Cartographie du risque épidémiologique : le défi des données fortement déséquilibrées

Rodrigue Govan*, Romane Scherrer*, Cyrille Goarant**,***
Arnaud Cannet**** Philippe Fournier-Viger♯, Nazha Selmaoui-Folcher*

*ISEA, Université de la Nouvelle-Calédonie, BP R4, 98851 Nouméa Cedex
rodrigue.govan@gmail.com, nazha.selmaoui@unc.nc
**UREL, Institut Pasteur de Nouvelle-Calédonie
***Communauté du Pacifique Sud, Nouvelle-Calédonie
****Direction des Affaires Sanitaires et Sociales, Nouvelle-Calédonie
♯College of Computer Science and Software Engineering, Shenzhen University, Chine

Résumé. L'émergence de la collecte de données massives a contribué au développement de méthodes d'extraction de connaissances, mais elle a également apporté son lot de défis. Parmi ces défis, on retrouve le problème des données fortement déséquilibrées, notamment dans les classes à prédire d'une tâche de classification catégorielle. Cet article présente une stratégie complète développée afin de pallier la problématique de données déséquilibrées dans le cadre d'une étude épidémiologique spatio-temporelle de la leptospirose. L'approche est évaluée sur des données réelles pour une tâche de classification binaire de "présence de risque" de contamination de la bactérie associée à la leptospirose, où la classe majoritaire représente environ 95% des classes à prédire. Par le sous-échantillonnage, l'entraînement de 200 modèles d'apprentissage supervisé et une prédiction pondérée, notre stratégie a atteint une justesse équilibrée de 76,19%, un AUC de 83,29% et un rappel de 83,93%.

1 Introduction

L'émergence de la collecte massive de données a suscité un intérêt exponentiel pour le développement de méthodes d'analyse et de la science des données visant à en extraire des connaissances. Appliquée dans divers domaines telles que l'économie, la finance, la biologie et l'épidémiologie, elle a cependant amené son lot de défis, notamment celui des données déséquilibrées (Fernández et al., 2017). La notion de données déséquilibrées se retrouve lorsque l'on cherche à résoudre un problème de classification dans lequel une ou plusieurs classes à prédire dans un jeu de données ne représentent qu'une très faible proportion par rapport à l'ensemble du jeu de données. Dans ce type de situation, un modèle de classification aura tendance à favoriser la classe majoritaire durant l'apprentissage affectant ainsi les performances de prédiction (Longadge et Dongre, 2013). Ce biais introduit par le déséquilibre des données peut conduire à des résultats erronés ou inefficaces, particulièrement lorsque les classes minoritaires revêtent une importance critique, comme dans le cas de la détection de maladies rares ou de fraudes financières.

Pour répondre à cette problématique, de nombreuses méthodes ont été proposées. Ces solutions se concentrent principalement sur la phase de préparation des données, avec des approches comme le sous-échantillonnage, le sur-échantillonnage, et les méthodes d'échantillonnage hybride (Haixiang et al., 2017). Le sous-échantillonnage consiste à réduire la classe majoritaire en éliminant certaines de ses instances, tandis que le sur-échantillonnage cherche à augmenter artificiellement la taille de la classe minoritaire, en créant de nouvelles instances. Deux techniques de sur-échantillonnage sont communément utilisées : la duplication aléatoire d'instances de classe minoritaire et l'approche SMOTE (*Synthetic Minority Over-sampling Technique*) proposée par Jian et al. (2016), combinant les avantages des deux approches. En appliquant le SMOTE à la classe minoritaire et un sous-échantillonnage à la classe majoritaire, elles permettent de créer plusieurs jeux de données rééquilibrés. Cependant, ces méthodes de sur-échantillonnage reviennent à dupliquer des données existantes ce qui résulte de l'existence de doublons dans les jeux de données. Dans cet article, afin de pallier le problème des données déséquilibrées, nous présentons une stratégie complète basée sur une combinaison de sous-échantillonnage et de techniques d'*ensemble learning* (apprentissage ensembliste).

Parmi les domaines où les données fortement déséquilibrées sont particulièrement courantes, l'épidémiologie se distingue, et plus spécifiquement la cartographie du risque de leptospirose. La leptospirose est une maladie bactérienne présente dans de nombreuses régions tropicales et subtropicales, causée par la bactérie *Leptospira*. Cette zoonose touche plus d'un million de personnes par an, entraînant près de 60 000 décès, bien qu'elle soit encore largement sous-estimée et négligée (Goarant et al., 2018). La transmission se fait par l'urine des mammifères, principalement des rongeurs, qui contaminent les sols et les eaux, créant ainsi un environnement propice à la propagation de la bactérie. Les humains peuvent alors contracter la maladie, avec des conséquences graves, telles que l'insuffisance rénale ou pulmonaire, ou encore la maladie de Weil, une forme sévère de la leptospirose (Bharti et al., 2003).

