Modélisation spatio-temporelle des déclarations de piqûres de tiques dans le cadre du projet CiTIQUE : exploration des facteurs explicatifs

Linh Nguyen Phuong

Master Mathématiques et Applications

M2 DATA SCIENCE

Année 2021-2022

Rapport de stage

Modélisation spatio-temporelle des déclarations de piqûres de tiques dans le cadre du projet CiTIQUE : exploration des facteurs explicatifs

---

Etudiante :
Linh Nguyen Phuong

Tuteurs de stage :
Thomas Opitz
Julien Papaïx
Karine Chalvet-Monfray

Remerciements

Dans un premier temps, je tiens à remercier chaleureusement l'ensemble des personnes étant intervenues dans mon travail de recherche durant ces derniers mois, et plus particulièrement Karine Chalvet-Monfray, Julien Papaïx et Thomas Opitz, pour toutes les connaissances qu'ils ont pu m'apporter, leur soutien, et surtout leur gentillesse. Ce fut pour moi un véritable plaisir de travailler auprès de personnes aussi bienveillantes. Jongler avec l'ensemble de ces chercheurs m'aura permis de pouvoir apprendre sur divers sujets, tant bien mathématiques qu'écologiques. Je tiens également à les remercier pour leur disponibilité et leur implication, notamment sur ces derniers jours de rédaction. Leur présence aura été un réel soutien et m'aura permis d'éviter de perdre pied dans ces moments de rush.

Je remercie également l'ensemble du corps enseignant du département mathématiques de la Faculté des Sciences de l'université d'Angers pour toutes les connaissances qu'ils ont pu me transmettre durant ces cinq années d'études.

En parallèle, un grand merci à toute l'équipe de BioSP pour leur accueil, je n'aurais pu imaginer meilleure intégration. Une mention spéciale est donnée à l'ensemble des jeunes (doctorants et ingénieurs). Merci à eux pour cette découverte de la Provence, du festival d'Avignon, pour ces moments de convivialité, ces Mölkky perdus et pause *Cémantix/Pédantix* les midi.

Merci aussi aux Zygomatiques et à la Corpo pour ces merveilleuses soirées entre Angers, Marseille et Madrid ces 5 derniers mois. Enfin, je remercie Romain de me supporter au quotidien (eh oui cela relève de l'exploit !) et surtout ces deux dernières semaines pour son soutien dans les moments de stress.

Table des matières

Table des figures

Introduction

0.1 INRAE PACA et l'unité Biostatistique et Processus Spatiaux

INRAE, Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement, est un établissement public français de recherche scientifique et technologique (EPST), dépendant du Ministère de l'Enseignement, de la Recherche et de l'Innovation (MESRI) et du Ministère de l'Agriculture et de l'Alimentation (MAA). INRAE a été créé en 2020, suite à la fusion de l'Institut National de Recherche agronomique (INRA), créé en 1946, et l'Institut National de Recheche en Sciences et Technologies pour l'Environnement et l'Agriculture (IRSTEA), créé en 2012. Cet institut rassemble pour la première fois les trois domaines de l'agriculture, l'alimentation et l'environnement. Ainsi, son objectif est de pouvoir répondre à différents enjeux sociétaux grâce à sa pluridisciplinarité, tels que la sécurité alimentaire et nutritionnelle, la transition des agricultures, la gestion des ressources naturelles et des écosystèmes, l'érosion de la biodiversité, l'économie circulaire et les risques naturels, ainsi que des enjeux sociétaux régionalisés. Ainsi, INRAE Provence-Alpes-Côte-d'Azur, centre dans lequel a été réalisé ce stage, axe ses travaux sur les objets et enjeux notamment liés au territoire méditerranéen.

L'unité Biostatistique et Processus Spatiaux (BioSP) est implantée sur le site Agroparc d'Avignon et fait partie du département MathNum (mathématiques et informatique appliquées) d'INRAE. Elle développe des travaux en statistique, en systèmes dynamiques, en écologie et épidémiologie, ainsi qu'aux interfaces entre ces disciplines, avec un intérêt spécifique pour les questions spatiales et spatio-temporelles. Elle est composée d'une équipe de recherche et d'une équipe Opérationnelle pour la Plateforme d'Epidémiosurveillance en Santé Végétale. L'unité fait intervenir 7 chargés de recherche, 7 directeurs de recherche, 2 chargés de mission, 7 ingénieurs d'étude, 6 ingénieurs de recherche, 3 techniciens de la recherche, 4 doctorants, 5 contractuels et enfin 2 stagiaires.

Ce stage s'est également déroulé dans le cadre du réseau Statistiques pour les Sciences Participatives (CiSStats). Le réseau CiSStats est financé par le département Mathnum et a pour but de rassembler des statisticiens appliqués, des écologues modélisateurs et des porteurs d'enjeux (associations, gestionnaires d'espaces naturels...) souhaitant développer des méthodes statistiques pour mieux valoriser les jeux de données issus des sciences participatives.

0.2 Les tiques et maladies vectorielles

Les *Ixodida* (McCoy et al., 2015a), appelées également tiques dures, sont des acariens ectoparasites (parasites vivant à la surface corporelle d'autres êtres vivants) hématophages (se nourrissant du sang d'autres êtres vivants). En Europe, les tiques sont les vecteurs les plus importants pour les humains comme pour les animaux, c'est-à-dire qu'elles sont susceptibles de transmettre des agents infectieux d'un organisme à un autre. Il existe environ 900 espèces de tiques dans le monde, dont une quarantaine présentes en France. Ces espèces sont classées en quatre genres différents : *Ixodes* et *Hyalomma* (s'attaquant à l'Homme), *Dermacentor* et *Rhipicephalus* (s'attaquant principalement aux ongulés). *Ixodes* et *Dermacentor* sont présents sur l'ensemble du territoire de France métropolitaine à l'exception de la côte méditerranéenne et de la Corse, où *Hyalomma* et *Rhipicephalus* sont majoritairement présents. Leur présence est un enjeu primordial en matière de santé publique et vétérinaire car elles transmettent bactéries, virus et parasites.
De nombreux instituts de recherche se consacrent à l'étude d'*Ixodes ricinus*, espèce très fréquente en France (plus de 95% des piqûres chez l'Homme) et dont environ 10% des nymphes et 20% des adultes sont contaminés par la bactérie *Borrelia* et autres pathogènes (Chalvet-Monfray, 2022). Ces instituts se concentrent également sur l'étude d'*Hyalomma marginatum*, peu présente actuellement mais en expansion depuis le bassin méditerranéen.

Les tiques évoluent durant une grande partie de leur vie au sol, où les oeufs éclosent, les larves se métamorphosent durant tout leur développement, et où elles cherchent des hôtes à tout stade. Durant deux à trois étapes

de son existence, elle se fixe à la peau d'un hôte (variant selon l'espèce de tique) grâce à un rostre (appendice buccal) afin de se nourrir de son sang et par la suite se reproduire. Durant ces phases, la tique peut transmettre à l'hôte de nombreux pathogènes. L'évolution d'*Ixodes ricinus* est illustrée dans la Figure 1. Cette espèce possède une large gamme d'hôtes, dont l'Homme fait partie.

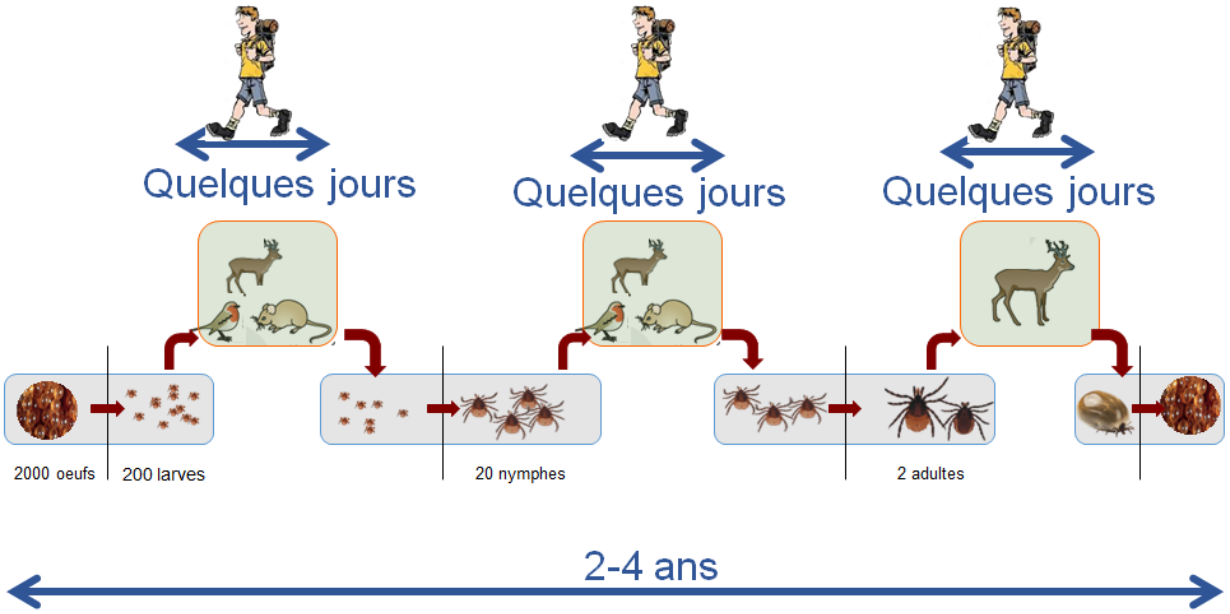


FIGURE 1 – Développement d'*Ixodes Ricinus* (source : K. Chalvet-Monfray, INRAE Vet'Agro Sup)

La présence des différentes espèces dépend du climat, du type de végétation, de l'altitude, et de la présence d'hôtes. Ainsi, l'activité d'*Ixodes ricinus* est plus élevée au printemps et en automne dans les zones européennes tempérées, en été au Nord de l'Europe et en hiver dans les montagnes du Mahgreb.

À l'état de nymphe ou d'adulte, la tique peut transmettre différents agents infectieux, tels que des virus comme l'*encéphalite à tiques*, des parasites comme le *Babesia*, ou bien des bactéries du type *Borrelia*, responsable de la maladie de Lyme. Celle-ci a été identifiée en 1975 suite à une épidémie d'arthrite à Lyme, dans l'État du Connecticut aux Etats-Unis. Elle est non contagieuse et est principalement reconnaissable par l'apparition d'un érythème migrant (plaque rouge s'étendant en cercle autour de la piqûre) quelques jours à un mois après la piqûre de tique, s'accompagnant parfois de fatigue et de maux de tête/articulations/muscles. Des traitements antibiotiques efficaces permettent de vaincre la maladie de manière très favorable. En revanche, en l'absence de traitement, des complications neurologiques, cutanées, rhumatologiques ou cardiaques peuvent être observées durant ce qui est appelé la phase secondaire. Celle-ci correspond à une dispersion bactérienne dans le corps, suivie d'une focalisation tissulaire, fréquemment traduite en Europe par une réaction neurologique appelée neuroborréliose. Cette phase se manifeste notamment par des troubles sensitifs, pouvant aller jusqu'à une paralysie partielle des membres. L'infection par une tique porteuse de la bactérie n'est pas systématique. D'un côté, la transmission de la bactérie demande un certain temps, il est donc important de pouvoir retirer la tique (dans son entièreté) au plus vite après morsure. De l'autre, le système immunitaire de certaines personnes permet d'inhiber l'infection.

Entre 2009 et 2020, le nombre de cas de borrélioses de Lyme diagnostiquées a été estimé entre 26 000 et 68 530 (SPF, 2021), avec une tendance à la hausse. En 2020, nous dénombrons 710 hospitalisations. Ces hospitalisations sont principalement observées en juin et octobre. À l'heure actuelle, la maladie de Lyme est un problème de santé publique. En effet, il n'existe pas de vaccin contre la maladie ; seule l'étude du risque d'exposition permet de lutter contre, grâce à la connaissance des tiques dans l'environnement, ainsi que des facteurs d'exposition des hôtes et des bactéries que les tiques portent. C'est pourquoi, en 2016, ont été lancés en parallèle deux projets de lutte contre la maladie de Lyme et les maladies transmises par les tiques :

- le plan national de lutte contre la maladie de Lyme, initié par le Ministère des Affaires Sociales et de la Santé ;
- le programme CiTIQUE, initié par INRAE et autres partenaires.

C'est dans ce dernier que s'inscrit ce travail.

0.3 Projet CiTIQUE : collecte et modélisation des déclarations de piqûres

Le projet CiTIQUE est un programme de recherche en sciences participatives qui se concentre sur les tiques et maladies transmissibles par les tiques. Il a été initié en 2017. Le projet permet ainsi à toute personne, chercheur comme citoyen, de contribuer aux travaux, à toute étape de la recherche, que ce soit pour l'observation, l'interprétation des résultats, mais aussi le travail en laboratoire. L'objectif de cette collaboration citoyenne est de permettre de faire avancer les connaissances scientifiques plus vite grâce à l'implication de tout type d'acteurs. Le programme est porté par INRAE, l'université de Lorraine, l'Agence Nationale de Sécurité Sanitaire de l'alimentation, de l'environnement et du travail (ANSES), le Centre Permanent d'Initiatives pour l'Environnement Nancy Champenoux, le LabEx de Recherches Avancées sur la Biologie de l'Arbre et des Ecosystèmes forestiers (ARBRE), et les laboratoires Tous Chercheurs.

Le projet CiTIQUE a plusieurs objectifs :

- la collecte d'informations de signalements de piqûres de tiques par le biais d'une application, permettant ainsi la création d'une base de données unique à l'origine de nouveaux travaux de recherche,
- la collecte massive de tiques grâce à l'aide du grand public et la constitution d'une tiquothèque,
- la mise en place de protocoles d'échantillonnage, d'identification morphologique des tiques et d'analyse ADN lors de stages *Tous chercheurs* ouverts aux élèves et citoyens à partir de 10 ans,
- l'analyse du risque d'exposition,
- l'amélioration des stratégies de prévention par le développement de nouveaux outils et la création d'un réseau de distribution de kits de collecte,
- l'organisation de débats et la mise en évidence d'axes de recherche par différents acteurs : chercheurs, professionnels, grand public.

À travers l'ensemble de ces objectifs, une application smartphone nommée "Signalement Tique" a été mise en place le 18 mai 2017 afin de permettre à tout citoyen de signaler une piqûre (sur l'Homme ou l'animal). L'application a été développée par le site INRAE E-phytia et financée par la Direction Générale de la Santé dans le cadre du plan Lyme. Toute personne peut alors indiquer un ensemble d'informations liées à une piqûre, telles que la date, les coordonnées géographiques et leur précision, ainsi que des détails sur l'environnement du lieu de piqûre, tels que le type de paysage ou la raison de la présence sur ce point géographique. L'application a depuis enregistré plus de 74 000 signalements (à la date de mars 2022).
Il devient alors intéressant de pouvoir analyser l'ensemble des données spatio-temporelles acquises en les croisant avec des facteurs pouvant potentiellement expliquer ces déclarations, représentant par exemple la dynamique des populations d'hôtes et de tiques. Cependant, une telle modélisation rencontre plusieurs difficultés, notamment par la taille parfois très importante des jeux de données d'intérêt. De plus, la dimension spatio-temporelle amène une corrélation entre les événements, et donc la dépendance des observations étudiées à travers l'espace et le temps. Enfin, une déclaration de tique peut être observée grâce à la rencontre de différents processus :

- la dynamique spatio-temporelle des tiques ;
- la dynamique d'exposition des hôtes, ici l'Homme ;
- la rencontre entre ces deux dernières dynamiques, manifestée par une piqûre de la tique sur l'Homme ;
- le rapport de la piqûre sur l'application « Signalement Tique ».

Or, l'ensemble des données de l'application ne nous permet pas de pouvoir observer ces quatre processus de manière distincte. En effet, l'observation de nombreuses déclarations dans le Grand-Est est notamment corrélée avec l'origine du projet CiTIQUE et donc une meilleure connaissance de l'application dans cette région. En revanche, un tel taux de signalements ne signifie pas forcément que le risque y est plus élevé qu'ailleurs. À l'inverse, dans certaines régions forestières, il n'est observé que très peu de signalements. Pour autant, il s'agit d'un type de paysage très favorable aux tiques. Ce faible taux de signalement est alors dû à l'absence de population humaine sur ce type de territoire. Le danger y est grand mais le risque faible en raison d'une faible exposition de l'Homme aux tiques.

0.3.1 Objectifs du stage

Ce stage de six mois a permis de démarrer un travail de recherche autour de la modélisation et prédiction spatio-temporelles du risque de piqûre de tique sur la base des données collectées par CiTIQUE. Notamment, le stage s'est focalisé sur deux objectifs : la collecte et le nettoyage des jeux de données multi-sources et multi-échelles afin de générer une base de données homogène ; l'exploration de facteurs explicatifs de l'évolution spatio-temporelle des déclarations de piqûres de tiques en France métropolitaine grâce à des modélisations stochastiques avancées via des modèles de régression additive. Les résultats des recherches sont discutés tout

au long de ce rapport. Nous commencerons par présenter l'ensemble des matériels utilisés pour cette recherche, puis les méthodes nous ayant permis d'aboutir à une modélisation. Les résultats sont ensuite présentés puis discutés.

Les implémentations informatiques de ce stage ont été réalisées sous forme de codes réutilisables et documentées dans le logiciel libre R afin d'assurer la reproductibilité des résultats. Un rapport technique (sous forme de document R Markdown) détaillant les étapes de traitement des données et de modélisation, à diffuser aux ingénieurs et chercheurs du projet CiTIQUE, a également fait partie des livrables de ce stage.

Matériels

La partie des matériels décrit l'ensemble des bases de données utilisées dans ce travail, détaillant leur origine et leur résolution dans l'espace. Les systèmes de projection sont ensuite détaillés.

1.1 Ensembles de données

L'étude se porte principalement sur l'application "Signalement Tique", qui a enregistré exactement **74 034** déclarations de piqûres entre le 13 juillet 2017 et le 11 mars 2023. L'ensemble des données se compose également de facteurs de risques spatialisés, sous forme de rasters (matrices de pixels) de résolutions différentes selon la base de données.

1.1.1 Base de données CiTIQUE

Dans un premier temps, nous détaillons la manière dont les données originales ont été acquises et de quels types d'informations elles sont constituées. Il est ensuite précisé les différentes mises à jour qui y ont été apportées depuis son lancement.

1.1.1.1 Acquisition des données

L'application est dérivée sur quatre supports différents, pouvant ainsi entraîner des différences de formats ou de précision dans les signalements : version Androïd, version Apple, formulaire web, formulaire papier.

Dans les trois premiers cas, le signalement est envoyé directement sur la base de données, tandis que dans le cas où le signalement est rendu sous format papier, un intermédiaire du programme se charge d'enregistrer numériquement le signalement.

id	**origine**	**datetime**	**...**	**x**	**y**	**id_user**
38155	iOS 13.3.1	2020-05-25 20 :59 :16	**...**	48.41631	2.559698	12420
58247	Android 8.1.0	2021-04-08 18 :09 :18	**...**	46.29467	5.680467	18200
45829	iOS 13.5.1	2020-06-16 07 :47 :08	**...**	47.10309	6.655513	16930
68672	Web 1.0.4	2021-06-20 08 :46 :18	**...**	48.81834	2.170028	12420

Table 1.1 – Base de données brutes CiTIQUE

Chaque signalement possède dix-neuf informations, décrites en détail en annexes 3.3.3. Le Tableau 1.1 illustre la base de données et les attributs utilisés dans ce travail de recherche sont décrits ci-dessous :

- Coordonnées géographiques → x, y : longitude et latitude du point de piqûre ;
- Dates → *datetime* : date et heure d'envoi du signalement sur l'application/ *bite_date* date de piqûre du signalement concerné ;
- Description de l'hôte :
  - > *for_human* : variable binaire précisant si l'hôte piqué est un animal ou un Homme ;
  - > *age* : age en années de l'hôte ;
- Identifiant d'utilisateur → *user_id* : identifiant de compte attribué au déclarant.

1.1.1.2 Mises à jour et reformatage

L'application connaît de nombreux changements au fil des années, notamment lors de la migration de l'application d'E-phytia vers l'application propre. Les principales mises à jour concernent son ergonomie et des formats de données. Un principal changement est réalisé en mai 2020, dans lequel les citoyens peuvent désormais entrer directement la date de naissance de l'hôte plutôt que la catégorie d'âge : il s'agit de la variable *birth_date*, de format **jour/mois/année**. Finalement, il n'est plus demandé une adresse localisant le lieu de la piqûre (version E-phytia) mais directement en le localisant sur une carte. L'ensemble des localisations précisées par une adresse, soit environ 18 000, sont converties en coordonnées géographiques directement par Jonas Durand, ingénieur responsable de la gestion de la base CiTIQUE. Des inversions de longitude et latitude, et d'autres erreurs restantes, ont été corrigées grâce au logiciel R.

1.1.2 Potentiels facteurs explicatifs

Les facteurs tels que le type d'occupation des sols, le climat, les données météorologiques et la démographie, facteurs pouvant décrire les différents processus cités en introduction, sont intégrés à la base de données afin d'être explorés.

1.1.2.1 Occupation des sols : base de données Corine Land Cover

L'occupation des sols représente les différents types d'usages faits par l'Homme d'un territoire. Elle permet ainsi de distinguer les surfaces minérales, l'eau, les forêts, les zones urbaines et autres. Différentes bases de données existent, dont la base Corine Land Cover (CLC) (CGDD/SOeS, 2009), utilisée ci-dessous pour l'exploration de facteurs explicatifs.

Cette base a été créée en 1985 dans le cadre de l'étude européenne de l'information sur l'environnement du projet CORINE. L'objectif est d'obtenir une base européenne unifiée permettant de décrire les différents usages des sols de façon homogène sur l'ensemble du territoire européen.

Il existe 3 niveaux de nomenclature CLC : 5 classes de niveau 1, 15 classes de niveau 2 et 44 classes de niveau 3 (figures 1.1 et 1.2). Ces niveaux sont hiérarchisés permettant l'utilisation du niveau de précision requis par l'analyse. Par exemple, le type biophysique de niveau 3 n°321, représentant les pelouses et pâturages naturels, est compris dans la catégorie de niveau 2 n°32, représentant les milieux à végétation arbustive et/ou herbacée, elle-même incluse dans le niveau 1 n°3, représentant les forêts et milieux semi-naturels.

