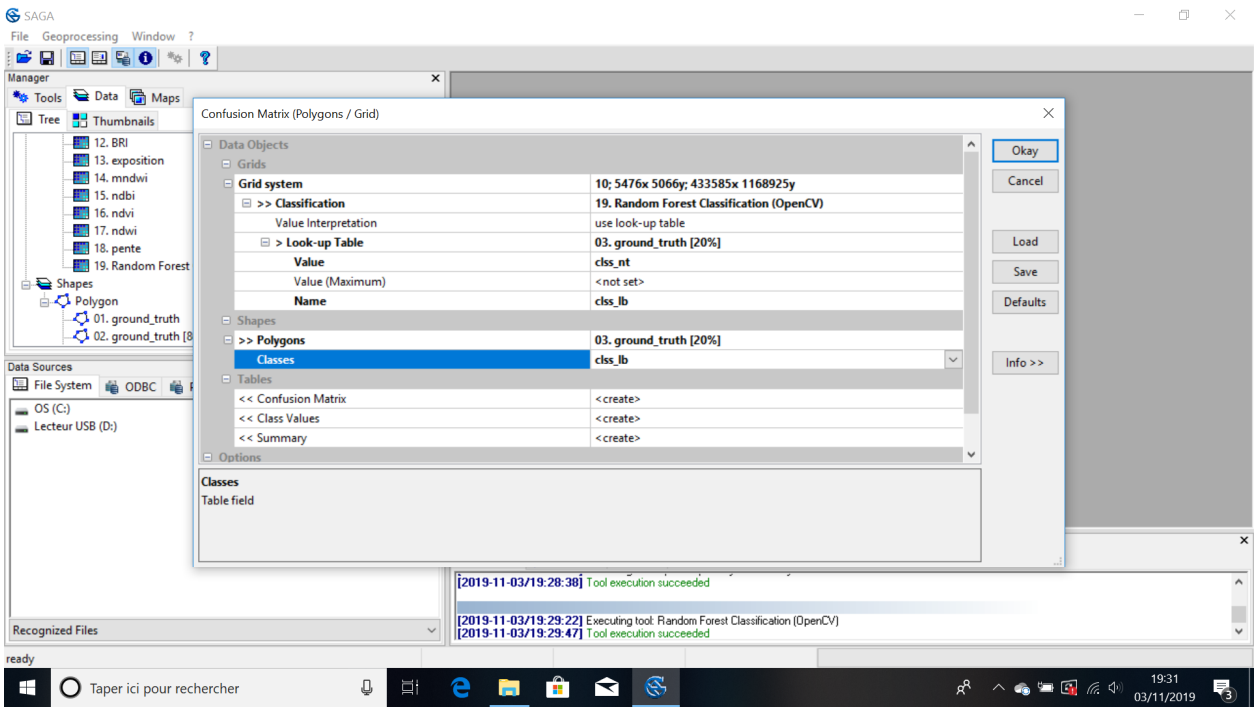


Figure 13: SAGA - Créer la matrice de confusion

- Cliquez sur `Okay` pour lancer la validation. A la fin du processus, les tables `Confusion Matrix`, `Class Values` et `Summary` sont créées.

5. Sauvegarder la classification et les indices de classification

- Dans le panneau `Data` à gauche, cliquez-droit sur la couche “Random Forest Classification (OpenCV)” puis sur `Save as`. Sauvez la couche dans le repértoire de votre choix. La couche est enregistrée dans un format raster lisible par QGIS.
- Faites de même pour les matrices de confusion “Confusion matrix”. Les matrices sont enregistrées au format txt.

Avez-vous réussi ? :

- Quelle est la valeur du coefficient Kappa de votre modèle de classification ?
- Selon la matrice de confusion, quelles sont les classes les mieux classées ? Les moins bien classées ? Quelles classes se confondent le plus entre elles ? A votre avis, pourquoi ?
- Ouvrez le raster contenant la sortie de la classification dans QGIS
- Attribuez lui une symbologie pertinente (classe “eau permanente” en couleur bleue, etc) et faites-en une carte (avec titre, échelle, etc) avec le composeur d’impression de QGIS

7 Note finale et notions non abordées

Vous avez généré votre première carte d’occupation du sol, félicitations ! Que faire avec maintenant ? Dans le cadre de notre étude fictive sur la caractérisation des habitats favorables aux vecteurs du paludisme (voir la section Présentation du cas d’études), vous pourriez maintenant l’utiliser pour extraire les surfaces relatives de chaque classe d’occupation du sol au voisinage des villages (par exemple, dans une zone tampon de 2 km), ou bien d’autres métriques paysagères pertinentes. En croisant cette information avec les données issues

de comptages de vecteurs, vous pourriez alors tenter d'expliquer les habitats favorables à la présence et abondance des vecteurs.

Ce tutoriel de fait qu'effleurer les possibilités offertes par la télédétection, et nombre de domaines n'ont pas été abordés... Par exemple, le traitement des zones sous nuages est essentiel. On peut utiliser d'autres images satellites, prises à des dates proches, pour combler les zones sous nuages et ombres de nuages sur notre image principal. Pour améliorer la qualité de la classification, on pourrait procéder de plusieurs manières, par exemple :

- en post-classification : appliquer des méthodes de filtrage de traitement d'image, par exemple un filtre majoritaire avec l'outil `SAGA > Majority filter`, pour atténuer l'effet "poivre et sel" ,
- paramétriser plus finement l'algorithme de classification (Random Forest), ou tester d'autres algorithmes de classification ,
- réaliser une classification par approche orientée objet
- intégrer davantages de variables explicatives, par exemple, des textures

Vous noterez aussi à travers ce tutoriel que la télédétection implique un nombre d'étapes important. Afin de rendre l'ensemble du processus transparent et reproductible, il peut être intéressant d'utiliser des outils dédiés, par exemple le Modeleur graphique de QGIS. Mieux encore, pour les utilisateurs de langages de programmation tels `Python` ou `R`, vous pourriez scripter l'ensemble des traitements. Toutes les librairies utilisées dans ce tutoriel (SAGA, GRASS, GDAL, etc.) existent sous R par exemple, et bien davantage encore. Vous pouvez tirer profit des capacités de ces logiciels à réaliser des figures et autres graphiques pour systématiser la production de sortie graphiques ou cartographiques, tester plusieurs modèles, etc.

8 Annexe 1 : Télécharger, installer et configurer les logiciels et extensions

Avant de travailler sur nos images, il faut installer les logiciels qui nous permettent de le faire... et dans le monde des logiciels libres et gratuits, cette étape n'est pas à négliger ! Elle peut représenter une part significative du temps de travail... (eh oui, ça fait partie des petits défauts des logiciels libres, ils ne sont pas toujours ausssi simples à installer que les logiciels propriétaires...!)

8.1 Télécharger et installer les logiciels

Nous utiliserons les logiciels QGIS, SAGA GIS et Orfeo Toolbox au cours de cette formation.

Nous utiliserons également la chaîne de traitement Sen2cor qui permet de faire les pré-traitements des images Sentinel 2 de niveau 1C (corrections atmosphériques et géométriques) pour les amener au niveau 2A, ainsi que l'extension QGIS Sen2Cor Adapter qui permet d'intégrer cette chaine de traitement dans l'interface de QGIS.

Objectifs : Installer les logiciels et extensions qui serviront pendant la formation

Etapes :

- Téléchargez et installez le logiciel QGIS (v3.10) (attention à choisir la bonne version, 32 bits ou 64 bits)
- Téléchargez le logiciel SAGA (v7.4.0).
- Téléchargez la chaîne de traitement Sen2cor (v2.5.5) (sous forme de fichier compressé .zip) (*attention : ne pas télécharger la dernière version v2.8 mais bien la version v2.5.5*). Notez le lien en local vers le fichier .zip pour l'étape suivante. Attention : Sen2Cor est disponible sous Windows uniquement en 64 bits.
- Téléchargez l'extension Sen2Cor Adapter de QGIS et notez le lien en local vers le fichier .zip pour l'étape suivante (sous forme de fichier compressé .zip). Notez le lien en local vers le fichier .zip pour l'étape suivante.

Avez-vous réussi ?

- Ouvrez QGIS sur votre ordinateur. La fenêtre suivante doit apparaître :

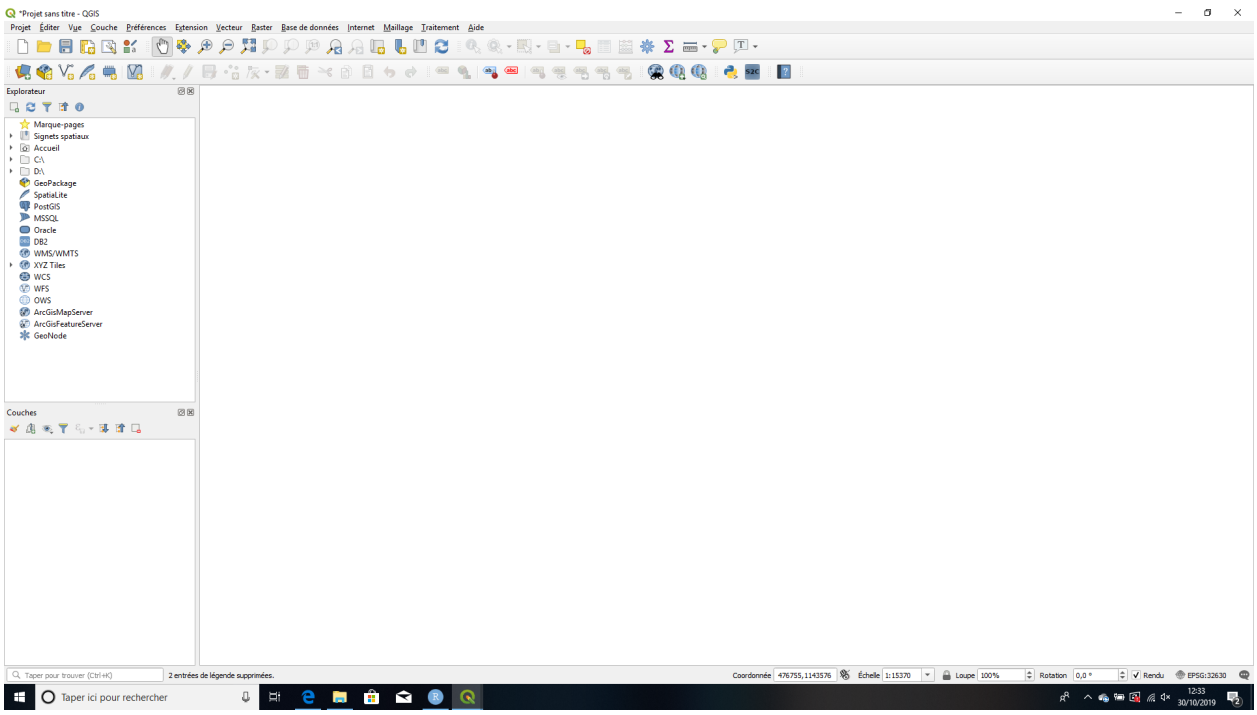


Figure 14: QGIS - fenêtre d'accueil

- Ouvrez SAGA GIS sur votre ordinateur. La fenêtre suivante doit apparaître :

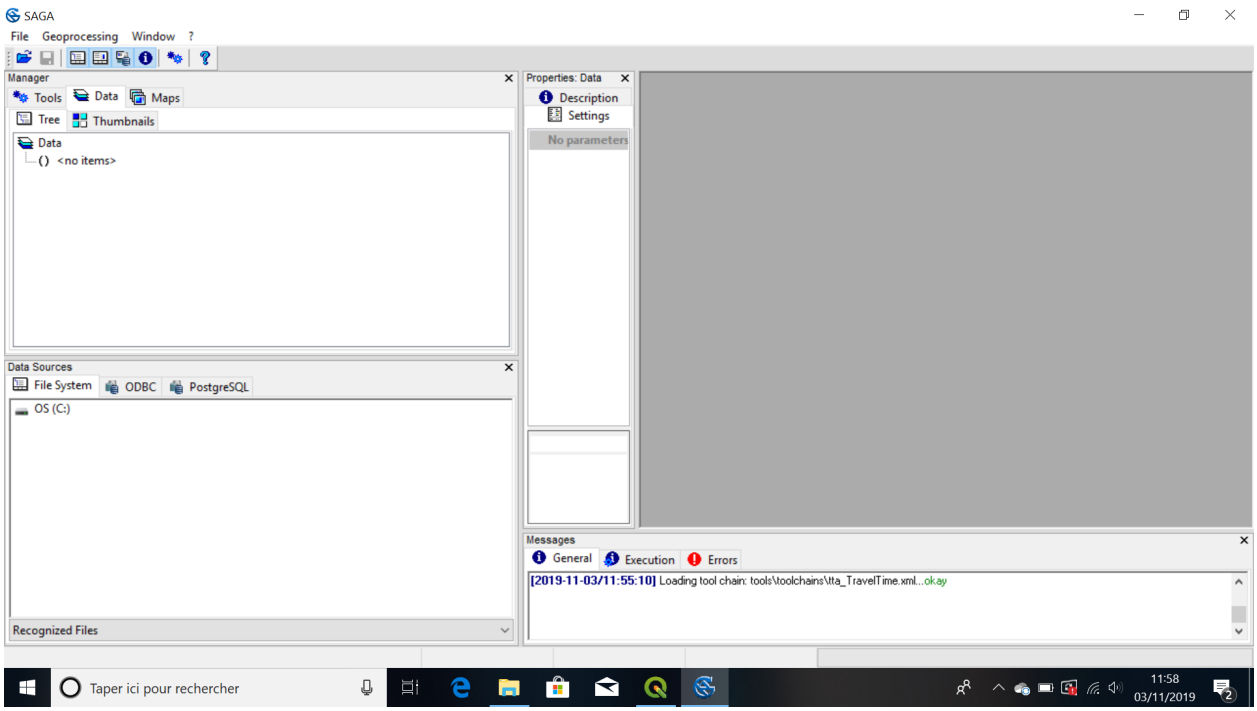


Figure 15: SAGA GIS - fenêtre d'accueil

8.2 Configurer les applications tierces et plugins dans QGIS

QGIS dispose d'un ensemble de traitements (algorithmes) qui sont installés par défaut avec le logiciel. Les opérations spatiales classiques sur les données géographiques vecteur et raster sont assurées par cette bibliothèque de traitements. Cependant, en télédétection, nous utilisons des algorithmes complexes qui ne sont pas disponibles dans la version de base de QGIS. La bonne nouvelle, c'est qu'ils sont disponibles gratuitement via d'autre logiciels ou bibliothèques SIG libres et gratuites, et qu'ils peuvent être incorporés dans QGIS (ainsi, pas besoin de passer sans cesse d'un logiciel à l'autre). La moins bonne nouvelle, c'est que l'intégration de ces logiciels/bibliothèques dans QGIS demande un peu de travail de configuration. C'est l'objectif de cette section : intégrer les logiciels et bibliothèques de traitements tiers dans QGIS.

QGIS offre deux moyens d'intégrer des applications tierces :

- Intégrer des bibliothèques de traitement développées en dehors de QGIS. Ces bibliothèques de traitement sont développés dans un contexte complètement externe à QGIS (par exemple par l'armée américaine, ou bien le Centre National d'Etudes Spatiales français), et existent parfois depuis plus longtemps même que QGIS. Il s'agit de SAGA, GRASS, GDAL et OTB. Ces bibliothèques peuvent être téléchargées et exécutées en dehors de QGIS (en tant que logiciel proprement dit) mais il est possible de les intégrer directement dans l'interface graphique QGIS. Les librairies SAGA, GDAL et GRASS sont automatiquement disponibles et configurées dans QGIS. Il faut par contre configurer OTB.
- Installer des extensions QGIS. Ces extensions sont aussi appelées *plugins* en anglais. Les plugins sont des librairies de codes Python développées par les utilisateurs de QGIS et mis à disposition de tous. Ces extensions permettent d'executer des traitements variés, qui ne sont pas présents dans les algorithmes de base de QIGS. Les plugins sont dotés d'une interface graphique qui permet d"executer les algorithmes sans avoir à"mettre les mains" dans le code Python.