Face à ce défi sanitaire, plusieurs études ont cherché à mieux comprendre la répartition géographique du risque de leptospirose afin de développer des outils d'aide à la décision pour les services de santé publique. Dans le sud du Brésil (Teles et al., 2023), des chercheurs ont exploré la saisonnalité de la maladie à l'aide d'un modèle SARIMA (*Seasonal Autoregressive Integrated Moving Average*) pour prédire le taux d'incidence. Cette approche, bien que pertinente, se limite aux zones où des cas ont déjà été recensés, ce qui réduit la portée de la cartographie à ces seules régions. Cette limitation est particulièrement problématique lorsque l'objectif est de prévenir la maladie dans des zones où les cas ne se sont pas encore manifestés. En Indonésie (Dhewantara et al., 2022) dont le climat est tropical, les chercheurs ont opté pour l'utilisation du modèle d'entropie maximale (MaxEnt). Ce modèle permet notamment de générer des pseudo-absences dans les zones où aucun cas de leptospirose n'a été signalé, offrant ainsi une vision plus étendue de la zone étudiée du risque de leptospirose. Cependant, en générant aléatoirement un nombre de pseudo-absences de cas de leptospirose, l'entropie maximale contraint la cartographie du risque aux seules zones où des points sont existants (présences et pseudo-absences). Ainsi, pour obtenir une cartographie complète du risque de leptospirose, il devra être nécessaire de générer un très grand nombre de pseudo-absences, ce qui reviendrait à créer un déséquilibre dans le jeu de données. Notre étude s'inscrit dans la continuité de ces travaux, avec l'objectif de proposer un outil de prévention pour les autorités de santé publique. Cet outil, fondé sur des données réelles de leptospirose recensées dans un archipel, vise à fournir une cartographie du risque précise, puisqu'elle sera réalisée à l'échelle

temporelle du mois et à une échelle spatiale plus fine que celle de la commune. Pour cela, nous formulons le problème comme une tâche de classification binaire visant à prédire l'absence ou la présence de risque de contamination par la bactérie liée à la leptospirose. En effet, le déséquilibre dans cette analyse spatio-temporelle provient de la granularité temporelle, étant menée sur un pas de temps mensuel. Une analyse réalisée sans cette échelle temporelle fine pourrait atténuer ce déséquilibre, mais notre objectif ici est de démontrer la faisabilité d'une analyse spatio-temporelle à cette résolution. Cela permettra non seulement de produire une cartographie précise du risque de leptospirose pour anticiper les recrudescences de cas, mais également d'identifier les facteurs les plus déterminants dans la transmission de la bactérie à une échelle globale. Cependant, en raison du nombre limité de pages, dans cet article, nous nous focaliserons sur la cartographie du risque de leptospirose.

L'article est structuré comme suit : la section 2 décrit la zone d'étude et les données collectées, avec un focus sur le traitement des données déséquilibrées dans cette étude épidémiologique. Cette section 2 détaille également comment la tâche de classification binaire est abordée et quels modèles de *machine learning* sont utilisés (section 2.3). La section 3 présente les résultats obtenus, confirmant la validité de notre approche. Enfin, la section 4 discute les conclusions de l'étude et propose des perspectives pour des recherches futures.

2 Méthodologie

2.1 Zone d'étude

Notre analyse se porte sur la cartographie du risque de leptospirose en Nouvelle-Calédonie. Situé dans le Pacifique Sud, la Nouvelle-Calédonie est un territoire français, composé de 271 407 habitants au dernier recensement (2019) répartis sur 33 communes, pour une superficie totale de 18 576 km^2 (Rivoilan, 2020). Le climat y est subtropical, avec deux principales saisons : chaude et pluvieuse de Novembre à Avril ; froide et sèche de Mai à Octobre (Lefèvre et al., 2010).

L'analyse spatiale est effectuée à l'échelle de l'IRIS (Ilôts Regroupés pour l'Information Statistique) développée par l'Institut National de la Statistique et des Études Économiques (INSEE). Il s'agit d'un découpage infra-communal permettant d'obtenir des zones résidentielles bien plus homogènes (en termes de population) que le découpage communal. Cette analyse à l'échelle de l'IRIS nous permettra d'obtenir une cartographie bien plus précise qu'à l'échelle de la commune. En Nouvelle-Calédonie, on recense 162 IRIS, décrits par des polygones (Figure 1). Concernant l'aspect temporel, l'analyse sera réalisée sur un pas de temps mensuel, que nous détaillons dans la section 2.2.

Afin de réaliser notre cartographie du risque de leptospirose, nous avons collecté plusieurs types de données spatialisées provenant de différentes sources que nous détaillons dans la sous-section suivante.