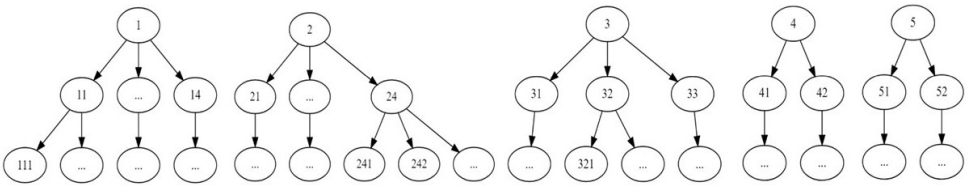


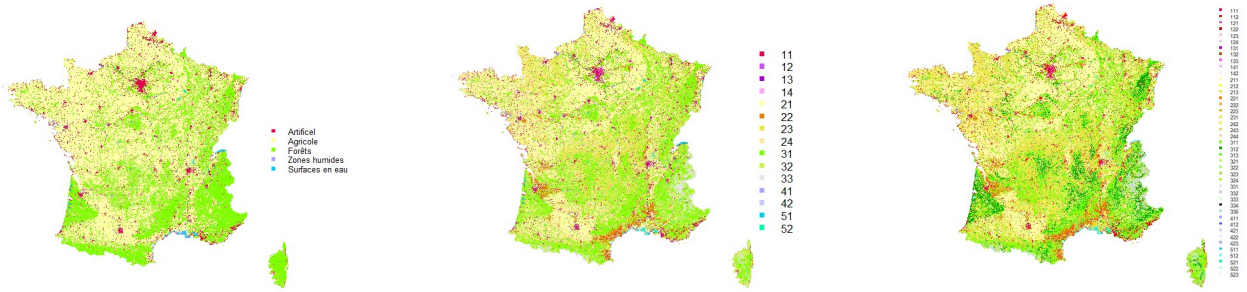
FIGURE 1.1 – Hiérarchie de la nomenclature Corine Land Cover. De haut en bas : niveau 1, niveau 2, niveau 3

Cette nomenclature a été réalisée par photo-interprétation d'images satellites. L'échelle finale est de $1/100\,000^e$ pour la catégorie 3, soit une représentation de pixels de 100m par 100m. Le projet CORINE étudie également les changements d'occupation des sols au cours du temps en publiant au fur et à mesure les bases de modification et la cartographie mise à jour des 39 pays européens concernés. Dans le cas de notre étude, un raster du niveau 3 de CLC dans sa version 2018 est mis à disposition. Le facteur d'occupation des sols est donc invariant dans le temps pour notre étude. L'ensemble des nomenclatures CLC est détaillé en annexes à la Table 4.1.

1.1.2.2 Climat

D'après de précédentes recherches (Gray et al., 2009; Estrada-Peña, 2008), il a été démontré que le climat influe sur la présence de tiques sur un territoire. Dans cette recherche, nous avons utilisé les bases de données LANMAP, DRIAS et Climatick, décrites ci-après.

Base de données LANMAP De nombreux facteurs environnementaux permettent d'expliquer une grande part des caractéristiques phénologiques d'espèces, tels que les variables météorologiques et climatiques. La base de données LANMAP a été développée par Mücher et al. (2010). Tout comme la base de données CLC, cette base a été développée afin d'offrir un nouvel outil d'étude environnementale. Il s'agit de données obtenues



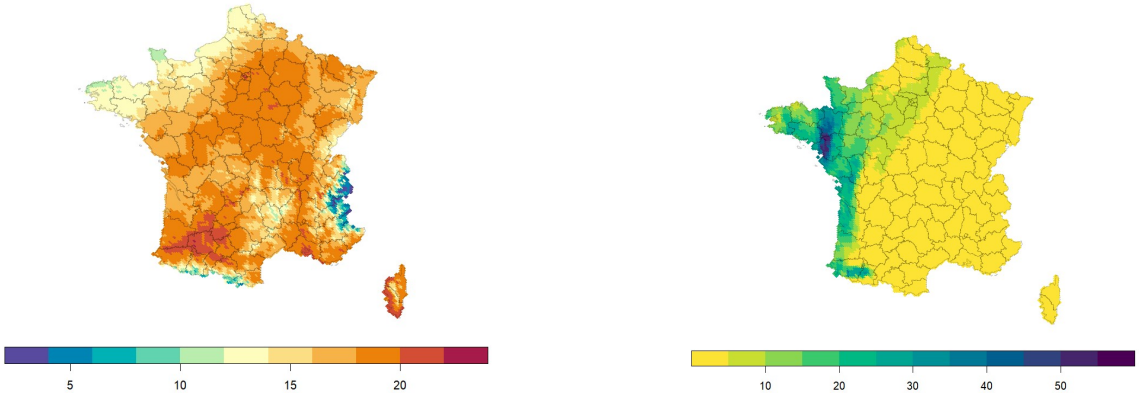
Classes du CLC de niveau 1 | Classes du CLC de niveau 2 | Classes du CLC de niveau 3

FIGURE 1.2 – Cartographie de la France métropolitaine sur la base CLC

par intégration de deux autres bases, la *European Environmental Zones* (EnZ) (Metzger et al., 2005) et la *Biogeographical Regions Map of Europe* (BRME) (Roekaerts, 2002).

L'EnZ a été développée par une classification non supervisée à partir de vingt des variables environnementales les plus pertinentes (pertinence déterminée par une analyse en composantes principales (ACP)). La BRME a été développée afin de répondre aux critères de la directive de l'Union Européenne sur la conservation des habitats naturels, puis complétée afin d'étendre la cartographie au reste des pays européens. Elle se base sur l'interprétation des travaux de Noirfalise (1987). Finalement, la base climatique LANMAP est composée de quinze catégories, dont neuf présentes en France métropolitaine, illustrées à la Figure 1.5a. Il s'agit ici de données vectorielles (pavage de l'espace sous forme d'un ensemble de polygones), avec la taille minimale des polygones fixée à 11 km^2.

Base de données DRIAS Dans cette recherche, certains attributs de simulation météorologique de la base DRIAS (Lémond, 2010) sont utilisés afin de construire une classification climatique (voir prochain paragraphe) explicative de la présence de tiques sur le territoire de la France métropolitaine. Suite à de nombreuses demandes de données climatiques régionalisées, les données DRIAS ont été simulées par des laboratoires français, à partir de différents modèles climatiques existants. Ces informations sont journalières et projetées sur une grille de résolution 8km par 8km, et pour trois scénarios climatiques différents : RCP 2.6, RCP 4.5 et RCP 8.5. Dans le prochain paragraphe sont utilisées, sur la période 1991 - 2020 inclus, la température moyenne en $°C$ (Figure 1.3a) et les précipitations totales journalières en mm (Figure 1.3b) du scénario de changement climatique RCP 2.6 (soit le scénario le plus favorable, prenant en compte une politique de diminution des émissions). Le modèle climatique utilisé est le modèle CNRM-CM5/ALADIN63, modèle plus adapté pour sa précision sur les données terrestres du territoire que nous considérons.



(a) Température moyenne (b) Précipitations totales

FIGURE 1.3 – Données météorologiques simulées DRIAS pour le 15 mai 2020

Base de données Climatick Wongnak (2022) travaille actuellement dans le cadre de sa thèse sur le projet Climatick et étudie l'influence des changements climatiques sur la phénologie de la tique *Ixodes ricinus*. Une classification a alors été réalisée afin de délimiter les espaces géographiques favorables à *Ixodes ricinus*. Ainsi, à la Figure 1.5b, nous distinguons trois sous-ensembles de l'espace : les Hautes Montagnes et le territoire méditerranéen, climats trop extrêmes pour permettre à *Ixodes ricinus* d'y vivre, et le reste du territoire, favorable

à la présence de l'espèce. Cette classification a été réalisée à partir de quatre critères climatiques calculés entre 1991 et 2020 inclus sur la base de données météorologiques DRIAS (présentée dans la section précédente) :

- $t_{hot} = \frac{1}{30}\sum_{y=1991}^{2020} t_{hot}^{(y)}$: moyenne de température du mois le plus chaud ;
- $t_{cold} = \frac{1}{30}\sum_{y=1991}^{2020} t_{cold}^{(y)}$: moyenne de température du mois le pus froid ;
- $t_{winter} = \frac{1}{30}\sum_{y=1991}^{2020} (t_{dec}^{(y)} + t_{jan}^{(y)} + t_{feb}^{(y)})$: moyenne des températures hivernales (décembre/janvier/février) ;
- $ombr = \sum_{y=1991}^{2020} (2\bar{t}^{(y)} - \bar{pr}^{(y)})_+$, où $\bar{t}^{(y)}$ est la température moyenne en y et $\bar{pr}^{(y)}$ la précipitation totale moyennées en y : somme des ombrothermiques positives.

Lorsque $t_{hot} > 22°C$, $t_{cold} > 4°C$ et $ombr > 305$, un territoire est considéré comme ayant un climat méditerranéen et donc défavorable à la présence de tiques. Lorsque $t_{winter} < 0°C$, un territoire est considéré de hautes montagnes et donc défavorable à la présence de tiques. Sinon, le territoire est considéré comme favorable à la présence de tiques.

Bien que la classification Climatick se concentre sur l'espèce *Ixodes ricinus*, celle-ci est majoritairement présente sur le territoire de France métropolitaine et représente 95% des piqûres de tique chez l'Homme. Nous la considérons donc comme suffisamment pertinente pour la présence générale des tiques sur ce territoire.

1.1.2.3 Démographie

Un phénomène de piqûre de tique nécessite la présence d'une tique mais également d'un hôte, et dans le cas présent d'un Homme. Le recensement de la population humaine en France métropolitaine en 2017 est exploitée afin de représenter la présence humaine sur le territoire. Les données sont tirées de l'Institut national de la statistique et des études économiques (Insee) (Figure 1.4). Il est attribué à chaque commune française un nombre d'habitants recensé en 2017 (les données étant parues en janvier 2020 et corrigées en décembre 2020).

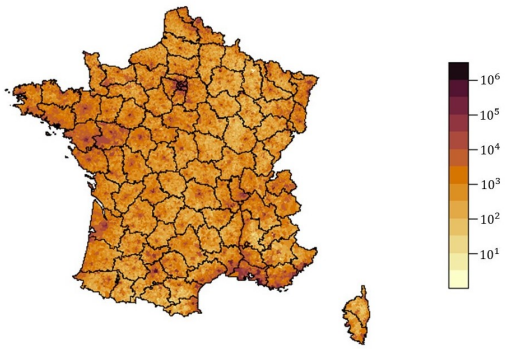


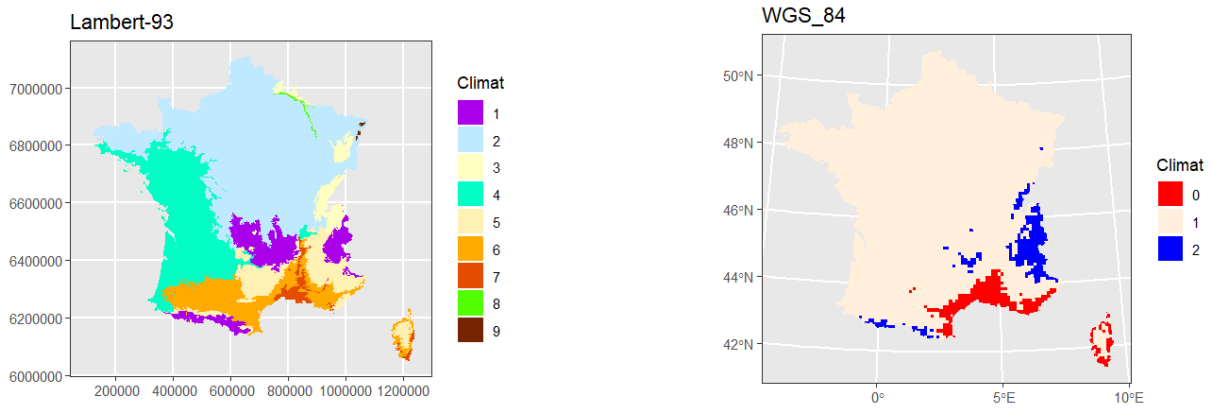
FIGURE 1.4 – Population communale en 2017 en France métropolitaine (échelle log)

1.2 Systèmes de projection géographiques

Il existe différents systèmes de coordonnées géographiques de référence permettant de localiser des objets à la surface de la Terre. Cette surface terrestre est décrite par un *géoïde*, soit une surface équipotentielle dont la pesanteur terrestre est le référentiel (Janssen, 2009). Cette surface approche le niveau moyen de la mer. Ce *géoïde* est lui-même simplifié par un ellipsoïde de révolution (aussi appelé sphéroïde), généré par la rotation d'une ellipse autour de sa largeur. Cet ellipsoïde est paramétré par la longueur de ces deux demi-axes. Ainsi de nombreux ellipsoïdes aux paramètres variants ont été créés, dans un premier temps de façon à correspondre localement aux territoires concernés, puis de manière globale grâce à l'arrivée des technologies spatiales. Un système géodésique appelé *datum* est déterminé par les paramètres de l'ellipsoïde choisie. Un objet spatial est alors localisé par un système de coordonnées comprenant longitude (angle par rapport au méridien dans l'axe Ouest-Est), latitude (angle par rapport à l'équateur dans l'axe Sud-Nord) et altitude (distance par rapport au niveau de la mer).

Le système de référence mondiale actuel est le « World Geodetic System 1984 »(WGS84), dans lequel les bases CiTIQUE, DRIAS et Climatick (Figure 1.5b) sont repérées. Ce système utilise un système décrit par la longitude et la latitude. Cependant, il existe un second type de coordonnées spatiales : il s'agit de coordonnées projetées (projection planaire). Ce type de référentiel projette une partie ou l'intégralité de la Terre sur une surface plane, sur une grille de coordonnées. De nombreux référentiels de projection ont été développés afin de

répondre à des propriétés de vraisemblance locale, en fonction du territoire d'intérêt – les distorsions ne pouvant être totalement évitées. Parmi les propriétés de distance, de surface, ou d'angles, seules une voire deux peuvent être conservées après projection. Ainsi, le système de coordonnées Lambert-93 est le référentiel de projection d'usage pour l'étude de la France métropolitaine. Sa projection est équivalente (conserve les surfaces), conforme (conserve les angles) et aphylactique (conserve les distances au méridien). Les bases de données CLC, LANMAP (Figure 1.5a) et Insee utilisent ce dernier référentiel.



(a) Base LANMAP

(b) Base Climatick

FIGURE 1.5 – Systèmes de coordonnées selon les bases de données

Les référentiels des différentes bases doivent alors être accordées afin d'homogénéiser les données pour la modélisation à suivre. Le référentiel choisi dans ce travail de recherche est le référentiel Lambert-93, plus adapté à l'analyse de données géolocalisées en France métropolitaine. Les conversions, précisées en 2.1.2, sont réalisées sur R via la librairie *sf* ; il s'agit des transformations avec les coordinate reference systems (CRS) suivants : $EPSG = 4326 \rightarrow EPSG = 2154$ en utilisant les codes EPSG (European Petroleum Survey Group) faisant référence dans le domaine.

Méthodes

Ci-dessus ont été présentées les données utilisées pour l'ensemble de l'étude, avec leur origine, leur résolution spatiale et leur dimension temporelle. Il s'agit maintenant de pouvoir modéliser le phénomène de déclaration des piqûres de tiques en France métropolitaine et d'explorer les facteurs explicatifs potentiels. Pour ce faire, une réflexion se tourne dans un premier temps autour de la représentation de ces données dans l'espace et le temps, puis du type de modèle pouvant représenter un tel événement, les lois de probabilités potentielles décrivant le comportement spatio-temporel des déclarations, puis les outils de validation et de sélection de performance des modèles.

2.1 Agrégation des données

L'ensemble des traitements suivants (nettoyage de données, homogénéisation spatiale, création de covariables) est réalisé sur le logiciel R en version 4.1.1. La Table 2.1 répertorie l'ensemble des covariables créées dans cette partie.

Variable	**Nom**	**Type**	**Description**
Climat LANMAP	*lanmap*	Catégoriel	Catégorie de climat majoritaire
Climatick	*climatick*	Catégoriel	Indice de favorabilité de présense de tiques
Climat interactions	*new_clim*	Catégoriel	Interactions LANMAP & Climatick
DRIAS	*ombr*, *TM_max*, *TM_min*, *TM_winter*	Quantitatif	Sur 30 ans : ombrothermique, température moyenne du mois le plus chaud, température moyenne du mois le plus froid, température moyenne hivernale
	ombr_check, *TM_max_check*, *TM_min_check*, *TM_winter_check*	Catégoriel	$\mathbb{1}_{\{ombr>305\}}$, $\mathbb{1}_{\{TM\_max>22\}}$, $\mathbb{1}_{\{TM\_min>4\}}$, $\mathbb{1}_{\{TM\_winter<0\}}$
Occupation des sols CLC	*niv*1, *niv*2, *niv*3, *niv*4	Catégoriel niv. 1-2-3-4	Catégorie d'occupation des sols majoritaire
	urban, *agric*, *forest*, *wet*, *water*	Quantitatif niv. 1	Proportion de chaque catégorie d'occupation des sols
	clc_i	Quantitatif niv. 2-3	Voir nomenclature CLC annexe 4.1
	incompl_art, *compl_art*, *agric*, *forest*, *open_vg*, *open_non_vg*, *h2o*	Quantitatif niv. 4	
Indices de diversité	$hill_{ij}$	Quantitatif	Nombre de Hill d'ordre i appliqué au niveau j, $0 \leq i \leq 2$, $1 \leq j \leq 4$

Présence d'eau	*pres.wet*	Catégoriel : Zones humides	Classification binaire des pixels par présence de zones humides
	pres.water	Surfaces en eau	" " par présence de surfaces en eau
	pres.all_w	Terrains aqueux	" " par présence de terrains aqueux
Temporel	*mois*, *annee*	Catégoriel et Quantitatif	$mois \in \{1, ...12\}$ $année \in \{Jan, ..., Dec\}$
	date	Quantitatif	$date \in \{1, ..., 63\}$
Spatial	x, y	Quantitatif	x coordonnée Ouest-Est et y coordonnée Sud-Nord

TABLE 2.1 – Variables sélectionnées pour la modélisation du nombre d'occurrences de signalements

2.1.1 Nettoyage des données brutes CiTIQUE

Dans un premier temps, un processus de nettoyage des données est réalisé sur les 74 034 signalements originaux (Table 2.2) :

- les signalements non localisés sont supprimés,
- les signalements localisés en dehors de la France continentale sont supprimés ;
- seuls les signalements dont la piqûre est datée entre 2017 et 2022 inclus sont conservés ;
- les attributs *age* et *birth_date* sont agrégés afin de ne conserver qu'une colonne *age* (calcul d'âge en cas d'absence réalisé à partir de la date de naissance lorsque celle-ci était disponible, valeur manquante sinon) et les dates de naissance aberrantes sont remplacées par une valeur manquante ;
- les signalements dont la date de signalement est inférieure à la date de piqûre sont jugés intraitables et donc supprimés ;
- seules les déclarations de piqûres faites sur l'Homme sont conservées ;
- les duplications de signalements sont supprimées. Un phénomène de duplication (pouvant aller jusqu'à 21 copies du signalement original) peut être dû à l'envoi multiple du formulaire lors de problèmes de connexion Internet.

	Nb. de déclarations	**% de l'ensemble de départ**	**Total restant**
Localisation manquante	16 748	22.62%	57 286
Localisation hors Fr. cont.	843	3.46%	56 443
$bite\_date \notin [2017, 2022]$	2 558	0.04%	53 855
$bite\_date < datetime$	30	9.32%	46 955
$for\_human = 0$	6 900	2.09%	45 407
Copies de signalements	1 548	0.03%	45 383

TABLE 2.2 – Nombre de signalements non exploitables pour l'étude et total restant après suppression

2.1.2 Homogénéisation spatiale

2.1.2.1 Occurrences des déclarations de piqûres de tiques

Les signalements sont géolocalisés de manière ponctuelle sur l'ensemble de la France (Figure 2.1a). Par la suite, nous transformons cette information ponctuelle en des comptages par unité spatiale. Grâce aux librairies *raster* et *sf* du logiciel R, une grille $\mathcal{G}$ de résolution 20 km est créée, dans laquelle chaque pixel est décrit par les coordonnées en km de son centroïde $s_k = (s_k^{(1)}, s_k^{(2)})$, où $s_k^{(1)}$ correspond à la coordonnée Ouest-Est et $s_k^{(2)}$ à la coordonnée Sud-Nord, et par le nombre de piqûres localisées dans celui-ci. Nous obtenons alors la carte de la Figure 2.1b.

La résolution de la grille $\mathcal{G}$ a été choisie afin d'éviter au plus le biais lié à l'imprécision de la localisation indiquée dans les formulaires de signalement, allant jusqu'à plus de 5 km. De plus, cette résolution permet d'obtenir une complexité algorithmique abordable lors de la modélisation du phénomène de comptage des

(a) Information ponctuelle

(b) Information matricielle (échelle log)

FIGURE 2.1 – Transformation d'échelle des données

occurrences. La France métropolitaine est alors divisée en 1529 pixels. La résolution peut être par la suite affinée afin d'améliorer la précision spatiale.

La Figure 2.1b présente le nombre d'occurrences par pixel P_k, il s'agit donc d'un espace en deux dimensions. La dimension temporelle est alors ajoutée, à l'échelle mensuelle, c'est-à-dire que chaque pixel représente cette fois-ci le nombre d'occurrences dans le pixel de centroïde s_k et au temps $t_k = (m_k, y_k)$, où $m_k \in \{1, ..., 12\}$ correspond au mois de l'année et $y_k \in \{2017, ..., 2022\}$ à l'année. Un pixel défini par le temps et l'espace est appelée voxel et notée V_k.