Objectifs : Configurer les application tierces dans QGIS

En savoir plus : https://docs.qgis.org/3.4/fr/docs/user_manual/processing/3rdParty.html
296

Etapes : Intégrer Sen2cor Adapter dans QGIS

- Dans QGIS, allez dans `Extensions > Installer/gérer les extensions > Installer depuis un ZIP`
- Dans la case `Fichier ZIP`, allez chercher le lien vers le fichier compressé Sen2Cor Adapter téléchargé en section Télécharger les logiciels et plugins puis cliquer sur `Installer le plugin`

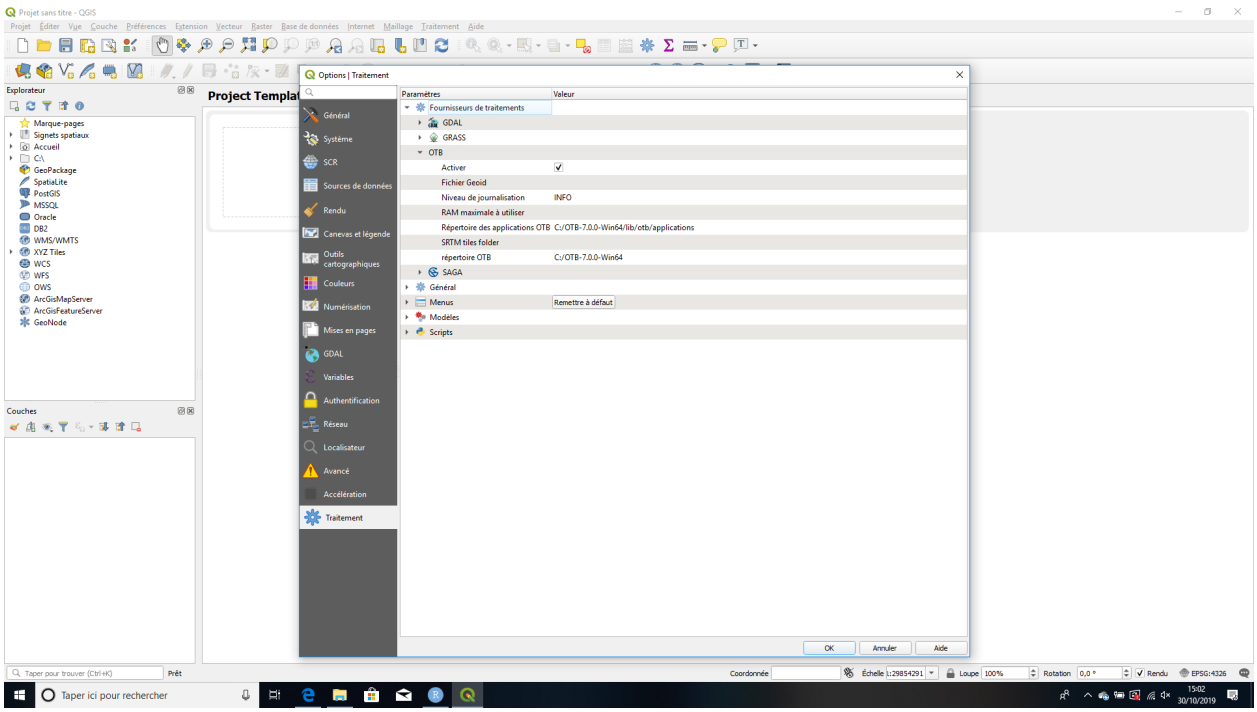


Figure 16: Configuration de l'extension Sen2Cor Adapter dans QGIS

Avez-vous réussi ?

- Vérifiez que le plugin Sen2Cor Adapter a bien été installé en ouvrant QGIS puis en allant dans `Raster` et en vérifiant que l'option `Sen2cor Adapter` est bien disponible. Ouvrez `Sen2cor Adapter`. L'écran suivant doit apparaître :

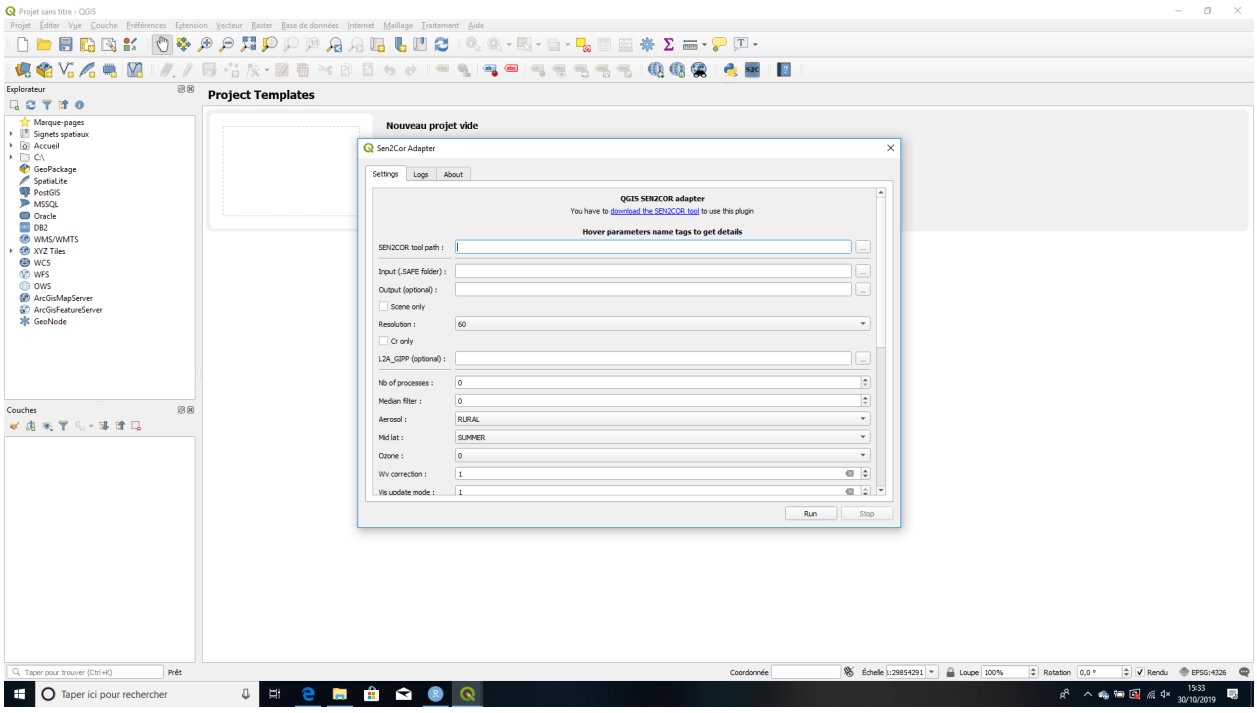


Figure 17: Extension Sen2Cor Adapter dans QGIS

9 Annexe 2 : Identifier et télécharger des produits satellitaires

Les portails web listés ci-dessous permettent de visualiser des images satellites sans avoir à les télécharger. Cela peut-être intéressant de les consulter pour, par exemple, évaluer la qualité d'une image satellite avant de la télécharger (afin d'identifier précisément les zones sous nuages), visualiser des indices spectraux (NDVI, NDWI, etc.) ou encore s'informer sur les produits satellites existants.

1. **Sentinel-hub EO-Browser** pour les données des missions Sentinel de l'ESA (Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-5P)

Disponible à l'adresse suivante : https://apps.sentinel-hub.com/

Le Sentinel-hub EO-Browser permet de visualiser les données des missions Sentinel de l'Agence Spatiale Européenne (Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-5P à ce jour). L'utilisateur filtre sa recherche par source de données (mission Sentinel), date d'acquisition de l'image et zone géographique. Les produits disponibles sont alors proposés pour visualisation. L'utilisateur peut choisir la composition de bandes parmi une vaste panoplie compositions disponibles pour chaque source (par exemple pour Sentinel-2 : couleurs réelles, NDVI, NDWI, SAVI, etc.).

Les produits Sentinel sont téléchargeables en utilisant le Copernicus Open Access Hub disponible à l'adresse suivante : https://scihub.copernicus.eu/.

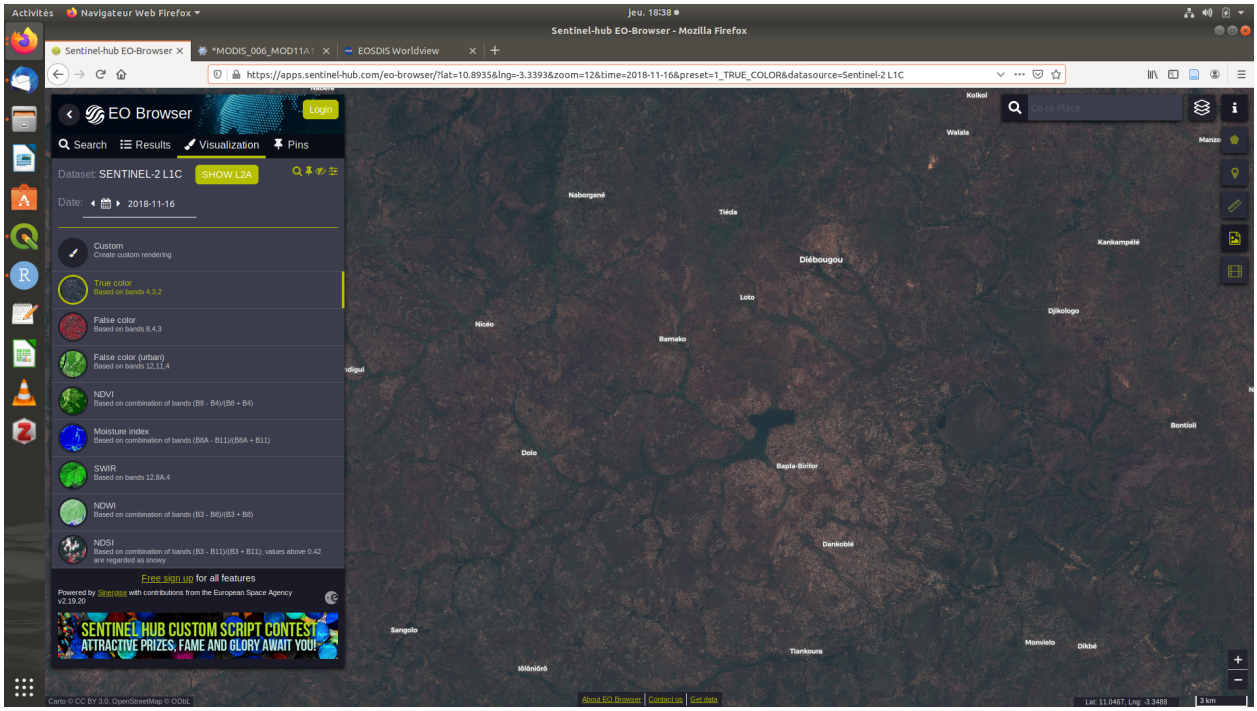


Figure 18: Sentinel-hub EO-Browser - image Sentinel-2 en couleurs réelles sur l'aire de Diébougou

2. **NASA EOSDIS Worldview** pour les données de la NASA

Disponible à l'adresse suivante : https://worldview.earthdata.nasa.gov

Le portail EOSDIS Worldview de la NASA permet de visualiser nombre de jeux de données d'observation de la Terre générés par l'agence spatiale américaine (MODIS, SMAP, VIIRS, etc.). L'interface permet de séléctionner la couche à visualiser et de filtrer spatialement et temporellement (grâce à une barre intuitive) les jeux de données. Il est possible, comme dans un SIG bureau, de superposer des couches, de leur apporter de la transparence ou encore de régler la palette de couleurs utlisée. Il est aussi possible d'exporter de courtes vidéos montrant l'évolution d'une série temporelle.

Les données d'observation de la Terre de la NASA sont pour la pluspart libres d'accès et téléchargeables sur le portail des données d'observation de la Terre de la NASA disponible à l'adresse suivante : https://search.earthdata.nasa.gov/ .

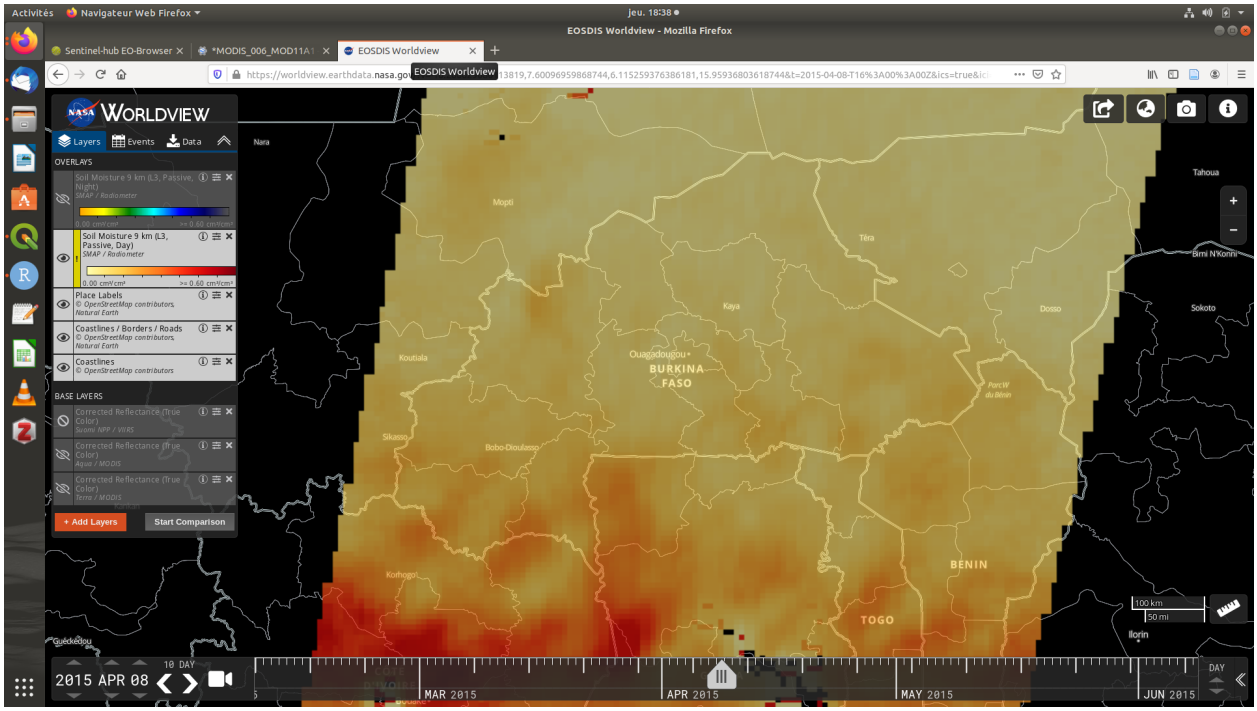


Figure 19: NASA Worldview - image SMAP (humidité du sol) au dessus du Burkina Faso

3. **Google Earth Engine** pour une très large panoplie de données d'observation de la Terre

Disponible à l'adresse suivante : https://developers.google.com/earth-engine/datasets

Le projet Google Earth Engine (GEE) est mené par Google. L'initiative a été motivée par le constat que la diversité des sources et des formats des données d'observation de la Terre générées par les divers organismes partout dans le monde est un frein aujourd'hui à leur utilisation et leur analyse conjointe. L'objectif de GEE est de centraliser un ensemble de jeux de données d'observation de la Terre provenant de très nombreuses sources (NASA, ESA, etc.) et couvrant des thématiques variées (climatologie, imagerie, géophysique, etc.). Ainsi, les données des missions Landsat, MODIS, Sentinel, etc. sont toutes disponibles via GEE. L'interface propose une manière commune à tous les jeux de données pour les visualiser en ligne, les télécharger, les croiser, effectuer des analyses spatiales. Par ailleurs, les éventuels traitements de données définis par les utilisateurs sont effectués sur les serveurs de Google, rendant ainsi les calculs particulièrement rapides (comparé à un traitement en local) sur des jeux de données volumineux.

300

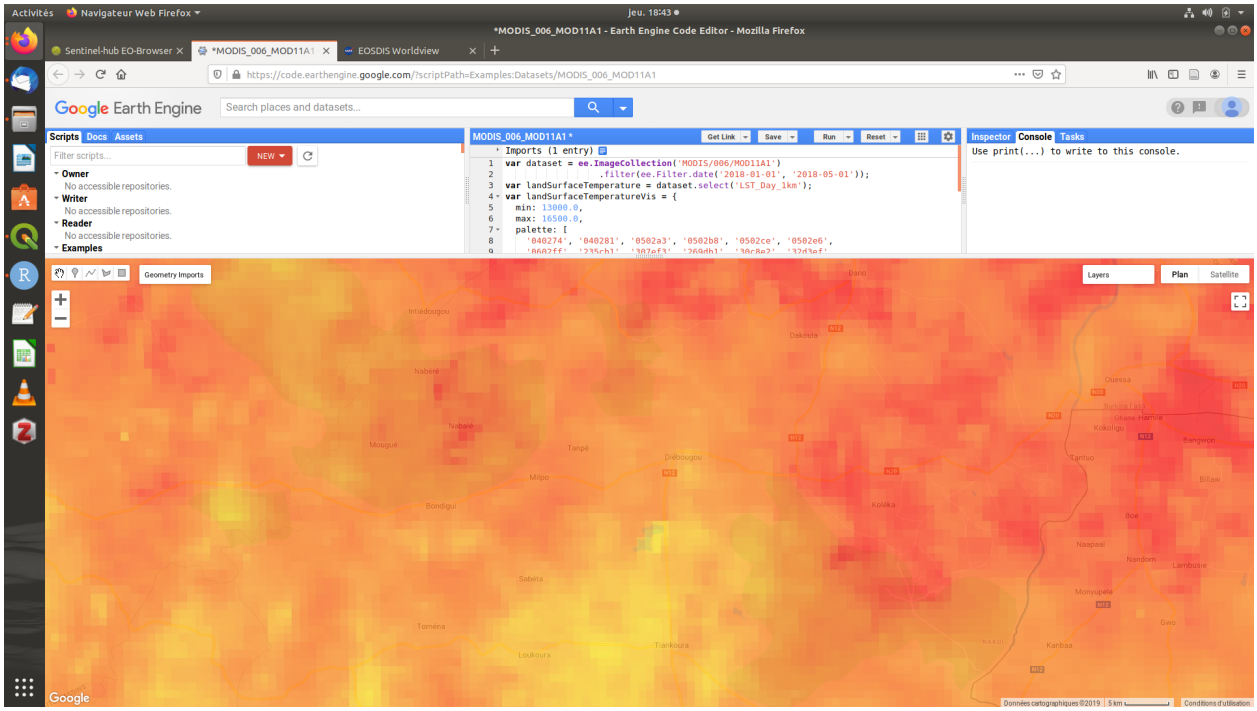


Figure 20: Google Earth Engine - image MODIS LST (température du sol) sur l'aire de Diébougou

Annexe E

Texte intégral de l'article complémentaire n°3

La référence de la publication est la suivante :

Soma, D.D., Zogo, B., ***Taconet, P****. et al. Quantifying and characterizing hourly human exposure to malaria vectors bites to address residual malaria transmission during dry and rainy seasons in rural Southwest Burkina Faso. BMC Public Health 21, 251 (2021).* https://doi.org/10.1186/s12889-021-10304-y

BMC Public Health

RESEARCH ARTICLE Open Access

Quantifying and characterizing hourly human exposure to malaria vectors bites to address residual malaria transmission during dry and rainy seasons in rural Southwest Burkina Faso

D. D. Soma[1,2,3*†], B. Zogo[3,4,5†], P. Taconet[1,3], A. Somé[1], S. Coulibaly[1], L. Baba-Moussa[5], G. A. Ouédraogo[2], A. Koffi[4], C. Pennetier[3,4], K. R. Dabiré[1] and N. Moiroux[1,3]

Abstract

Background: To sustain the efficacy of malaria vector control, the World Health Organization (WHO) recommends the combination of effective tools. Before designing and implementing additional strategies in any setting, it is critical to monitor or predict when and where transmission occurs. However, to date, very few studies have quantified the behavioural interactions between humans and *Anopheles* vectors in Africa. Here, we characterized residual transmission in a rural area of Burkina Faso where long lasting insecticidal nets (LLIN) are widely used.