2.2 Collecte et pré-traitement de données

Cas de Leptospirose. La Direction des Affaires Sanitaires et Sociales (DASS) de Nouvelle-Calédonie a fourni l'ensemble des cas individuels de leptospirose recensés entre 2011 et 2022 inclus. Ces données contiennent notamment le lieu de résidence de la personne infectée, le

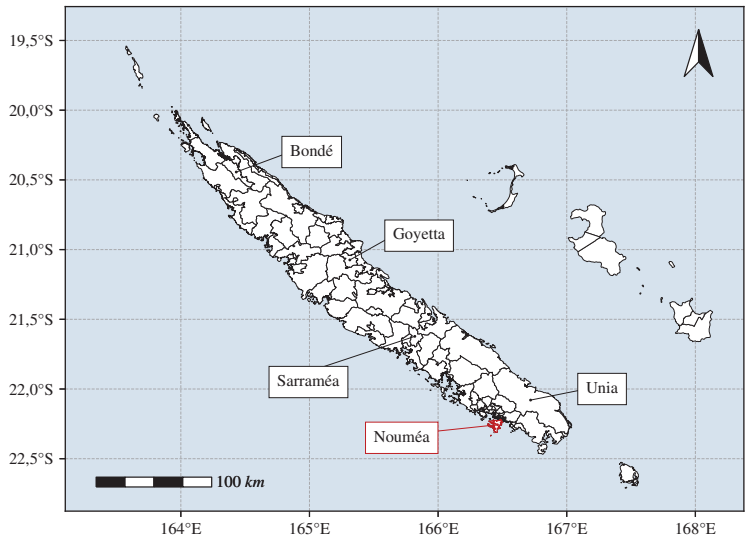

FIG. 1 – *Zone d'étude comprenant le découpage par IRIS.*

lieu de contamination le plus probable, le mois ainsi que l'année de contamination. Pour cette étude, nous avons utilisé le lieu de contamination le plus probable en tant qu'information géographique concernant les cas de leptospirose. Cette information est donnée à l'échelle du quartier. À partir de cela, nous avons associé chaque cas de leptospirose à l'IRIS correspondant.

Cependant, une grande proportion des cas recensés dans la capitale Nouméa (Figure 1), ne possède pas une précision géographique aussi fine que celle du quartier. Nous avons donc fusionné l'ensemble des IRIS localisés dans la ville de Nouméa en un seul IRIS, réduisant ainsi le nombre d'IRIS, passant de 162 à 114.

Entre 2011 et 2022 inclus, l'archipel a recensé plus de 1 200 cas de leptospirose, avec des taux d'incidence allant de 0 à 93,98 cas pour 10 000 habitants, le taux le plus élevé ayant été enregistré en Janvier 2021 dans l'IRIS de Sarraméa (Figure 1). Par respect aux personnes concernées, les cas ont été anonymisés par la DASS. Ainsi, aucune information tels que le nom, l'adresse de résidence et l'âge permettant d'identifier la personne, n'est fournie.

Puis, à partir de ces cas de leptospirose par mois et par IRIS, nous avons binarisé le nombre de cas à ces échelles. Ainsi, pour un mois donné, lorsqu'un IRIS n'a recensé aucun cas de leptospirose, la classe "absence de risque" lui est affectée. *A contrario*, lorsqu'un IRIS a recensé au moins un cas durant un mois donné, la classe "présence de risque" lui est affectée. Alors que notre étude cherche à établir une cartographie du risque de leptospirose à partir de données déséquilibrées, nous avons considéré qu'à partir du moment où un IRIS a recensé au moins un cas de leptospirose à une période donnée, le risque y est présent. Nous sommes donc dans une tâche de classification binaire.

Pour cette étude, nous obtenons un nombre total de 16 416 instances ($114\,\text{IRIS} \times 12\,\text{mois} \times 12$ ans) qui composent notre jeu de données. Le déséquilibre de ce jeu de données réelles se retrouve donc dans les cas de leptospirose mensuels par IRIS. En effet, sur nos 16 416 IRIS

mensuels, seulement 909 d'entre eux recensent au moins un cas de leptospirose, ce qui ne représente que 5,54% de notre jeu de données réelles. Dans le cas d'une classification binaire, où la classe 0 représente une "absence de risque" et la classe 1 représente une "présence de risque", nous avons donc une répartition des classes (0–1) de 94,46–5,54%.

Afin de réaliser l'apprentissage pour une classification binaire à partir de données fortement déséquilibrées, nous avons collecté des données environnementales et socio-démographiques pouvant expliquer cette présence de cas de leptospirose.

Données météorologiques. À travers le service Météo-France de Nouvelle-Calédonie, nous avons collecté l'ensemble des cumuls de précipitations, les minimales, moyennes et maximales de température, ainsi que les moyennes d'humidité. L'ensemble de ces données a été récolté sur un pas de temps mensuel entre 2010 et 2022 et est disponible en libre accès [1]. Cependant, ces variables (précipitations, température et humidité) ont été mesurées via des stations météorologiques. On recense environ une cinquantaine de stations à travers la Nouvelle-Calédonie qui mesure ces variables à un pas de temps mensuel. Malgré la cinquantaine de stations installées, elles ne permettent pas d'obtenir une représentation précise de la météorologie en Nouvelle-Calédonie (i.e., les IRIS ne possèdent pas tous au moins une station installée).