2.1.2.2 Covariables

Dans un premier temps, l'ensemble des bases de données énumérées dans la section 1.1.2 sont agrégées afin de décrire le territoire et l'écologie de chaque pixel (et non voxel, ces données étant invariantes dans le temps). Leurs systèmes de coordonnées sont reprojetés sur le système Lambert-93 (voir section 1.2) et les résolutions de chacune sont adaptées à la grille $\mathcal{G}$. Pour ce faire, $\mathcal{G}$ est intersectée avec les rasters ou polygones de chacune de ces bases de données. Pour chaque pixel de $\mathcal{G}$, l'aire de chacune des modalités des covariables CLC, LANMAP et climatick, est calculée. Ensuite, la modalité majoritaire de chaque pixel est conservée dans une variable catégorielle. Pour les covariables CLC, il est également conservé la proportion d'aire de chaque modalité. Lorsqu'un pixel n'intersecte pas une base de données (dû à un manque d'information, notamment sur le littoral où les délimitations ne sont pas précisément les mêmes d'une base à l'autre), la donnée est imputée à partir des k-plus proches voisins géographiques avec $k = 5$. La modalité majoritaire parmi les cinq pixels les plus proches est alors conservée. L'algorithme est notamment réalisé pour les données qualitatives LANMAP et Climatick.

Pour les variables quantitatives, deux méthodes différentes sont utilisées. Pour les variables telles que $t_{hot}, t_{cold}, t_{winter}, ombr$, la moyenne pondérée par les proportions d'aire est calculée en chaque pixel $P_k \in \mathcal{G}$. Ainsi, pour le pixel P_{1397} illustré à la Figure 2.2 et d'aire totale égale à 353.075 km^2, la température moyenne du mois le plus chaud serait la suivante :

$$t_{hot}^{P_{1397}} = \frac{\mathcal{A}_A \times 24.79 + \mathcal{A}_B \times 24.58 + \mathcal{A}_C \times 24.06 + \dots}{353.075}$$

où $\mathcal{A}_j$ représente l'aire du pixel j de la grille DRIAS.

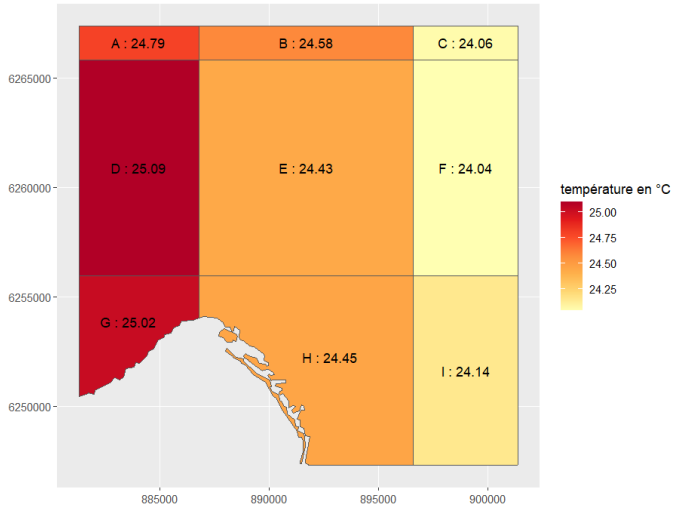


FIGURE 2.2 – Intersection de la grille DRIAS de température moyenne du mois le plus chaud avec le pixel P_{1397}

Le calcul diffère pour la covariable quantitative *pop* du nombre d'habitants. Cette fois-ci, les poids accordés au nombre d'habitants de chaque commune présente sur un pixel ne sont pas calculés de la même manière. Si une commune chevauche plusieurs pixels, il faut répartir la population totale de cette commune aux différents pixels concernés. Ces poids correspondent à la proportion d'aire par rapport à la commune concernée, et non la proportion d'aire par rapport au pixel. Par exemple, considérons la ville de Marseille à la Figure 2.3, d'aire totale égale à 237.420km^2. Les polygones P_{1397}, P_{1398}, P_{1426} et P_{1427} intersectent la ville. Alors la population $pop_{P_{1397}}$ du polygone P_{1397} peut être calculée comme suit :

$$pop_{P_{1397}} = \frac{pop_{Marseille} \times \mathcal{A}_A}{237.420} + \dots \tag{2.1}$$

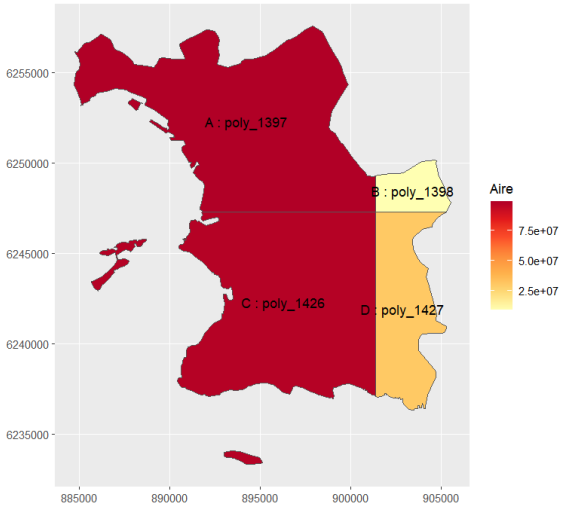


FIGURE 2.3 – Intersection de la ville de Marseille avec la grille $\mathcal{G}$.

2.1.3 Création de nouvelles covariables

A partir des données agrégées vues en section 2.1.2.2, de nouvelles covariables sont construites, permettant ainsi d'en simplifier certaines ou de définir l'interaction entre plusieurs covariables. L'ensemble des covariables utilisées est finalement regroupé en Table 2.1.

2.1.3.1 Interactions climatiques

L'interaction des bases climatiques LANMAP et Climatick est créée à la Figure 2.4 afin de pouvoir discriminer les territoires au sein des climats LANMAP en fonction de critères météorologiques de favorabilité de présence de tiques.

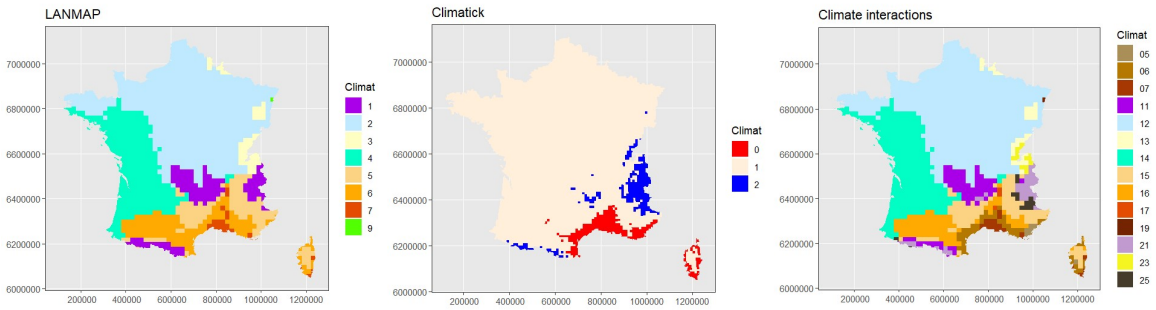


FIGURE 2.4 – Interactions de classes climatiques. La nouvelle variable d'interaction est catégorielle : le chiffre des unités correspond à la classe climatique majoritaire de la base LANMAP et le chiffre des dizaines au niveau de favorabilité décrit par la base Climatick.

2.1.3.2 Interfaces écologiques spatiales

L'interface écologique est la zone de contact entre deux types d'espaces aux caractéristiques écologiques distinctes. La zone de transition est également appelée écotone, et possède des caractéristiques particulière faisant d'elle une zone riche, comprenant sa propre biodiversité ainsi que celles des écosystème adjacents. L'étude de cette interface permet l'observation de certains phénomènes liés à l'échange entre les deux espaces concernés. Dans cette recherche, nous nous intéressons à la lisière entre les différentes catégories d'occupation des sols. En effet, d'après (Wongnak et al., 2022), il a été observé une plus forte présence de tiques autour de points d'observation dont le paysage est décrit par une fragmentation de forêts, partageant une interface avec des prairies et zones agricoles. De plus, les zones péri-urbaines, à la lisière de zones vertes (forêts), présentent une plus forte densité de tiques, induite par la présence d'hôtes tels que l'Homme mais aussi de mammifères (Rizzoli et al., 2014).

Une manière de quantifier l'interface écologique est de calculer plusieurs indices de diversité. Ces indices permettent de décrire l'interface de façon simplifiée, sans pour autant devoir représenter la totalité des interfaces. Marcon (2015) présente le nombre de Hill développé par Scheiner (2012). Soient R le nombre de catégories (des espèces dans le cas classique) et p_i la proportion de la catégorie i dans la cellule considérée. Le nombre de Hill d'ordre q est défini par :

$$ {}^qD = \left(\sum_{i=1}^{R} p_i^q \right)^{\frac{1}{1-q}} $$

A l'ordre 1, ${}^qD$ n'est pas défini. Le nombre de Hill est alors donné par sa limite, soit

$$ {}^1D = \exp\left(-\sum_{i=1}^{R} p_i \log(p_i) \right) $$

Plus q augmente, moins il est donné de poids aux catégories rares. En particulier :

- ${}^0D = R$. Cet ordre correspond à la richesse, c'est à dire au nombre de catégorie. Ainsi, un type d'occupation des sols de proportion très faible a le même poids qu'un type a la proportion élevée ;
- ${}^1D = \exp(H)$, où H est l'indice de Shannon défini par $H = -\sum_{i=1}^{R} p_i log(p_i)$, représentant la quantité d'information. Lorsque les différents types de couverture de sols sont équirépartis, le nombre de Hill d'ordre 1 est égal à la richesse. Pour un nombre de types de sol fixé, ce nombre diminue et tend vers 1 lorsque nous sommes en présence de catégories rares.
- ${}^2D = \frac{1}{1-E}$, où E est l'indice de Simpson défini par $E = \sum_{i=1}^{R} p_i^2$. Cet indice représente la probabilité que deux points tirés aléatoirement dans l'espace considéré appartiennent à la même catégorie. Le nombre de Hill vaut 1 en présence d'une unique catégorie et tend vers 0 lorsque le nombre de catégories augmente et que celles-ci sont équiréparties. Pour un nombre fixé de catégories présentes, ce nombre tend vers 1 lorsque nous sommes en présence de catégories rares et lorsque la proportion d'une catégorie tend vers 1.

Les trois indices précédents sont calculés sur les trois niveaux de catégories de sols CLC présentés précédemment, ainsi que sur un quatrième niveau. Il s'agit d'une classification différente, regroupant les catégories CLC de niveaux 1, 2 et 3 en sept nouvelles catégories décrites par la Figure 2.5, selon la favorabilité de présence de tiques. Ainsi la catégorie D est fortement favorable à la présence de tiques, les catégories B et E moins favorables et les autres catégories ne le sont pas.

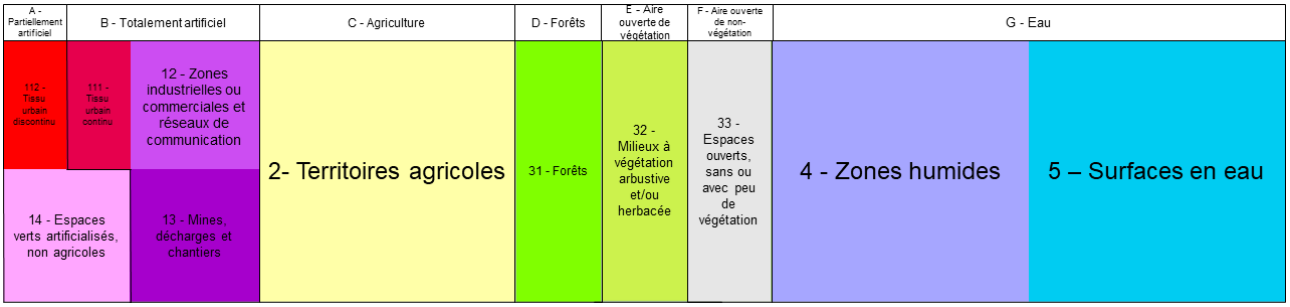


FIGURE 2.5 – Occupation des sols – Niveau 4

2.1.3.3 Présence d'eau

Les proportions des catégories CLC reflétant la présence d'eau étant trop faibles, il est donc décidé de simplifier ces variables potentiellement explicatives en les remplaçant par des vecteurs binaires de présence ou non d'eau dans un pixel donné. Cette classification est réalisée à partir des proportions en zones humides, des proportions en surfaces en eau, puis sur des proportions de ce que nous appelons territoires aqueux, soit la somme des proportions en zones humides et surfaces en eau. Afin de classifier correctement les pixels, nous réalisons des arbres binaires de régression tels que ceux montrés dans la Figure 2.6 grâce au logiciel R (Chesneau, 2020).

Un arbre binaire de régression permet de déterminer la valeur du nombre d'occurrences de déclarations en fonction d'une variable quantitative. A chaque noeud de l'arbre, l'ensemble est divisé à partir d'un seuil, permettant d'augmenter au maximum la variance inter-groupes des deux feuilles créées. Cette méthode nous permet alors successivement de trouver le seuil discriminant au mieux le nombre d'occurrences en fonction des variables *wet* (zones humides), *water* (surfaces en eau) et *all_water* (territoires aqueux). Un booléen sur les seuils déterminés par les arbres nous permettent alors d'identifier les pixels ayant assez d'eau libre pour expliquer la présence de tiques.

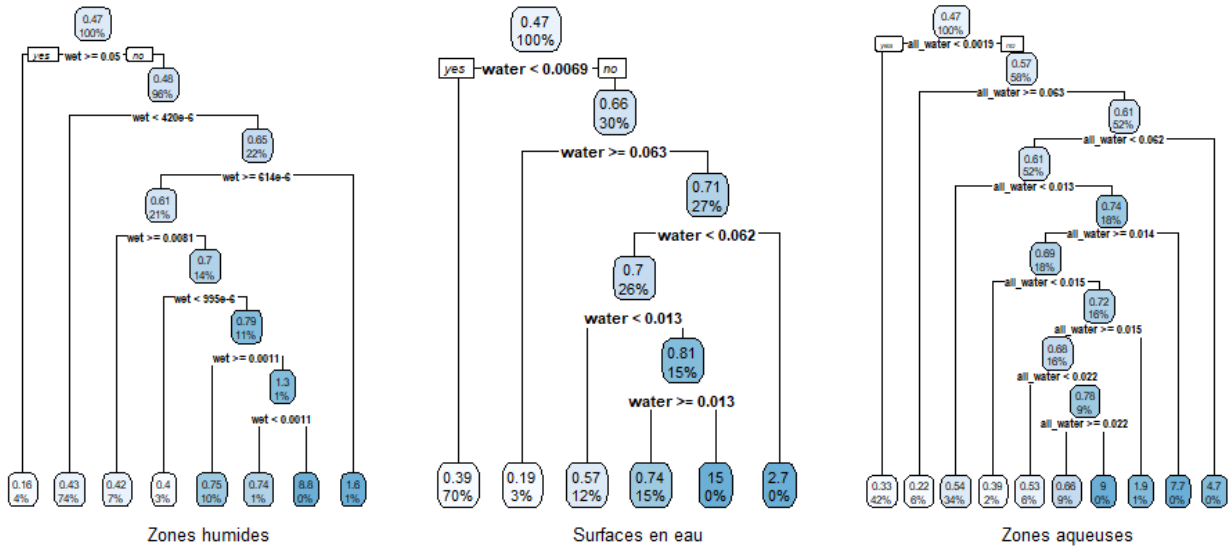


FIGURE 2.6 – Arbres binaires de régression – Proportions de zones humides, surfaces en eau, terrains aqueux. Dans chaque noeud se trouve le pourcentage d'observations respectant la condition précédente ainsi que la moyenne de cette part d'observations.

2.1.3.4 Variable temporelle

Le temps est décrit initialement par le mois et l'année. Il est également crée une covariable quantitative faisant interagir les deux, soit le facteur $date \in \{1, ..., 63\}$, représentant alors chaque mois de chaque année :

- janvier 2017 $\rightarrow$ 1 ;
- ... ;
- décembre 2017 $\rightarrow$ 12 ;

- janvier 2018 $\rightarrow$ 13 ;
- ... ;
- décembre 2018 $\rightarrow$ 24 ;
- ... ;
- mars 2022 $\rightarrow$ 63.

2.2 Modèles Additifs Généralisés

Nous définissons comme variable réponse à expliquer et à prédire le nombre d'occurrences de déclarations de piqûres de tiques, et nous la modélisons à l'aide d'un Modèle Additif Généralisé (GAM, Hastie and Tibshirani, 1987; Wood, 2006). Un GAM est une généralisation non linéaire (par rapport à la forme des contributions des covariables dans le prédicteur) des Modèles Linéaires Généralisés (GLM) et permet de modéliser des relations non linéaires entre la variable réponse et les variables explicatives, ce qui est souvent le cas pour des variations spatiales. Ici nous donnons des généralités sur les GAM qui nous permettrons de justifier les choix de modélisation réalisés par la suite. La structure du modèle utilisé est spécifiquement présentée dans la partie 3.

2.2.1 Introduction

Soit un échantillon de n variables aléatoires indépendantes $Y = (Y_1, ..., Y_n)^T$ à modéliser (^{T} signifie transposée, ainsi Y est un vecteur colonne), tel que Y_i (pour $i = 1, \ldots, n$) suit une certaine loi $\mathcal{L}(\mu_i, \tau)$, avec $\mathbb{E}[Y_i] = \mu_i$, et τ un paramètre de forme généralement supposé constant à travers l'échantillon. Soit également un jeu de p covariables $\boldsymbol{X}_i \in \mathbb{R}^p$. L'idée principale de toute régression est de pouvoir modéliser l'espérance de Y_i conditionnellement à $\boldsymbol{X}$; en général, le paramètre de loi de Y_i dépendant des covariables peut être différent de l'espérance, cependant les cas abordés dans ce travail se limitent à l'espérance. Il existe différentes manières de réaliser ces régressions. L'une des plus simples est la régression linéaire, qui permet la modélisation linéaire d'une variable suivant une loi normale. Soit alors $Y_i \sim \mathcal{N}(\mu_i, \sigma)$, avec μ_i la moyenne et σ^2 la variance. Dans ce cas,

$$\mu = \boldsymbol{X}\boldsymbol{\beta}$$

où $\boldsymbol{X}$ est la matrice $(X_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ où chaque ligne i est le vecteur de covariables décrivant l'observation y_i. $\boldsymbol{\beta}$ est un vecteur de paramètres à estimer.

Les modèles linéaires généralisés GLM (GLM, Wood, 2006) font intervenir une fonction de lien afin de relier le paramètre μ_i au prédicteur $\boldsymbol{X}_i\boldsymbol{\beta}$ et permettent ainsi de généraliser aux modèles suivant une loi autre que normale. Dans ce travail, la fonction de lien est exclusivement la fonction logarithme pour les lois considérées. La structure des modèles GLM est la suivante :

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}, \qquad \forall i \in \{1, ..., n\}$$

où la fonction g, appelée fonction de lien, est strictement monotone et appartient à la classe $\mathcal{C}^1$ des fonctions dont la dérivée est continue, $\mathbf{X}_i$ est le vecteur des covariables de la i-ème observation et $\boldsymbol{\beta}$ est un vecteur de paramètres (appelés coefficients) à estimer. De plus, Y_i suit une loi appartenant à la famille exponentielle (par exemple, la loi normale, la loi Poisson, la loi Gamma), réunissant l'ensemble des distributions dont la fonction de densité peut se réécrire de la façon suivante :

$$f_\theta(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

où a, b et c sont certaines fonctions, ϕ est un paramètre de forme et θ un paramètre d'échelle. Par exemple, pour la loi de Poisson, ce paramètre est égal au logarithme de la moyenne μ, et donc $\theta_i = g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$ pour la i-ème observation. Pour illustration, la démonstration d'appartenance à la famille exponentielle est faite pour la loi de Poisson en section 2.3.

Le prédicteur étant linéaire (c'est-à-dire une combinaison linéaire des covariables via les coefficients $\boldsymbol{\beta}$), de nombreux concepts et idées de la régression linéaire se généralisent ici. Cependant, l'algorithme d'estimation devient approximatif et l'ajustement doit être réalisé itérativement. Soient la vraisemblance et la log-vraisemblance

de $\boldsymbol{\beta}$ comme suit (en tenant compte de la relation entre $\boldsymbol{\beta}$ et θ susmentionnée) :

$$
\begin{aligned}
& L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\theta_i}(y_i) \\
\iff & l(\boldsymbol{\beta}) = \sum_{i=1}^{n} log\left[f_{\theta_i}(y_i)\right] \\
\iff & l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b_i(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right)
\end{aligned}
$$

La log-vraisemblance est maximisée en différenciant celle-ci en fonction des éléments de $\boldsymbol{\beta}$, puis en résolvant les équations annulant le gradient suivant en fonction du vecteur $\boldsymbol{\beta}$ (Wood, 2006) :

$$
\sum_{i=1}^{n} \frac{y_i - \mu_i}{b''(\theta_i)/\omega} \frac{\partial \mu_i}{\beta_j} = 0, \forall j
$$

où ω est une constante connue telle que la variance $\mathbb{V} = \frac{b''(\theta)\phi}{\omega}$. Dans la solution itérative de ce système d'équations, les équations à résoudre forment un problème de moindres carrés généralisés.

Il est également possible de généraliser la structure du modèle encore plus en autorisant un certain degré de non-linéarité dans le prédicteur linéaire. Par exemple, l'espérance de Y_i peut être modélisée en fonction d'une covariable par une relation polynomiale ou même une spline (terme explicité dans la section suivante). Nous sortons finalement du cadre des GLM lorsque les effets des prédicteurs modélisés par des splines sont pénalisés par un paramètre de contrôle de la courbure. Cette pénalisation est explicitée à la section suivante.

2.2.2 Formalisme des modèles additifs généralisés

Les GAM (Wood, 2006) sont des extensions des GLM. Dans le cas des GAM, la forme de la structure reste additive, cependant la variable réponse est expliquée par une somme de fonctions de classe au moins $\mathcal{C}^1$ des covariables. Ces modèles approchent plus précisément les effets des covariables grâce à l'utilisation de fonctions plus complexes que des fonctions polynomiales. Chacune de ces fonctions peut être représentée comme la combinaison linéaire d'un ensemble de fonctions de base ; ces fonctions de base sont appelées fonctions splines. Ces splines peuvent en général également être utilisées dans le cas des GLM, cependant, les modèles GAM comportent un paramètre de lissage permettant le contrôle de ces splines. Un exemple d'ajustement selon ces plusieurs méthodes est présenté à la Figure 2.7.