Methods: We analysed data on both human and malaria vectors behaviours from 27 villages to measure hourly human exposure to vector bites in dry and rainy seasons using a mathematical model. We estimated the protective efficacy of LLINs and characterised where (indoors vs. outdoors) and when both LLIN users and non-users were exposed to vector bites.

Results: The percentage of the population who declared sleeping under a LLIN the previous night was very high regardless of the season, with an average LLIN use ranging from 92.43 to 99.89%. The use of LLIN provided > 80% protection against exposure to vector bites. The proportion of exposure for LLIN users was 29–57% after 05:00 and 0.05–12% before 20:00. More than 80% of exposure occurred indoors for LLIN users and the estimate reached 90% for children under 5 years old in the dry cold season.

Conclusions: LLINs are predicted to provide considerable protection against exposure to malaria vector bites in the rural area of Diébougou. Nevertheless, LLIN users are still exposed to vector bites which occurred mostly indoors in late morning. Therefore, complementary strategies targeting indoor biting vectors in combination with LLIN are expected to be the most efficient to control residual malaria transmission in this area.

Keywords: Diébougou, LLIN, *Anopheles*, Humans, Behaviours, Residual transmission

* Correspondence: dieusoma@yahoo.fr
†D. D. Soma and B. Zogo contributed equally to this work.
[1]Institut de Recherche en Sciences de la Santé (IRSS), Bobo-Dioulasso, Burkina Faso
[2]Université Nazi Boni (UNB), Bobo-Dioulasso, Burkina Faso
Full list of author information is available at the end of the article

Background

Massive distribution of long-lasting insecticidal nets (LLINs) is a core intervention for malaria control in Burkina Faso. Scaling-up of coverage with LLIN in sub-Saharan Africa has been very successful between 2000 and 2015 during which malaria morbidity and mortality have dropped considerably [1]. Unfortunately, this significant progress is stalling or even reversing in some countries. Burkina Faso is indeed one of the sixteen (16) in the world that documented an increase in malaria burden from 2016 to 2017 [2]. This trend might be attributed to the recent increases in prevalence and strength of pyrethroid resistance in malaria vectors [3–5]. Another possible cause is the development of behavioural resistance in vector populations [6–8]. In sub-Saharan Africa, there have been many reports of changes in vector species and/or vector biting behaviours to avoid contact with LLIN [6–8]. Such changes in vector populations are considered by many specialists as an important threat for indoor control strategies such as LLIN [9, 10].

To sustain the efficacy of vector control, the WHO recommends the combination of effective tools [11]. At present, there are a number of recommended tools available and many others under development that can potentially be combined with LLIN [12, 13]. However, national malaria control programs (NMCPs) are now facing challenges to design effective control strategies due to high variations in malaria epidemiology between and even within countries [14]. To do so, NMCP must be able to monitor or predict when and where transmission occurs and to characterize residual transmission (i.e. the transmission that escapes vector control by LLINs).

In order to compare the impact of LLINs on human exposure to malaria vectors bite among sites, Killeen et al. [15] developed an approach that quantify behavioural interactions between mosquitoes and humans. The approach use measures of indoor and outdoor vector biting as well as the distribution of people outdoors, indoors and under LLINs for each hour of the night. It produces average hourly and nightly weighted estimates of exposure occurring indoors and outdoors as well as estimates of prevented exposure. The analytical model developed by Killeen et al. and extended by Geissbühler et al., [16] is therefore a useful tool to estimate protective efficacy of LLINs and to characterize residual transmission. Indeed, it allows to identify where (indoors vs. outdoors) and during which hours LLIN users are exposed to anopheles bite, i.e. where and when residual transmission is expected to occur. Numerous studies have used this model in Africa [15–30]. However until now, only one of these studies has reported exposure estimates for sites located in Burkina Faso [18].

The present study aims to provide and discuss up-to-date estimates of human exposure to *Anopheles* bite and to characterise residual malaria transmission in an area of Burkina Faso where malaria vectors shows high levels of pyrethroid resistance [31]. Results of entomological surveys previously reported [31] were used in combination with human behavioural data to quantify, through the Killeen's model, the behavioural interactions between humans and *Anopheles* mosquitoes during both dry and rainy seasons in the Diébougou area, southwest Burkina Faso. Data were collected during the pre-intervention stage of a large randomized control trial designed to investigate whether the combination of LLINs with other vector control tools can provide additional protection over malaria cases and transmission.

Methods

This study was conducted in 27 villages located in the Diébougou health district, southwest Burkina Faso (Fig. 1). These villages were selected based on geographical (distance between two villages higher than 2 km and accessibility during the rainy season) and demographic (a population size ranging from 200 to 500 inhabitants) criteria [31] to participate in a randomized controlled trial. The climate in the study area is tropical with one dry season from October to April (including a cold period from December to February and a hot period from March to April) and one rainy season from May to September. Average daily temperature amplitudes are 18–36 °C, 25–39 °C and 23–33 °C in dry cold, dry hot and rainy season, respectively. The mean annual rainfall is 1200 mm. The natural vegetation is dominated by wooded savannah dotted with clear forest gallery. The main economic activity is agriculture (cotton growing and cereals) followed by artisanal gold mining and production of coal and wood [32, 33]. In the study area as in the whole country, a mass distribution of LLINs (PermaNet 2.0) was carried out by the NMCP in July 2016. No LLINs were distributed by our teams.

The study involved the conduct of three entomological surveys and two human behavioural surveys. Figure 2 shows the timeline of the study. We conducted three entomological surveys in the dry cold (January 2017), dry hot (March 2017) and rainy seasons (June 2017). During each survey, we collected mosquitoes using the standard method of human landing catch (HLC). Mosquitoes were sampled both indoors and outdoors from 17:00 h to 09:00 h in 4 houses per villages during one night [31]. In each study village, two teams of eight collectors were deployed, with the first team collecting from 17:00 h to 01:00 h and the second from 01:00 h to 09:00 h. All the collected mosquitoes were morphologically identified [34, 35] and *Anopheles spp.* mosquitoes were subsequently identified to the species level by polymerase

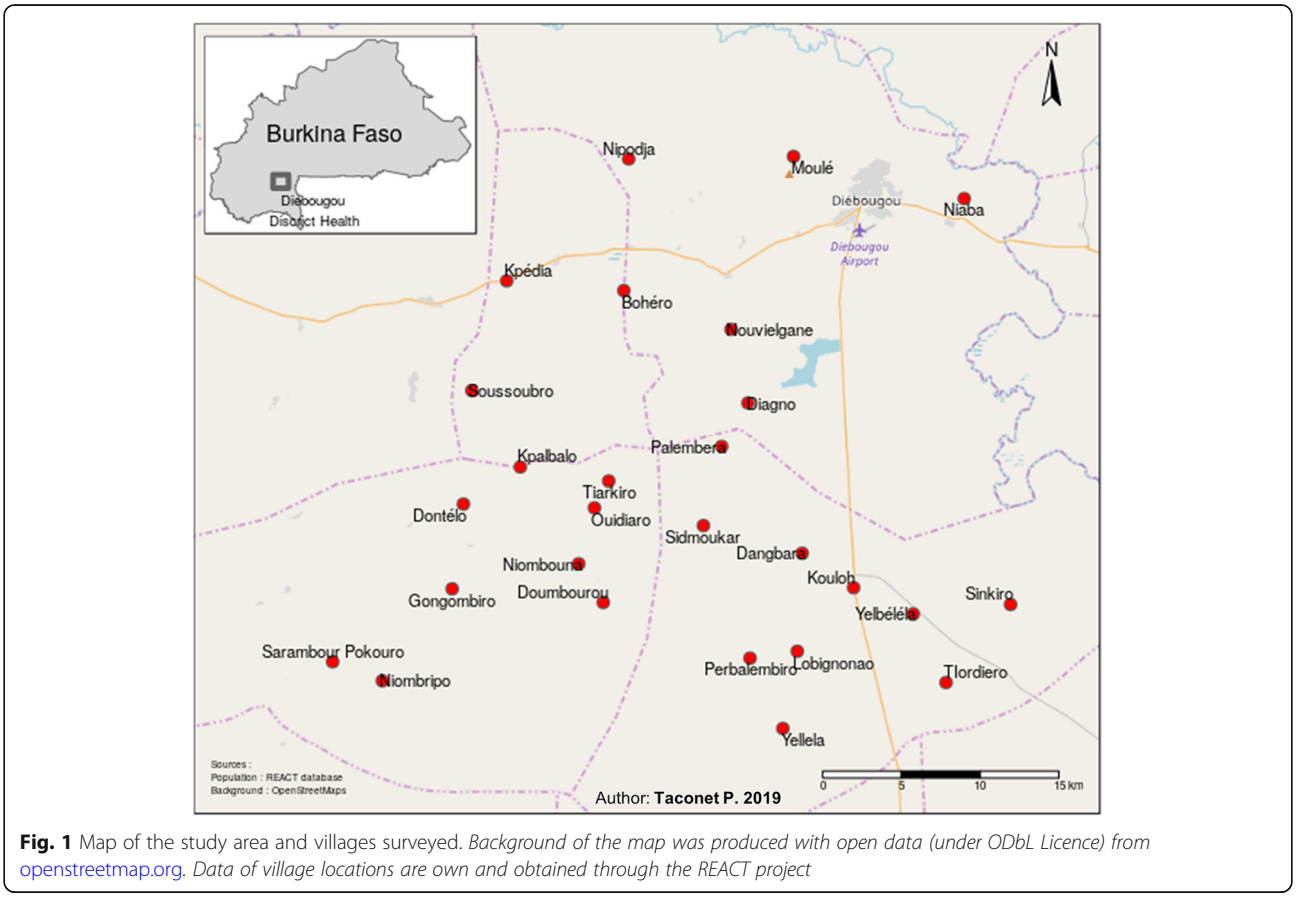


Fig. 1 Map of the study area and villages surveyed. *Background of the map was produced with open data (under ODbL Licence) from* openstreetmap.org. *Data of village locations are own and obtained through the REACT project*

chain reaction [36–38]. Detailed descriptions of the methods used are provided in our previous publication [31]. In the current work, we aggregate data for all species belonging to the *Anopheles* genus (*Anopheles spp*) in order to have appropriate data regarding malaria vectors behaviour. Overall, *Anopheles funestus s.s* was the main malaria vector in the study area during the dry cold season [31]. During the dry hot and rainy seasons, *Anopheles coluzzii* and *Anopheles gambiae s.s* were the dominant species. The mean endophagy rate (ER) of malaria vectors was 63.23, 50.18 and 57.18% during the dry cold, dry hot and rainy seasons, respectively [31].

In order to obtain appropriate data regarding relevant human behaviours, we surveyed 401 and 339 randomly selected households in dry (end of February to April 2017) and rainy (September 2017) seasons, respectively

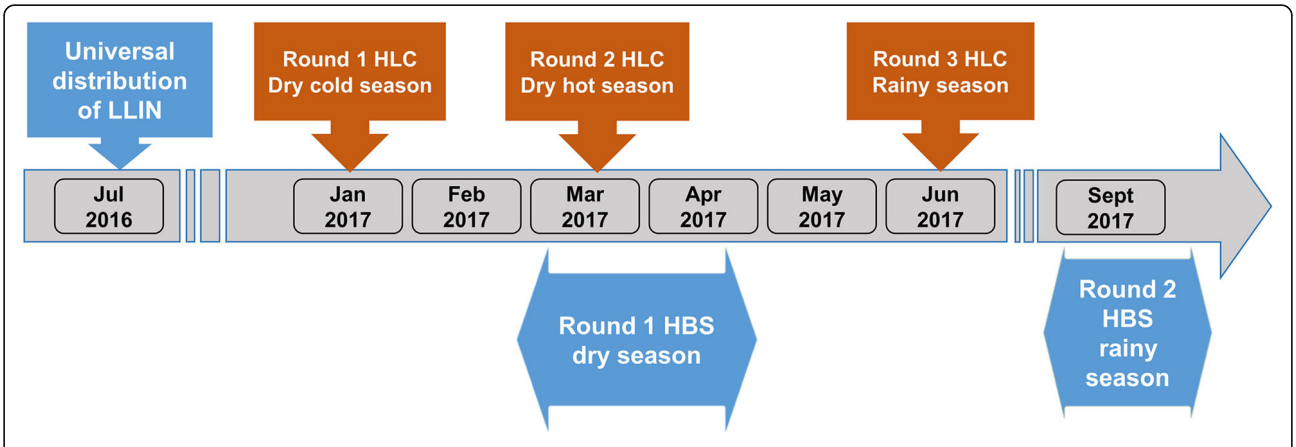


Fig. 2 Timeline of long-lasting insecticidal impregnated net (LLIN) distribution campaigns, human behaviours surveys (HBS), and entomological surveys. *HLC: Human Landing Collection*

(corresponding to an average of 15 and 13 households per village). Among people usually leaving in each selected household, we randomly selected 3 persons (maximum) belonging to each of the 3 following age groups: 0–5 years old, 6–17 years old and ≥ 18 years old. We asked the head of the household the time at which each selected person (1) entered and exited his own house the night preceding the survey and (2) the time each LLIN user entered and exited his sleeping space the night preceding the survey (the questionnaire was previously published in supplementary Text S1 of [20]). In order to know the relative weight of each age group in the population, we recorded the number of individuals belonging to these groups in each household. A total of 3045 and 2880 individuals were surveyed in dry and rainy seasons, respectively, representing 35.08 and 33.17% of the 27 villages' population according to a census carried out by our team in 2016 [31]. The human behavioural surveys were carried out in the same villages where mosquitoes were collected. The selections of households for human behavioural surveys and of houses for entomological survey were independent. Data were recorded using tablets running Open Data Kit (ODK) forms.

From data of each entomological survey, we calculated indoor and outdoor hourly biting rates (i.e. the number of *Anopheles* mosquitoes collected per human per hour) at the village level and for the whole study area. At the same scales, we calculated from data of each human behavioural survey the hourly proportions of people being indoors or under an LLIN. Hourly biting rates and hourly distribution of people were combined to calculate estimates of human exposure to *Anopheles spp.* bites in the dry season (both cold and hot) and the rainy season using an extension of the Killeen's model [15] as previously described in Geissbühler et al. [16] and Moiroux et al. [20] and detailed in Additional file 1.

Since only one survey of human behaviour was carried out in dry season, we used the same human behaviour data to model human exposure to *Anopheles* bite during both dry cold and dry hot seasons.

We estimated the average *true* personal protection (P^*) of using an LLIN (i.e. the proportion of exposure to all bites occurring both indoors and outdoors that is prevented by using an LLIN) as well as the proportion of exposure which occurred indoors for LLIN users either accounting for the personal protection provided by net use ($\pi_{i,n}$) or ignoring it to compare with available estimates for unprotected people (π_i) [16] (Additional File 1). Exposure when sleeping under an LLIN was assumed to be reduced by 92% [20]. Moreover, to characterize residual transmission, we calculated the proportion of exposure occurring before 20:00 ($\pi_{e,n}$) and after 5:00 ($\pi_{m,n}$) that are the times respectively preceding and following the period when most (> 50%) of LLIN users are protected (Additional File 1).

All the exposure estimates (i.e. P^*, $\pi_{i,n}$, π_i, $\pi_{e,n}$, $\pi_{m,n}$) were calculated at the village and study area levels, for each age group as well as for the whole population. The relative weight of each age classes in the population was taken into account when calculating exposure values at the population level (see Additional File 1). For these calculation and to produce figures, we developed an R [39] package named "biteExp" (https://github.com/Nmoiroux/biteExp).

Results

The average declared LLIN use rate was very high in the study population ranging from 95.49% in the dry season to 99.67% in the rainy season (Table 1). The declared LLIN use rate was higher in the 0–5 years old age group (97.87% in the dry season to 100% in the rainy season) compared to children aged 6–17 years old (95.36% in the dry season to 99.79% in the rainy season) and adults (92.45% in the dry season to 99.19% in the rainy season) (Table 1). However, we found that the LLIN use rate varied among villages (see Additional file 2) with the lowest rates observed in Kpédia (68.42%), Palembera (71.73%) and Diagnon (78.78%) in the adults group during the dry season. In the other villages LLIN use rates ranged from 80 to 100% whatever the season (see Additional file 2). Figure 3 shows humans and *Anopheles* behaviour profiles as well as average hourly exposure and prevented exposure to bites for LLIN users in our study area.

The majority of the population was indoors from 20:00 in both dry and rainy seasons (Fig. 3a, b and c). These populations woke up around 05:00 in the early morning in all seasons (Fig. 3a, b and c). Most of the total exposure to *Anopheles* bites occurred indoors (> 94% for non-users, Table 1) but was largely preventable by using of LLIN (Fig. 3d, e and f). Indeed, LLIN were estimated to provide average 'true' personal protection against 84.93, 80.89 and 82.82% of exposure in dry cold season, dry hot season and rainy season, respectively (Table 1, Additional file 3). The peak of exposure for users occurred indoors between 05:00 and 06:00 just before sunrise whatever the season (Fig. 3d, e and f). On average, between 33 and 57% of residual exposure of LLIN users occurred after wake up (after 5:00) depending on age groups. Early bites (before 20:00) represented less than 12% of the residual exposure of LLIN users (Table 1).