Afin d'obtenir une carte représentative de la météorologie en Nouvelle-Calédonie, nous avons appliqué une méthode d'interpolation. En l'occurrence, il s'agit de la méthode AURELHY (Analyse Utilisant le RELief pour l'HYdrométéorologie), proposée par Bénichou et Le Breton (1987). La méthode AURELHY est une technique d'interpolation spatiale qui utilise à la fois des données météorologiques et des caractéristiques du relief (altitude, pente, exposition) pour modéliser les variations locales du climat. Elle est particulièrement adaptée aux zones à relief complexe, comme les régions montagneuses. En intégrant le relief, cette méthode améliore la précision des cartes climatiques. Par le biais de cette méthode, nous avons donc pu interpoler nos points associés aux cumuls de précipitations (mm), aux minimales, moyennes et maximales de température (°C, Figure 2), ainsi qu'aux moyennes d'humidité (%). De plus, à partir de cette méthode, il est possible de mesurer la qualité d'interpolation par le biais du coefficient de corrélation (R^2) entre les données météorologiques existantes et les caractéristiques du relief. À titre indicatif, les interpolations appliquées aux précipitations, températures et humidités sur un pas de temps mensuel, ont obtenu un coefficient de corrélation de 0,5 $\pm$ 0,14, 0,76 $\pm$ 0,19 et 0,81 $\pm$ 0,13, respectivement.

Afin de réaliser l'apprentissage sur les absences et présences de risque de contamination, nous avons intégré ces données interpolées. Pour cela, nous avons, pour chaque IRIS à un mois donné, récupéré les valeurs météorologiques de pixels obtenus par AURELHY qui sont contenus dans le polygone associé à chaque IRIS. Concernant les cumuls de précipitations, nous avons pour un mois m donné, agrégé par la moyenne, la médiane et la variance, les précipitations des mois $m-1$, $m-2$ et $m-3$, notamment car une étude en Nouvelle-Calédonie (Thibeaux et al., 2024) a démontré la resuspension de la bactérie *Leptospira* dans les sols causée par les précipitations jusqu'à 3 mois avant les prélèvements et ainsi le risque de contamination. Puis, concernant les données de température, nous les avons agrégées par le minimum, la moyenne, le maximum ainsi que son écart ($T^{\circ}_{\max} - T^{\circ}_{\min}$) au mois $m-1$, pour chaque IRIS à un mois m donné. Enfin, les données interpolées d'humidité au mois $m-1$ ont été agrégées par la moyenne des pixels contenus dans chaque IRIS, pour un mois m donné.

1. `https://meteo.data.gouv.fr/`

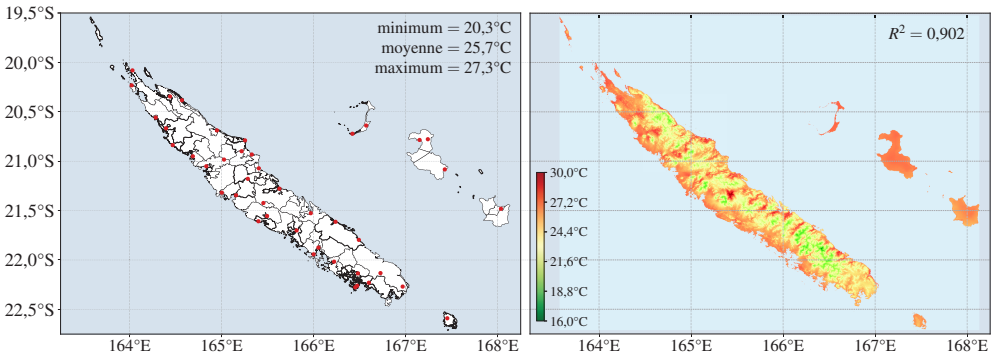

FIG. 2 – *Exemple d'une interpolation par AURELHY (à droite) réalisée sur les températures moyennes de Décembre 2022 à partir des stations météorologiques (à gauche).*

Données environnementales et topographiques. Concernant les données environnementales et topographiques, nous avons récolté les types de sols recouvrant la Nouvelle-Calédonie, ses modes d'occupation du sol (MOS), les zones d'élevage, le recensement des feux de forêt, ainsi que le relief.

Les types de sols recouvrant l'archipel ont été fournis par l'Institut de Recherche pour le Développement (IRD). Ces types, décrits par des polygones, sont divisés en six groupes principaux : "*Regosols et Leptosols*", "*Vertisols*", "*Cambisols*", "*Acrisols*", "*Ferrasols et Plinthosols*", et "*Fluvisols*". Puis, ces types sont sous-divisés respectivement en 4, 2, 5, 2, 4 et 3 sous-catégories, comptabilisant un nombre total de 20 catégories de sols. Cette sous-division des types de sols a été réalisée selon le *World Reference Base for Soil Resources* (WRB, 2007) puis adaptée aux spécificités de la Nouvelle-Calédonie (Bonvallot et al., 2012).