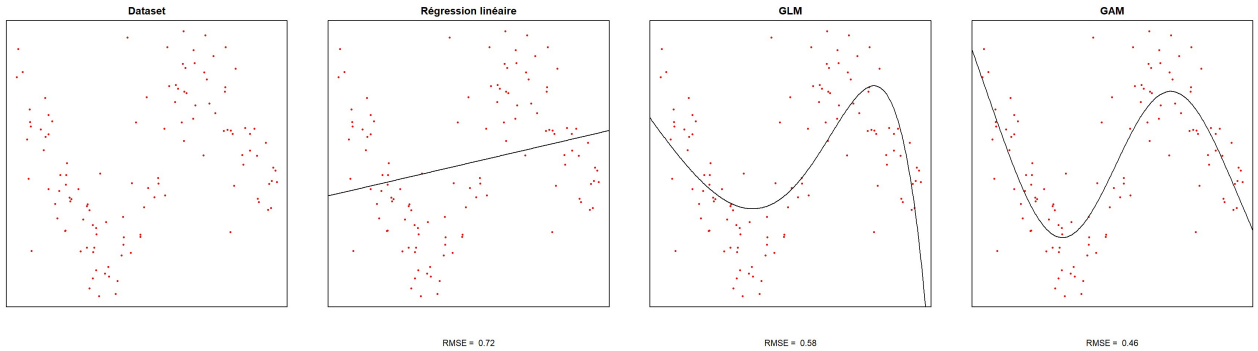


FIGURE 2.7 – Ajustement de données par différents modèles. En rouge, les données à modéliser. En noir, les courbes de régression.

Les GAM ont l'avantage de pouvoir contrôler la précision des fonctions (notamment leur caractère lisse) par différents paramètres afin d'éviter le sur-apprentissage ou le sous-apprentissage (Figure 2.8).

La structure la plus complète possible pour un GAM à p covariables (comprenant les coordonnées spatiales x et y et temporelles *mois* et *année*) est la suivante ; typiquement, seulement une partie des termes inclus dans l'équation suivante sont utilisées dans les modèles implémentées en pratique :

$$
\forall i \in \{1, ..., n\}, \qquad g(\mu_i) = \mathbf{X}_i^* \boldsymbol{\beta}^* + \sum_{j=1}^{p} f_j(x_{ij}) + \sum_{1 \leq j_1 < j_2 \leq p} f_{j_1, j_2}(x_{ij_1}, x_{ij_2}) \tag{2.2}
$$

où :

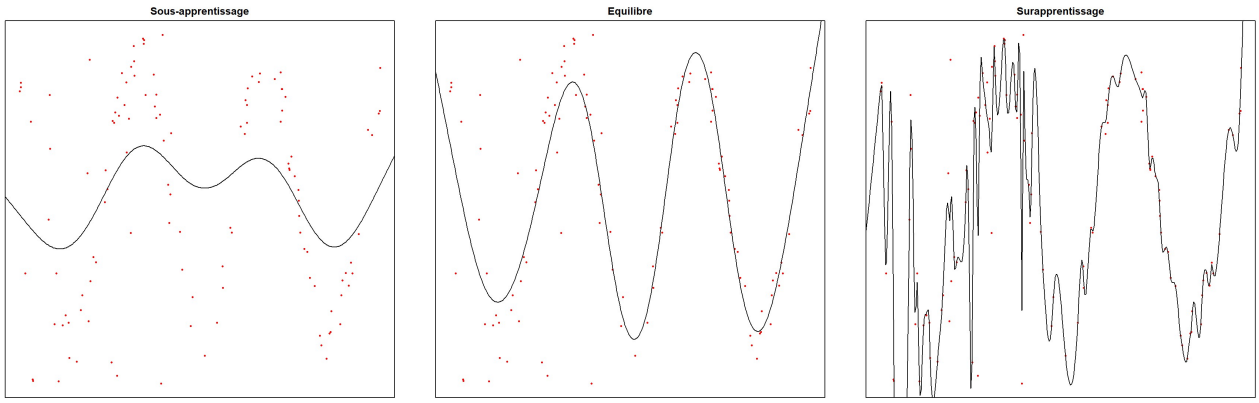


FIGURE 2.8 – GAM : variations de paramètres régissant le caractère lisse du prédicteur.

- g est une fonction de lien de classe $\mathcal{C}^1$ strictement monotone ;
- $\mu_i = \mathbb{E}[Y_i]$, $Y_i \sim$ *loi de famille exponentielle*, les réponses Y_i sont indépendantes conditionnellement aux paramètres μ_i ;
- $\mathbf{X}_i^*$ est le vecteur de $0 \leq n_1 \leq p$ covariables (une sélection parmi les p covariables disponibles) de la i-ème observation, et $n_1 < n$ (comme dans notre application) ;
- $\boldsymbol{\beta}^\star$ est un vecteur de coefficients à estimer de taille n_1 ;
- $f_j(x)$ et $f_{j_1,j_2}(x_1, x_2)$ sont des fonctions d'une ou deux covariables, de classe au moins $\mathcal{C}^1$. Ces fonctions sont des fonctions splines. Nous pouvons par exemple avoir $f_j(urban)$ ou $f_{j_1,j_2}(x, y)$ où x et y sont les coordonnées spatiales.

En pratique, lorsqu'un nombre relativement grand de covariables est disponible, le modèle complet est rarement ajusté, pour des raisons d'identifiabilité, de complexité et d'interprétabilité du modèle. Certaines composantes du modèle sont alors annulées afin d'estimer uniquement les composantes d'intérêt. Par exemple, le prédicteur linéaire classique $\mathbf{X}_i^*\boldsymbol{\beta}^*$ contient typiquement seulement une partie des p covariables, notamment dans le cas où la contribution de certaines covariables est modélisée via une des fonctions non linéaires f_j ou f_{j_1,j_2}. La partie $\mathbf{X}_i^*\boldsymbol{\beta}^*$ est aussi appelée prédicteur paramétrique, car elle comporte au maximum un paramètre par covariable, contrairement aux fonctions f_j et f_{j_1,j_2} dont les degrés de liberté seront déterminés de façon semi-paramétrique grâce aux fonctions splines.

2.2.3 Fonctions splines

Afin de préserver une structure linéaire dans le prédicteur $g(\mu_i)$, plus facilement exploitable lors de l'estimation des paramètres en grande dimension dans l'équation (2.2), nous choisissons des fonctions splines pour les fonctions $f(.)$.

2.2.3.1 Cas univarié

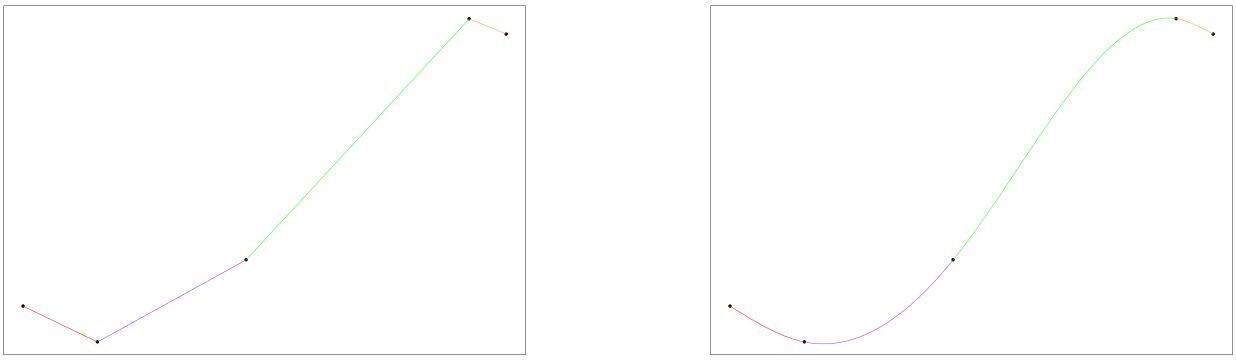
Une fonction spline (notée f de façon générique par la suite) de dimension $r = k$ peut être définie comme suit :

$$f(x) = \sum_{\ell=1}^{k} \beta_\ell b_\ell(x) \tag{2.3}$$

Il s'agit de la combinaison linéaire de certaines fonctions de base $b_\ell(x)$, c'est-à-dire des éléments de l'espace vectoriel des fonctions auquel f appartient. Autrement dit, un nombre k de points distincts est fixé dans le support de la j-ème covariable. Ces points sont appelés noeuds et utilisés pour ancrer les fonctions de base b_ℓ. En fonction de la distribution des valeurs x_{ij} de cette covariable, les points sont souvent choisis soit de façon uniforme, soit comme certains quantiles de cette distribution (par exemple, des quantiles « équiprobables »). Par défaut, l'implémentation du package *mgcv* de R utilise des noeuds équidistants. Différents choix sont possibles pour la base b_ℓ et déterminent les propriétés de la fonction spline, comme le nombre de dérivées continues de la fonction f.

Deux exemples sont illustrés dans les figures 2.9a et 2.9b : la spline linéaire (par morceaux), et la spline cubique (c'est-à-dire ayant le comportement local d'un polynôme cubique), ici dans un contexte d'interpolation où la fonction spline doit passer par les points des données. Soit l'ensemble $E \in \mathbb{R}^2$ de cinq points (trois points

intérieurs et deux points terminaux). Une fois un coefficient β_ℓ associé à chaque fonction de base b_ℓ, $\ell = 1, \ldots, 5$, nous obtenons les fonctions splines des figures. Sur la figure 2.9a, les fonctions de base (et la fonction spline qui en résulte) sont des fonctions linéaires par morceau et continues. Sur la figure 2.9b, les fonctions de base sont des polynômes de degré trois, et la fonction spline est continue en sa deuxième dérivée. Les splines cubiques sont caractérisées par l'utilisation de polynômes de degré trois, ce qui permet d'obtenir une fonction spline 2.9b plus lisse et de courbure continue.



(a) Spline linéaire

(b) Spline cubique

FIGURE 2.9 – Exemples de fonctions splines

Les splines cardinales (Farouki, 2008) et les splines dites à « plaque mince »(TPS, *Thin Place Spline* en anglais) sont encore d'autres possibilités pour définir la base, et elles possèdent des propriétés intéressantes, par exemple pour imposer un certain comportement aux noeuds. Dans le cas des splines cardinales, nous avons $f(x_\ell) = \beta_\ell$ si x_ℓ est le ℓ-ième noeud utilisé pour construire la base, et cette propriété simplifie l'interprétation des coefficients $\boldsymbol{\beta}$ estimés. Différentes bases peuvent représenter le même espace vectoriel de fonctions, notamment si l'ensemble des noeuds coïncide entre ces bases.

La Figure 2.10 illustre la construction d'une spline cubique avec la base cardinale. À gauche est représentée la quatrième fonction de base. À droite, en pointillés sont représentées les sept fonctions de base correspondant aux sept noeuds équidistants. En ligne continue est représentée la spline cubique une fois chaque base multipliée par un coefficient puis sommées.

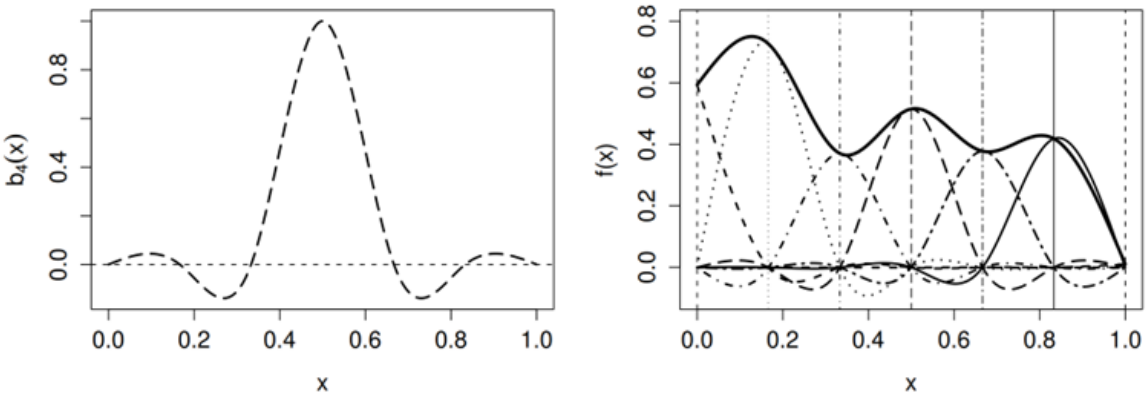


FIGURE 2.10 – Construction d'une fonction spline avec base cardinale(Wood, 2006). La fonction $b_4(x)$ est la 4ème spline sur 6 de la fonction (courbe continue) du graphe de droite. $b_4(x)$ est centrée sur le noeud $x = 0.5$.

Afin de mieux contrôler la courbure aux noeuds terminaux, il est possible d'imposer des contraintes supplémentaires. Si $x_{(0)} < x_{(n)}$ sont les deux noeuds terminaux, les splines dites naturelles annulent la courbure en ces extrémités : $f''(x_{(0)}) = f''(x_{(n)}) = 0$. Les bases cardinales facilitent l'implémentation de cette propriété.

Finalement, contraindre la fonction à être périodique aux noeuds terminaux (c'est-à-dire avec l'égalité des valeurs de f en $x_{(1)}$ et $x_{(k)}$ jusqu'à ses dérivées secondes) est utile dans certains cas. Par exemple, cela permet de tenir compte du comportement saisonnier pour une covariable indiquant le jour, la semaine ou le mois de l'année. Dans le package mgcv de R, cette contrainte est disponible pour la classe des splines cubiques cycliques. La Figure 2.11 illustre la construction d'une spline cubique cyclique, comme à la Figure 2.10.

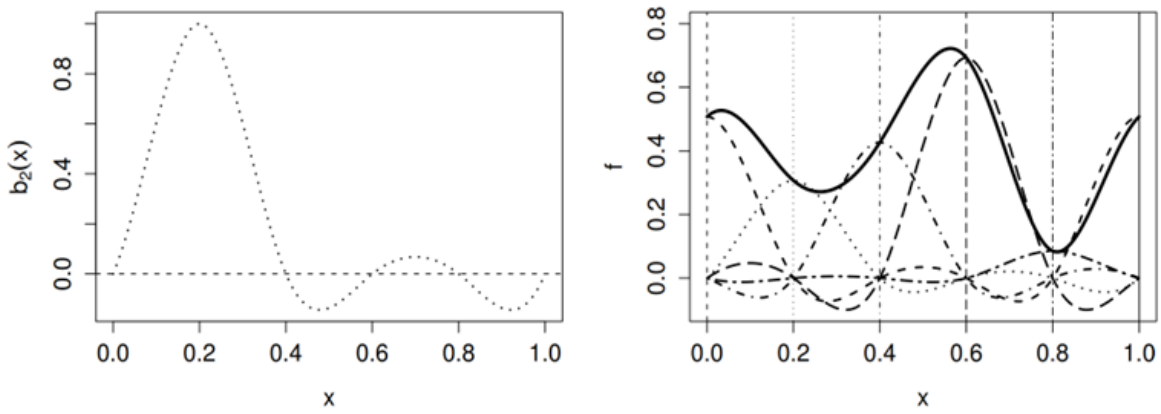


FIGURE 2.11 – Construction d'une fonction spline cubique cyclique (Wood, 2006). La fonction $b_2(x)$ est la 2ème spline sur 6 de la fonction (courbe continue) du graphe de droite. $b_2(x)$ est centrée sur le noeud $x = 0.2$.

Pénalisation pour contrôler le caractère lisse Dans notre cadre de la régression statistique (et non de l'interpolation exacte entre les points illustrée dans les figures 2.9a et 2.9b), l'utilisation d'un ensemble de n noeuds devient trop fastidieux dans les cas où n est très grand, et nous sélectionnons plutôt un ensemble de k noeuds avec $k < n$. De plus, afin d'éviter le sur-apprentissage, une nouvelle problématique entre en jeu : il s'agit de « *prioriser une variation globale plutôt que les variations locales* »(Côté, 2016). Cela signifie que la fonction spline doit capter la forme générale de la fonction qu'on cherche à modéliser sans pour autant y intégrer les erreurs de mesure (il s'agit d'un concept semblable aux moindres carrés en régression linéaire). Pour ce faire, il est nécessaire de pouvoir contrôler le lissage de f, en pénalisant les fonctions trop irrégulières.

Le caractère oscillatoire d'une fonction est décrit par sa dérivée seconde. Au carré, la dérivée seconde permet de quantifier l'ampleur de la courbure de f :

$$\int_{x_{(1)}}^{x_{(n)}} (f''(x))^2 dx. \tag{2.4}$$

Ainsi, l'équation (2.4) est utilisée comme pénalité de la fonction objectif (ici, la log-vraisemblance du modèle), elle-même ajustée par un paramètre de lissage, λ. Considérons un modèle relativement simple avec une seule fonction spline dans le prédicteur : $g(\mu_i) = f(x_i)$. En notant $\boldsymbol{\beta}$ le vecteur des k coefficients de la fonction spline et $\ell(\boldsymbol{\beta})$ la log-vraisemblance du modèle, l'objectif est alors de minimiser :

$$(\boldsymbol{\beta}, \lambda) \mapsto -\ell(\boldsymbol{\beta}) + \lambda \int_{x_{(1)}}^{x_{(n)}} (f''(x))^2 dx.$$

L'hyperparamètre de pénalité $\lambda \geq 0$ permet ici de contrôler le lissage de la fonction $f(.)$. Ainsi, si λ est nul, $f(.)$ peut osciller et va tendre à sur-apprendre. Quand λ tend vers l'infini, $f(.)$ tend à être constante ou linéaire. Si le prédicteur comprend plusieurs fonctions splines, on additionne leurs termes de pénalisation respectifs, et chaque terme est contrôlé par son propre hyperparamètre λ.

Grâce à la pénalisation, le choix du nombre de fonctions de base k est facilité en pratique. En général, un k relativement petit va limiter les différentes formes possibles que la fonction estimée $\hat{f}$ peut prendre, mais pour une valeur de k modérée voire élevée (par exemple s'approchant de n), les fonctions estimées pour différentes valeurs de k vont fortement se ressembler en raison de la pénalisation. Afin de limiter le coût et la complexité des calculs numériques, le choix d'une valeur relativement modérée de k est pertinent dans la plupart des cas.

2.2.3.2 Cas bivarié

Différentes bases de splines permettent de modéliser un effet d'interaction de deux covariables scalaires (par exemple, les deux coordonnées spatiales). La modélisation des interactions entre plus de deux variables est également possible mais ne sera pas abordée ici.

Un premier choix consiste à utiliser les splines dites à « plaque mince »ou Thin Plate Splines (TPS). Une propriété importante des TPS est leur isotropie, c'est-à-dire que les fonctions de base sont invariantes à une rotation de l'espace. Or, dans le cas d'une modélisation spatio-temporelle, l'utilisation de fonctions lisses ne permet pas de capter les différences d'échelle, par exemple entre l'espace et le temps. Ce changement d'échelle est plus facilement capté par les produits tensoriels, au contraire des splines à plaque mince. De plus, les produits tensoriels permettent d'appliquer une pénalité pour chacune des dimensions.

Produit tensoriel « lisse » Nous nous limitons à nouveau à la modélisation de l'interaction de deux covariables. Soient les deux splines cubiques de régression suivantes, représentant les covariables x_1 et x_2 :

$$f_1(\boldsymbol{x}_1) = \sum_{\ell_1=1}^{k_1} \beta_{\ell_1} b_{\ell_1}(\boldsymbol{x}_1) \qquad \text{et} \qquad f_2(\boldsymbol{x}_2) = \sum_{\ell_2=1}^{k_2} \beta_{\ell_2} b_{\ell_2}(\boldsymbol{x}_2)$$

où k_1 et k_2 sont les nombres de noeuds respectifs des splines. Afin d'obtenir une fonction lisse dépendant des deux prédicteurs, nous pouvons utiliser produit tensoriel des deux covariables, et nous obtenons alors la fonction suivante, également appelée tenseur :

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{\ell_1=1}^{k_1} \sum_{\ell_2=1}^{k_2} \beta_{\ell_1 \ell_2} b_{\ell_1}(\boldsymbol{x}_1) b_{\ell_2}(\boldsymbol{x}_2)$$

La fonction objective du tenseur peut également être pénalisée, et de manière individuelle pour chacun des prédicteurs. Ainsi, l'estimation d'un modèle avec une telle composante dans $g(\mu_i)$ est réalisée par minimisation de la fonction objectif suivante :

$$(\boldsymbol{\beta}, \lambda_1, \lambda_2) \mapsto -\ell(\boldsymbol{\beta}) + \int_{x_1, x_2} \left(\lambda_1 \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + \lambda_2 \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \right) dx_1 dx_2,$$

où $\lambda_1, \lambda_2 \geq 0$.

Précisons qu'il est possible d'utiliser différents types de bases univariées dans cette construction, et il est même possible de choisir des bases différentes pour les deux prédicteurs univariés.