Discussion

The average declared LLIN use rate was very high (> 95%) in all age groups of our study population. The LLIN use rate was slightly higher in children under five years of age than the rest of the population. This finding is consistent with results from a multi-country analysis that revealed that the most vulnerable groups are

Table 1 Average LLIN use rates, true average protection efficacy of LLINs against exposure to vector bites and proportions of indoors, "before bed" and "after bed" exposure to *Anopheles* bites for both LLIN users and non-users in 27 villages of the Diébougou area, Burkina Faso

Season	Age (years)	LLIN use rate (%[min-max])	[a]True average LLIN personal protection efficacy (% [min-max])	Exposure indoors (%[min-max])		Exposure before 20:00 h (%[min-max])		Exposure after 05:00 h (%[min-max])	
				LLIN users	Non-users	LLIN users	Non-users	LLIN users	Non-users
Dry cold season	**18+**	92.45 [68–100]	83.44 [0–92]	79.92 [0–100]	96.67 [0–100]	0.07 [0–0.13]	0.04 [0–0.34]	44.99 [0–100]	8.16 [0–100]
	6 to 17	95.36 [71–100]	83.79 [0–92]	85.44 [0–100]	97.64 [0–100]	0.58 [0–1]	0.12 [0–0.73]	48.93 [0–100]	9.01 [0–100]
	0 to 5	97.87 [81–100]	86.73 [0–92]	90.52 [0–100]	98.74 [0–100]	3.93 [0–100]	0.62 [0–100]	40.23 [0–100]	12.20 [0–100]
	population	95.49 [77–100]	84.93 [0–92]	85.62 [0–100]	97.83 [0–100]	1.66 [0–100]	0.31 [0–100]	44.50 [0–100]	10.11 [0–100]
Dry hot season	**18+**	92.45 [68–100]	78.00 [0–92]	69.57 [19–100]	93.31 [75–100]	3.38 [0–26]	0.82 [0–1]	57.20 [0–100]	13.19 [0–100]
	6 to 17	95.36 [71–100]	79.88 [2–92]	82.70 [21–100]	96.52 [72–100]	4.57 [0–5]	0.99 [0–2]	56.20 [0–100]	12.27 [0–100]
	0 to 5	97.87 [81–100]	83.63 [13–92]	88.73 [29–100]	98.15 [82–100]	11.30 [0–20]	2.13 [0–3]	43.95 [0–100]	12.32 [0–100]
	population	95.49 [77–100]	80.89 [5–92]	80.54 [24–100]	96.28 [78–100]	6.56 [0–30]	1.41 [0–2]	52.19 [0–100]	12.55 [0–100]
Rainy season	**18+**	99.19 [92–100]	79.13 [53–92]	75.61 [11–100]	94.91 [62–100]	10.08 [0–23]	2.17 [0–5]	42.90 [0–90]	9.81 [0–44]
	6 to 17	99.79 [94–100]	81.83 [51–92]	83.28 [45–100]	96.96 [91–100]	10.24 [0–25]	2.22 [0–8]	48.59 [0–91]	10.42 [0–50]
	0 to 5	100.00	87.00 [72–92]	89.21 [69–100]	98.60 [96–100]	11.33 [0–19]	2.31 [0–11]	33.88 [0–85]	10.55 [0–50]
	population	99.67 [97–100]	82.82 [58–92]	81.93 [27–100]	96.90 [82–100]	10.47 [0–23]	2.23 [0–9]	42.40 [0–89]	10.27 [0–48]

Min and max reported in brackets give the value recorded in the village with the lower and the higher average value, respectively
[a]True average LLIN personal protection efficacy: estimated proportion of *Anopheles* bites prevented by the use of a LLIN

preferentially protected by LLIN in sub-Saharan Africa [40]. At the village level, the use rate rarely fall under 80%, being consistently higher than the nationwide LLIN use value of 67% published by WHO in 2017 [41]. This may be explained by the fact that the study was conducted approximately 6 months after a wide LLIN distribution. However, our reported LLIN use may be overestimated because it was based on self-reported survey questions, the most commonly used method to assess bednet use [42]. To more accurately estimate LLIN use, future studies quantifying human exposure to mosquito bites should consider using other measurement methods such as electronic monitoring devices [43, 44].

This study shows that the overall protective efficacy of LLINs against vector bites in the rural area of Diébougou was high (80–85%) during the three seasons. Our estimates for LLIN personal efficacy were comparable with those found in Benin (80 and 87%) [20] but were higher than those reported elsewhere such as in Kenya (51%) [21] and Tanzania (70, 59 and 38%) [15, 16]. Our results support strongly the use of LLIN as a primary malaria vector control tool in the area. Nevertheless, such a protection level (85% in average) has to be put into perspective with the high malaria transmission and endemicity [31] in order to measure/realize the importance of malaria residual transmission in the area.

We estimated that 33–57% of residual exposure to *Anopheles* bites of LLIN users occurred after 5:00 and 0.07–12% occurred before 20:00 when most of users are awake. The proportion of exposure for LLIN users has been higher in the late part of the morning than in the early part of the evening in some settings while the opposite trend has been observed in other settings [15, 20, 23, 45]. In our study area, over 80% of human exposure to vector bites occurred indoors for LLIN users. For children under 5 years who use LLINs, the exposure rate occurring indoors reached 90%. Therefore, these results suggest that adding other indoor intervention such as indoor residual spraying (IRS) to LLINs would be relevant to reduce malaria transmission in the rural area of

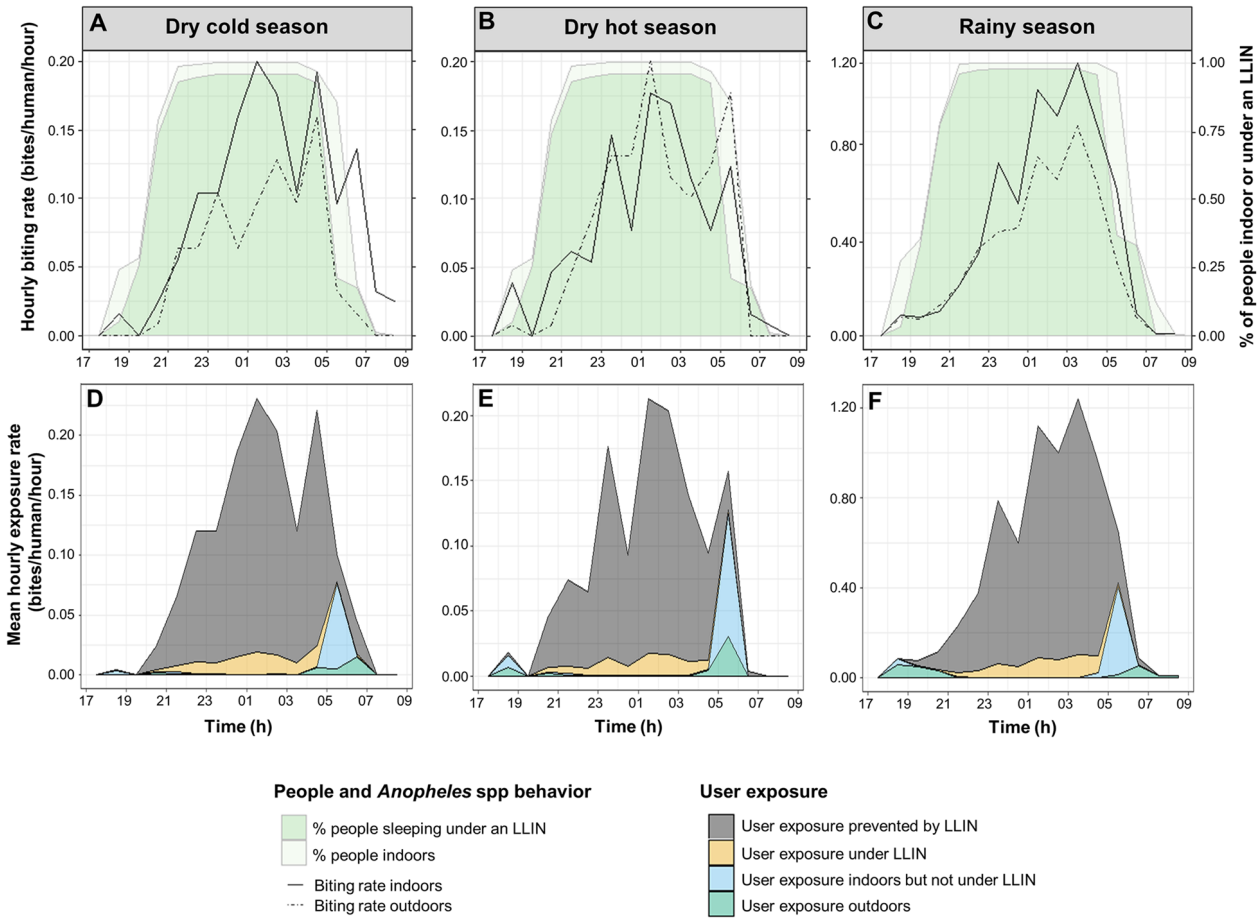


Fig. 3 Hourly human and *Anopheles* spp. behavior (**a, b, c**) and hourly exposure to bites of LLIN users (**d, e, f**), Burkina Faso. *Human behavioural data in panel A and B are the same (only one dry season survey) but plotted with different entomological data*

Diébougou. In 2017, 28 countries in the world have implemented IRS in combination with LLINs to combat malaria [2]. IRS contributed to an estimated 10 (5–14)% of the reduction in malaria burden achieved recently [1]. When used together, IRS and LLINs are expected to target vectors at different stages of their gonotrophic cycle using insecticides with different mode of action. However, trials assessing the impact of the combination IRS + LLIN over LLIN use alone have yielded conflicting results [46–51]. House improvement is another indoor measure which needs careful consideration and deep investigations. Indeed, house improvement has been strongly associated with reduced malaria transmission and disease in many studies [52–54]. The main house improvement interventions studied are closed eaves, closed ceilings, window screens and metal-roof houses as opposed to eaves, ceilings, windows openings and thatched-roof houses. Such improvements protect against malaria by providing physical barriers that prevent vectors from entering houses and can reduce vector survivorship [52, 55]. Nonetheless, there is compelling evidence that even a full coverage of effective measures within houses would not be sufficient to suppress transmission of malaria in Africa [56].

In this study, we evidenced that a significant proportion of LLIN users exposure to vector bites occurred outdoors (ranging from 9.48 to 30.43%), with the highest estimate recorded in adults (over the age of 18 years old) during the dry hot season. Many studies conducted in various areas of Africa reported similar or even higher estimates of exposure occurring outdoors [15, 16, 18, 45]. Recently, a systematic review categorized Burkina Faso along with Eritrea, Ethiopia, Gabon, and Tanzania as countries with high levels of outdoor vector biting [10]. However, our results do not fully support this categorization since we show that both LLIN users and LLIN non users are far more exposed to vector bites indoors than outdoors in the study area. Nevertheless, strategies targeting outdoor bites would probably be required to achieve malaria elimination in the area.

Almost all the existing indoor vector control strategies face two important evolutive challenges. First, they

induce a strong selective pressure on physiological resistance in vector populations because they almost all rely on synthetic chemicals [57]. Second, they also induced selective pressure for behavioral changes in vector populations resulting in a reduced contact with interventions [57]. In this context, there is a crucial need to monitor these resistance mechanisms, as well as residual transmission, after the deployment of strategies to inform decision makers in order to allow them to adapt their strategic plans.

Conclusions

This study showed that the use of LLINs prevented more than 80% of *Anopheles* bite exposure. Nevertheless, LLIN users are still exposed to vector bites which occurred mostly indoors in late morning. Therefore, complementary strategies targeting indoor biting vectors in combination with LLIN are expected to be the most efficient to control residual malaria transmission in this area.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-021-10304-y.

Additional file 1 Model specification. *Formulae used to calculate mean exposure to bite, true average personal protection efficacy of LLINs (P*), proportions of indoor (π_i and $\pi_{i,n}$), "before bed" (π_e and $\pi_{e,n}$) and "after bed" (π_m and $\pi_{m,n}$) exposure to bite.*

Additional file 2 LLIN Use rate per village. *N: number.*

Additional file 3 True average protection efficacy of LLINs against transmission and Proportions of indoors, early evening and late morning exposure to *Anopheles* bites per village. *NA: Not Applicable.*

Abbreviations

HBS: Human behaviours surveys; HLC: Human landing catch; ER: Endophagy Rate; IRS: Indoor Residual Spraying; LLIN: Long-Lasting Insecticidal Nets; ODK: Open Data Kit; NMCP: National Malaria Control Programs; WHO: World Health Organization

Acknowledgements

We acknowledge the Burkina Faso Ministry of Health, particularly Dr. Dembélé Henri and local medical team who facilitated the data collection. We thank all the villagers and local authorities for their kind collaboration throughout the study. Special thanks are due to Mr. Maiga Issouf for his strong commitment during human behavioural surveys. We are very grateful to Mr. Dahounto Amal for their substantial contributions and collaboration. We thank all the IRSS field staff for their assistance and the "Laboratoire Mixte International sur les Maladies à Vecteurs" (LAMIVECT) for their technical support. We also thank Mr. Ouattara Adama and Mr. Zoumenou Felix for their administrative support.

Authors' contributions

NM, RKD and DDS conceived and designed the study. DDS and SC collected the data. DDS and NM analyzed the data. DDS and BZ drafted the manuscript. NM, CP, PT, AS, LMB, GAO, AK and RKD reviewed the manuscript; all authors read and approved the final manuscript.

Funding

This work was part of the REACT project, funded by the French Initiative 5% – Expertise France (No. 15SANIN213).
The funding had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

The protocol of this study was reviewed and approved by the Institutional Ethics Committee of the Institut de Recherche en Sciences de la Santé (IEC-IRSS) and registered as N°A06/2016/CEIRES. We received community agreement before the beginning of human and *Anopheles* spp. behavioral surveys. Behavioral surveys did not involve participants under 16 years old. Indeed, questionnaires were administrated only to the heads of households and information relative to children under 16 years old were therefore directly collected from either a parent or guardian. Mosquito collectors were over 16 years old. All participants (to behavioral surveys or mosquito collections) gave their written informed consent. Mosquito collectors and supervisors received a vaccine against yellow fever as a prophylactic measure. Collectors were treated free of charge for malaria according to WHO recommendations.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institut de Recherche en Sciences de la Santé (IRSS), Bobo-Dioulasso, Burkina Faso. [2]Université Nazi Boni (UNB), Bobo-Dioulasso, Burkina Faso. [3]MIVEGEC, Univ. Montpellier, CNRS, IRD, Montpellier, France. [4]Institut Pierre Richet (IPR), Bouaké, Côte d'Ivoire. [5]Université d'Abomey Calavi, Abomey-Calavi, Benin.

Received: 21 November 2019 Accepted: 21 January 2021
Published online: 30 January 2021

References

1. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. Nature. 2015;526:207–11.
2. WHO. World malaria report 2018; 2018. p. 1–238. Available from: http://www.who.int/malaria/publications/world-malaria-report-2018
3. Dabiré KR, Diabaté A, Djogbenou L, Ouari A, Guessan RN, Ouédraogo J, et al. Dynamics of multiple insecticide resistance in the malaria vector. Malar J. 2008;9:1–9.
4. Toé KH, Jones CM, N'fale S, Ismai HM, Dabiré RK, Ranson H. Increased pyrethroid resistance in malaria vectors and decreased bed net effectiveness Burkina Faso. Emerg Infect Dis. 2014;20:1691–6.
5. Toé KH, N'Falé S, Dabiré RK, Ranson H, Jones CM. The recent escalation in strength of pyrethroid resistance in Anopheles coluzzi in West Africa is linked to increased expression of multiple gene families. BMC Genomics. 2015;16:1–11.
6. Ojuka P, Boum Y, Denoeud-Ndam L, Nabasumba C, Muller Y, Okia M, et al. Early biting and insecticide resistance in the malaria vector *Anopheles* might compromise the effectiveness of vector control intervention in southwestern Uganda. Malar J. 2015;14:1–8.
7. Moiroux N, Gomez MB, Pennetier C, Elanga E, Djènontin A, Chandre F, et al. Changes in *Anopheles funestus* biting behavior following universal coverage of long-lasting insecticidal nets in Benin. J Infect Dis. 2012;206:1622–9.
8. Fornadel CM, Norris LC, Glass GE, Norris DE. Analysis of *Anopheles arabiensis* blood feeding behavior in southern zambia during the two years after introduction of insecticide-treated bed nets. Am J Trop Med Hyg. 2010;83:848–53.
9. Ranson H, Lissenden N. Insecticide resistance in African Anopheles mosquitoes: a worsening situation that needs urgent action to maintain malaria control. Trends Parasitol. 2016;32:187–96 Available from: https://doi.org/10.1016/j.pt.2015.11.010.
10. Sherrard-Smith E, Skarp JE, Beale AD, Fornadel C, Norris LC, Moore SJ, et al. Mosquito feeding behavior and how it influences residual malaria transmission across Africa. Proc Natl Acad Sci U S A. National Academy of Sciences. 2019;116:15086-95.
11. WHO. Global strategy for dengue prevention and control 2012-2020. 2012.