La carte des MOS a été réalisée par le Gouvernement de la Nouvelle-Calédonie et l'Observatoire de l'Environnement en Nouvelle-Calédonie (OEIL). Dans l'archipel, les modes d'occupation du sol sont proposés sous trois niveaux de précision de classification des modes, contenant 5, 22, et 26 catégories, respectivement. Pour cette analyse, nous avons intégré le niveau contenant 22 catégories, puisque parmi elles, nous retrouvons la catégorie "*Décharges, chantiers, extraction de matériaux*" qui nous donne indirectement une information sur la présence de rongeurs, un des vecteurs principaux de la bactérie *Leptospira* (Guernier et al., 2018). La carte des zones d'élevage, décrite par des polygones, a été réalisée et fournie par l'OEIL. Les cartes des types de sols, des MOS et des zones d'élevage ont été intégrées dans l'analyse de la même manière. Pour chacune des variables et chacune des catégories (si la variable en possède), nous y avons calculé la couverture par rapport à chaque IRIS. Ainsi pour chaque IRIS mensuel, ces variables sont représentées par des valeurs décrivant leur proportion de couverture, bornée entre 0 et 1. La cartographie des feux de forêt a été également réalisée et fournie par l'OEIL. Pour cela, l'observatoire a mis en place un traitement en temps réel d'imagerie satellitaire, en particulier les données *Moderate-Resolution Imaging Spectroradiometer* (MODIS) et *Suomi National Polar-Orbiting Partnership* (NPP), afin de détecter les feux de forêt. Ces feux de forêt sont décrits par des points répartis sur tout le territoire. Ainsi, pour intégrer ces points dans notre analyse, nous avons recensé les occurrences de points contenus dans

chacun des IRIS, pour un mois donné.

Enfin, concernant le relief, il s'agit d'un raster d'une résolution de 10 mètres fourni par le Gouvernement de la Nouvelle-Calédonie. Afin de l'intégrer dans notre analyse, nous avons appliqué la même procédure que pour les rasters obtenus par les interpolations de données météorologiques. Pour chaque IRIS, les pixels qui y sont contenus ont été agrégés afin que l'IRIS soit décrit par la moyenne, la médiane et la variance de l'altitude (exprimées en mètres).

Données socio-démographiques. L'ensemble des données socio-démographiques que nous avons récoltées ont été produites et fournies par l'Institut de la Statistique et des Études Économiques (ISEE) de Nouvelle-Calédonie. Elles ont été recoltées lors de chaque recensement de la population (i.e., 2009, 2014 et 2019). De plus, les variables que nous allons énumérer ont déjà été découpées par IRIS. Il n'y a donc eu aucun pré-traitement spatial réalisé.

Selon la littérature (Mayfield et al., 2018; Baquero et Machado, 2018), les indicateurs démographiques tels que la densité de population, l'indice de pauvreté et le mouvement interne des populations apparaissent comme des variables influençant le risque de contamination. Ainsi, les variables socio-démographiques par IRIS intégrées dans l'analyse sont les suivantes : la population totale, la densité de population (hab/km^2), le nombre de personnes dont le mode de transport principal est la marche à pied, le nombre de personnes dont le mode de transport principal est le transport public, le nombre de personnes vivant en tribu, le nombre de personnes agricoles, le nombre de ménages sans accès individuel à l'eau potable, le nombre de ménages sans accès individuel à l'électricité, le nombre d'anciens logements (nous avons considéré qu'une maison est ancienne dès lors que sa fin de construction fut dix ans avant le mois analysé), le nombre de personnes travaillant dans une zone autre que celle de résidence, le nombre de personnes de la classe ouvrière et le nombre de personnes au foyer.

Cependant, même si ces variables ont été récoltées durant les recensements de population, ces derniers ne sont réalisés qu'une fois tous les cinq ans, en l'occurrence 2009, 2014 et 2019. Afin d'éviter de répéter les mêmes valeurs de variables socio-démographiques pour chaque IRIS tous les mois durant cinq années consécutives, nous y avons appliqué une interpolation linéaire. Ainsi, pour un mois d'une année y donnée, les variables socio-démographiques ont été estimées de la manière suivante : si y se retrouve entre deux recensements (y_i, y_j), nous avons appliqué une simple fonction affine de la variable v tel que $v(y) = a \times y + b$, où $a = \frac{y_j - y_i}{v_j - v_i}$ et $b = y_i - a \times v_i$, avec v_i et v_j les valeurs de la variable v aux recensements y_i et y_j, respectivement. Dans le cas où y ne se situe pas entre deux recensements disponibles, en l'occurrence lorsque $y \geq 2019$, nous avons retenu les mêmes valeurs associées aux variables socio-démographiques récoltées durant le recensement 2019.

Pour récapituler, dans le cadre de notre analyse spatio-temporelle du risque de leptospirose en Nouvelle-Calédonie par le biais d'une classification binaire, la variable à expliquer est l'absence/présence de cas de leptospirose par mois et par IRIS, comptabilisant 16 416 instances ; et nous avons un nombre total de 73 variables explicatives.