2.2.4 Ajustement de modèles additifs généralisés

Revenons à la structure du modèle (2.2). Les fonctions $f(\cdot)$ sont construites d'après la description des splines précédentes. Notons $\boldsymbol{x} \in \mathbb{R}^{n \times \ell}$, $\ell \in \{1, 2\}$, le vecteur de covariables en entrée d'une fonction spline. Pour chacune des $R \geq 1$ fonctions, nous obtenons alors un vecteur de ses valeurs pour les n observations des covariables :

$$f^{(r)} \equiv f_r(\boldsymbol{x}) = \sum_{\ell=1}^{k_r} \beta_\ell^{(r)} b_\ell^{(r)}(\boldsymbol{x}) = \tilde{\boldsymbol{X}}^{(r)} \tilde{\boldsymbol{\beta}}^{(r)}, \quad r \in \{1, 2, \ldots, R\},$$

où $\tilde{\boldsymbol{X}}^{(r)}$ est la matrice dont les entrées sont données par $b_\ell^{(r)}(x_i)$ en ligne i et colonne ℓ. Selon la structure de l'équation des prédicteurs globaux μ_i (par exemple, comprenant une constante de régression β_0 et une ou plusieurs fonctions splines), le modèle n'est pas identifiable, à moins de contraintes de centralité. Par exemple, pour la plupart des fonctions splines f, un certain choix des coefficients permet de représenter une constante de régression c quelconque ($c \in \mathbb{R}$ et $f(x) = c$), ce qui entraîne un problème de confusion entre la constante de régression β_0 et f, et aussi entre différentes fonctions splines, s'il y en a plusieurs. En effet, il existe différents vecteurs $\tilde{\boldsymbol{\beta}}^{(r)}$ tels que la somme des fonctions $f(\cdot)$ donne le même résultat dans le vecteur des prédicteurs globaux $\boldsymbol{\mu}$. En cas de non identifiabilité, il est donc nécessaire d'imposer la contrainte suivante, pour toute fonction f_r,

$$\mathbf{1}_n^T f^{(r)} = 0;$$

cette contrainte linéaire enlève un degré de liberté de la fonction spline. Sur le plan algorithmique, le modèle est alors re-paramétrisé , et le vecteur des valeurs de la spline re-paramétrisée est alors réécrite de la façon suivante : $f^{(r)} = \boldsymbol{X}^{(r)} \boldsymbol{\beta}^{(r)}$. Chaque vecteur $\boldsymbol{\beta}^{(r)}$ ne possède plus que $k_r - 1$ paramètres à estimer, et la structure du modèle (2.2) se réécrit simplement comme celle d'un GLM :

$$g(\mu_i) = \boldsymbol{X}_i \boldsymbol{\beta}$$

où $\boldsymbol{X}_i = [\boldsymbol{X}^*, \boldsymbol{X}_i^{(1)}, \ldots, \boldsymbol{X}_i^{(R)}]$ et $\boldsymbol{\beta} = [(\boldsymbol{\beta}^*)^T, (\boldsymbol{\beta}^{(1)})^T, ..., (\boldsymbol{\beta}^{(R)})^T]^T$. En principe, le vecteur de coefficients $\boldsymbol{\beta}$ pourrait alors être estimé par maximisation de la fonction de vraisemblance, comme vu pour les GLM. En plus, pour éviter le sur-apprentissage, la fonction de vraisemblance est pénalisée. Comme avant, notons $ll(\boldsymbol{\beta})$ la log-vraisemblance définie comme la somme des logarithmes des densités évaluées en chacune des n observations, selon la loi de réponse choisie. Pour estimer les paramètre du GAM, on minimise la fonction objectif suivante incluant les pénalités des courbures des fonctions splines avec le vecteur des hyperparamètres de pénalité $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_R)$:

$$(\boldsymbol{\beta}, \boldsymbol{\lambda}) \mapsto -ll_P(\boldsymbol{\beta}; \boldsymbol{\lambda}) = -ll(\boldsymbol{\beta}) + \frac{1}{2} \sum_{r=1}^{R} \lambda_r \left(\boldsymbol{\beta}^{(r)} \right)^T S_r \left(\boldsymbol{\beta}^{(r)} \right),$$

où

$$\left(\boldsymbol{\beta}^{(r)}\right)^T S_r \left(\boldsymbol{\beta}^{(r)}\right) = \int (f_r''(x))^2 dx,$$

et S_r est la matrice de pénalité de la fonction spline f_r. En raison de la haute complexité de ce problème d'optimisation en grande dimension, on procède souvent par des techniques d'estimation en deux étapes (d'abord $\boldsymbol{\lambda}$, ensuite $\boldsymbol{\beta}$). Les valeurs des λ_j peuvent alors être estimées selon une méthode comme l'Estimation Restreinte du Maximum de Vraisemblance (REML) ou la validation croisée généralisée (GCV) (voir Wood (2006)). Finalement, les dérivées partielles de $ll_P(\boldsymbol{\beta}; \boldsymbol{\lambda})$ sont calculées. Les équations $\frac{\partial ll_P}{\partial \beta_k} = 0$ sont résolues par la méthode *Penalized Iteratively Re-weighted Least Squares*, méthode décrite par Wood (2006).

2.3 Lois de probabilités

Différentes distributions sont utilisées pour la loi de probablité de la réponse dans ces régressions de type GAM et permettent ainsi d'exprimer la log-vraisemblance à maximiser en (2.2.4). Celles-ci sont brièvement décrites ci-dessous.

2.3.1 Loi de Poisson

La loi de Poisson est la loi de probabilité fondamentale pour les comptages. Cette loi discrète décrit le nombre d'occurrences d'un événement dans un intervalle de temps ou d'espace donné, dont la fréquence d'apparition moyenne est connue et pour lequel chaque événement est indépendant du précédent. Soit la variable aléatoire $Y \sim \mathcal{P}(\mu)$. La probabilité que Y prenne la valeur y, c'est-à-dire la fonction de masse de la loi, est de :

$$\begin{aligned} f_\mu(y) &= \frac{\mu^y}{y!} \exp(-\mu), \qquad y \in \mathbb{N}, \qquad \mu \in]0, +\infty[\\ &= \exp(-\mu) \exp\left[\log(\mu^y) - \log(y!)\right] \\ &= \exp\left[y \log(\mu) - \mu - \log(y!)\right] \end{aligned}$$

Ainsi, la distribution de Poisson appartient à la famille exponentielle, et l'on a $\theta = \log(\mu)$, $\phi = 1$, $a(\phi) = \phi = 1$, $b(\theta) = \exp(\theta) = \exp(\log(\mu)) = \mu$ et $c(y, \phi) = -\log(y!)$.

L'espérance et la variance de la variable Y sont égales et ont pour valeur μ.

2.3.2 Loi binomiale négative

La loi binomiale négative est une loi discrète ayant un rôle fondamentale en combinatoire car elle réalise le dénombrement d'échecs à l'issue d'épreuves de Bernoulli de probabilité p, avant l'obtention de n succès. Dans le contexte de la régression pour données de comptage, elle est appréciée pour sa propriété de sur-dispersion la rendant plus flexible que la la loi de Poisson. Par contraste avec celle-ci, sa variance peut être supérieure à son espérance, traduisant ainsi une plus forte stochasticité.

De façon constructive, une loi binomiale négative peut être obtenue à partir d'une loi de Poisson en supposant que l'espérance μ de celle-ci suive une loi Gamma avec espérance μ et variance $\mu/\sqrt{\nu}$, où $\mu, \nu > 0$ (Lord et al., 2010). Alors la loi binomiale négative a pour espérance μ et pour variance $\mu + \mu^2/\nu > \mu$. Lorsque $\nu \to \infty$, la loi binomiale négative tend vers une loi de Poisson d'intensité μ. Soit la variable aléatoire $Y \sim \mathcal{BN}(\mu, \nu)$. La probabilité que Y prenne une valeur $y \in \mathbb{N}$ est de :

$$\begin{aligned} f_{BN}(y) &= \frac{\Gamma(y+\nu)}{\Gamma(y+1)\Gamma(\nu)} \left(\frac{\nu}{\mu+\nu}\right)^\nu \left(\frac{\mu}{\mu+\nu}\right)^y \\ &= \exp\left(y \log\left(\frac{\mu}{\mu+\nu}\right) + \dots\right) \end{aligned} \tag{2.5}$$

Il peut être démontré que la loi binomiale négative appartient à la famille exponentielle avec $\theta = \log \frac{\mu}{\mu+\nu}$, mais seulement si son paramètre ν est fixé. En pratique, il faut estimer ν, et l'algorithme d'estimation des GAMs avec réponse de type binomiale négative procède par une estimation itérative alternant entre ν et $\boldsymbol{\beta}$.

2.3.3 Zero-inflated Poisson

Le modèle Zero-inflated Poisson (ZIP) est utilisé dans le cas où la fréquence d'apparition de zéros est trop élevée par rapport à une loi de Poisson. Ce modèle, introduit par Lambert (1992) possède deux processus de

génération de zéros. Soit la variable aléatoire Y décrivant le comptage d'événements, tel que la distribution de Poisson, et pour laquelle la fréquence de zéros est élevée. Alors,

$$\begin{aligned} &Y \equiv 0, \quad \text{avec probabilité } \pi \in [0,1], \\ &Y \sim \mathcal{P}(\lambda) \quad \text{avec probabilité } 1-\pi. \end{aligned}$$

Ainsi, nous avons :

$$\begin{aligned} &\mathbb{P}(Y=0) = \pi + (1-\pi)e^{-\lambda}, \\ &\mathbb{P}(Y=y) = (1-\pi)\frac{\lambda^y}{y!}e^{-\lambda}, \quad \forall y \in \mathbb{N}^* \end{aligned}$$

Il est également possible de modéliser une forte fréquence de zéros dans le cas où la variable suit une loi binomiale négative de mêmes paramètres qu'en (2.5). Soit $Y \sim \mathcal{ZINB}(\pi; \mu, \nu)$. Y a les probabilités suivantes :

$$\begin{aligned} &\mathbb{P}(Y=0) = \pi + (1-\pi)f_{BN}(0) \\ &\mathbb{P}(Y=y) = (1-\pi)f_{BN}(y), \quad y \in \mathbb{N}^* \end{aligned}$$

2.4 Critères de validation et sélection de modèles

Il est important de pouvoir sélectionner le meilleur modèle parmi un ensemble de modèles et ensuite de savoir si le modèle sélectionné est valide ou non, par exemple en vérifiant si la distribution des résidus de régression (c'est-à-dire les différence entre observations et valeurs prédites) sont cohérents avec la loi choisie pour la variable de réponse. Certains critères numériques, dont seulement l'interprétation relative des valeurs est pertinente, se prêtent principalement pour la sélection et non pour la validation ; un exemple est le critère AIC. L'étude de la qualité des prédictions d'un modèle grâce divers critères permet de comparer et de valider des modèles. Dans la suite, quelques critères courants utilisés dans cette étude sont rappelés.

2.4.1 Erreur quadratique moyenne

L'écart quadratique moyen (RMSE, pour Root Mean Squared Error) permet dans un premier temps de donner une idée de la qualité de prédiction d'un modèle. Il s'agit de l'écart-type de l'erreur de prévision. Ainsi, pour un modèle dont la fonction de prédiction est $h(\boldsymbol{x})$, alors le RMSE est donné par :

$$RMSE(Y) = \sqrt{\sum_{i=1}^{n}(y_i - h(\boldsymbol{x}_i))^2}$$

Dans nos modèles, $h(\boldsymbol{x}) = \mu(\boldsymbol{x})$, avec μ l'espérance de la loi de réponse du modèle de régression. Le RMSE est calculé pour chaque modèle. Le modèle le plus performant est alors le modèle dont l'erreur quadratique est la plus faible.

2.4.2 Courbes ROC et aire sous la courbe AUC

Finalement, il est également intéressant de pouvoir vérifier la performance d'un modèle à prédire certains types de données, telles que les données binaires. Dans ce travail de recherche, nous nous intéressons notamment à la prédiction de zéros mais également de comptages élevés de signalements par unité spatio-temporelle, i.e. les valeurs supérieures au 95$^{\text{ème}}$ percentile, soit les comptages supérieurs à 2. Une formulation binaire est alors obtenue en considérant la fonction indicatrice d'un dépassement de seuil, comme les seuils de 0 ou de 2, respectivement, pour les deux types de prédiction cités avant. La courbe ROC (*Receiving Operator Characteristic*) est une méthode permettant de quantifier ces performances.

Prenons l'exemple des valeurs nulles. Toute observation et toute prédiction est classifiée : elle vaut 1 si elle est nulle, 0 sinon. La matrice de confusion est alors données par le Tableau 2.3. Dans ce tableau, nous avons les vrais/faux positifs/négatifs :

- $VP = \sum_i \mathbb{1}_{\{obs_i=1, pred_i=1\}}$;
- $FN = \sum_i \mathbb{1}_{\{obs_i=1, pred_i=0\}}$;

- $FP = \sum_i \mathbb{1}_{\{obs_i=0, pred_i=1\}}$;
- $VN = \sum_i \mathbb{1}_{\{obs_i=0, pred_i=0\}}$.

Observation \ Prédiction	**1**	**0**
1	$VP \rightarrow$ *Vrai Positif*	$FN \rightarrow$ *Faux Négatif*
0	$FP \rightarrow$ *Faux Positif*	$VN \rightarrow$ *Vrai Négatif*

TABLE 2.3 – Matrice de confusion

Soit la fonction $h(\boldsymbol{x})$ renvoyant la prédiction des observations. Soit un seuil s déterminant pour chacune des prédictions sa classification :
- Si $h(\boldsymbol{x}_i) > s$, alors l'observation i est classée en 1 ;
- Sinon, l'observation est classée en 0.
La sensibilité $Se(s)$ et la spécifité $Sp(s)$ sont alors calculées pour différents seuils s :

$$Se(s) = \frac{VP(s)}{VP(s) + FN(s)} = \frac{VP(s)}{P(s)}$$
$$Sp(s) = \frac{VN(s)}{VN(s) + FP(s)} = \frac{VN(s)}{N(s)}$$

La courbe ROC représente alors l'évolution de la sensibilité (taux de vrais positifs) en fonction de 1− la spécificité (taux de faux positifs) lorsque le seuil s varie. La courbe ROC est une courbe nonnegative et croissante entre les points $(0, 0)$ et $(1, 1)$. Lorsqu'une prédiction est peu informative (et donc très aléatoire), la courbe est proche de la bissectrice. Une prédiction parfaite correspond à l'union du point $(0, 0)$ et de la droite $Se = 1$ pour tout point de l'antispécifité appartenant à $]0, 1]$. Ci-dessus, le modèle dont la courbe ROC est noire est plus performant que les deux autres modèles dont les courbes sont rouge ou verte. Une manière de résumer la qualité de la courbe en une seule valeur est de calculer l'aire sous celle-ci (AUC : *Area Under Curve*). Ainsi, plus un modèle est performant en terme de prédiction, plus l'AUC tend vers 1. Si le modèle se rapproche d'un modèle aléatoire, alors son AUC tend vers 0,5.

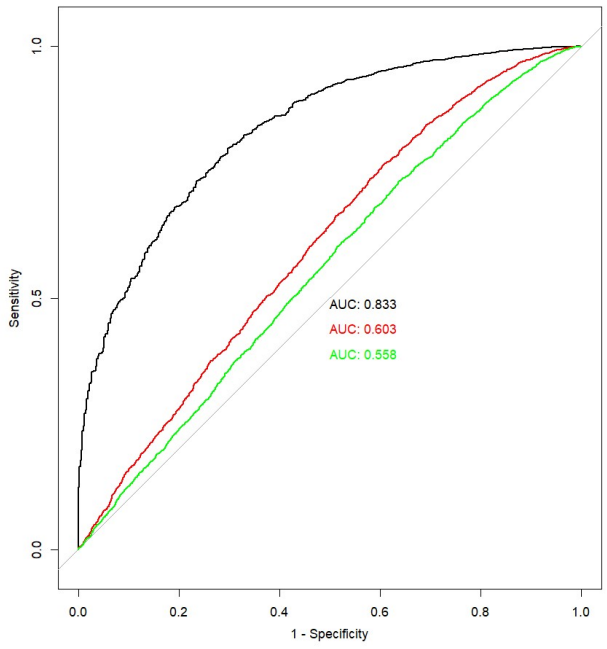


FIGURE 2.12 – Exemples de courbes ROC

2.4.3 Critère AIC

Le critère d'information Akaike permet de calculer la vraisemblance d'un modèle tout en la pénalisant sa complexité exprimée par le nombre k de paramètres à estimer. En effet, augmenter le nombre de paramètres augmente la qualité d'ajustement mais aussi sa variance.

$$AIC(\mathcal{M}) = 2k - 2\log(ll(\mathcal{M}))$$

Ainsi, le modèle ayant l'AIC le plus faible est considéré comme le plus performant. L'AIC permet de sélectionner le modèle le plus performant tout en pénalisant les modèles trop complexes, i.e. les modèles dont la dimension de l'espace des paramètres se rapproche fortement du nombre d'observations. Ainsi, l'AIC permet d'éviter le sur-apprentissage. Dans le cas des GAMs, faisant déjà intervenir une pénalisation de la log-vraisemblance, nous calculons pour k une valeur correspondant à un nombre de coefficients effectif.

2.4.4 Implémentation

L'ensemble des modélisations a été réalisé sur le logiciel R version 4.1.1 grâce à la librairie *mgcv*. De plus, afin d'accélérer l'exécution des modèles, un tirage aléatoire de 30 000 des 96 327 observations est réalisé. Il est vérifié que l'ensemble des pixels et l'ensemble des mois de la période considérée sont représentés dans ce sous-échantillon. L'ensemble des modélisations de la section 3.1 est réalisé uniquement sur ces 30 000 données. Les trois distributions précédentes sont testées pour différents modèles, c'est-à-dire pour différentes combinaisons de covariables. La fonction *gam* de la librairie *mgcv* librairie est utilisée pour l'estimation des paramètres. La fonction *gam.check*, retourne un diagnostique des résidus. Dans un premier temps, la table retourne les résultats de tests statistiques réalisés sur les fonctions splines et tenseurs. Le test fait l'hypothèse H0 : « Les résidus sont aléatoires ». Ainsi, une *p-value* inférieure à 0.05 implique que la distribution des résidus n'est pas aléatoire, et qu'une tendance, qui n'a pas été captée par la spline ou le tenseur, persiste. Cela s'explique généralement par une dimension de la spline trop faible. Il est alors conseillé d'augmenter le nombre k de fonctions de base. Dans un second temps, la fonction *gam.check* retourne un diagnostique graphique des résidus.

Enfin, la fonction *concurvity* fournit la « concordance »de modèles GAM, c'est-à-dire permet de révéler des relations entre covariables autres que linéaires. L'analyse se fait notamment à la colonne *worst*, précisant à quel point les termes lisses sont déterminés par les autres. Lorsqu'une valeur est supérieure à 0.8, il est important de regarder de plus près les relations entre covariables et de faire attention à toute interprétation.

Résultats

Dans ce chapitre, nous commencerons par présenter un premier modèle de base implémenté sur le logiciel R. Puis, nous détaillerons toutes les démarches ayant permis de construire le modèle final. Enfin, nous analyserons ce dernière modèle.

3.1 Construction des modèles

3.1.1 Choix des lois de réponse

L'objectif est ici de pouvoir modéliser le comptage d'occurrences, $y_i \in \mathbb{N}$, de déclarations de piqûres de tiques par voxel, dont les effectifs sont illustrés à la Figure 3.1. Dans un premier temps, la loi de Poisson est un candidat naturel et pourrait permettre de décrire raisonnablement ce type d'événements. Cependant, ce type de données est souvent sur-dispersé (avec une variance des observations supérieure à leur espérance), ce qui n'est pas modélisable par une loi de Poisson. La loi de Poisson modélise des événements dont l'espérance est égale à la variance. Ainsi, une alternative courante est l'utilisation de la loi binomiale négative. En effet, comme vu en section 2.5, la loi binomiale négative tend vers une loi de Poisson lorsque son paramètre de sur-dispersion ν tend vers l'infini, et la plupart des implémentations de modèle de régression (GLM, GAM) proposent cette loi. Finalement, il peut y avoir un excès de zéros dans la distribution des comptages, non explicable par les covariables disponibles et par la structure de la loi de réponse. Par conséquent, une seconde alternative est de modéliser Y par une loi de Poisson zéro-enflée (ZIP ; voir section 2.3.3). Un premier processus est alors en charge de générer les zéros, tandis que le second gère une distribution de Poisson.

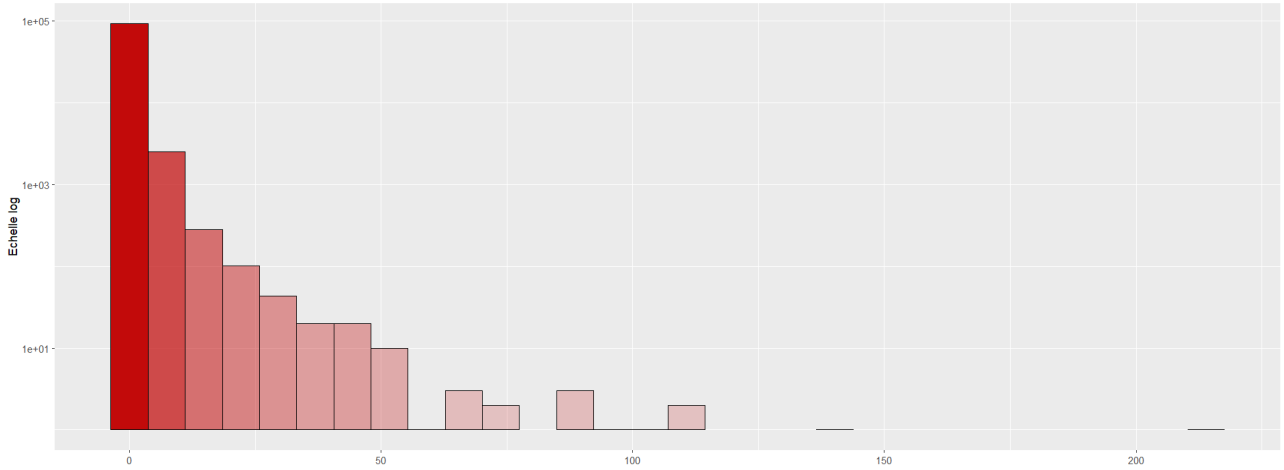


FIGURE 3.1 – Distribution des observations y_i, fréquences à l'échelle logarithmique

3.1.2 Un premier modèle de base

Dans un premier temps, une recherche manuelle de multiples modèles assez simples est effectuée. Ainsi, nous souhaitons estimer les modèles ayant la structure suivante dans leur prédicteur linéaire :

$$\begin{aligned}\mathcal{M}_0 : g(\mu) = {} & lanmap + te(urban, forest) \\ & + s(mois, k = 3, bs = "cc") + annee + te(x, y) + offset(log(area)) + offset(log(pop))\end{aligned} \tag{3.1}$$

où g est la fonction de lien correspondant à la famille choisie, et chaque terme additif représente une covariable décrivant un processus spécifique. Le terme *lanmap* est le facteur climatique LANMAP à 8 niveaux. La fonction $s()$ est une fonction spline cubique de régression (et cyclique d'après le paramètre *bs*), et $te()$ est le produit tensoriel de deux fonctions splines cubiques. Ces notations sont conservées pour toute la suite des modélisations.