12. Killeen GF, Tatarsky A, Diabate A, Chaccour CJ, Marshall JM, Okumu FO, et al. Developing an expanded vector control toolbox for malaria elimination. BMJ Glob Heal. 2017;2:1–9.
13. Barreaux P, Barreaux AMG, Sternberg ED, Suh E, Waite JL, Whitehead SA, et al. Priorities for broadening the malaria vector vontrol tool kit. Trends Parasitol. 2017;33:763–74.
14. Kelly-Hope LA, FE MK. The multiplicity of malaria transmission: A review of entomological inoculation rate measurements and methods across sub-Saharan Africa. Malar. J; 2009. p. 19.
15. Killeen GF, Kihonda J, Lyimo E, Oketch FR, Kotas ME, Mathenge E, et al. Quantifying behavioural interactions between humans and mosquitoes : evaluating the protective efficacy of insecticidal nets against malaria transmission in rural Tanzania. BMC Infect Dis. 2006;6:1–10.
16. Geissbühler Y, Chaki P, Emidi B, Govella NJ, Shirima R, Mayagaya V, et al. Interdependence of domestic malaria prevention measures and mosquito-human interactions in urban Dar Es Salaam. Tanzania Malar J. 2007;6:1–17.
17. Seyoum A, Sikaala CH, Chanda J, Chinula D, Ntamatungiro AJ, Hawela M, et al. Human exposure to anopheline mosquitoes occurs primarily indoors , even for users of insecticide- treated nets in Luangwa Valley, South-East Zambia. Parasit Vectors. 2012;5:1–10.
18. Huho B, Briët O, Seyoum A, Sikaala C, Bayoh N, Gimnig J, et al. Consistently high estimates for the proportion of human exposure to malaria vector populations occurring indoors in rural Africa. Int J Epidemiol. 2013;42:235–47.
19. Bayoh MN, Walker ED, Kosgei J, Ombok M, Olang GB, Githeko AK, Killeen GF, Otieno P, Desai M, Lobo NF, Vulule JM, Hamel MJ, SK and JEG. Persistently high estimates of late night, indoor exposure to malaria vectors despite high coverage of insecticide treated nets. Parasit Vectors. 2014;7:1–13.
20. Moiroux N, Damien GB, Egrot M, Djenontin A, Chandre F, Corbel V, et al. Human exposure to early morning *Anopheles funestus* biting behavior and personal protection provided by long-lasting insecticidal nets. PLoS One. 2014;9:8–11.
21. Cooke MK, Kahindi SC, Oriango RM, Owaga C, Ayoma E, Mabuka D, et al. "A bite before bed": exposure to malaria vectors outside the times of net use in the highlands of western Kenya. Malar J. 2015;14:1–15.
22. Bradley J, Lines J, Fuseini G, Schwabe C, Monti F, Slotman M, et al. Outdoor biting by *Anopheles* mosquitoes on Bioko Island does not currently impact on malaria control. Malar J. 2015;14:1–8.
23. Kamau A, Mwangangi JM, Rono MK, Mogeni P, Omedo I, Midega J, et al. Variation in the effectiveness of insecticide treated nets against malaria and outdoor biting by vectors in Kilifi, Kenya. Wellcome Open Res. 2017;2:1–56.
24. Finda MF, Moshi IR, Monroe A, Limwagu AJ, Nyoni AP, Swai JK, et al. Linking human behaviours and malaria vector biting risk in South-Eastern Tanzania. PLoS One. 2019;14:e0217414.
25. Pollard EJM, MacLaren D, Russell TL, Burkot TR. Protecting the peri-domestic environment: the challenge for eliminating residual malaria. Sci Rep. 2020; 10:1–9.
26. Monroe A, Moore S, Koenker H, Lynch M, Ricotta E. Measuring and characterizing night time human behaviour as it relates to residual malaria transmission in sub - Saharan Africa : a review of the published literature. Malar J. 2019;18:1–12. Available from. https://doi.org/10.1186/s12936-019-2638-9.
27. Monroe A, Msaky D, Kiware S, Tarimo BB, Moore S, Haji K, et al. Patterns of human exposure to malaria vectors in Zanzibar and implications for malaria elimination efforts. Malar J. 2020;19:1–14. Available from. https://doi.org/10.1186/s12936-020-03266-w.
28. Thomsen EK et al. Mosquito behaviour change after distribution of bednets results in decreased protection against malaria exposure. J Infect Dis. 2017; XX:1–8.
29. Govella NJ, Maliti DF, Mlwale AT, Masallu JP, Mirzai N, Johnson PCD, et al. An improved mosquito electrocuting trap that safely reproduces epidemiologically relevant metrics of mosquito human-feeding behaviours as determined by human landing catch. Malar J. 2016;15:1–17.
30. Sougoufara S, Thiaw O, Cailleau A, Diagne N, Harry M, Bouganali C, et al. The impact of periodic distribution campaigns of long-lasting insecticidal-treated bed nets on malaria vector dynamics and human exposure in Dielmo, Senegal. Am J Trop Med Hyg. 2018;98:1343–52.
31. Soma DD, Zogo BM, Somé A, Tchiekoi BN, de Sales Hien DF, Pooda HS, et al. *Anopheles* bionomics, insecticide resistance and malaria transmission in Southwest Burkina Faso: a pre-intervention study. PLoS One. 2020;15:1–21.
32. INSD. Tableau de bord économique et social 2014 de la région du Sud Ouest. 2015.
33. INSD. Enquête nationale sur le secteur de l'orpaillage (ENSO). 2017.
34. Mattingly P, F P, Rageau J. Contributions de la faune des moustiques du Sud-Est Asiatique: 12. Clés illustrées des genres de moustiques. Contrib Am Entomol Inst. 1971;7:1–86.
35. Gillies M, Coetzee M. A supplement to the Anophelinae of Africa South of the Sahara (Afrotropical Region). South African Inst Med Res. 1987;143:1-143.
36. Koekemoer LL, Kamau L, Hunt RH, Coetzee M. A cocktail polymerase chain reaction assay to identify members of the *Anopheles funestus* (*Diptera: Culicidae*) group. Am J Trop Med Hyg. 2002;66:804–11.
37. Cohuet A, Simard F, Berthomieu A, Raymond M, Fontenille DWM. Isolation and characterization of microsatellite DNA markers in the malaria vector *Anopheles funestus*. Mol Ecol Notes. 2002;2:498–500.
38. Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z. Della Torre a. insertion polymorphisms of SINE200 retrotransposons within speciation islands of *Anopheles gambiae* molecular forms. Malar J. 2008;7:1–10 Available from: http://www.malariajournal.com/content/7/1/163.
39. The R Development Core Team. R : A Language and Environment for Statistical Computing; 2008. p. 1–2547. Available from: http://www.gnu.org/copyleft/gpl.html
40. Olapeju B, Choiriyyah I, Lynch M, Acosta A, Blaufuss S, Filemyr E, et al. Age and gender trends in insecticide-treated net use in sub-Saharan Africa: a multi-country analysis. Malar J. 2018;17:423.
41. WHO. World malaria report 2017. World heal. Organ. 2017. Available from: http://www.who.int/malaria/publications/world-malaria-report-2017
42. Krezanoski PJ, Bangsberg DR, Tsai AC. Quantifying bias in measuring insecticide-treated bednet use: meta-analysis of self-reported vs objectively measured adherence. J Glob Health. Edinburgh University Global Health Society. 2018;8:1–11.
43. Koudou BG, Malone D, Hemingway J. The use of motion detectors to estimate net usage by householders, in relation to mosquito density in central Cote d'Ivoire: Preliminary results. Parasites and Vectors. 2014;7:1-6.
44. Krezanoski PJ, Santorino D, Agaba A, Dorsey G, Bangsberg DR, Carroll RW. How are insecticide-treated bednets used in ugandan households? A comprehensive characterization of bednet adherence using a remote monitor. Am J Trop Med Hyg. 2019;101:404–11.
45. Russell TL, Govella NJ, Azizi S, Drakeley CJ, Kachur SP, Killeen GF. Increased proportions of outdoor feeding among residual malaria vector populations following increased use of insecticide-treated nets in rural Tanzania. Malar J. 2011;10:1–10 Available from: http://www.malariajournal.com/content/10/1/80.
46. West PA, Protopopoff N, Wright A, Kivaju Z, Tigererwa R, Mosha FW, et al. Indoor residual spraying in combination with insecticide-treated nets compared to insecticide- treated nets alone for protection against malaria : a cluster randomised trial in Tanzania. PLoS Med. 2014;11:1–12.
47. Protopopoff N, Mosha JF, Lukole E, Charlwood JD, Wright A, Mwalimu CD, et al. Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two fact. Lancet. 2018;391:1577–88. Available from:. https://doi.org/10.1016/S0140-6736(18)30427-6.
48. Kafy HT, Ismail BA, Mnzava AP, Lines J, Abdin MSE, Eltaher JS, et al. Impact of insecticide resistance in *Anopheles arabiensis* on malaria incidence and prevalence in Sudan and the costs of mitigation. Proc Natl Acad Sci. 2017; 114:E11267–75.
49. Corbel V, Akogbeto M, Damien GB, Djenontin A, Chandre F, Rogier C, et al. Combination of malaria vector control interventions in pyrethroid resistance area in Benin: a cluster randomised controlled trial. Lancet Infect Dis. 2012;12:617–26.
50. Pinder M, Jawara M, Jarju LBS, Salami K, Jeffries D, Adiamoh M, et al. Efficacy of indoor residual spraying with dichlorodiphenyltrichloroethane against malaria in Gambian communities with high usage of long-lasting insecticidal mosquito nets: a cluster-randomised controlled trial. Lancet. 2015;385:1436–46.
51. Loha E, Deressa W, Gari T, Balkew M, Kenea O, Solomon T, et al. Long-lasting insecticidal nets and indoor residual spraying may not be sufficient to eliminate malaria in a low malaria incidence area: results from a cluster randomized controlled trial in Ethiopia. Malar J. 2019;18:1–15.
52. Tusting LS, Ippolito MM, Willey BA, Kleinschmidt I, Dorsey G, Gosling RD, et al. The evidence for improving housing to reduce malaria: A systematic review and meta-analysis. Malar. J. 2015;14:1-12.

53. Rek JC, Alegana V, Arinaitwe E, Cameron E, Kamya MR, Katureebe A, et al. Rapid improvements to rural Ugandan housing and their association with malaria from intense to reduced transmission: a cohort study. Lancet Planet Heal. 2018;2:83-94.
54. Killeen GF, Govella NJ, Mlacha YP, Chaki PP. Suppression of malaria vector densities and human infection prevalence associated with scale-up of mosquito-proofed housing in Dar Es Salaam, Tanzania: re-analysis of an observational series of parasitological and entomological surveys. Lancet Planet Heal. 2019;3:e132–43.
55. Lindsay SW, Jawara M, Mwesigwa J, Achan J, Bayoh N, Bradley J, et al. Reduced mosquito survival in metal-roof houses may contribute to a decline in malaria transmission in sub-Saharan Africa. Sci Rep. Nat Publ Group. 2019;9:7770.
56. WHO. Control of residual malaria parasite transmission. WHO Media Cent. 2014;11:1–5 Available from: http://www.who.int/malaria/publications/atoz/technical-note-control-of-residual-malaria-parasite-transmission-sep14.pdf.
57. Lies D, Marc C. Residual transmission of malaria: an old issue for new approaches. Anopheles mosquitoes - New insights into malaria vectors. Intech Open Science. 2013; 21: 671–704. Available from: http://www.intechopen.com/books/anopheles-mosquitoes-new-insights-into-malaria-vectors.

Publisher's Note

Annexe F

Texte intégral de l'article complémentaire n°4

La référence de la publication est la suivante :

Bationo, C.S., Gaudart, J., Dieng, S., Cissoko, M., ***Taconet, P****. et al. Spatio-temporal analysis and prediction of malaria cases using remote sensing meteorological data in Diébougou health district, Burkina Faso, 2016–2017. Sci Rep 11, 20027 (2021).* https://doi.org/10.1038/s41598-021-99457-9

scientific reports

Check for updates

OPEN

Spatio-temporal analysis and prediction of malaria cases using remote sensing meteorological data in Diébougou health district, Burkina Faso, 2016–2017

Cédric S. Bationo[1,2,6], Jean Gaudart[3,4]✉, Sokhna Dieng[1], Mady Cissoko[1,4], Paul Taconet[2,6], Boukary Ouedraogo[5], Anthony Somé[6], Issaka Zongo[6], Dieudonné D. Soma[2,6,7], Gauthier Tougri[8], Roch K. Dabiré[6], Alphonsine Koffi[9], Cédric Pennetier[2,6,9] & Nicolas Moiroux[2,6]

Malaria control and prevention programs are more efficient and cost-effective when they target hotspots or select the best periods of year to implement interventions. This study aimed to identify the spatial distribution of malaria hotspots at the village level in Diébougou health district, Burkina Faso, and to model the temporal dynamics of malaria cases as a function of meteorological conditions and of the distance between villages and health centres (HCs). Case data for 27 villages were collected in 13 HCs. Meteorological data were obtained through remote sensing. Two synthetic meteorological indicators (SMIs) were created to summarize meteorological variables. Spatial hotspots were detected using the Kulldorf scanning method. A General Additive Model was used to determine the time lag between cases and SMIs and to evaluate the effect of SMIs and distance to HC on the temporal evolution of malaria cases. The multivariate model was fitted with data from the epidemic year to predict the number of cases in the following outbreak. Overall, the incidence rate in the area was 429.13 cases per 1000 person-year with important spatial and temporal heterogeneities. Four spatial hotspots, involving 7 of the 27 villages, were detected, for an incidence rate of 854.02 cases per 1000 person-year. The hotspot with the highest risk (relative risk = 4.06) consisted of a single village, with an incidence rate of 1750.75 cases per 1000 person-years. The multivariate analysis found greater variability in incidence between HCs than between villages linked to the same HC. The time lag that generated the better predictions of cases was 9 weeks for SMI1 (positively correlated with precipitation variables) and 16 weeks for SMI2 (positively correlated with temperature variables. The prediction followed the overall pattern of the time series of reported cases and predicted the onset of the following outbreak with a precision of less than 3 weeks. This analysis of malaria cases in Diébougou health district, Burkina Faso, provides a powerful prospective method for identifying

[1]INSERM, IRD, SESSTIM, UMR1252, Institute of Public Health Sciences, ISSPAM, Aix Marseille Univ, 13005 Marseille, France. [2]CNRS, IRD, MIVEGEC, Univ. Montpellier, Montpellier, France. [3]INSERM, IRD, SESSTIM, UMR1252, Institute of Public Health Sciences, ISSPAM, APHM, Hop Timone, BioSTIC, Biostatistic & ICT, Aix Marseille Univ, 13005 Marseille, France. [4]Malaria Research and Training Center—Ogobara K. Doumbo (MRTC-OKD), FMOS-FAPH, Mali-NIAID-ICER, Université des Sciences, des Techniques et des Technologies de Bamako, Bamako 1805, Mali. [5]Direction des Systèmes d'information en Santé, Ministère de la Santé du Burkina Faso, Ouagadougou, Burkina Faso. [6]Institut de Recherche en Sciences de la Santé (IRSS), Bobo Dioulasso, Burkina Faso. [7]Institut Supérieur des Sciences de la Santé, Université Nazi Boni, Bobo-Dioulasso, Burkina Faso. [8]Programme National de Lutte Contre le Paludisme, Ministère de la Santé du Burkina Faso, Ouagadougou, Burkina Faso. [9]Institut Pierre Richet (IPR), Institut National de Santé Publique (INSP), Bouaké, Côte d'Ivoire. ✉email: jean.gaudart@univ-amu.fr

and predicting high-risk areas and high-transmission periods that could be targeted in future malaria control and prevention campaigns.

Abbreviations

WHO	World Health Organization
CI	Confidence interval
IPT	Intermittent preventive treatment
RDT	Rapid diagnostic test
SMC	Seasonal malaria chemoprevention
LLIN	Long-lasting insecticidal net
IRS	Indoor residual spraying
HC	Health Centre
SMI	Synthetic meteorological indicator
PCA	Principal component analysis
GAM	Generalized additive model
GLM	Generalized linear model
UBRE	Unbiased risk estimator
GAMM	Generalized additive mixed model
SIR	Standardized incidence ratio
RR	Relative risk
SD	Standard deviation

Malaria is one of the most life-threatening diseases and poses a great socio-economic burden worldwide[1]. According to World Health Organization (WHO) estimates, the global number of malaria cases was 229 million in 2019 compared to 251 million in 2010 and 214 million in 2015[1]. Although the estimated number of cases decreased by 23 million from 2010 to 2018, data for the period 2015–2018 highlight the lack of significant progress during this period. In 2018, the WHO African Region accounted for most cases (200 million or 93% of all cases), far ahead of the WHO South-East Asian region (3.4%) and the WHO Eastern Mediterranean Region (2.1%)[1]. At the time, nearly 80% of global malaria deaths were concentrated in 17 countries of the WHO African Region and in India. The WHO estimates that Burkina Faso carries about 6% of the global malaria burden[1]. Statistical data from the Ministry of Health of Burkina Faso for the year 2018 show that malaria is the second reason for consultation (31.7%), and that pregnant women and children under 5 years are the most at risk of contracting malaria[2]. According to those data, the average parasite prevalence in children under 5 years was 17% in 2017–2018[2,3]. In 2018, the number of confirmed cases reported in health facilities was 11,624,595 of which 4.14% were severe forms and 2.8% resulted in death.

The National Malaria Control Program in Burkina Faso recommends the following control strategies[4]: early case management in health facilities and at the community level, with a particular focus on children aged 3 to 59 months[5]; intermittent preventive treatment (IPT) for pregnant women; universal access to rapid diagnostic tests (RDTs) and artemisinin-based combination therapies; seasonal malaria chemoprevention (SMC) for children under 5 years; and vector control using long-lasting insecticidal nets (LLINs), indoor residual spraying (IRS), larval control, and environmental sanitation.

For strategic reasons or lack of resources, not all of these strategies are optimally implemented everywhere and all the time. Thus, in 2018, 25% of households reported not owning an LLIN (with coverage varying between 58 and 87% depending on the region) and 42% of pregnant women did not receive the recommended three doses of IPT, as reported by the Burkina Faso Malaria Indicator Survey[6].