2.3 Ensemble learning

Dans notre classification binaire, le jeu d'apprentissage regroupe tous les IRIS des mois de 2011 à 2020 inclus, et le jeu de test regroupe tous les IRIS des mois de 2021 et 2022. Afin de pallier le problème des classes fortement déséquilibrées, nous avons appliqué plu-

sieurs méthodes. En premier lieu, nous avons normalisé notre jeu de données complet via la normalisation Min-Max qui consiste à normaliser nos données entre 0 et 1 afin que nos variables soient dans le même ordre de grandeur. Puis, nous avons appliqué la méthode du sous-échantillonnage. En effet, entre 2011 et 2020, nous ne comptabilisons que 573 sur 13 680 IRIS qui ont recensé au moins un cas de leptospirose. De ce fait, nous avons utilisé la méthode du sous-échantillonnage de la manière suivante : nous avons effectué un tirage aléatoire sur les IRIS n'ayant recensé aucun cas de leptospirose afin d'en sélectionner autant d'IRIS ayant recensé au moins un cas, c'est-à-dire 573. Cela résulte d'un nouveau jeu de données parfaitement équilibré contenant 1 146 IRIS. Pour la suite, afin d'obtenir une meilleure représentativité des données et affiner nos résultats, nous avons effectué ce tirage 50 fois dans le but de créer 50 jeux de données qui, certes, contiennent les mêmes IRIS ayant recensé au moins un cas de leptospirose, mais ne contiennent pas les mêmes IRIS n'ayant recensé aucun cas.

Pour chacun de ces 50 jeux de données, nous y avons appliqué un *ensemble learning* dans lequel quatre modèles d'apprentissage supervisé sont utilisés : la régression logistique (LR), la machine à vecteurs de support (SVM), l'extreme gradient boosting (XGB) et la forêt aléatoire (RF). Puis, pour chacun de ces quatre modèles, nous y avons appliqué un *Grid Search* couplé à une validation croisée 5-fold qui nous permettra notamment de diviser les 1 146 IRIS en cinq sous-ensembles égaux. Pour chacune des cinq itérations, un sous-ensemble sera utilisé comme jeu de validation tandis que les quatre restants serviront à l'entraînement du modèle. En classification binaire, plusieurs métriques sont disponibles. Pour notre étude épidémiologique, puisqu'il nous est crucial d'évaluer notre stratégie à partir des données réellement observées, nous avons utilisé les métriques suivantes : le *rappel* (aussi appelé la *sensibilité*), la *spécificité*, le score du *Brier loss*, l'*accuracy*, la *justesse équilibrée* (en anglais la *Balanced Accuracy* et notée *Bal Acc*) et l'aire de la courbe ROC (AUC-ROC) (Berrar, 2024).

Pour réaliser le *Grid Search*, nous avons fourni une large plage de paramètres associés à chaque modèle que nous avons fait varier et les meilleurs modèles seront sélectionnés selon la *justesse équilibrée*. Finalement, à partir des 50 jeux de données et des 4 modèles d'apprentissage supervisé, nous obtenons 200 modèles entraînés. Cependant, afin de réaliser une prédiction optimisée et rapide, nous avons fait le choix de ne retenir que les 5% meilleurs modèles parmi les 200 modèles entraînés. Ces modèles ont été retenus selon la même métrique utilisée durant le *Grid Search*, en l'occurrence la *justesse équilibrée*. Cela résulte de 10 modèles retenus. À partir de ces 10 modèles retenus, une prédiction pondérée est réalisée. En effet, pour chaque IRIS x_i prédit du jeu de test, la prédiction pondérée sur n modèles est définie par :

$$P(x_i) = \max_{p \in \{A,B\}} \sigma(p) \tag{1}$$

$$A = \frac{2}{n} \times \sum_{j=1}^{n} \textit{Specificité}_j \times \left| p_j(x_i) - \frac{1}{2} \right| \times \mathbb{1}_{p_j(x_i) < \frac{1}{2}} \tag{2}$$

$$B = \frac{2}{n} \times \sum_{j=1}^{n} \textit{Rappel}_j \times \left| p_j(x_i) - \frac{1}{2} \right| \times \mathbb{1}_{p_j(x_i) \geq \frac{1}{2}} \tag{3}$$

Pour résumer, si un modèle prédit une "absence de risque", alors la probabilité est pondérée par rapport à la *spécificité*. À l'inverse, si un modèle prédit une "présence de risque", alors la probabilité est pondérée par rapport au *rappel*. Afin d'assurer que la somme des probabili-

tés pondérées obtenues soit égale à 1, une fonction softmax, notée $\sigma(\cdot)$, est appliquée. Cette dernière est définie par :

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \ \forall\, j \in \{1, ..., K\} \text{ et } z = (z_1, ..., z_K) \in \mathbb{R}^K \tag{4}$$

Avec K correspondant au nombre de classes présentes dans le jeu de données. Dans notre cas, nous avons 2 classes (“absence de risque” et “présence de risque”).