Pour chaque type de covariable (climatique, couverture de sol...), nous considérons différentes façons d'estimer son effet sur le nombre de déclarations. Les facteurs et covariables sont d'abord estimés linéairement (avec un seul coefficient par niveau de facteur ou par covariable continue). Ensuite, afin de capter de potentiels effets non linéaires, les covariables quantitatives sont estimées par une spline cubique de régression d'un certain nombre k de fonctions de base (qui peut être adapté selon la covariable considérée), tel que pour la covariable *mois* du modèle $\mathcal{M}_0$. Lorsque nous souhaitons modéliser une interaction entre covariables, par exemple entre longitude x et latitude y ou entre *urban* et *forest* dans le modèle $\mathcal{M}_0$, nous utilisons un tenseur de nombre k_1 de fonctions de base pour la première covariable et de nombre k_2 de fonctions de base pour la seconde. Par défaut, $k_1 = k_2 = 5$. Dans le cas où ces valeurs sont modifiées, celles-ci sont précisées. Lorsque le type de fonction utilisé pour chacune des covariables n'est pas précisé, nous utilisons des splines cubiques de régression.

Les variables d'aire (*area*) et de population (*pop*) sont précisées comme étant des termes « offset », c'est-à-dire des termes d'échelle pour le nombre de déclarations, déterministes et fixés (sans coefficient à estimer). Par exemple, deux voxels observant chacun trois déclarations mais dont l'aire est différente n'auront pas le même poids dans la modélisation. En effet, la quasi-totalité des pixels limitrophes ne possèdent pas la même aire que les pixels intérieurs. Il s'agit de la même chose pour la covariable démographique. L'inclusion de ces termes « offset »permet alors de rétablir l'équilibre entre ces différences en définissant un comportement réaliste en absence d'autres prédicteurs, mais également de prédire le nombre de déclarations à l'échelle du pixel.

Les premiers modèles ont été validés puis comparés grâce aux outils décrits en Section 2.4 et ont permis d'aboutir au premier modèle $\mathcal{M}_0$, dont la structure est donnée en 3.1.

Choix de la loi de réponse Dans cette recherche d'un premier modèle de base, les trois lois énoncées en Section 3.1.1 sont testées comme potentielle distribution de l'échantillon. Les AIC relatifs au premier modèle pour lequel $Y \sim \mathcal{P}$ sont répertoriés à la Figure 3.1.

Loi	**Poisson**	**Binomiale négative**	**ZIP**
AIC relatif	0	-5038.542	-2089.816

TABLE 3.1 – AIC relatifs au modèle de loi de Poisson

Ainsi, la loi binomiale négative semble la plus adaptée à notre échantillon, si la structure du prédicteur est choisie comme en (3.1). De plus, dans ce modèle de base, le paramètre de sur-dispersion ν estimé est de 0.86. Or, la distribution tend vers une loi Poisson lorsque ν tend vers l'infini. La valeur estimée de ν confirme alors l'hypothèse selon laquelle notre variable réponse est plus dispersée qu'une loi de Poisson. Dans la suite des résultats, nous supposerons donc que nos observations sont distribuées selon une loi binomiale négative. Pour les structures proposées pour le prédicteur linéaire, nous avons systématiquement vérifié que la loi de réponse de type binomiale négative donne toujours un meilleur ajustement en termes d'AIC.

Usage des sols Dans l'équation 3.1, seules les catégories *urban* et *forest*, catégories CLC de niveau 1, ont été conservées. Les niveaux 2 et 3 de CLC ont également été considérés tour à tour mais n'ont pas été conservés pour une question d'interprétabilité. De plus, la catégorie *agric* a été retirée en raison de sa forte corrélation avec la catégorie *forest* (-0.896), ainsi que les catégories *wet* et *water* au vu de leur non-significativité. Les paramètres k (et, par extension, les noeuds utilisés pour construire les fonctions splines) ont été sélectionnés afin d'optimiser l'AIC tout en évitant le sur-apprentissage. Ce phénomène est également contrôlé par l'optimisation du paramètre de lissage λ. La valeur de k varie entre 3 et le nombre de valeurs distinctes prises par le prédicteur concerné. Finalement, nous estimons l'effet de l'interaction entre *urban* et *forest* par un tenseur.

Démographie Dans un premier temps, la covariable démographique a été intégrée au modèle par une spline modélisant l'influence de sa proportion au sein d'un pixel. Cependant, les variables *pop* et *urban* étant très fortement liées, comme illustré à la Figure 3.2, l'intégration des deux covariables simultanément dans un modèle peut créer des problèmes d'identification. Le nombre de déclarations dépendant fortement de la population sous-jacente, en terme d'échelle, la covariable démographique est conservée seulement en *offset*.

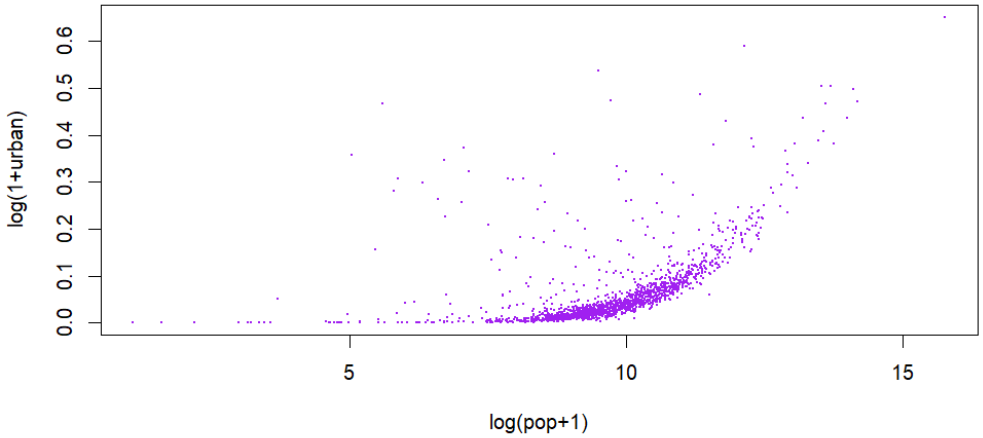


FIGURE 3.2 – Proportion log-transformée de zones urbaines en fonction de la population log-transformée

Prédicteurs temporels L'année étant un facteur, celle-ci est ajouté au modèle comme un prédicteur catégoriel avec un coefficient estimé pour chaque niveau. En ce qui concerne le mois, nous nous attendons à un effet cyclique étant donné sa nature. L'effet est alors modélisé par une spline cubique cyclique. L'effet cyclique est visible à la Figure 3.3. De la même manière que précédemment, le paramètre k est optimisé en retenant la valeur de k minimisant le critère AIC parmi plusieurs valeurs possibles de k, mais également afin de simplifier dans un premier temps les interprétations.

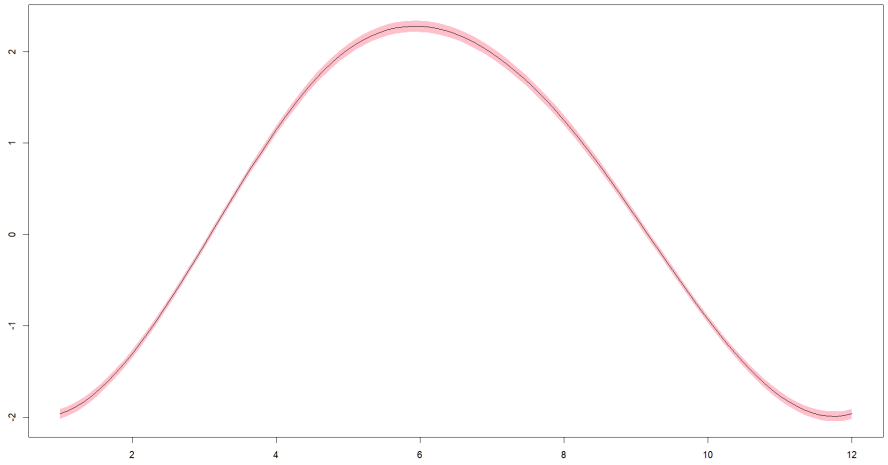


FIGURE 3.3 – Effet partiel du mois sur Y, estimé par fonction spline

3.1.3 Recherche de facteurs et covariables explicatifs

Une fois ce premier modèle validé, nous explorons l'ensemble des covariables créées en Section 2 afin de complexifier le modèle de base par optimisation de l'AIC et pouvoir expliquer au mieux le phénomène de déclarations de piqûres. Les recherches sont expliquées ci-dessous, tandis que chacun des modèles validés est répertorié à la Table 3.5 afin d'être comparé aux autres en termes d'erreur et d'AIC.

3.1.3.1 Présence d'eau

Précédemment, les différentes modélisations ont pu montrer que les représentations des effets des zones humides et des plans d'eau par des splines n'étaient pas significative au seuil de 5%. La Section 2.1.3.3 nous permet de simplifier les covariables *wet* et *water* en créant trois nouvelles covariables :

- $pres.wet = \mathbb{1}_{\{wet>0.05\}}$
- $pres.water = \mathbb{1}_{\{water>0.0069\}}$
- $pres.all\_water = \mathbb{1}_{\{all\_water>0.0019\}}$

Le choix des valeurs seuils a été guidé par des analyses exploratoires, principalement via les arbres de régression. En intégrant chacune à leur tour les covariables précédentes, nous obtenons les AIC relatifs au modèle $\mathcal{M}_0$ à la Table 3.2.

Covariable	***pres.wet***	***pres.water***	***pres.all_water***
AIC relatif	-124.34	-98.31	-102.46

TABLE 3.2 – AIC relatifs au modèle $\mathcal{M}_0$

Le modèle intégrant linéairement la covariable indiquant une présence significative de zones humides est sélectionné par optimisation de l'AIC. Le facteur d'humidité est le plus pertinent au seuil de 5%. Ainsi, le modèle $\mathcal{M}_1$ est défini par

$$\begin{aligned}\mathcal{M}_1 : g(\mu) = lanmap + te(urban, forest) + pres\_wet \\ + année + s(mois, k = 3, bs = "cc") + te(x, y) + offset\end{aligned}$$

où $offset = offset(log(area)) + offset(log(pop))$ (nous conserverons cette notation pour les prochaines équations).

3.1.3.2 Indices de diversité

Plusieurs indices de diversité sont calculés à la section 2.1.3.2 à partir des quatre niveaux de catégories CLC. Ainsi, la covariable $hill_{ij}$, $0 \leq i \leq 2$, $1 \leq j \leq 4$, correspond au nombre de Hill à l'ordre i calculé pour le niveau j. L'effet de chacune de ces covariables est représenté par une spline cubique de régression. Pour chacun des indices, nous testons différentes valeurs de k (nombre de fonctions de base), comprises entre 5 et 10 inclus. Il est important de noter que la richesse des niveaux 1 et 4 ne peuvent être représentés respectivement par plus de 5 et 7 fonctions de base, c'est-à-dire au maximum par le nombre de valeurs uniques. Les modèles testés ont pour structure l'équation (3.3). Pour chacun des indices, nous sélectionnons le modèle à l'AIC le plus faible selon la valeur de k. Les AIC relatifs au modèle $\mathcal{M}_1$ des modèles alors sélectionnés sont récapitulés à la Table 3.3.

$$\begin{aligned}g(\mu) = lanmap + te(urban, forest) + pres.wet + s(hill_{ij}) \\ + année + s(mois, k = 3, bs = "cc") + te(x, y) + offset\end{aligned} \tag{3.2}$$

Ordre \ **Niveau**	**1**	**2**	**3**	**4**
0	-15.57	-19.75	-12.16	-9.08
1	8.42	2.20	-19.18	-5.09
2	9.24	2.02	-0.03	-14.25

TABLE 3.3 – AIC relatifs au modèle $\mathcal{M}_1$ selon différentes approches pour rajouter les indices de diversité dans ce modèle.

Nous pouvons voir que l'indice d'ordre 0, soit la richesse, appliquée au niveau 2 (détaillée en Annexe 4.1) de couverture des sols améliore le plus l'AIC.

La structure du modèle 2 est alors

$$\begin{aligned}\mathcal{M}_2 : g(\mu) = lanmap + te(urban, forest) + pres.wet + s(hill_{02}) \\ + année + s(mois, k = 3, bs = "cc") + te(x, y) + offset\end{aligned} \tag{3.3}$$

3.1.3.3 Climat et interactions

Il peut être intéressant d'observer l'effet de la saison en fonction du climat. En effet, les facteurs climatiques sont invariants dans le temps et caractérisent non pas les voxels mais les pixels. Or, un climat dans l'espace n'apporte pas la même information selon le mois de l'année. Nous pouvons alors modéliser l'interaction entre le facteur climatique et la covariable mensuelle en réalisant une spline cubique cyclique du prédicteur mensuelle

différente par climat, c'est-à-dire que la base de données est divisée en sous-ensembles de données en fonction de leur catégorie climatique. L'AIC relatif au modèle $\mathcal{M}_2$ est alors de -182.272, pour un nombre de fonctions de base par spline égal à 3. En posant $k = 10$ afin de pouvoir mieux capter les variations mensuelles par climat, nous obtenons un AIC relatif de -493.6283. Ce dernier modèle est décrit à l'équation suivante 3.4 :

$$\begin{aligned}\mathcal{M}_3 : g(\mu) = & \, te(urban, forest) + pres.wet + s(hill_{02}) + année \\ & + s(mois, bs = "cc", by = lanmap) + te(x, y) + offset\end{aligned} \tag{3.4}$$

Le facteur climatique est représenté par les catégories de la base LANMAP. Cependant, comme vu à la Figure 1.5a, le climat nord-méditerranéen ne distingue pas le sud-ouest de la côte méditerranéenne. Or, nous avons vu précédemment que le climat méditerranéen semble être très défavorable à *Ixodes ricinus*. La base Climatick et les covariables DRIAS à l'origine de cette base peuvent permettre de distinguer ces deux zones climatiques. Il peut alors être intéressant d'estimer le comptage des déclarations grâce à une nouvelle covariable issue de l'interaction entre LANMAP et Climatick. Les interactions entre le temps et le climat peuvent alors être remplacées par les fonctions suivantes :

(A) $s(mois, climatick)$;

(B) $s(mois, new\_clim)$;

(C) $te(mois, ombr, bs = ("cc", "cr"), k_1 = 12, k_2 = 12)$, où $bs = ("cc", "cr"))$ indique que le prédicteur mensuel est estimé par une spline cubique cyclique de régression et le facteur climatique une spline cubique de régression. Le nombre de fonctions de base est fixé à 12 afin d'obtenir une spline dont les noeuds sont l'ensemble des mois de 1 à 12 ;

(D) $te(mois, TM\_min, bs = ("cc", "cr"), k = 12)$;

(E) $te(mois, TM\_max, bs = ("cc", "cr"), k = 12)$;

(F) $te(mois, TM\_winter, bs = ("cc", "cr"), k = 12)$.

Les AIC relatifs au modèle $\mathcal{M}_3$ sont répertoriés en Table 3.4.

Interaction	(A)	(B)	(C)	(D)	(E)	(F)
AIC	243.84	-26.87	-103.74	6.73	-131.59	-6.66

TABLE 3.4 – AIC relatifs au modèle $\mathcal{M}_4$

Le modèle (E) optimise l'AIC en l'améliorant substantiellement. La structure du nouveau modèle de comparaison $\mathcal{M}_4$ est donnée en (3.5) :

$$\begin{aligned}\mathcal{M}_4 : g(\mu) = & \, te(urban, forest) + pres.wet + s(hill_{02}) + année \\ & + te(mois, TM\_max, bs = ("cc", "cr"), k = 12) + te(x, y) + offset\end{aligned} \tag{3.5}$$

3.1.4 Prédicteurs temporels et interactions spatio-temporelles

Finalement, des alternatives persistent dans la représentation temporelle du modèle. En effet, les deux covariables *mois* et *année* ont permis d'expliquer la variabilité temporelle jusqu'alors. Cependant, il est également possible d'estimer l'effet du temps sur les comptages de déclarations grâce à la covariable *new_date* créée à la section 2.1.3.4 (correspondant au mois compris entre 1 et 63) ou bien en estimant l'interaction climatique et temporelle année par année. Le tenseur $te(mois, TM\_max)$ peut alors être remplacé par l'une des deux options suivantes :

(A) $te(new\_date, TM\_max)$

(B) $te(mois, TM_{max}, by = annee, bs = ("cc", "cr"), k = 12)$

Dans l'option (B), l'effet de l'année est modélisé via une catégorie avec plusieurs niveaux, et il est donc possible d'estimer une autre forme du tenseur pour chaque année. Nous obtenons respectivement les AIC relatifs au modèle $\mathcal{M}_4$ suivants : 247.97 et -815.13. Nous préférons alors la structure (B), décrite dans la nouvelle version du modèle en (3.6) :

$$\begin{aligned}\mathcal{M}_5 : g(\mu) = & \, te(urban, forest) + pres.wet + s(hill_{02}) + année + te(x, y) \\ & + te(mois, TM\_max, bs = ("cc", "cr"), k = 12, by = année), + offset\end{aligned} \tag{3.6}$$

L'interaction ajoutée représente un effet temporel hétérogène dans l'espace, et la variabilité spatiale du prédicteur dans l'espace est caractérisée par la donnée climatique. Ici, les variances longitudinale et transversale sont décrites par les variances de température de TM_max. Une autre manière de représenter cette interaction est d'intégrer un tenseur à trois covariables, prenant en compte la longitude x, la latitude y ainsi qu'une covariable temporelle. Pour tester cette possibilité, le modèle (3.7) suivant est estimé :

$$\begin{aligned} g(\mu) = {} & te(x,\ y,\ mois,\ bs = c(\text{"}cr\text{"},\text{"}cr\text{"},\text{"}cc\text{"}),\ d = c(2,1),\ by = année) + s(TM\_max,\ bs = \text{"}cr\text{"}) \\ & + te(urban, forest) + pres.wet + s(hill_{02}) + année + te(x,y) + offset \end{aligned} \tag{3.7}$$

Son AIC relatif au modèle $\mathcal{M}_5$ est de 28.2. Nous ne le conservons donc pas ici. Cependant, son interprétation, bien que fastidieuse par sa complexité, peut être également intéressante.

3.2 Comparaison de modèles

La Table 3.5 résume les critères de validation et de sélection des six modèles construits à la Section 3.1 :

- ***AIC*** relatif au modèle de base ;
- ***RMSE*** calculé sur la totalité des données (et non uniquement sur le sous-échantillon) ;
- $\boldsymbol{AUC_0}$, l'aire sous la courbe calculant la performance du modèle à prédire les comptages nuls de déclarations ;
- $\boldsymbol{AUC_{quant0.95}}$, l'aire sous la courbe calculant la performance du modèle à prédire les valeurs extrêmes de comptages, supérieures à 2.

Modèle	***AIC*** **relatif au modèle de base**	***RMSE***	$\boldsymbol{AUC_0}$	$\boldsymbol{AUC_{quant0.95}}$
$\mathcal{M}_0$	0.000	1.760	0.873	0.943
$\mathcal{M}_1$	−70.335	1.736	0.873	0.944
$\mathcal{M}_2$	−111.327	1.746	0.874	0.944
$\mathcal{M}_3$	−177.762	1.739	0.874	0.945
$\mathcal{M}_4$	−248.492	1.725	0.875	0.945
$\mathcal{M}_5$	−468.304	1.728	0.878	0.947

TABLE 3.5 – Comparatif des critères de validation et de sélection

Nous pouvons voir que le modèle $\mathcal{M}_5$ a une meilleure qualité que les autres, comme vu précédemment lors de sa construction. De plus, sa qualité prédictive globale (RMSE) est plus forte que celle des autres modèles. Enfin, lorsque nous comparons les aptitudes à prédire les valeurs nulles ou extrêmes, le modèle $\mathcal{M}_5$ est également plus performant.

(Il a également été calculé les critères ci-dessus – excepté l'AIC – dans un premier temps sur le sous-échantillon, puis sur les données hors données d'entraînement, afin de pouvoir mettre en avant un potentiel sur-apprentissage. Les classements sont alors identiques et les différences marginales entre les valeurs des critères sur les données d'ajustement et celles des données test).

Finalement, les résultats ont été confortés grâce à une validation croisée spatiale divisant l'ensemble de données complet en cinq groupes permettant d'ajuster les modèles sur la totalité des pixels et la totalité des points temporels. Ainsi, chaque point de l'espace (x et y) ainsi que chaque point temporel (*mois* et *année*) est utilisé quatre fois en donnée d'ajustement et une fois en donnée test. Les six modèles présentés sont ajustés pour chaque ensemble d'entraînement composé de quatre groupes parmi les cinq créés. Les RMSE (calculés sur le dernier groupe test) et RMSE sont optimisés par l'application du modèle $\mathcal{M}_5$.

3.3 Modèle final

3.3.1 Analyse du modèle

Cette section présente les différents résultats des fonctions d'analyse de modèle proposés par la librairie *mgcv* et présentés en Section 2.4.4, telles que les fonctions *summary*, *gam.check* et *concurvity*.