At the same time, new tools and strategies are being developed, including administration of ivermectin, bi-impregnated nets, transmission-blocking vaccines, and conventional vaccines[4,7,8]. In Burkina Faso, the REACT project ("Insecticide resistance management in Burkina Faso and Côte d'Ivoire: A study on vector control strategies") conducted in 2016–2018 aimed to evaluate the efficacy of strategies designed to complement LLINs, namely pirimiphos methyl-based IRS, enhanced communication, and administration of ivermectin to domestic animals.

Malaria control and prevention programs are more efficient and cost-effective when they target high-risk spatial clusters (hotspots)[9] or when they select the best times of year[10] to initiate interventions (e.g. SMC or LLIN distribution). Indeed, as numerous studies have shown, malaria incidence at the local level is heterogeneous and associated with spatio-temporal clusters[11–13] that are likely to maintain transmission during low-risk periods and, consequently, to increase transmission during high-risk periods[14–16]. Identifying these clusters can therefore help to improve the fight against malaria and to anticipate future outbreaks.

This study aimed to identify the spatial distribution of malaria hotspots at the village level in Diébougou health district, Burkina Faso, and to model the temporal dynamics of malaria cases as a function of meteorological conditions and of the Euclidean distance between villages and their corresponding health centres (HCs). Data on malaria cases were obtained through HC-based passive case detection for the 27 villages included in the REACT project.

Materials and methods

Study area. The study was conducted in 27 villages of Diébougou health district that were included in the REACT project. All included villages met two criteria: a population between 200 and 500 inhabitants and a Euclidean distance of at least 2 km from the nearest village. A population census carried out in July 2016 by our

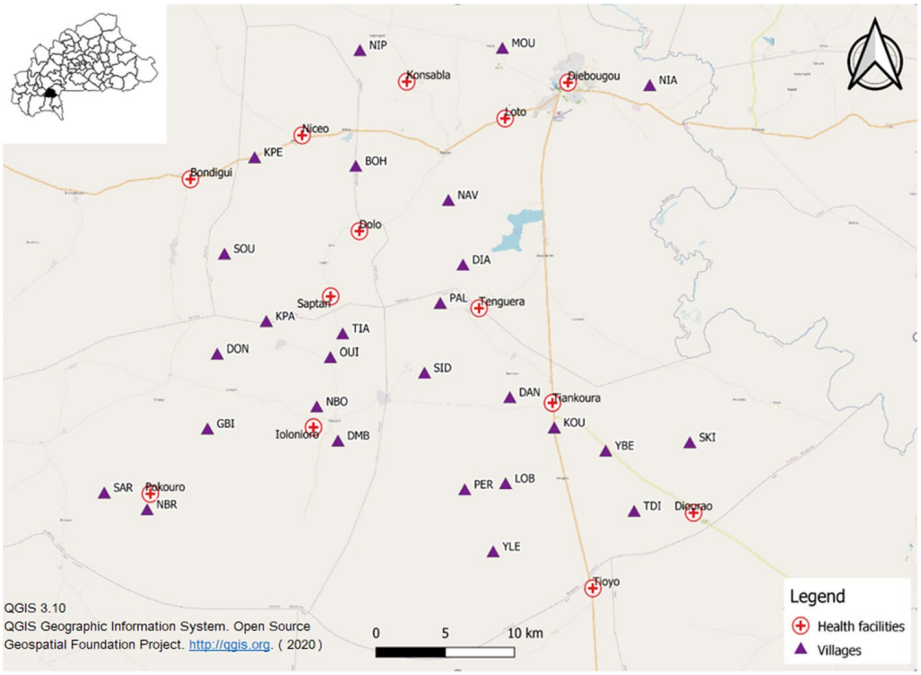


Figure 1. Map of the study area showing the location of villages (triangles) and health centres (red crosses). Background: OpenStreetMaps. *MOU* moule, *NIA* niaba, *NIP* nipodja, *BOH* bohero, *NBR* niombripo, *SAR* sarambour, *NBO* niombouna, *GBI* gongombiro, *KPA* kpalbalo, *DON* dontelo, *SID* sidmoukar; *DMB* dombouro, *OUI* ouidiaro, *TIA* tiakiro, *NAV* nouvielgane, *DIA* diagnon, *PAL* palembera, *KPE* kpedia, *SOU* soussoubro, *TDI* tordiero, *YLE* yellela, *YBE* yelbelela, *SKI* sinkiro, *DAN* dangbara, *KOU* kouloh, *LOB* lobignonao, *PER* perglembiro.

research team found that the 27 villages were home to 7408 inhabitants. The villages were linked to 13 HCs. Villages and HCs were geo-referenced using *Global Positioning System* (GPS) (Fig. 1).

Diébougou health district is located in South-Western Burkina Faso, a region characterized by a tropical climate with a dry season from October to April and a rainy season from May to September. The dry season is divided into a cold dry season lasting from December to February and a hot dry season lasting from March to April. Average daily minimum and maximum temperatures in the cold dry, hot dry, and rainy seasons are 18 and 36 °C, 25 and 39 °C, and 23 and 33 °C, respectively. Average annual rainfall is 1200 mm. The natural vegetation is dominated by wooded savannah dotted with clear forest gallery[17,18]. Let's remind that Burkina Faso is spread over 3 climatic zones: in the north, (Sahelian zone), rainfall is less than 600 mm/year. While the centre (or northern Sudanese) zone receives 600–900 mm/year, rainfall in the southern (or southern Sudanese, where Diebougou is located) zone exceeds 900 mm/year.

Passive case detection. Case data for the 27 villages included in the REACT project were collected using continuous HC-based passive case detection during 2016 and the first 36 weeks of 2017, which corresponded to the period preceding the implementation of the interventions studied (i.e. pirimiphos methyl-based IRS, enhanced communication, and administration of ivermectin to domestic animals). Specifically, consultation data for village residents were retrieved from each HC registries and recorded using tablets equipped with Open Data Kit collect forms. A malaria case was defined as a person who presented with fever and had a positive RDT result.

Study period. Of the 88 weeks of data collection, 52 weeks corresponding to an epidemic year (a complete malaria epidemic) were considered for spatio-temporal analysis. The epidemic year ran from week 20 (in May) of 2016 to week 19 (in May) of 2017 (Fig. 2).

Meteorological data. The meteorological data used in this study were drawn from the Era-5 dataset from 2016 to 2017[19] published by the European Centre for Medium-Range Weather Forecasts, which provides hourly estimates of several atmospheric and land parameters at a spatial resolution of 0.25°[20]. These data were aggregated into weekly counts. The meteorological variables included in the analysis were: Weekly rainfall (mm), number of rainy days per week, weekly mean of daily average temperature (°C), weekly mean of daily minimum temperature (°C), weekly mean of daily maximum temperature (°C), weekly mean of daily average wind speed (km/h), weekly mean of daily average relative humidity (%), weekly mean of daily average atmospheric pressure (hPa), weekly mean of daily average cloud cover (%), and weekly mean of daily thermal amplitude (°C) (Table 1).

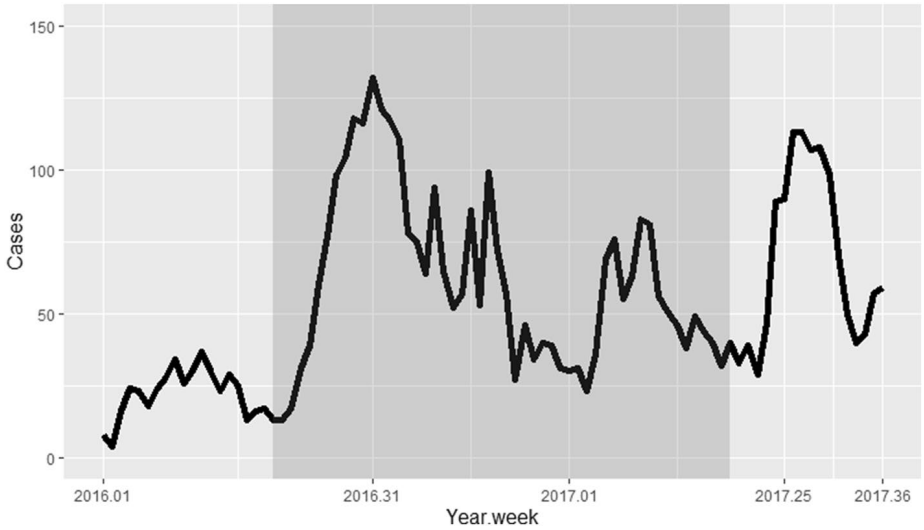


Figure 2. Time series of the number of malaria cases collected through passive case detection. The shaded (dark) area represents the epidemic year considered for analysis.

Var.	Mean	St. dev	Min	Pctl (25)	Median	Pctl (75)	Max
Tmean: weekly mean of daily average temperature (° C)	28.7	2.2	25.7	26.6	28.8	30.2	33.9
Re: number of rainy days per week	4.8	2.9	0	2	7	7	7
P: weekly mean of daily average atmospheric pressure (hPa)	975.3	1.6	972.3	974.0	975.5	976.7	978
Cl: weekly mean of daily average cloud cover (%)	0.6	0.2	0.1	0.4	0.6	0.7	0.9
W: weekly mean of daily average wind speed (km/h)	7.6	3.6	2.8	4.7	7.2	9.5	17.4
H: weekly mean of daily average relative humidity (%)	0.5	0.2	0.1	0.3	0.6	0.8	0.8
Tmax: weekly mean of daily maximum temperature (°C)	34.4	3.5	28.6	31.2	35.4	36.6	40.7
Tmin: weekly mean of daily minimum temperature (°C)	23.3	1.8	19.1	22.4	23.0	24.2	27.0
Rc: weekly rainfall (mm)	14.4	18.3	0.0	0.005	5.0	24.5	69.8
tvar: weekly mean of daily thermal amplitude (°C)	11.1	3.4	5.9	7.7	10.6	14.4	17.1

Table 1. List of meteorological variables with their abbreviations and descriptive statistics. *Var* variables, *St. Dev* standard deviation, *Min* minimum, *Pctl(25)* first quartile, *Pctl(75)* third quartile, *Max* maximum.

Data analysis and statistical methods. *Meteorological data.* To reduce the number of variables and avoid collinearity, we constructed synthetic meteorological indicators (SMIs) using a principal component analysis (PCA) of weekly meteorological variables. Different cofactors, affected by collinearity (such as rainfall, vegetation and temperature), are known to impact the parasitological cycle at different steps. Using PCA allowed keeping the main environmental characteristics without losing part of the environmental cofactors associated with malaria. Principal components that met Kaiser's criterion[21] were selected as SMIs and included in the temporal analysis.

Spatial analyses. Hotspots, i.e. high-risk clusters, were detected using the Kulldorf scanning method[22] with a Monte Carlo algorithm in a purely spatial analysis. The Kulldorf scanning method helps to identify spatial clusters based on geographical coordinates and to avoid the problem of multiple non-independent tests[22]. We defined clusters as aggregates of cases with observed values higher than expected (i.e. unlikely to have been obtained by chance). The p-value (i.e. the probability, under the null hypothesis, that the expected number of cases is the same or higher than the observed number of cases) was calculated for each cluster.

Scan parameters were: elliptical window, non-overlapping clusters, maximum cluster size set at 50% of the population at risk, Monte Carlo replication number set at 999.

Temporal analyses. Lagged SMI selection. Several studies have observed a lag between malaria time series and meteorological data time series[14–16]. In view of this, we decided to investigate the time lag (in weeks) between the time series of weekly malaria cases and the time series of SMIs. Using a generalized additive model (GAM) with a negative binomial distribution and a smoothing spline function, we modelled the time series of total malaria cases (for all villages) as a function of each SMI for time lags ranging from 1 to 30 weeks (thus generating 30 models per SMI). The GAM is an extension of the generalized linear model (GLM): while it includes random effects in the predictor like the GLM does, it can be used with nonparametric smoothing terms instead of

constant parameters[23–25]. The usefulness of the GAM lies in the fact that it provides a flexible method to identify the effects of non-linear covariates in exponential family distributions and in likelihood-based methods[26–28]. However, instead of estimating a single parameter, the GAM provides an unspecified (non-parametric) general function that compares predicted response values to predictor values.

We compared the 30 models generated for each SMI using the unbiased risk estimator (UBRE), i.e. an unbiased estimate of the mean square error of a non-linear biased estimator. For each SMI, the time lag associated with the best model (i.e. with the lowest UBRE) was selected for the multivariate analysis.

Multivariate time analysis. To account for the non-linearity of the relationship between the response and predictor variables, we analysed the time series of weekly cases reported in all villages during the epidemic year using a generalized additive mixed model (GAMM). To account for the non-independence of data from the same village or HC, we fitted this model with nested random intercepts for villages and HCs. The GAMM accounted for space and time processes, by using a Gaussian field and an auto-regressive model. The Gaussian field with a negative exponential variogram accounted for the spatial auto-correlation. The first-order auto-regressive temporal auto-correlation structure was introduced, within the random part of the mixed model, to account for the temporal auto-correlation of malaria incidence. We analysed the time series of cases using selected lagged SMIs (with a smoothing spline function) and of the Euclidean distance between villages and their corresponding HCs as predictors. For each predictor, the standardized incidence ratio (SIR) was estimated by modelling the log-transformed population as the offset.

To account for the non-linearity of the relationship between the response and predictor variables, we calculated SIRs according to the deciles of the distribution of values for each predictor. Indeed, SIRs cannot be calculated with GAMs as they are with GLMs, because when the relationship between the response and the predictor is non-linear, SIRs are not constant across the range of values of the predictor[24,28].

Lastly, we tested the multivariate model fitted with data from the epidemic year to predict the number of cases in both 2016 and 2017.

Software and packages. Statistical analyses were performed using R software (version 3.6.1)[29]. The PCA was performed using the PCA function in the *FactoMineR* package[30]. The GAMs and the GAMM were generated using the "gam" and "gamm" functions in the *mgcv* package, respectively[26–28]. Data overdispersion was tested using the "dispersiontest" function in the *AER* package[31]. The spatial analysis was performed using SatScan™ software (version 9.6). Maps were produced using QGIS software (version 3.10)[32].

Ethics approval and consent to participate. The protocol of this study was reviewed and approved by the Institutional Ethics Committee of the Institut de Recherche en Sciences de la Santé (IEC-IRSS) and registered as No A06/2016/CEIRES and all the methods were performed in accordance with the guidelines and regulations stated in the protocol. Informed consent was obtained from all subjects and/or their legal guardian(s).

Results

Descriptive analysis. A total of 3179 malaria cases were reported in HCs during the epidemic year, corresponding to an incidence of 429.13 cases per 1000 person-years. On average, 61.13 cases per week were reported, with a peak of 132 cases in week 31 of 2016 (week 1 of August; Fig. 2). The curve of cases over the epidemic year shows two peaks (Fig. 2): a very pronounced peak between weeks 27 and 45 of 2016 (August to November), which accounted for 60% of cases, and a less pronounced peak between weeks 7 and 11 of 2017 (mid-February to the end of March), which accounted for 12% of cases.

Synthetic meteorological indicators. The PCA conducted using Kaiser's criterion led us to construct and retain two SMIs that explained 85.4% of the total inertia (Fig. 3A).

The first SMI (i.e. the first principal component) explained 52.9% of the total inertia. The variables that most contributed to this SMI, henceforth called SMI1, were mainly correlated with precipitation variables: weekly mean of daily thermal amplitude (18.24%, correlation coefficient $r = -0.98$), weekly mean of daily average relative humidity (17.74%, $r = 0.96$), weekly rainfall (14. 5%, $r = 0.8$), weekly mean of daily average cloud cover (13.32%, $r = 0.83$), number of rainy days per week (12.47%, $r = 0.81$), and weekly mean of daily maximum temperature (12.03%; $r = -0.79$) (Fig. 3B). The second SMI (i.e. the second principal component) explained 32.5% of the total inertia. The variables that most contributed to this SMI, henceforth called SMI2, were mainly correlated with temperature variables: weekly mean of daily minimum temperature (25.83%; $r = 0.91$), weekly mean of daily average temperature (24.72%; $r = 0.89$), weekly mean of daily maximum temperature (10.51%; $r = 0.58$), and weekly mean of daily average atmospheric pressure (19.6%, $r = -0.79$) (Fig. 3C).

The values of SMI1 were positive between late June and early October, which corresponds to the rainy season (Fig. 4). The values of SMI2 were positive between mid-February and mid-June, which corresponds to the hot dry season (March–June), and between October and November, which corresponds to the transition period between the rainy season and the dry season. Both SMIs were negative throughout the cold dry season (December–mid-February) (Fig. 4).

