

Namur, le 16 mai 2022

Sujet: Rapport sur les travaux de thèse de Paul Taconet

Chers collègues,

J'ai lu la thèse de M. Taconet avec grand intérêt et je tiens à le féliciter pour la qualité de son manuscrit et son engagement. Son travail de thèse met en valeur les recherches robustes et approfondies qu'il a effectuées sur les dynamiques spatio-temporelles du risque de transmission résiduelle du paludisme, par une approche holistico-inductive. Travaillant moi-même dans le domaine de la modélisation holistico-inductive appliquée au paludisme, je partage un grand nombre des idées, propositions et recommandations de la thèse. La rédaction en français et en anglais est claire et soignée. La thèse comprend un chapitre publié en premier auteur dans *Parasites & Vectors*, un chapitre prêt à être soumis ainsi que deux publications en co-auteur.

Globalement, je trouve que cette thèse est de grande qualité et j'approuve donc sa soutenance. Dans la suite de ce rapport, je ferai, pour chaque chapitre, un commentaire général, suivi de commentaires plus spécifiques qui permettront éventuellement d'améliorer le manuscrit avant publication. Les commentaires et questions soulevées ont pour but de pousser la discussion plus loin et d'envisager des perspectives de recherches futures, mais n'ont pas pour objectif de remettre en question la qualité du travail, qui est déjà très abouti.

Chapitre 1

Le chapitre 1 présente le contexte scientifique de la thèse, et introduit le lecteur à différents éléments essentiels tels que le paludisme, son mode de transmission, le système vectoriel, et les enjeux liés à la lutte anti-vectorielle. Le chapitre se termine naturellement par une présentation des objectifs de la thèse. Ce chapitre est très clair et bien structuré. Dans la dernière partie (objectifs), j'ai particulièrement apprécié la distinction qui est faite entre la caractérisation, la compréhension et la prédiction du risque.

Commentaires spécifiques :

- La transmission résiduelle est définie comme celle qui persiste après avoir atteint une couverture universelle complète en MIILDA et/ou PID (p. 28). Je me questionne sur ce qu'on entend par « couverture universelle complète ». Quelle échelle considère-t-on pour estimer qu'une

couverture est universelle, le niveau national ? Par ailleurs, la notion de « couverture complète » me semble toute relative, et peut fortement varier d'une région à l'autre.

- Dans les facteurs humains influençant la probabilité de contact hommes-vecteurs, la qualité de l'habitat (notamment pour éviter que les moustiques entrent dans les habitations) est très peu mentionnée et discutée. Ce facteur pourrait être ajouté dans les facteurs de vulnérabilité p. 32. Par ailleurs, une approche holistique voudrait qu'on inclût ces facteurs humains dans l'analyse. Pourquoi ne pas les avoir ajoutés dans les analyses des chapitres 3 et 4 ?
- P. 8 (fin) : j'aurais aimé un peu plus d'explications/hypothèses sur la façon dont l'épidémie de covid-19 a pu influencer les cas et décès liés au paludisme. Est-ce uniquement lié à la perturbation des services sanitaires ? N'y a-t-il pas eu aussi un impact potentiel sur le reporting des cas et décès (une sous-estimation des cas lié au fait que les malades ne se sont pas nécessairement rendu dans un centre de santé déjà débordé ou, au contraire, une surestimation liée à la difficulté de diagnostic (par ex. fièvres qui ont pu être attribuées au covid plutôt qu'au palu) ? Cela ouvre aussi la discussion sur les données, leur fiabilité et leurs limites.
- La figure 1.1 manque de lisibilité dans le message. Le travail cartographique pourrait être amélioré en termes de couleurs et légende.
- P.25, dernier paragraphe : « on observe une corrélation forte... ». Où peut-on observer cette corrélation. Elle n'est pas visible sur la figure 1.13 référencée. Y a-t-il une autre source ?

Chapitre 2

Ce chapitre présente le contexte méthodologique de la thèse. Il pose les bases essentielles à la compréhension de la thèse et justifie les choix méthodologiques qui ont été posés. Ce chapitre est d'une grande qualité et montre le souci du détail du candidat. Les méthodes ont été choisies de façon minutieuse, argumentée et pertinente. Même s'il est assez courant de distinguer des objectifs d'explication et de prédiction pour les modèles, il est moins courant d'y ajouter la 3ème catégorie « modéliser pour décrire », dont les nuances avec la « modélisation pour expliquer » sont subtiles mais néanmoins pertinentes.