Coefficients paramétriques				
	Estimate	**Std.Error**	**z-value**	**Pr(>|z|)**
$(Intercept)$	-11.85061	0.0547	-217.146	$< 2e-16***$
$pres.wet1$	0.56821	0.10546	5.388	$7.12e-08***$
$annee2017$	-1.41281	0.10761	-13.129	$< 2e-16***$
$annee2018$	-0.79774	0.09481	-8.414	$< 2e-16***$
$annee2019$	-0.65342	0.08202	-7.966	$1.64e-15***$
$annee2021$	0.06804	0.07204	0.944	0.34494
$annee2022$	-3.62834	1.30872	-2.772	$0.00556 **$
Critères de significativité approximative des termes splines				
	edf	**Red.df**	**Chi.sq**	**p-value**
$te(x, y)$	18.996	20.54	592.83	$< 2e-16***$
$te(urban, forest)$	18.100	19.85	1288.13	$< 2e-16***$
$s(hill_{02})$	5.796	6.51	25.51	$0.00175 **$
$te(mois, TM\_max30) : annee2020$	35.513	45.54	2851.94	$< 2e-16***$
$te(mois, TM\_max30) : annee2017$	24.950	33.28	677.39	$< 2e-16***$
$te(mois, TM\_max30) : annee2018$	30.979	40.89	1395.11	$< 2e-16***$
$te(mois, TM\_max30) : annee2019$	27.998	36.97	1024.24	$< 2e-16***$
$te(mois, TM\_max30) : annee2021$	30.292	39.81	1566.94	$< 2e-16***$
$te(mois, TM\_max30) : annee2022$	8.602	11.32	21.63	$0.03215*$

TABLE 3.6 – Résultats de la fonction *summary* du modèle final

L'ensemble des prédicteurs du modèle final (le modèle $\mathcal{M}_5$) est significatif, comme le montre les valeurs des *p-values* données par la fonction *summary* à la Table 3.6, à l'exception de l'intercept de l'année 2021, qui n'est alors pas significativement différente de l'année de référence, soit 2020. De plus, l'année 2022 ne comportant que très peu d'observations, comprises entre janvier et mars inclus, l'effet mois appliquée à cette année ne peut être interprété correctement. Ensuite, nous retrouvons un paramètre de sur-dispersion ν égal à 1.33 pour la distribution binomiale négative. Enfin, la déviance est expliquée à 63.1%.

	k'	**edf**	**k-index**	**p-value**
$te(x, y)$	24	19.0	0.82	$< 2e-16***$
$te(urban, forest)$	24	18.1	0.83	$< 2e-16***$
$s(hill_{02})$	8	5.8	0.90	0.955
$te(mois, TM\_max30) : annee2020$	131	35.5	0.85	$0.020*$
$te(mois, TM\_max30) : annee2017$	131	24.9	0.85	$0.010 **$
$te(mois, TM\_max30) : annee2018$	131	31.0	0.85	$0.005 **$
$te(mois, TM\_max30) : annee2019$	131	28.0	0.85	$0.015*$
$te(mois, TM\_max30) : annee2021$	131	30.3	0.85	$0.005 **$
$te(mois, TM\_max30) : annee2022$	131	8.6	0.85	$< 2e-16***$

TABLE 3.7 – Analyse résiduelle du modèle final. La colonne *edf* indique le degré de liberté effectif (en tenant compte de la pénalisation de la log-vraisemblance). La colonne k' indique l'*edf* maximal possible. Le *k-index* est un rapport de variance locale et globale de la réponse permettant de détecter de l'auto-corrélation locale non prise en compte par le modèle, par exemple en raison d'un nombre de noeuds k trop faible dans une fonction spline.

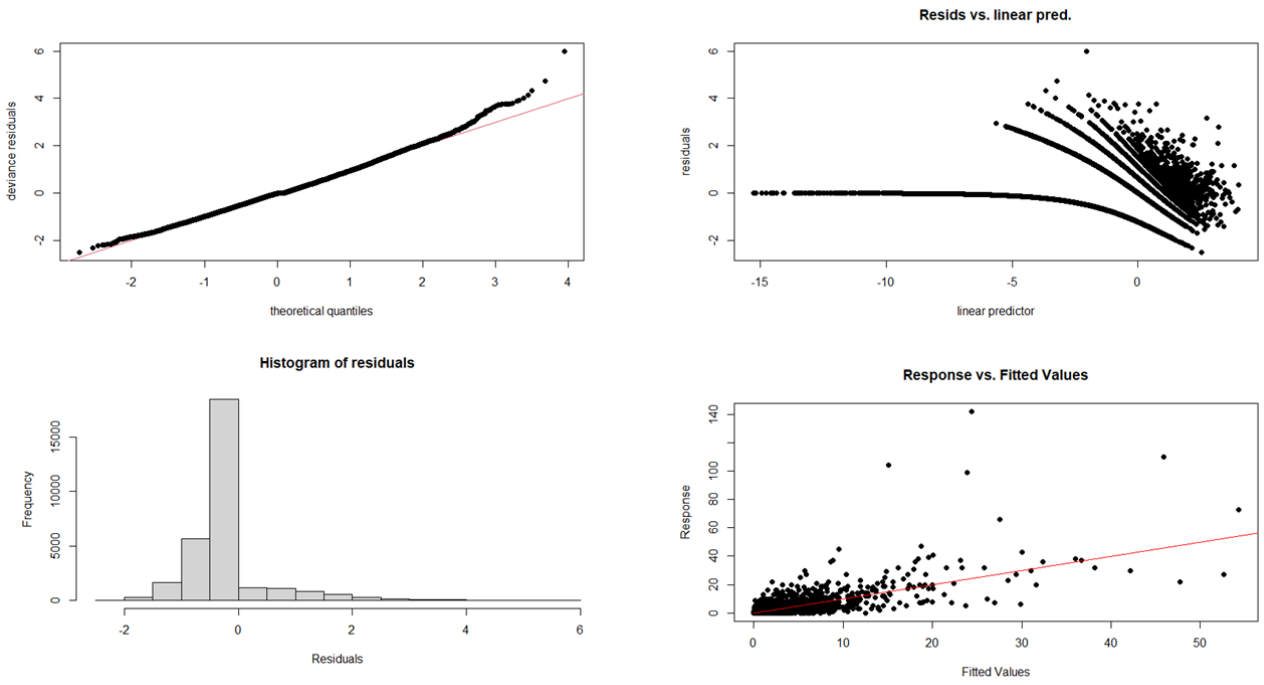


FIGURE 3.4 – Analyse graphique des résidus du modèle final

La Table 3.7 présente les résultats de la fonction *gam.check* de notre modèle. Ici, l'hypothèse H0 selon laquelle la fonction spline estimée explique parfaitement la variabilité locale de la réponse due à la covariable en question, est rejetée pour l'ensemble des covariables, à l'exception du prédicteur de diversité. Cependant, nous éviterons d'augmenter leurs nombres de fonctions de base pour éviter des instabilités numériques des modèles mais aussi tout sur-apprentissage et toute difficulté d'interprétation.
La Figure 3.4 présente l'analyse graphique des résidus. Nous pouvons alors voir que le modèle a plus de difficulté à prédire les valeurs extrêmes que les valeurs centrales (un comportement assez typique pour la plupart des modèles prédictifs dans diver domaines). De plus, le modèle semble la plupart du temps sous-estimer le nombre de déclarations, bien que dans ce cas, la sous-estimation soit assez faible (comprise entre 0 et 2) ; encore une fois, il s'agit d'un comportement typique des modèles de prédiction qui ont tendance à fournir des prédictions plus lisses que les observations, surtout en absence de prédicteurs permettant de faire des prédictions "parfaites". En revanche, lorsque le modèle surestime le nombre de déclarations, celui-ci a tendance à obtenir des erreurs plus élevée, de 0 à 6).

	worst	*observed*	*estimate*
$te(x, y)$	0.88	0.67	0.49
$te(urban, forest)$	0.82	0.71	0.24
$s(hill_{02})$	0.66	0.43	0.27
$te(mois, TM\_max30) : annee2020$	0.53	0.05	0.05
$te(mois, TM\_max30) : annee2017$	0.54	0.04	0.06
$te(mois, TM\_max30) : annee2018$	0.65	0.06	0.06
$te(mois, TM\_max30) : annee2019$	0.56	0.06	0.07
$te(mois, TM\_max30) : annee2021$	0.59	0.07	0.08
$te(mois, TM\_max30) : annee2022$	5.66	0.88	0.03

TABLE 3.8 – Concordance du modèle final

La Table 3.8 présente les résultats de la fonction *concurvity* explorant des comportements de "colinéarité" (ou de "concordance") entre les courbes splines ajustées. Ici, la plupart des covariables semble avoir une concordance relativement faible. Pour les covariables *urban*, *forest*, la valeur est assez proche de 0.8. Les relations entre ces

deux valeurs et les autres prédicteurs lisses sont alors analysées. Cette analyse ne révèle pas de liens déterministes empêchant l'identifiabilité des différentes composantes, cependant, il nous semble important de prêter attention à l'interprétation finale de l'effet de cette interaction. En ce qui concerne la concordance du tenseur climatico-mensuel de l'année 2022, le résultat semble biaisé par le faible nombre d'observations durant cette année. Finalement, la concordance du prédicteur spatial semble assez élevé, cependant, aucune relation déterministe directe avec les autres composantes du prédicteur linéaire semble claire après analyse des relations entre les différents prédicteurs.

3.3.2 Effets partiels

Dans cette section, nous analysons les effets de chacun des termes lisses du modèle de façon partielle.

3.3.2.1 Effet partiel des zones forestières et urbaines

La Figure 3.5 illustre l'effet partiel du prédicteur linéaire d'interaction entre les zones forestières et les zones urbaines, modélisé par un tenseur dont les deux dimensions sont des splines cubiques de régression. Nous pouvons alors voir que les zones à forte proportion de zone forestière ont un effet positif sur les comptages de déclarations. Plus la proportion de zone forestière augmente, plus le nombre de déclarations augmente. À l'inverse, une proportion de forêts inférieure à $\sim 40\%$ diminue le nombre de déclarations de piqûres. En ce qui concerne les zones urbaines, les déclarations diminuent lorsque la proportion de zones urbaines augmente et lorsque la proportion de zones forestières est pratiquement nulle. Cependant, l'effet devient positif et augmente lorsque les proportions de zones forestières et de zones urbaines tendent toutes les deux vers 50%.

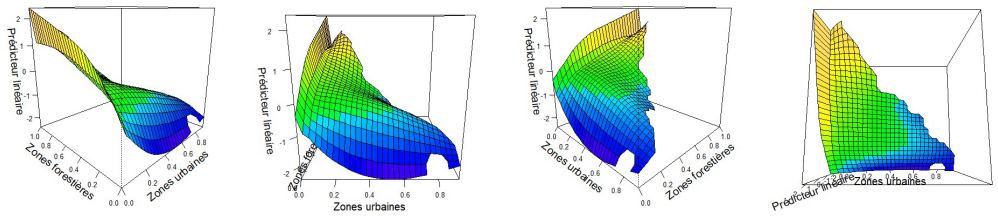


FIGURE 3.5 – Effet partiel des zones forestières et urbaines

3.3.2.2 Effet partiel de la richesse de niveau 2

La Figure 3.6 illustre l'effet partiel du prédicteur linéaire de la richesse de niveau 2, modélisée par une spline cubique de régression. Nous pouvons voir qu'un nombre trop faible de couvertures de sol différentes ($\sim< 4$) ou trop élevé (14 ou 15) est défavorable au nombre de déclarations de piqûres. Entre 5 et 11 catégories différentes, l'effet semble approximativement nul, tandis que l'effet est relativement plus élevé pour 12 et 13 catégories. L'interprétation épidémiologique de ce facteur reste difficile et nécessitera d'autres analyses supplémentaires au delà des travaux de ce stage.

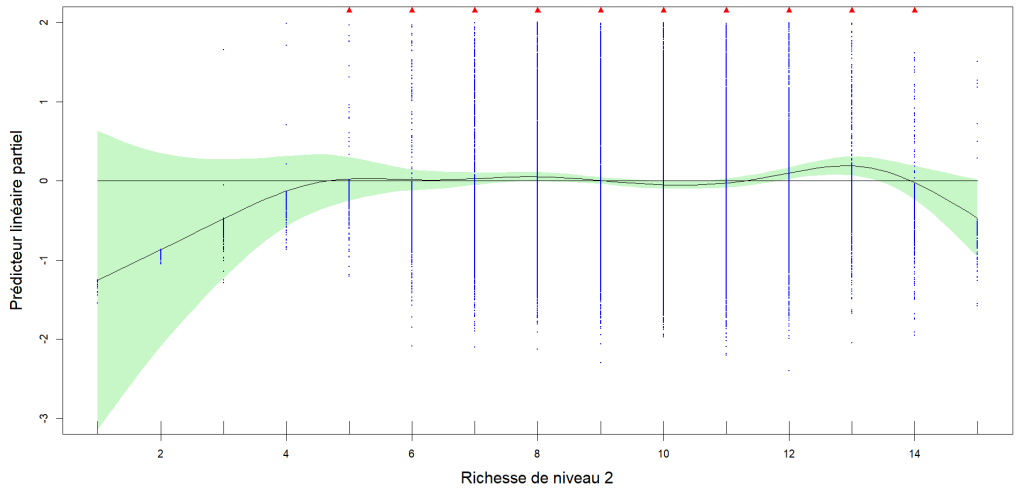


FIGURE 3.6 – Effet partiel de la richesse de niveau 2 (En bleu : résidus partiels ; En rouge : résidus partiels > 2)

3.3.2.3 Effet partiel climatique et temporel

La Figure 3.7 représente le prédicteur linéaire tensoriel du mois et de la température moyenne du mois le plus chaud en fonction de l'année. Trois angles de vue sont proposés par année. Seule l'année 2022 n'est pas représentée, celle-ci n'ayant pas assez d'observations pour être correctement modélisée. L'ensemble des prédicteurs varie dans le négatif, à l'exception de l'interaction en 2020 et 2021 (que nous rappelons ne pas être significativement différentes).

Nous pouvons voir que les variations mensuelles varient selon les années. De plus, sur une même année, les pics de déclarations de piqûres varient selon la température moyenne du mois le plus chaud. Ainsi, de 2018 à 2021, lorsque la température est inférieure à 15°C, un pic de déclarations est observé en juillet-août, tandis que pour les températures supérieures, ce pic est observé en mai-juin. Cependant, en 2018, un second pic est observé pour les températures les plus basses aux alentours d'avril-mai, ainsi qu'un second pic en septembre pour 2019 et 2021. En outre, 2020 semble touchée par des perturbations engendrant différents pics, dont le premier est observé en mars pour les températures les plus basses. Finalement, l'effet négatif hivernal paraît amoindri en 2021 pour les températures hautes. En ce qui concerne l'année 2017, les résultats précédant juillet doivent être interprétés avec prudence, sachant que la grosse majorité des déclarations ne sont observées qu'à partir du 13 juillet 2017.

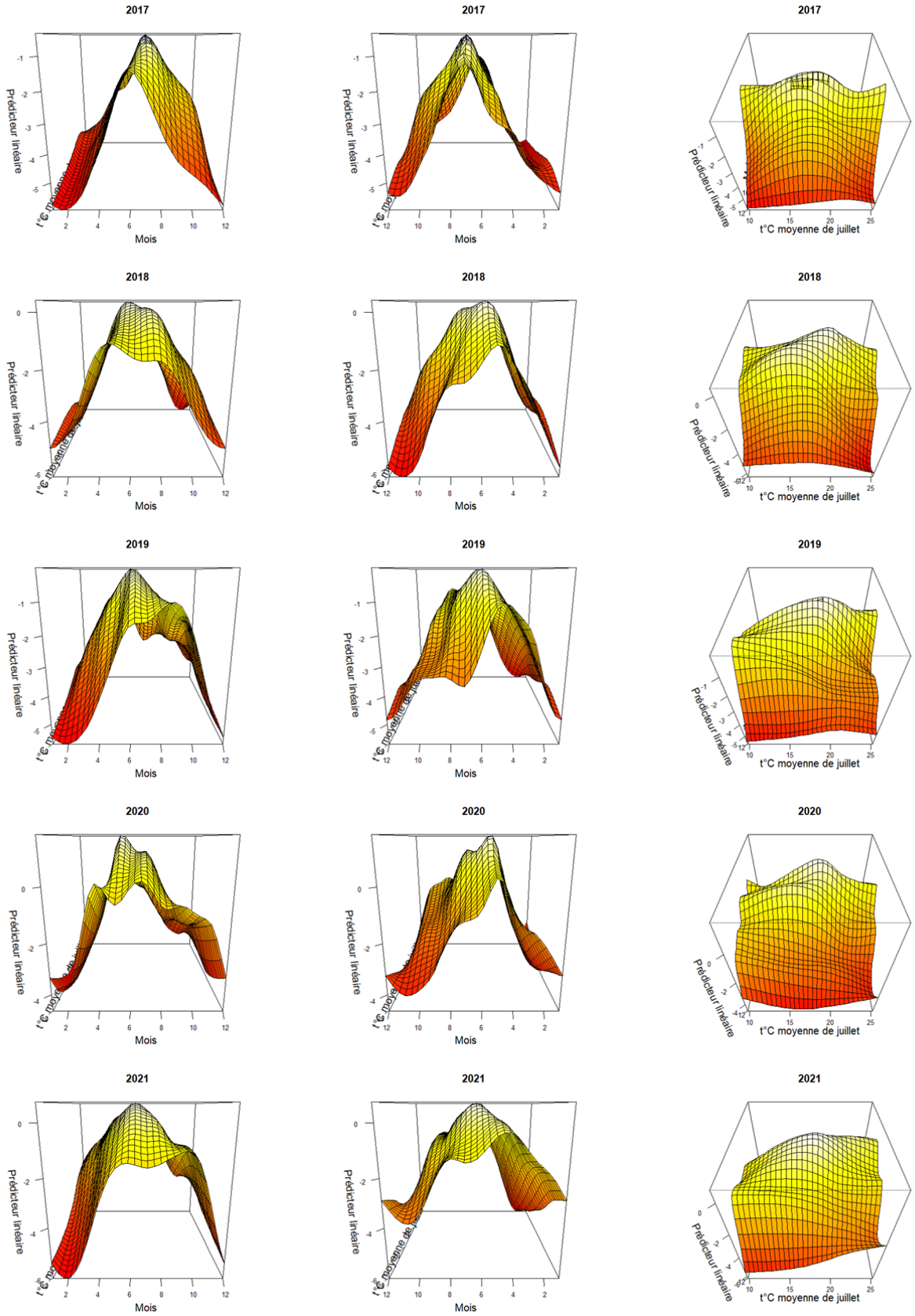


FIGURE 3.7 – Effet partiel climatique et temporel en fonction des années. La colonne de gauche représente l'angle de vue le plus proche des évolutions mensuelles des climats les plus froids (c'est-à-dire lorsque la température moyenne du mois le plus chaud est faible, soit proche de 10). À l'inverse, la colonne centrale présente l'évolution mensuelle du nombre de déclarations pour des climats plus chauds. Attention à l'échelle des mois, qui est inversée pour cette colonne. Les mois vont de gauche à droite de décembre à janvier.

3.3.2.4 Effet partiel du tenseur spatial

Finalement, la Figure 3.8 illustre le tenseur spatial, permettant de capter le reste des variabilités spatiales non expliquées par le reste des prédicteurs. Ainsi, il semble que l'ensemble des prédicteurs ne puissent pas assez discriminer la côte méditerranéenne du reste du territoire. De plus, nous pouvons voir que le Sud-Ouest et le Nord-Est sont plus favorables aux déclarations de piqûres.

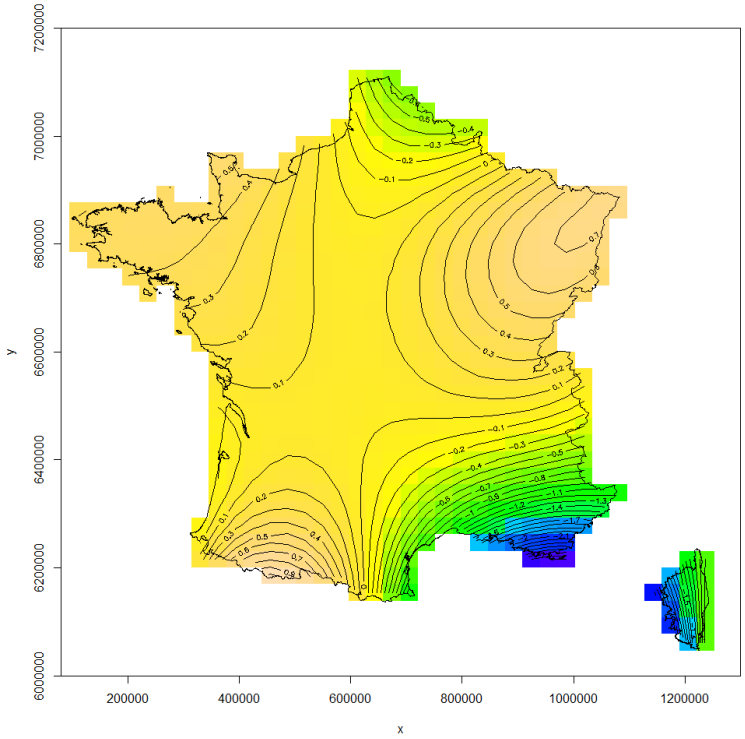


FIGURE 3.8 – Effet partiel du tenseur spatial

3.3.3 Prédictions

La Figure 3.9 montre les erreurs de prédiction dans l'espace géographique par mois pour l'année 2020 (cette année est choisie en raison de ses observations élevées). Nous pouvons alors voir que les erreurs sont assez faibles, voire inexistantes, entre septembre et mars inclus. Cela peut s'expliquer par le nombre assez faible de déclarations durant ces mois, allant de 0 (pour 94% des observations) jusqu'à 12, et le modèle prédit bien ces valeurs relativement faibles. En revanche, les valeurs extrêmes printanières paraissent plus difficile à prédire pour le modèle.

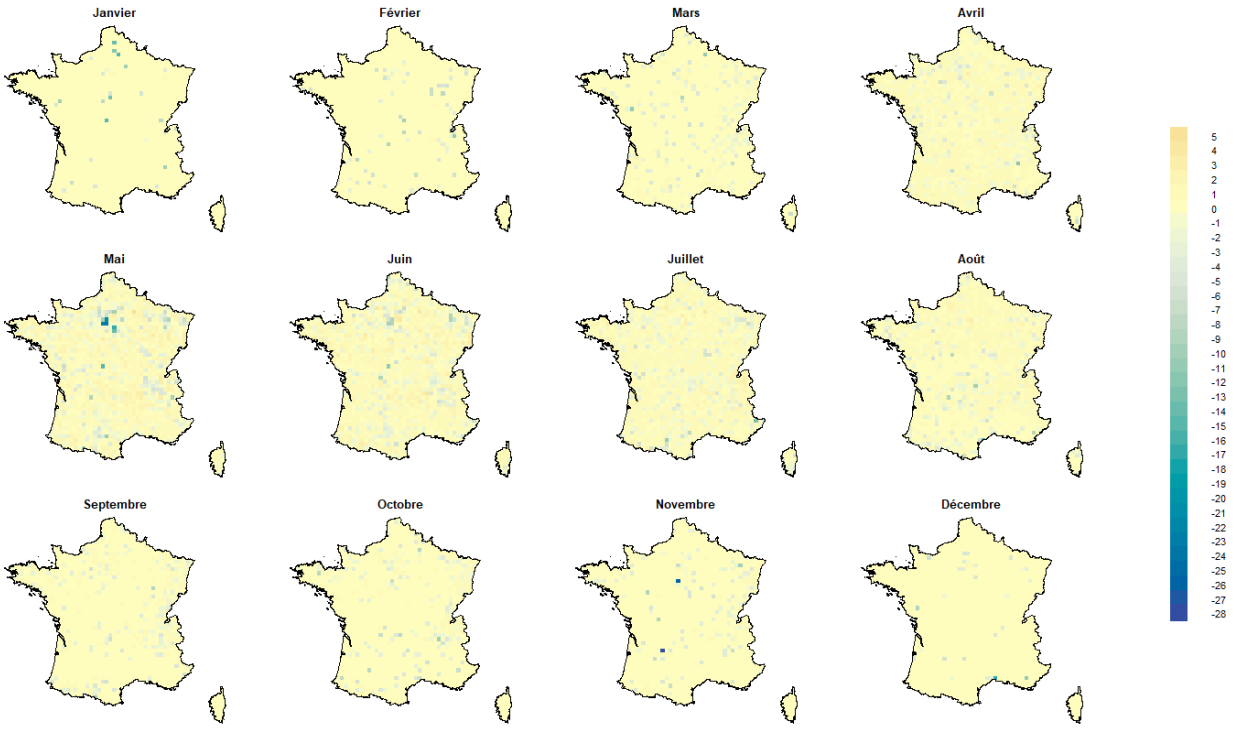


FIGURE 3.9 – Résidus r_i de Pearson par mois en 2020. $r_i = \frac{pred_i - obs_i}{\sqrt{pred_i}}$

À la Figure 3.10, les différences entre observations et prédictions sont assez visibles en région parisienne, région pour laquelle de nombreuses valeurs extrêmes sont observées. Plusieurs pixels contenant plus de 110 déclarations ne sont prédits qu'entre 20 et 60 déclarations. Bien que l'erreur soit forte, ce comparatif observation-prédiction nous montre que le modèle arrive à prédire les différences d'ordre de grandeur.