Spatial analysis. The spatial analysis allowed us to identify and map malaria hotspots for the epidemic year from week 20 (in May) of 2016 to week 19 (in May) of 2017. Four hotspots were detected that accounted for 1685 cases in 1973 inhabitants, i.e. an average incidence rate of 854.02 cases per 1000 person-years (Fig. 5). These hotspots were mainly located in the southern and central parts of the study area. The hotspot with the highest

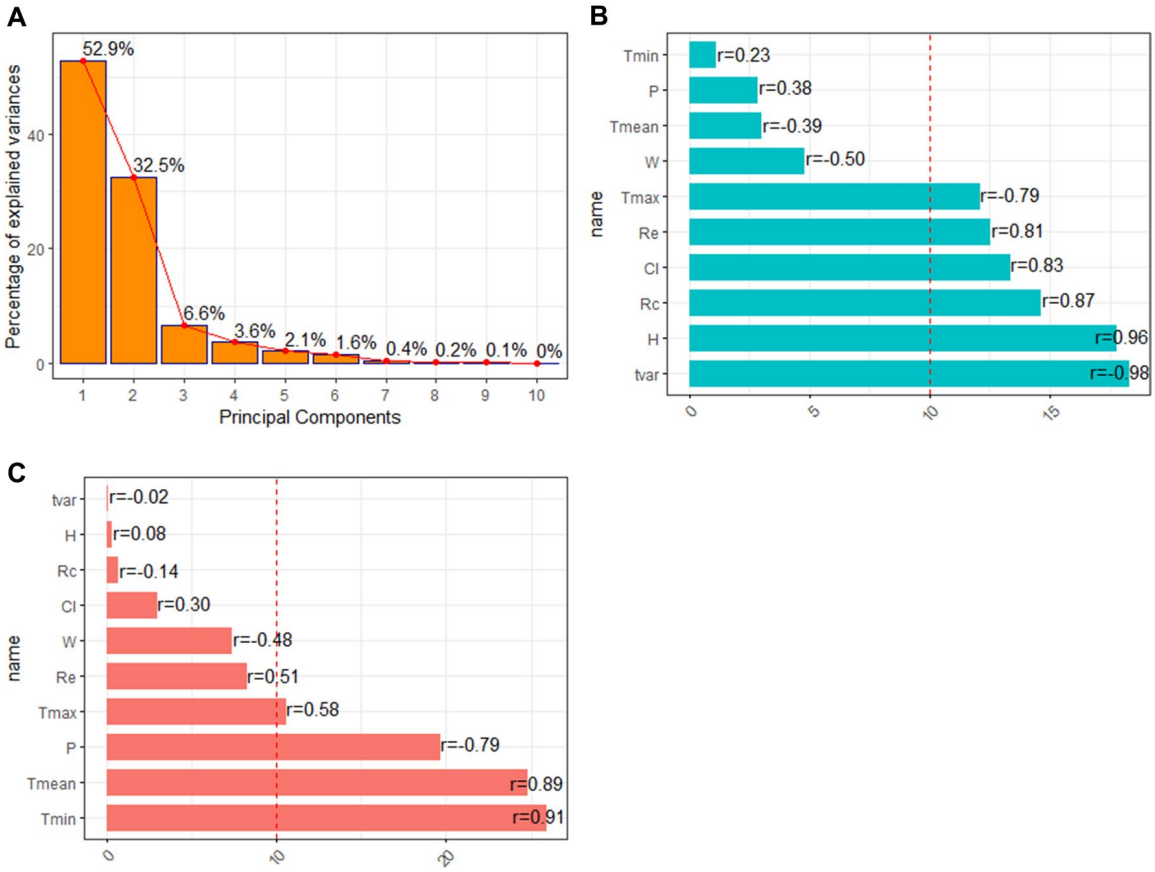


Figure 3. Principal Component Analysis of meteorological variables. Percentage of inertia explained by each principal component (**A**). Contribution of meteorological variables to the first principal component (SMI1; (**B**)) and the second principal component (SMI2; (**C**)). *SMI* synthetic meteorological indicator, *r* correlation coefficient between the meteorological variable and the SMI. Abbreviations of variable names are detailed in Table 1. The dashed line represents the contribution that would have been expected if all variables had contributed equally to the SMI.

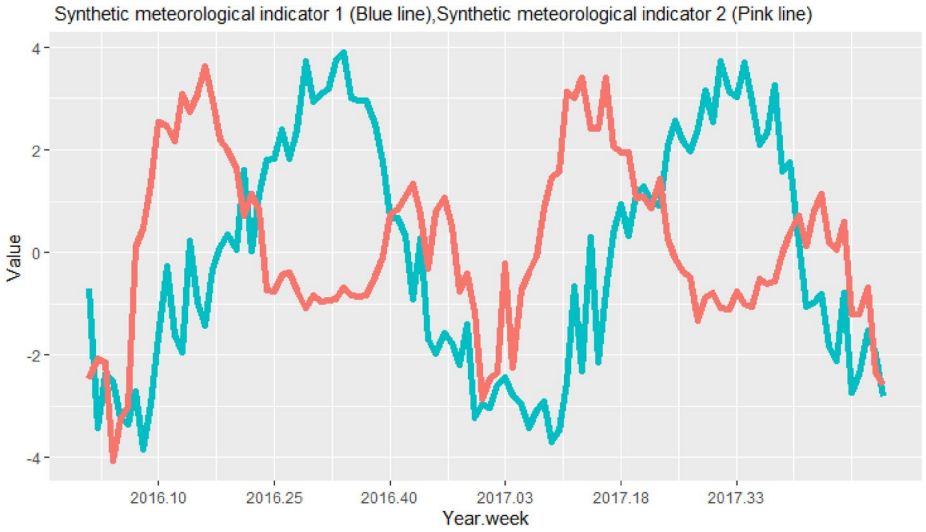


Figure 4. Time series of synthetic meteorological indicator 1 (mainly correlated with precipitation variables) and synthetic meteorological indicator 2 (mainly correlated with temperature variables) from 2016 to 2017.

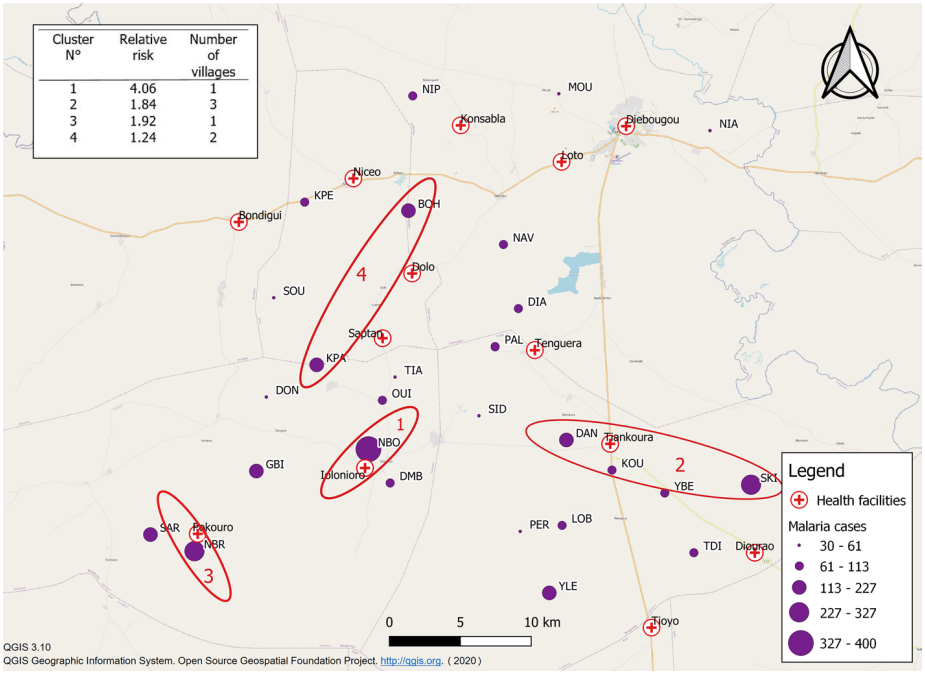


Figure 5. Map of malaria cases detected in 27 villages of Diébougou health district, Burkina Faso, for the epidemic year 2016–2017; hotspots identified with the Kulldorf scanning method. Background: OpenStreetMaps.

risk (hotspot 1; relative risk (RR) = 4.06, $p < 0.0001$) consisted of a single village (Niombouna) and accounted for 400 cases for 228 inhabitants, i.e. an incidence rate of 1,750.75 cases per 1000 person-years. The second hotspot (hotspot 2; RR = 1.84, $p < 0.0001$) was made up of three villages (Sinkiro, Yelbelela, and Dangbara) and accounted for 604 cases for 753 inhabitants, i.e. an incidence rate of 802.12 cases per 1000 person-years. The third hotspot (hotspot 3; RR = 1.92, $p < 0.0001$) was made up of a single village (Niombripo) and accounted for 326 cases for 376 inhabitants, i.e. an incidence rate of 867.02 cases per 1000 person-years. The fourth hotspot (hotspot 4; RR = 1.24, $p = 0.04$) consisted of two villages (Bohero and Kpalbalo) and accounted for 355 cases for 616 inhabitants, i.e. an incidence rate of 576.2 cases per 1000 person-years.

Temporal analysis. The time lag that generated the model with the lowest UBRE was 9 weeks for SMI1 and 16 weeks for SMI2.

The multivariate analysis found greater variability in incidence between HCs (standard deviation (SD) = 5.74) than between villages linked to the same HC (SD = 0.69). The coefficient of the temporal autocorrelation structure indicated the presence of temporal autocorrelation between cases (Phi = 0.32, 95% CI [0.20, 0.38]).

In the multivariate model, lagged SMI1 and lagged SMI2 were significantly associated with the number of malaria cases at the village level ($p < 0.001$ and $p < 0.001$, respectively). The relationship between the number of cases and SMI1 (consisting mainly of precipitation variables) was positive and almost linear (Fig. 6A) across the range of values. A positive non-linear relationship was observed for SMI2 (consisting mainly of temperature variables), with the number of cases increasing for SMI2 values above zero (Fig. 6C). Below zero, changes in SMI2 values did not influence the number of cases (Fig. 6C). The Euclidean distance between villages and their corresponding HCs was not correlated to the recorded malaria incidence ($p = 0.78$).

The evolution of SIRs as a function of SMI values is presented in Fig. 6. For SMI1, risk was constant over deciles 1 to 3 (SIR = 1.07, 95% CI [1.03, 1.10], [1.05, 1.08], and [1.06, 1.08], respectively), increased from decile 4 to 7, and then reached a plateau from decile 8 to 10 (SIR = 1.14 [1.14, 1.14]) (Fig. 6B). For SMI2, risk was constant over deciles 1 to 2 (SIR = 0.99, 95% CI [0.93, 1.05], and [0.98, 1.00], respectively), increased from decile 4 to 8 (SIR = 1.37 [1.36, 1.37]), and then decreased from decile 9 to 10 (Fig. 6D). For the distance, no evolution in the risk was displayed (Fig. 6F).

Prediction. The multivariate model generated for the epidemiological year was used to predict the number of cases in the 27 villages for all of 2016 and for the first 36 weeks of 2017. The resulting prediction was superimposed on the time series of reported cases for graphical analysis (Fig. 7). The prediction followed the overall pattern of the time series of reported cases but with a tendency for underestimation, especially during the second peak in early 2017. In addition, the model predicted the onset of the malaria outbreak for the 2017–2018 epidemic year with a delay of three weeks.

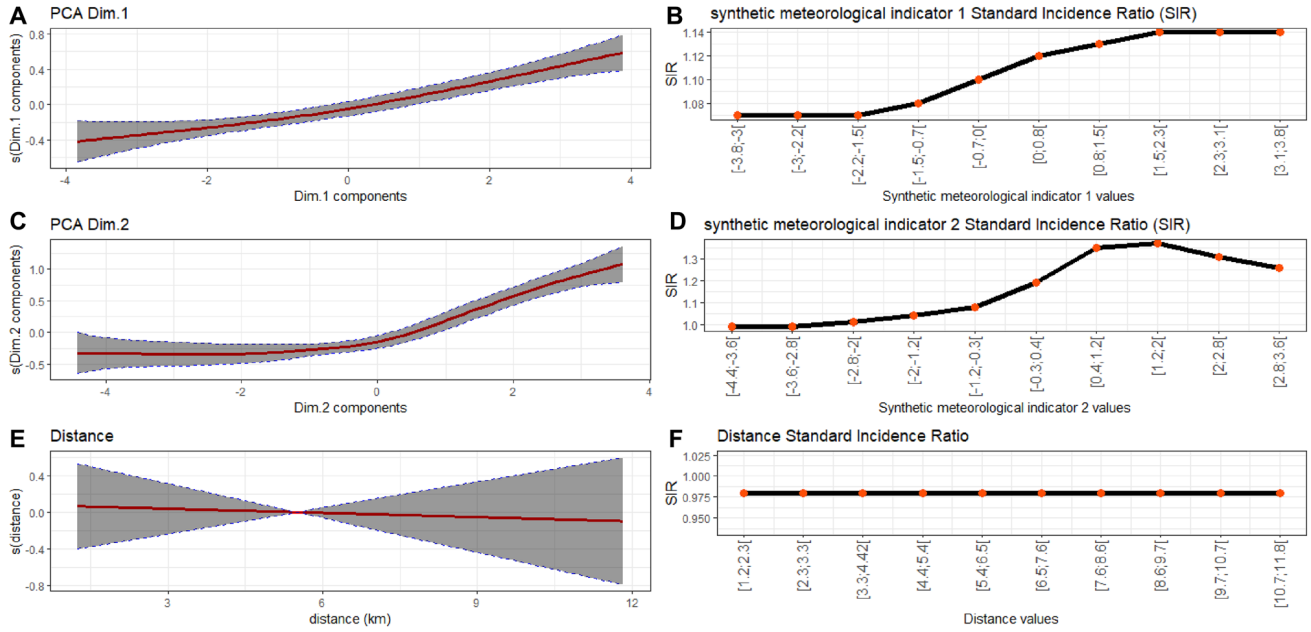


Figure 6. Relationship (red curve) between malaria cases and SMI1 (**A**), SMI2 (**C**), and Euclidean distance to health centre (**E**) with 95% confidence intervals (shaded area). Evolution of the standard incidence ratio (SIR) as a function of SMI1 (mainly correlated with precipitation variables) (**B**), SMI2 (mainly correlated with temperature variables) (**D**), and Euclidean distance to health centre (**F**).

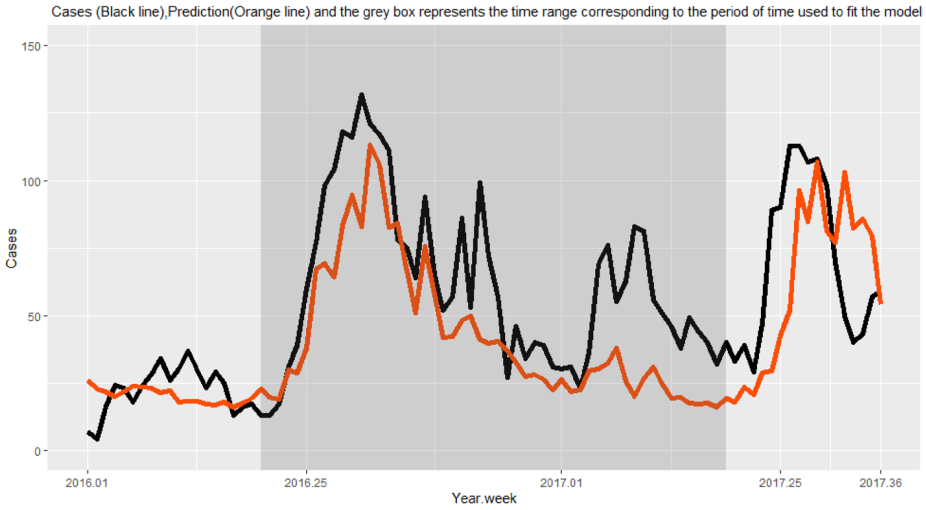


Figure 7. Cumulative number of reported cases (black line) and predicted cases (orange line) in 27 villages of South-Western Burkina Faso using a meteorological model.

Discussion

In this study, we analysed the spatio-temporal distribution of malaria cases in 27 villages of South-Western Burkina Faso.

The spatial analysis conducted using the Kulldorf scanning method helped to identify four malaria hotspots. The first three hotspots were located in the southern part of the study area and the last one was located in the central part, reflecting spatial heterogeneity in the distribution of cases. A comparison of the spatial distribution of these hotspots with that of mosquito vector density[33] showed no correlation between the two, leading us to conclude that the spatial heterogeneity of vector density does not explain the distribution of hotspots in our study area.

A number of studies have found an association between spatial inequalities in access to care and spatial heterogeneity of malaria incidence[34,35]. Yet, contrary to what has been reported elsewhere[36], we failed to found a correlation between the number of malaria cases and the Euclidean distance between villages and their corresponding HC. We used Euclidean distance because it is considered the simplest proxy for travel time, which is

considered a good measure of access to care. However, Euclidean distance may not have been the best option, as roads in our study area are in highly variable condition and some become impassable during the rainy season, with some villages left completely isolated. Future studies in the region should use better proxies for travel time in trying to explain the detected hotspots[36].

Since entomological factors and spatial inequalities in access to care failed to explain the distribution of hotspots in our study area, other potential explanatory factors should be investigated in the future, including socio-economic factors (level of education, income, professional activity, individual and societal behaviour, etc.)[37–40] and factors linked to LLIN usage[41–44]. Such investigations could help to explain in particular why the two hotspots composed of a single village (Niombripo and Niombouna) had a much higher incidence than neighbouring villages. Nevertheless, hotspot analyses like ours make it possible to identify, in a simple and cost-efficient manner, villages that can constitute priority areas for intervention. Indeed, studies conducted elsewhere have shown that targeting hotspots helps to reduce malaria transmission[45,46]. This strategy is appropriate in resource-limited countries like Burkina Faso as it allows for efficient allocation of prevention resources[14–16].

Our analysis of the temporal dynamics of malaria cases found a strong correlation between malaria incidence and two SMIs with specific time lags. These SMIs were constructed through a PCA of meteorological data derived from readily and rapidly available satellite imagery. The first SMI (SMI1: positively correlated with cumulative rainfall, humidity, cloud cover, and number of rainy days, and negatively correlated with thermal amplitude) corresponded to the rainy season, while the second (SMI2: positively correlated with temperature and negatively correlated with atmospheric pressure) corresponded to the warm periods preceding and following the rainy season. We found that SMI1 and SMI2 predicted the number of cases with a time lag of 9 and 16 weeks, respectively, which is consistent with studies carried out in Burkina Faso, Mali, and Ethiopia[14,47,48].