Commentaires spécifiques :

- Dans la note de la p. 49, les termes de « modélisation statistique » et « fouille de données » sont présentés comme des synonymes, ce qui ne me semble pas tout à fait être le cas. La fouille de données, dans le sens de « data mining » comme mentionné dans la préface permet d'extraire des connaissances à partir de bases de données volumineuses, mais pas nécessairement à l'aide de modèles statistiques selon moi.
- Dans ce chapitre, on oppose les modèles paramétriques aux modèles non-paramétriques, les premiers étant plus adaptés à la modélisation explicative et les seconds à la modélisation prédictive (p. 57). Je pense qu'il faut garder de la prudence quant à la capacité des modèles paramétriques à mettre en évidence des relations causales. L'existence d'une relation statistique entre 2 variables (illustrée par un coefficient de régression significatif) n'implique pas nécessairement une relation de causalité entre ces deux variables.

- Figure 2.4 : que veut dire EDA ?
- Où se situent les modèles bayésiens parmi les grandes familles de modèles statistiques ?

Chapitre 3

Le chapitre 3 présente la zone d'étude et les données environnementales produites dans le cadre du projet REACT et utilisées dans la thèse. Un des atouts non-négligeable de ce travail de thèse est la volonté de rendre toutes les données et codes produits accessibles à la communauté scientifique, ainsi que l'utilisation de logiciels open-source. Je félicite le candidat pour ces efforts, et en particulier pour le développement du package R « openapr ».

Commentaires spécifiques :

- Je trouve dommage que les détails de la procédure utilisée pour produire les cartes d'occupation du sol se trouvent en annexe. Si c'est bien le candidat qui a produit ces cartes, cela devrait faire partie intégrante de la thèse.
- Pourquoi utiliser des images Spot en plus des images Sentinel ? Sentinel a une meilleure résolution spatiale et une meilleure concordance temporelle avec les données de terrain. En annexe B, on explique pourquoi Sentinel a été utilisé en complément de Spot, mais pas le contraire.
- Il me semble important de préciser les résolutions spatiales des données en input et output dans le texte principal (et pas uniquement en annexe). Cela pourrait être précisé notamment dans les descriptions des figures 3.3 et 3.4.
- Quel est le procédé utilisé pour fusionner les données de différentes résolutions ? Ce n'est pas tout à fait clair pour moi.
- Pourquoi n'avoir utilisé qu'un modèle théorique pour les réseaux hydrographiques ? Vu l'importance de cette variable dans le travail qui suit, il aurait été intéressant de faire une validation de terrain et d'affiner éventuellement le modèle, par exemple en tenant compte des variations saisonnières.

Chapitre 4

Avec le chapitre 4, nous entrons dans le cœur du travail de thèse. Ce chapitre vise à modéliser les dynamiques spatio-temporelles de la présence et de l'abondance des vecteurs, sur base d'une large base de données de monitoring des piqûres de moustiques. Cette base de données, comprenant plus de 3500 nuits-homme de capture au total est certainement un point fort de ce chapitre. Il aurait été utile de comparer cela aux études existantes. Y a-t-il déjà eu d'autres campagnes de collecte de données d'une telle ampleur ? Par ailleurs, l'innovation en termes de méthode statistique est mise en avant, ce qui pour moi est moins le cas. Les modèles « random forest » sont de plus en plus utilisés en épidémiologie et la procédure en deux étapes (présence/abondance) est bien connue, notamment sous la terminologie « zero inflated models ». Une partie de ce chapitre est publiée dans *Parasites & Vectors*.

Commentaires spécifiques :

- Dans ce chapitre, les termes « abondance de vecteurs » et « contacts hommes-vecteurs » sont utilisés comme des synonymes. Ce n'est pourtant pas tout à fait la même chose, puisque le nombre de contacts (calculé ici sur un individu) dépend aussi de la présence d'autres hôtes dans les environs. Autrement dit, il y a une différence entre un nombre absolu des moustiques présents (l'abondance) et le nombre de contacts hommes-vecteurs, qui est plutôt un nombre relatif au nombre d'hôtes présents.
- Il y a pour moi une contradiction dans le fait d'utiliser des modèles « random forest » dans le but, notamment, de prendre en compte les relations non-linéaires et d'exclure les variables qui sont faiblement corrélées à la variable dépendante. Une faible corrélation peut justement être due à une relation non-linéaire.
- Il est dommage de ne pas avoir testé les capacités d'extrapolation du modèle d'une zone à l'autre, puisque toutes les données sont là pour le faire.
- Est-ce que des cartes de distribution prédite des taux de contact ont été produites ? Si pas, est-ce prévu ? Ce serait intéressant et utile d'aller jusqu'au bout du processus (y compris de sonder l'intérêt des décideurs locaux pour ces cartes).
- Les figures 4.4, 4.5, ainsi que les figures additionnelles p.120-121 bénéficieraient à être mise en format paysage afin d'en améliorer la lisibilité.
- Comment la collecte indoor/outdoor a-t-elle été contrôlée dans ce chapitre ?
- Pourquoi ne pas avoir utilisé les données collectées sur les conditions micro-climatiques utilisées dans le chapitre 5 (et décrites page 148) ?