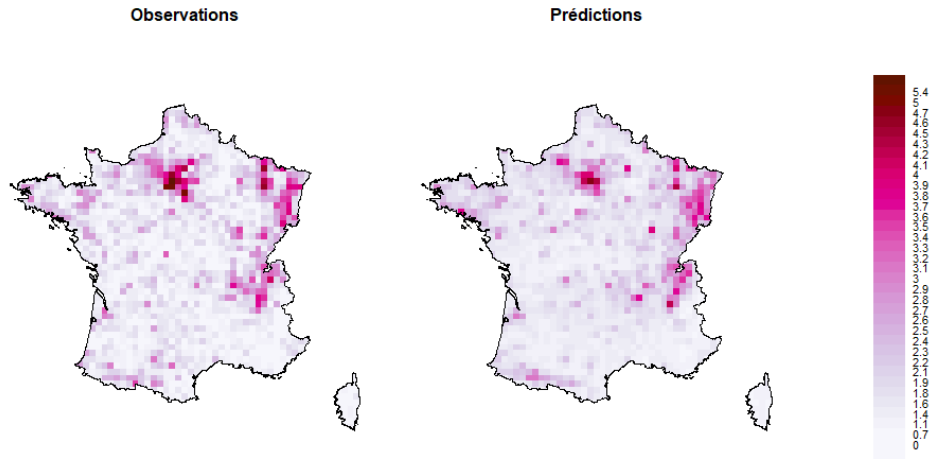


FIGURE 3.10 – Observations VS Prédictions du nombre de déclarations de piqûres par pixel en mai 2020, échelle $log(x+1)$

Discussion

Ce travail de recherche s'est concentré sur la modélisation du nombre de déclarations de piqûres de tiques par mois, par année et par pixel de résolution 20×20 km en France métropolitaine, afin de pouvoir mettre en avant différents facteurs de risque de piqûre. Les résultats mettent en évidence l'effet de différentes covariables en lien avec la dynamique spatio-temporelle des tiques, la dynamique d'exposition de l'Homme, ou l'interaction entre ces deux dynamiques.

Dans un premier temps, différentes covariables ont été agrégées afin de permettre de décrire la dynamique des tiques, comme les usages des sols ou le climat. En effet, ces facteurs permettent de définir la niche écologique des tiques ainsi que leur phénologie. Nos résultats montrent que les proportions de zones urbaines et les proportions de zones forestières ont des effets plus significatifs sur le nombre de déclarations que les proportions de plans d'eau et zones humides. De plus, l'interaction entre ces deux types de couverture de sol apporte d'autant plus d'information. Ainsi, les zones les plus urbaines sans aucune proportion de zones forestières tendent à dénombrer moins de déclarations. En effet, ce type de sol est notamment composé de terrains artificiels tels que des zones industrielles ou commerciales, routes et autres. Or, ce type de sol ne correspond pas à la niche écologique des tiques. Ensuite, lorsque la proportion de zones urbaines est quasi-nulle, les pixels ayant une forte part de zones forestières, comprenant forêts, milieux à végétation arbustive et herbacée et espaces ouverts sans ou avec peu de végétations et étant partiellement humide ($\approx 10\%$), semblent avoir un effet très positif sur le nombre de déclarations. En effet, McCoy et al. (2015b) rappelle que la population de tiques est plus élevée dans des biotopes caractérisés par leur forte densité végétale, notamment grâce à une plus forte densité de population d'hôtes tels que les mammifères. Enfin, un pixel tendant à n'être composé que de zones urbaines et forestières, et à parts égales, semble favorable aux déclarations de piqûres, mais reste moins favorable que pour des pixels très majoritairement composés de zones boisées. Rappelons également la forte corrélation négative entre les zones forestières et les zones agricoles. Nous pouvons penser que les pixels à faible proportion de zones forestières sont principalement des zones agricoles. L'effet reste alors positif bien que moins favorable que les zones forestières. En revanche, les zones aux proportions quasi-nulles de zones forestières, potentiellement entièrement agricoles, ont un effet négatif sur le nombre de déclarations. Cela peut alors s'expliquer par une présence de l'Homme plus faible sur ce type de sols moins propice aux activités récréatives.

Nous avons également mis en évidence par le facteur d'humidité *pres.wet* un effet positif de la présence de zones humides. McCoy et al. (2015b) explique également que la tique est fortement présente dans des biotopes humides. En outre, les résultats montrent que la richesse de niveau 2 des couvertures de sol a un effet négatif sur les déclarations de piqûres lorsque celle-ci est trop élevée ou trop faible. Ainsi, un nombre de catégories de niveau 2 inférieur à 5, correspondent à des milieux plutôt homogènes, voire trop pour les tiques. En ce qui concerne des richesses égales à 12 ou 13, les pixels sont alors plus hétérogènes, pouvant correspondre à des zones d'interface entre forêt et autres types de couvertures de sol. Ces pixels sont alors plus favorables aux déclarations de piqûres. Enfin, une richesse trop élevée (égale à 15) semble défavorable aux piqures de tiques. Il faut cependant prêter attention au faible nombre de pixels respectant cette condition et pouvant induire cet effet négatif, et une étude approfondie de cet effet serait de mise.

Par la suite, les résultats montrent que les variations climatiques, en particulier la température moyenne du mois le plus chaud calculée sur les trente années précédant 2020, influe fortement sur les cycles mensuels de déclarations de piqûres de tiques. Ce résultat, également établi par Wongnak et al. (2022) pour les nymphes de l'espèce *Ixodes Ricinus*, démontre des pics de déclarations en printemps-été, variant selon la température caractéristique du climat. Plus le climat est chaud, plus tôt dans l'année apparaîtront les tiques. Plus les températures estivales sont basses, plus les tiques apparaissent tardivement dans l'année.

Dans un second temps, différents facteurs permettent d'expliquer la dynamique d'exposition de l'Homme. Rappelons que la variable de population mise en *offset* est très fortement liée aux types de sols urbains. Ainsi, les zones urbaines, comprenant les tissus urbains, zones industrielles et commerciales, ou espaces verts artificialisés permettent de représenter approximativement les lieux de présence de l'Homme. Les très fortes proportions en zones urbaines semblent moins favorables aux déclarations de piqûres. Cependant, cet effet est dû à la

faible présence de tiques sur ce type de sol. Il faudrait alors pouvoir intégrer un facteur plus représentatif des déplacements et présences humaines en dehors de leur lieu d'habitation.

Enfin, le cycle saisonnier influe sur la dynamique humaine. Ainsi, les pics de déclarations entre mai et août inclus peuvent s'expliquer par les sorties plus fréquentes de la population, et notamment dans des milieux à végétation plus denses, tels que les forêts, dans le cadre par exemple de randonnées. Les variations annuelles peuvent également expliquer la dynamique d'exposition. En effet, nous avons observé certaines particularités en 2020 et 2021, deux années ponctuées par les confinements. En 2020, les résultats montrent deux pics pour les climats plus frais, en mars et juin. Le pic de mars correspond au premier confinement en France durant la pandémie de Covid-19. Au contraire des autres années, l'effet du prédicteur est positif durant cette période de confinement. Il semblerait alors que la population ait profité des jardins privés pour pouvoir s'aérer, impliquant une exposition aux tiques plus élevée. De plus, les résultats montrent que l'effet climatico-mensuel est moins faible aux alentours de novembre (~ -2 contre ~ -4 en 2019), période correspondant au 2^{nd} confinement.

Néanmoins, ces chiffres peuvent également s'expliquer par une température plus élevée durant l'année 2020, notamment courant novembre et début décembre, ayant alors permis d'adoucir les climats plus frais et donc d'influencer le cycle d'apparition des tiques. En effet, Météo-France (2021) fait état d'un écart à la moyenne saisonnière de référence entre 0.5 et 1°C pour les climats les plus frais, ainsi que des écarts supérieurs pour les climats plus chauds. À l'instar de 2020, 2021 a également connu des périodes de douceur en mars, juillet et septembre d'après le bilan Météo-France (2022), pouvant également expliquer les variations du nombre de déclarations.

Finalement, il serait intéressant de pouvoir intégrer des facteurs météorologiques telles que l'évolution de la température moyenne mensuelle ou l'humidité relative mensuelle, pouvant décrire plus finement les cycles d'apparition des tiques mais également les périodes favorables aux sorties extérieures pour la population humaine, et la variation inter-annuelle de ces propriétés. En outre, un facteur binaire représentant les périodes de confinement pourrait aider à mieux expliquer les variations inter-annuelles caractéristiques de la dynamique d'exposition. Pour finir, il a été discuté de l'intégration d'une covariable représentative la présence de certains types d'hôtes, notamment certains types de mammifères, pouvant alors détailler la dynamique des tiques.

Conclusion et perspectives

Ce travail de recherche a permis en grande partie de pouvoir confirmer et de mieux quantifier de nombreux résultats démontrés par d'autres recherches, grâce à un nouveau type de collecte de données et une modélisation spatio-temporelles du nombre de déclarations de piqûres de tiques. Cependant, il est important de prendre ces résultats avec précaution de par la méthode de collecte de donnés non protocolée. De plus, de nombreux facteurs évoqués en discussion restent à explorer. Ainsi, ce travail de recherche devrait être poursuivie par un approfondissement des niveaux inférieurs de CLC (niveaux 2, 3 et 4) afin de déterminer à un niveau plus précis les couvertures de sol les plus favorables ou défavorables aux piqûres de tiques. Ensuite, il pourrait être intéressant d'intégrer un nouveau facteur représentatif des déplacements de l'Homme sur le territoire en fonction des moments dans l'année. Finalement, la résolution de la grille $\mathcal{G}$ pourrait être affinée afin d'avoir une meilleur représentation des piqûres de tiques en fonction des écotones. Enfin, rappelons que la dynamique décrivant le fait de signaler sur l'application lorsqu'il y a eu piqûre n'est pas représentée dans cette modélisation.

Nous concluons que les travaux futurs devront se focaliser sur l'objectif d'améliorer l'identification des trois dynamiques spatio-temporelles suivantes et de démêler les interactions entre ces dynamiques : les processus biologiques des tiques ; les variations de l'exposition de l'Homme ; l'effort d'échantillonnage via la collecte des déclarations de piqûres. Parmi les pistes envisageables, citons l'intégration d'autres sources de données (par exemple, les connections d'appareils mobiles, ou les données du réseau sentinelle des médecins), la collecte d'autres types de données (par exemple, pour signaler l'absence de piqûre de tiques, et non seulement leur présence), et l'utilisation de méthodes d'apprentissage statistique encore plus sophistiquées (modèles à espace d'état, approche mécanistico-statistique, modèles multivariés pour les variables de réponse).

Bibliographie

CGDD/SOeS (2009). CORINE Land Cover France Guide d'utilisation.

Chalvet-Monfray, K. (2022). Le risque lié aux tiques, habitats et adaptations des tiques au climat et aux activités humaines.

Chesneau, C. (2020). Introduction aux arbres de décision (de type CART). Lecture.

Côté, S. (2016). Modèles additifs généralisés dans la modélisation de l'impact du kilométrage et de l'exposition au risque en assurance automobile.

Estrada-Peña, A. (2008). Climate, niche, ticks, and models : what they are and how we should interpret them. Parasitology Research, 103.

Farouki, R. T. (2008). Geometry and Computing, Pythagorean - Hodograph Curves : Algebra and Geometry Inseparable, chapter 15 Spline basis functions, pages 345–368. Springer edition.

Gray, J. S., Dautel, H., Estrada-Peña, Kahl, O., and Lindgren, E. (2009). Effects of climate change on ticks and tick-borne diseases in europe. Interdisciplnarity Perspective on Infectious Diseases.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models : some applications. Journal of the American Statistical Association, 82(398) :371–386.

Janssen, V. (2009). Understanding coordinate reference systems, datums and transformations. International Journal of Geoinformatics, 5.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics, 34 :1–14.

Lord, D., Park, B.-J., and Hart, J. D. (2010). Bias properties of bayesian statistics in finite mixture of negative binomial regression models in crash data analysis – negative binomial regression models and estimation methods.

Lémond, J. (2010). Le projet drias : premières études et documents.

Marcon, E. (2015). Mesures de la Biodiversité.

McCoy, K. D., Boulanger, and (Eds.), N. (2015a). Tiques et maladies à tiques : Biologie, écologie évolutive, épidémiologie, pages 53–84. IRD Éditions.

McCoy, K. D., Boulanger, and (Eds.), N. (2015b). Tiques et maladies à tiques : Biologie, écologie évolutive, épidémiologie - Facteurs environnementaux influant sur la dynamique des populations de tiques, pages 89–98. IRD Éditions.

Metzger, M., Bunce, R., Jongman, R., Mücher, C., and Watkins, J. (2005). A climatic stratification of the environment of europe. Global Ecologyand Biogeography, 14 :549–563.

Météo-France (2021). Bilan climatique de l'automne 2020. Technical report, Ministère de la Transition écologique et de la Cohésion des territoires.

Météo-France (2022). Bilan climatique de l'automne 2021. Technical report, Ministère de la Transition écologique et de la Cohésion des territoires.

Mücher, C., Klijnbdirk, J. A., Wascher, D. M., and Schaminée, J. (2010). A new European Landscape Classification (LANMAP) : A transparent, flexible and user-oriented methodology to distinguish landscapes. Ecological Indicators, 10 :87–103.

Noirfalise, A. (1987). Map of Natural Vegetations of the member countries of the European Community and of the Council of Europe.

Rizzoli, A., Silaghi, C., and Obiegala, A. (2014). Ixodes ricinus and its transmitted pathogens in urban and peri-urban areas in Europe : new hazards and relevance for public health. Frontiers in Public Health.

Roekaerts, M. (2002). The biogeographical regions map of europe - basic princliples of its creation and overview of its development.

Scheiner, S. M. (2012). A metric of biodiversity that integrates abundance, phylogeny, and function. OIKOS, Advancing Ecology, 121 - Issue 8 :1175–1334.

SPF (2021). Surveillance nationale : points clés 2020 en france métropolitaine.

Wongnak, P. (2022). Phd - modélisation de la dynamique de population de tiques par approche hiérarchique bayésienne prenant en compte des facteurs abiotiques (température et humidité relative) dans le cadre de l'adaptation au changement climatique [soutenance 2023].

Wongnak, P., Bord, S., and Jacquot, M. (2022). Meteorological and climatic variables predict the phenology of Ixodes ricinus nymph activity in France, accounting for habitat heterogeneity. Scientific reports, 12.

Wood, S. N. (2006). Generalize additive models : an introduction with R.

Annexes

Description des données brutes de l'application *Signalement Tiques*

Identifiant → *id*

Chaque signalement est identifié par son numéro *id*. Il s'agit d'une clé numérique unique permettant de distinguer chaque signalement.

Type de formulaire → *origine*

Il est précisé pour chaque signalement la source de l'enregistrement parmi les différentes versions Apple, Androïd et Web, et les formulaires papier.

Dates → *datetime, bite_ date*

Un signalement contient également une variable *datetime* sous format date et heure **1900-01-01 00 :00 :00** correspondant au moment où le signalement a été envoyé sur l'application, ains qu'une variable *bite_ date* sour format date **1900-01-01**, correspondant à la date de piqûre du signalement concerné.

Coordonnées géographiques → *x, y*

Chaque signalement est géolocalisé par une variable longitudinale y et transversale x.

Précisions spatio-temporelles→ *precis_ date, precis_ loc, precis_ coord_ gps*

Toute personne peut donner son degré de précision spatio-temporelle quant à la piqûres :
- *precis_ date* ∈ {date exacte, 2-3j pres, date inconnue} ;
- *precis_ loc* ∈ {loc connue, loc inconnue} ;
- *precis_ coord_ gps* ∈ {moins de 1 km, entre 1 et 5 km, plus de 5 km} ;

Description de l'hôte → *for_ human, who, age, type_ animal, sex_ animal*

Il est demandé plusieurs informations quant à l'hôte concerné par la piqûre :
- la nature de l'hôte *for_ human* : {1 : s'il s'agit d'un Homme ; 0 : s'il s'agit d'un animal} ;
- le sexe *who* ∈ {Femme, Homme, Ne se prononce pas} s'il s'agit d'un Homme ;
- le sexe *sex_ animal* ∈ {male, femelle, indetermine} s'il s'agit d'un animal ;
- l'âge *age* ∈ {0-5, 6-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, +70}.

Identifiant d'utilisateur → *nb_ ticks, environment, reason*

Finalement, tout utilisateur peut indiquer le nombre de piqûres observées à la date *bite_ date* déclarée et préciser le type de lieu *environment* où l'hôte a été piqué parmi Domicile/maison (intérieur), Forêt, Jardin privé, Jardin privé ou parc municipal, Par public/municipal, Prairie, Zone agricole cultivée ou Autre, ainsi que la raison d*reason* de la présence sr le lieu de piqûre parmi Acitivté professionnelle, Activité scolaire, Centre de loisirs/cap/colo, Chasse/pêche, Lieu de résidence, Loisir, Scoutisme, Sport ou Autre.

Identifiant d'utilisateur → *user_id*

Il est attribué à chaque utilisateur un identifiant d'utilisateur, attribué à l'identité numérique de toute personne qui signale une piqûre de tique. Ainsi, plusieurs signalements peuvent avoir le même identifiant d'utilisateur. Lorsqu'un utilisateur signale plus de 1000 fois, il s'agit soit d'une personne du projet CiTIQUE (enregistrant les données d'une autre personne, par exemple les données d'un formulaire papier), soit d'une personne n'ayant pas souhaité créer de compte.

Nomenclature CLC

Niveau	Nom de variable	Description
1	*urban*	Territoires artificialisés
	agric	Territoires agricoles
	forest	Forêts et milieux semi-naturels
	wet	Zones humides
	water	Surfaces en eau
2	clc_{11}	Zones urbanisées
	clc_{12}	Zones industrielles ou commerciales et réseaux de communication
	clc_{13}	Mines, décharges et chantiers
	clc_{14}	Espaces verts artificialisés, non agricoles
	clc_{21}	Terres arables
	clc_{22}	Cultures permanentes
	clc_{23}	Prairies
	clc_{24}	Zones agricoles hétérogènes
	clc_{31}	Forêts
	clc_{32}	Milieux à végétation arbustive et/ou herbacée
	clc_{33}	Espaces ouverts, sans ou avec peu de végétation
	clc_{41}	Zones humides intérieures
	clc_{42}	Zones humides côtières
	clc_{51}	Eaux continentales
	clc_{52}	Eaux maritimes
3	clc_{11}	Tissu urbain continu
	clc_{112}	Tissu urbain discontinu
	clc_{121}	Zones industrielles ou commerciales et installations publiques
	clc_{122}	Réseaux routier et ferroviaire et espaces associés
	clc_{123}	Zones portuaires
	clc_{124}	Aéroports
	clc_{131}	Extraction de matériaux
	clc_{132}	Décharges
	clc_{133}	Chantiers

clc_{141}	Espaces verts urbains
clc_{142}	Equipements sportifs et de loisirs
clc_{211}	Terres arables hors périmètres d'irrigation
clc_{212}	Périmètres irrigués en permanence
clc_{213}	Rizières
clc_{221}	Vignobles
clc_{222}	Vergers et petits fruits
clc_{223}	Oliveraies
clc_{231}	Prairies et autres surfaces toujours en herbe à usage agricole
clc_{241}	Cultures annuelles associées à des cultures permanentes
clc_{242}	Systèmes culturaux et parcellaires complexes
clc_{243}	Surfaces essentiellement agricoles, interrompues par des espaces naturels importants
clc_{244}	Territoires agroforestiers
clc_{311}	Forêts de feuillus
clc_{312}	Forêts de conifères
clc_{313}	Forêts mélangées
clc_{321}	Pelouses et pâturages naturels
clc_{322}	Landes et broussailles
clc_{323}	Végétation sclérophylle
clc_{324}	Forêt et végétation arbustive en mutation
clc_{331}	Plages, dunes et sable
clc_{32}	Roches nues
clc_{333}	Végétation clairsemée
clc_{334}	Zones incendiées
clc_{335}	Glaciers et neiges éternelles
clc_{411}	Marais intérieurs
clc_{412}	Tourbières
clc_{421}	Marais maritimes
clc_{422}	Marais salants
clc_{423}	Zones intertidales
clc_{511}	Cours et voies d'eau
clc_{512}	Plans d'eau
clc_{521}	Lagunes littorales
clc_{522}	Estuaires
clc_{523}	Mers et océans

4	$incompl\_art$	Partiellement artificiel
	$compl\_art$	Totalement artificiel
	$agric$	Territoires agricoles
	$forest$	Forêts
	$open_vg$	Aire ouverte de végétation
	$open_non_vg$	Aire ouverte de non-végétation
	$h2o$	Eau

TABLE 4.1 – Description de chacune des covariables CLC, pour les niveaux 1-2-3