In our study, the relationship between rainfall (SMI1) and the number of cases was quasi-linear, as was the case in a study performed in the Ouagadougou area of Burkina Faso[14]. By contrast, two studies conducted in the Sahel region—one in Mali (Niger River Valley, Timbuktu region) and the other in Senegal (Bambey and Fatick Health Districts)—found a monotonic non-linear relationship between rainfall and malaria incidence[15,16]. The drop in the number of cases above a certain level of cumulative rainfall observed in Mali and Senegal may be explained by the flushing out of larval breeding sites, which can lead to high mortality in Anopheles larval populations[49,50] and can reduce the human biting rate[50]. Vector populations are almost monospecific in these two countries: They are largely dominated by *An. arabiensis* in Senegal[51,52] and by *An. coluzzii* in Mali[53]. These two species are also present in our study area and in the Ouagadougou area of Burkina Faso[54]. However, in both these areas, they live in sympatry with both *An. gambiae s.s.* and *An. funestus*[33,55–57]. The quasi-linear relationship observed in our study between rainfall and the number of malaria cases may be explained by the fact that these species are not very susceptible to flushing out, due to rapid larval development in the case of *An. gambiae s.s.*[58,59] and to a preference for deeper environments in the case of *An. funestus*[60]. These species may therefore relay *An. coluzzii* when abundances of this later fall due to excessive rainfall.

In our study, the relationship between the number of malaria cases and temperature (SMI2) was non-linear. This is consistent with findings from two other studies conducted in the Sahel region (in Mali and in the Ouagadougou area of Burkina Faso)[14,15]. However, unlike these studies, we found no negative relationship between the number of malaria cases and temperature at higher temperature values. This discrepancy may be explained by the fact that temperatures can reach higher values in Mali and in the Ouagadougou area (> 34 °C) than in the Diébougou region, which is sufficient to inhibit the development of Anopheles larvae[61] and to reduce the survival of adult Anopheles[62,63]. In addition, we found that below a certain temperature, an increase in temperature had no effect on the number of cases (a finding also observed by Cissoko et al.[15]). Our hypothesis is that the increase in temperature, which should favour the development of *Anopheles*, is compensated by another phenomenon at low SMI2 values. While this phenomenon has yet to be clearly identified, high levels of LLIN usage during cooler periods may be a contributing factor[41].

Our spatio-temporal model fitted with two lagged SMIs and case data for a single epidemiological year helped to predict the start of the next outbreak nine weeks in advance, but with an error of three weeks (i.e. the actual outbreak began three weeks before the prediction). The prediction was good enough to make it possible to issue early warnings and to organize local prevention campaigns ahead of time. Our model could probably be improved with routine inclusion of new data and regular updated predictions. This study can be generalized to determine optimal periods and zones for prevention campaigns or interventions campaigns related to weather-related diseases such as dengue fever, malaria, etc.… Indeed, prioritizing a few numbers of areas and periods is helpful in strengthening malaria control programme in the context of lack of resources. This is even more important when countries will reach the pre-elimination phase: resources should then be concentrated on the most effective areas and periods. This is the principle of the "bottle neck" approach[11]. On the other hand, recent papers have shown that there is no conclusive evidence that targeted hotspot interventions accelerate malaria elimination. Therefore, a targeted approach to high-risk individuals that allows for a precise delineation of parasite transmission networks within and between households may be investigated[45,64]. For this purpose, data from HC consultations should be made available quickly, ideally at the same pace as ERA5 meteorological data (i.e. within 5 days). This can easily be achieved by using connected tablets for data entry.

Conclusion

In this study, a spatial analysis was conducted that highlighted the spatial heterogeneity of malaria cases and helped to identify four malaria hotspots in South-Western Burkina Faso. In the temporal analysis, an effective predictive model was built with data obtained through passive case detection and with simple and accessible meteorological data. Future studies should further investigate the detected hotspots to identify the local determinants of transmission. Our spatio-temporal analysis provides a powerful prospective method to identify high-risk

areas that may constitute priority areas during malaria prevention campaigns. Since our approach allowed us to determine the hotspots and to predict the start of the next annual epidemic, this approach should be cost-effective. Because of the scarcity of the resources in developing countries such as Burkina Faso, implementing the same policy at the same time through the whole health districts should be less cost-effective. Indeed, malaria epidemiology, at least onset, peak and length of malaria annual epidemic, is variable, as well as the environmental characteristics, through the country. Being able to determine areas and predict malaria onset dates can help policies makers to target actions at the right time and at the right places.

Data availability

The datasets analysed in this study may be available from the last author on reasonable request.

Received: 9 June 2021; Accepted: 27 September 2021
Published online: 08 October 2021

References

1. OMS. *Rapport sur le Paludisme dans le Monde* (2019). https://www.who.int/malaria/publications/world-malaria-report-2019/report/fr. Accessed on March 6, 2020.
2. Ministère de la santé du Burkina Faso. *Annuaire Statistique 2015* (2015). http://cns.bf/IMG/pdf/annuaire_ms_2015_signe.pdf. Accessed on March 6, 2020.
3. Institut National de la Statistique et de la Démographie, Burkina Faso. *Enquête sur les Indicateurs du Paludisme 2014* (2014). https://dhsprogram.com/pubs/pdf/MIS19/MIS19.pdf. Accessed on January 4, 2020.
4. Ministère de la santé. *Directives Nationales Pour la Prise en Charge du Paludisme dans les Formations Sanitaires du Burkina Faso* (2010). http://pdf.usaid.gov/pdf_docs/PA00JPHB.pdf. Accessed on March 13, 2020..
5. Ministère de la Sante. *Plan Stratégique National de lutte Contre le Paludisme 2016–2020* (2017). http://onspsante.bf/sites/default/files/publications/166/PSN%20%20%20%20%202016-2020_Paludisme_20_02_2017.pdf. Accessed on March 13, 2020.
6. INSD. *Enquête sur les Indicateurs du Paludisme 2017–2018* (2018). http://www.insd.bf/n/contenu/enquetes_recensements/enquete_palu/EIPBF_2018.pdf. Accessed on January 6, 2020.
7. Ouédraogo, A. L. *et al.* Efficacy and safety of the mosquitocidal drug ivermectin to prevent malaria transmission after treatment: A double-blind, randomized, clinical trial. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **60**, 357–365 (2015).
8. Stoute, J. A. *et al.* Long-term efficacy and immune responses following immunization with the RTS, S malaria vaccine. *J. Infect. Dis.* **178**(4), 1139–1144 (1998).
9. Selvaraj, P., Wenger, E. A. & Gerardin, J. Seasonality and heterogeneity of malaria transmission determine success of interventions in high-endemic settings: A modeling study. *BMC Infect. Dis.* **18**, 413 (2018).
10. Mogeni, P. *et al.* Effect of transmission intensity on hotspots and micro-epidemiology of malaria in sub-Saharan Africa. *BMC Med.* **15**, 121 (2017).
11. Landier, J., Rebaudet, S., Piarroux, R. & Gaudart, J. Spatiotemporal analysis of malaria for new sustainable control strategies. *BMC Med.* **16**, 226 (2018).
12. Ribeiro, J. M. C., Seulu, F., Abose, T., Kidane, G. & Teklehaimanot, A. Temporal and spatial distribution of anopheline mosquitos in an Ethiopian village: Implications for malaria control strategies. *Bull. World Health Organ.* **1996**(743), 299–305 (1996).
13. Bousema, T. *et al.* Hitting hotspots: Spatial targeting of malaria for control and elimination. *PLoS Med.* **9**, e1001165 (2012).
14. Ouedraogo, B. *et al.* Spatio-temporal dynamic of malaria in Ouagadougou, Burkina Faso, 2011–2015. *Malar. J.* **17**, 138 (2018).
15. Cissoko, M. *et al.* Geo-epidemiology of malaria at the health area level, dire health district, Mali, 2013–2017. *Int. J. Environ. Res. Public. Health* **17**, 3982 (2020).
16. Dieng, S. *et al.* Spatio-temporal variation of malaria hotspots in Central Senegal, 2008–2012. *BMC Infect. Dis.* **20**, 424 (2020).
17. INSD. *Tableau de bord économique et social 2014 de la région du Sud-Ouest* (2015). https://www.insd.bf/contenu/pub_periodiques/tableaux_de_bord/TBES/TBES_SO_2014.pdf. Accessed on March 15, 2020.
18. INSD. *Enquête Nationale sur le Secteur de l'orpaillage* (2017). http://www.insd.bf/contenu/enquetes_recensements/ENSO/Principaux_Resultats_ENSO.pdf. Accessed on March 15, 2020.
19. Copernicus Climate Change Service. *ERA5: Fifth Generation of ECMWF Atmospheric Reanalyses of the Global Climate* (Copernicus Climate Change Service, 2017).
20. Belmonte Rivas, M. & Stoffelen, A. Characterizing ERA-Interim and ERA5 surface wind biases using ASCAT. *Ocean Sci.* **15**, 831–852 (2019).
21. Yeomans, K. & Golder, P. The Guttman-Kaiser criterion as a predictor of the number of common factors. *The Statistician.* https://doi.org/10.2307/2987988 (1982).
22. SaTScan_Users_Guide.pdf (2021). https://www.satscan.org/cgi-bin/satscan/register.pl/SaTScan_Users_Guide.pdf?todo=process_userguide_download. Accessed on February 1, 2020.
23. Wood, S. N. *Generalized Additive Models: An Introduction with R* (CRC Press, 2006).
24. Guisan, A., Edwards, T. C. Jr. & Hastie, T. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecol. Model.* **157**, 12 (2002).
25. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **1**, 297–310 (1986).
26. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 3–36 (2011).
27. Wood, S. N. *Generalized Additive Models: An Introduction with R* 2nd edn. (Routledge and CRC Press, 2017).
28. Wood, S. N. Thin plate regression splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65**, 95–114 (2003).
29. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing, 2005). http://www.R-project.org. Accessed on February 1, 2020.
30. Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
31. Kleiber, C. & Zeileis, A. Applied Econometrics with R: Package Vignette and Errata. https://doi.org/10.1007/978-0-387-77318-6 (2015).
32. QGIS.org QGIS Geographic Information System. *Open Source Geospatial Foundation Project.* (2020). http://qgis.org. Accessed on February 1, 2020.
33. Soma, D. D. *et al.* Anopheles bionomics, insecticide resistance and malaria transmission in southwest Burkina Faso: A pre-intervention study. *PLoS ONE* **15**, e0236920 (2020).
34. Lombrail, P. & Pascal, J. Inégalités sociales de santé et accès aux soins. *Trib. Santé* **8**, 31–39 (2005).
35. Kadio, K., Ridde, V. & Mallé Samb, O. Les difficultés d'accès aux soins de santé des indigents vivant dans des ménages non pauvres. *Santé Publique* **26**, 89–97 (2014).
36. Ilboudo, S. D. O., Sombié, I., Soubeiga, A. K. & Dræbel, T. Facteurs influençant le refus de consulter au centre de santé dans la région rurale Ouest du Burkina Faso. *Santé Publique* **28**, 391–397 (2016).

37. Pierrat, C. Risque palustre: Appréhender la vulnérabilité des individus à l'échelle locale (Sud du Bénin). *VertigO Rev. Électron. Sci. Environ.* https://doi.org/10.4000/vertigo.11549 (2012).
38. Yonkeu, S., Maïga, A. H., Wethé, J., Mampouya, M. & Maga, G. P. Conditions socio-économiques des populations et risques de maladies: Le bassin versant du barrage de Yitenga au Burkina Faso. *VertigO Rev. Électron. En Sci. Environ.* https://doi.org/10.4000/vertigo.4778 (2003).
39. Baragatti, M. *et al.* Social and environmental malaria risk factors in urban areas of Ouagadougou, Burkina Faso. *Malar. J.* **8**, 13 (2009).
40. Berthélemy, J.-C. & Seban, J. Dépenses de santé et équité dans l'accès aux services de santé dans les pays en développement. *Rev. Déconomie Dév.* **17**, 33–71 (2009).
41. Kreuels, B. *et al.* Spatial variation of malaria incidence in young children from a geographically homogeneous area with high endemicity. *J. Infect. Dis.* **197**, 85–93 (2008).
42. Brooker, S. *et al.* Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. *Trop. Med. Int. Health* **9**, 757–766 (2004).
43. Van Der Hoek, W. *et al.* Towards a risk map of malaria for Sri Lanka: The importance of house location relative to vector breeding sites. *Int. J. Epidemiol.* **32**, 280–285 (2003).
44. el Samani, F. Z., Willett, W. C. & Ware, J. H. Nutritional and socio-demographic risk indicators of malaria in children under five: A cross-sectional study in a Sudanese rural community. *J. Trop. Med. Hyg.* **90**, 69–78 (1987).
45. Bousema, T. *et al.* The impact of hotspot-targeted interventions on malaria transmission in Rachuonyo South District in the Western Kenyan highlands: A cluster-randomized controlled trial. *PLoS Med.* **13**, e1001993 (2016).
46. Nesbitt, R. C. *et al.* Methods to measure potential spatial access to delivery care in low- and middle-income countries: A case study in rural Ghana. *Int. J. Health Geogr.* **13**, 25 (2014).
47. Sissoko, M. S. *et al.* Temporal dynamic of malaria in a suburban area along the Niger River. *Malar. J.* **16**, 420 (2017).
48. Kibret, S., Glenn Wilson, G., Ryder, D., Tekie, H. & Petros, B. Environmental and meteorological factors linked to malaria transmission around large dams at three ecological settings in Ethiopia. *Malar. J.* **18**, 54 (2019).
49. Paaijmans, K. P., Wandago, M. O., Githeko, A. K. & Takken, W. Unexpected high Losses of *Anopheles gambiae* larvae due to rainfall. *PLoS ONE* **2**, e1146 (2007).
50. Moiroux, N. *et al.* Modelling the risk of being bitten by malaria vectors in a vector control area in southern Benin, west Africa. *Parasit. Vectors* **6**, 71 (2013).
51. Robert, V. *et al.* Malaria transmission in the rural zone of Niakhar, Senegal. *Trop. Med. Int. Health* **3**, 667–677 (1998).
52. Sy, O. *et al.* Entomological impact of indoor residual spraying with pirimiphos-methyl: A pilot study in an area of low malaria transmission in Senegal. *Malar. J.* **17**, 64 (2018).
53. Sogoba, N. *et al.* Spatial distribution of the chromosomal forms of *Anopheles gambiae* in Mali. *Malar. J.* **7**, 205 (2008).
54. Ouédraogo, M. *et al.* Spatial distribution and determinants of asymptomatic malaria risk among children under 5 years in 24 districts in Burkina Faso. *Malar. J.* **17**, 460 (2018).
55. Costantini, C. *et al.* Living at the edge: Biogeographic patterns of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*. *BMC Ecol.* **9**, 16 (2009).
56. Fournet, F. *et al.* Diversity in anopheline larval habitats and adult composition during the dry and wet seasons in Ouagadougou (Burkina Faso). *Malar. J.* **9**, 78 (2010).
57. Guelbeogo, W. M. *et al.* Seasonal distribution of *Anopheles funestus* chromosomal forms from Burkina Faso. *Malar. J.* **8**, 239 (2009).
58. Diabate, A. *et al.* Larval development of the molecular forms of *Anopheles gambiae* (Diptera: Culicidae) in different habitats: A transplantation experiment. *J. Med. Entomol.* **42**, 548–553 (2005).
59. Paaijmans, K. P., Huijben, S., Githeko, A. K. & Takken, W. Competitive interactions between larvae of the malaria mosquitoes *Anopheles arabiensis* and *Anopheles gambiae* under semi-field conditions in western Kenya. *Acta Trop.* **109**, 124–130 (2009).
60. Hamon, J. *Biologie d'Anopheles funestus. 6 p. multigr.* (1955). http://www.documentation.ird.fr/hor/fdi:28874. Accessed on June 23, 2020
61. Bayoh, M. N. & Lindsay, S. W. Effect of temperature on the development of the aquatic stages of *Anopheles gambiae* sensu stricto (Diptera: Culicidae). *Bull. Entomol. Res.* **93**, 375–381 (2003).
62. Lyons, C. L., Coetzee, M. & Chown, S. L. Stable and fluctuating temperature effects on the development rate and survival of two malaria vectors, *Anopheles arabiensis* and *Anopheles funestus*. *Parasit. Vectors* **6**, 104 (2013).
63. Lyons, C. L., Coetzee, M., Terblanche, J. S. & Chown, S. L. Desiccation tolerance as a function of age, sex, humidity and temperature in adults of the African malaria vectors *Anopheles arabiensis* and *Anopheles funestus*. *J. Exp. Biol.* **217**, 3823–3833 (2014).
64. Stresman, G., Bousema, T. & Cook, J. Malaria hotspots: Is there epidemiological evidence for fine-scale spatial targeting of interventions? *Trends Parasitol.* **35**, 822–834 (2019).

Acknowledgements

The authors would like to thank the Ministry of Health of Burkina Faso, in particular Dr. Dembélé Henri, Médecin–chef de poste in Diebougou, and the local medical team, for facilitating data collection. Special thanks are due to Mr. Maiga Issouf for his strong involvement in data collection. We are also grateful to the "Laboratoire Mixte International sur les Maladies à Vecteurs" (LAMIVECT) for providing technical support. Lastly, we would like to thank Mr. Ouattara Adama and Mr. Félix Zoumènou for providing administrative support.

Author contributions

C.S.B., J.G., C.P., A.K., R.K.D. and N.M. designed the study; A.S. and D.D.S. conducted the field study; C.S.B., J.G. and N.M. designed the statistical analysis plan; C.S.B. conducted the statistical analysis under the supervision of J.G. and N.M.; C.S.B. performed the cartographic analysis with the participation of M.C. and P.T.; S.D. participated in the statistical analysis; C.S.B., J.G. and N.M. validated and interpreted the results; C.S.B., J.G. and N.M. wrote the manuscript; all authors read and approved the final manuscript.

Funding

This work was part of the REACT project, funded by the French Initiative 5%—Expertise France (No. 15SANIN213). C.S.B. received a grant from the *Revivre Développement* Endowment Fund through the NGO *Prospective et Coopération*, as well as a French Government Grant through the French Embassy in Burkina Faso. The funding was also provided by Excellence Initiative of Aix–Marseille University –A*MIDEX.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.