3 Résultats

La figure 3 illustre la cartographie du risque de leptospirose en 2021. En raison du nombre limité de pages, seuls les quatre premiers mois de l’année 2021 ont été affichés. Le tableau 1 synthétise les résultats obtenus des 5% meilleurs modèles retenus. Les métriques y ont été calculées à partir des données du jeu de test (les IRIS des mois de 2021 et 2022). À titre de comparaison, nous avons également affiché dans ce tableau 1, les résultats obtenus à partir d’une autre approche d’échantillonnage, en l’occurrence, nous avons expérimenté un échantillonnage hybride. En effet, nous avons appliqué un sur-échantillonnage sur les données observées, c’est-à-dire sur les IRIS de la classe minoritaire (“présence de risque”). Pour chacun de ces IRIS, nous les avons fusionnés avec leurs voisins respectifs afin de créer de nouvelles instances basées sur des couples d’IRIS adjacents. Puis, nous avons appliqué le sous-échantillonnage sur les IRIS de la classe majoritaire afin d’obtenir de plus grands jeux d’apprentissage équilibrés. Par la suite, nous avons appliqué exactement les mêmes étapes décrites dans la section 2 (normalisation, *ensemble learning*, *Grid Search* et prédictions pondérées).

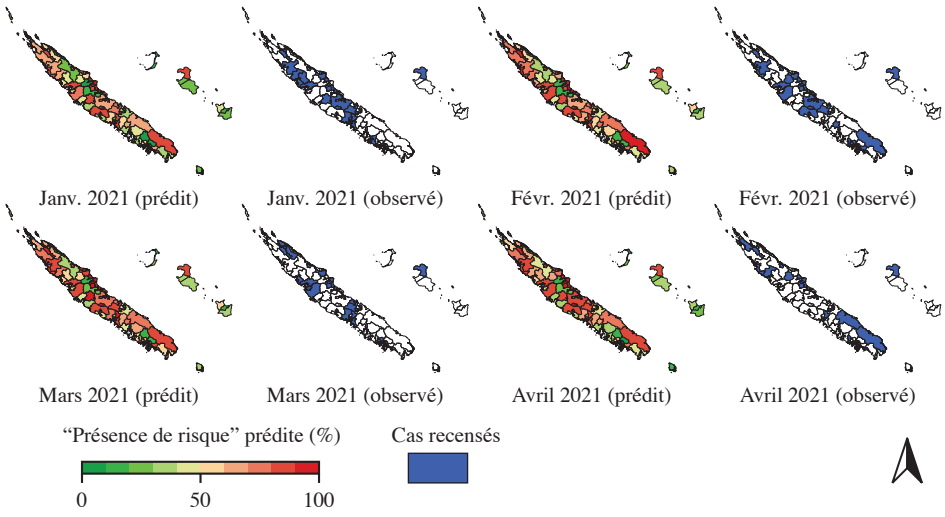

FIG. 3 – *Prédictions pondérées de la “présence de risque” sur les IRIS des mois de 2021.*

	Modèle		AUC		*Brier Loss*		*Accuracy*		*Rappel*		*Specificité*		*Bal Acc*	
–	Ensemble		**83,29**	**82,23**	20,18	**9,64**	70,36	87,54	83,93	36,61	68,46	94,67	**76,19**	65,64
01	SVM	SVM	80,97	81,25	21,06	10,06	68,35	87,13	83,33	33,63	66,25	94,62	74,79	64,13
02	RF	SVM	81,76	81	21,13	10,26	67,73	87,1	84,23	36,9	65,42	94,12	74,82	65,51
03	SVM	SVM	81,6	81,29	21,06	10,12	68,71	87,1	83,04	37,5	66,71	94,04	74,87	65,77
04	SVM	SVM	81,69	81	**19,16**	10,11	**71,42**	87,1	79,46	35,71	**70,29**	94,29	74,88	65
05	RF	LR	81,7	80,22	21,69	9,91	68,17	87,06	83,93	37,2	65,96	94,04	74,94	65,62
06	SVM	SVM	82,06	80,02	21,61	9,85	68,93	87,54	83,04	30,95	66,96	95,46	75	63,21
07	RF	SVM	81,34	80,2	20,99	9,75	69,88	87,65	81,85	30,36	68,21	**95,67**	75,03	63,01
08	RF	LR	81,73	80,52	21,21	9,93	67,95	87,24	**84,82**	**39,88**	65,58	93,88	75,2	**66,88**
09	RF	SVM	81,16	80,92	21,37	9,86	68,42	87,65	84,52	30,65	66,17	95,62	75,35	63,14
10	RF	SVM	80,64	81,23	20,54	9,75	69,37	**87,68**	83,33	34,52	67,42	95,12	75,38	64,82

TAB. 1 – *Résultats des 5% des meilleurs modèles retenus et du modèle (Ensemble) de prédictions pondérées sur le jeu de test. Les cellules divisées en deux colonnes correspondent aux scores obtenus à partir d'un apprentissage basé sur le sous-échantillonnage (à gauche, en bleu) et sur l'échantillonnage hybride (à droite, en rouge). Les valeurs sont exprimées en pourcentage (%).*