Chapitre 5

Le chapitre 5 utilise une méthodologie similaire au chapitre 4 afin de modéliser non plus l'abondance, mais les résistances physiologiques et comportementales des vecteurs. Ce chapitre est original et robuste, notamment grâce à la grande quantité de données collectées sur le terrain. Le chapitre est très dense et très long (notamment la description des résultats). Certaines parties gagneraient à être un peu plus synthétiques, en évitant notamment les répétitions. Ce chapitre pourrait aussi faire l'objet de 2 publications séparées.

Commentaires spécifiques :

- Dans l'optique d'une publication dans un journal scientifique, je ne trouve pas nécessaire d'introduire toutes les questions de recherche dans l'introduction, ce qui allonge fortement le papier. En parler dans la discussion (en synthétisant) me semble suffisant.
- Lisibilité de la table 1 (p.151) à améliorer
- Le papier montre que les résistances (notamment les mutations) varient dans le temps (entre les deux enquêtes). Serait-il possible de tester si cette variation est significative ou non ?

- Les figures 3 et 4 sont très complètes, mais assez difficile à digérer. J'ai mis un temps à comprendre ce qui est GLMM et RF. Il serait bien d'ajouter cette information dans la description de la figure, en plus de la légende. Par ailleurs, la couleur du GLMM dans la légende (orange/rouge) ne correspond pas à celle des figures (jaune). Aussi, pourquoi n'y a-t-il pas de ligne verte sur tous les graphes ? Ce n'est pas clair pour moi.
- P.165 : le pouvoir explicatif des modèles est qualifié de « very weak », « weak », « moderate », ... Comment ces seuils ont-ils été définis ?
- Je ne comprends pas pourquoi le GLMM est meilleur que le RF dans plusieurs cas (selon les figures pages 198-199 et dans le texte p. 165). Y a-t-il une explication ?
- Comme pour le chapitre 4, je ne trouve pas cohérent de sélectionner les variables sur base du coefficient de corrélation.
- P.170, 2ème paragraphe : il est dit au début du paragraphe que les résistances sont stables d'une saison à l'autre et à la fin du paragraphe qu'il n'y a pas eu d'enquête entomologique durant la saison des pluies. N'est-ce pas en contradiction ?

Chapitre 6

Le chapitre 6 résume deux papiers publiés (l'un dans BMC Public Health, l'autre dans Scientific reports), dont M. Taconet est co-auteur. Les versions complètes de ces papiers se trouvent en annexe. Ce chapitre est tout à fait complémentaire aux autres travaux de la thèse et s'intéresse plus particulièrement aux facteurs humains influençant les contacts hommes-vecteurs ainsi qu'à la capacité de modèles basés sur des données météorologiques à prédire un nombre de cas humains de paludisme. J'ai peu de commentaires sur ce chapitre étant donné le statut des papiers déjà publiés.

Chapitre 7

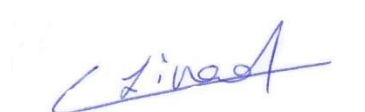
Le dernier chapitre reprend la discussion générale et est orienté autour de propositions de stratégies de réduction du risque de transmission résiduelle du paludisme dans les deux zones d'étude considérées. Dans une 2^{ème} partie sont discutées les limites, perspectives des méthodologies utilisées, notamment la science des données.

Les recommandations émises afin de prévenir le paludisme (section 7.1) sont parfois très générales et pas forcément basées sur les résultats de la thèse. Le dernier point soulevé (« cibler et prioriser le déploiement des stratégies complémentaires à la MIILDA ») est celui qui est le plus soutenu par les résultats de thèse.

Etant donné la richesse des données collectées dans le cadre du projet REACT, il serait dommage ne pas les exploiter jusqu'au bout. Tester la capacité d'extrapolation des modèles d'une zone géographique à une autre, croiser les données entomologiques aux données épidémiologiques, sont autant de perspectives mentionnées que j'encourage à poursuivre.

Pour conclure ce rapport, je réitère mes compliments à Paul pour son excellent travail. Je me tiens à votre disposition pour toute demande de précisions.

Veillez recevoir, chers collègues, l'expression de mes sentiments les meilleurs.



Professeur Catherine LINARD
Département de Géographie
Université de Namur, Belgique