Comme mentionné dans la section 2.2, l'IRIS de Sarraméa a recensé le plus haut taux d'incidence de leptospirose, avec 93,98 cas pour 10 000 habitants en Janvier 2021. Avec notre stratégie, notre modèle y a prédit une "présence de risque" de 85,98%. De plus, dans trois autres IRIS, Goyetta, Bondé et Unia (Figure 1), qui ont également enregistré des taux d'incidence pour 10 000 habitants élevés (69,24 cas en Janvier 2021, 61,35 cas en Avril 2021 et 43,73 cas en Avril 2021), notre modèle a prédit une "présence de risque" de 87,14%, 94,19% et 80,42%, respectivement à ces périodes. Comme nous le montre le tableau 1, notre modèle a bien détecté les IRIS où au moins un cas de leptospirose a été recensé, puisque la stratégie a résulté d'un AUC-ROC de 83,29%, d'un *rappel* de 83,93%, d'un *accuracy* de 70,36% et d'un score de *Brier loss* de 0,2. Cela montre notamment que malgré le fort déséquilibre dans notre jeu de données initial, notre stratégie a pu développer une capacité à discriminer les deux classes à prédire ainsi qu'un modèle final bien calibré compte tenu du score de *Brier Loss*.

De plus, notre stratégie a pu déterminer la saisonnalité du risque de leptospirose puisqu'elle a prédit une "présence de risque" médiane de 47% durant la saison dite chaude et pluvieuse (de Novembre à Avril) comme nous le montre la figure 3, tandis que cette médiane est de 31% durant la saison dite froide et sèche (de Mai à Octobre). Concernant l'année 2022 qui a été une année particulièrement touchée par le phénomène de La Niña, la "présence de risque" prédite est restée globalement élevée toute l'année, résultant d'une "présence de risque" médiane égale à 43%, tandis qu'en 2021 la "présence de risque" médiane prédite est égale à 35%. Le phénomène de La Niña est un phénomène climatique où la température de la surface de l'eau est plus faible que la moyenne, affectant la météorologie sur la région ouest du Pacifique et résultant de conditions climatiques plus humides et d'une augmentation des épisodes cycloniques.

Cependant, il est également notable que sur certains IRIS où aucun cas n'a été recensé, notre modèle a tout de même prédit une "présence de risque" et non une "absence de risque", ce qui a résulté d'une *spécificité* modérée de 68,46%. Cela est notamment dû au seuil de probabilité entre ces deux classes que nous avons fixé à 50%. Ce seuil est certes modifiable, selon le cas d'application. Dans une étude épidémiologique telle que la nôtre, il demeure difficile de trouver le bon compromis entre un modèle "rassurant" et un modèle "alarmant". Au final, il en revient aux experts métier tels que les épidémiologistes et les décideurs en santé publique, de

déterminer le seuil optimal afin d'y obtenir la cartographie la plus adaptée.

En comparaison, avec l'approche par l'échantillonnage hybride (Tableau 1), nous obtenons des résultats bien plus optimaux que ceux par le sous-échantillonnage. Cependant, nous avons obtenu un *rappel* de 36,61% et une *spécificité* de 94,67%, indiquant une prédiction quasi-systématique d'une "absence de risque". Par ces deux métriques obtenues, pour une étude épidémiologique telle que la nôtre, l'approche ne peut être validée, notamment lorsque l'objectif premier est la bonne capacité à détecter les IRIS avec une "présence de risque" observée.

4 Conclusion et Perspectives

Dans cet article, nous avons présenté une méthodologie complète afin de pallier au problème de données déséquilibrées. Nous avons décrit le processus complet de pré-traitement de données déséquilibrées, par le sous-échantillonnage couplé à l'*ensemble learning*.

Pour démontrer la faisabilité de notre stratégie, nous l'avons appliquée à une classification binaire d'une étude spatio-temporelle comprenant des données de leptospirose entre 2011 et 2022, domaine particulièrement propice à la présence de données fortement déséquilibrées. Malgré ces classes déséquilibrées, notre stratégie a pu atteindre des résultats très satisfaisants, avec une *justesse équilibrée* de 76,19%, un *rappel* de 83,93% et un AUC de 83,29%.

La prochaine étape de cette étude de données déséquilibrées sera de déterminer parmi les variables que nous avons intégrées, quelles sont celles qui contribuent le plus à la transmission de la bactérie et donc, du risque de contamination. Cette étape nous permettra notamment de vérifier si malgré les données déséquilibrées, nous arrivons à retrouver les mêmes variables influentes que dans d'autres régions du monde touchées par la leptospirose.

Remerciements

Ce travail a été financé dans le cadre du projet ANR SPIraL (ANR-19-CE35-0006-02). Nous remercions la Direction des Affaires Sanitaires et Sociales de Nouvelle-Calédonie pour la collecte, la consolidation et la mise à disposition des données relatives aux cas de leptospirose.

Références

Baquero, O. S. et G. Machado (2018). Spatiotemporal dynamics and risk factors for human leptospirosis in brazil. *Scientific Reports 8*(1), 15170.

Berrar, D. (2024). *Performance Measures for Binary Classification*, pp. 31. Elsevier.

Bharti, A., J. Nally, J. Ricaldi, M. Matthias, M. Diaz, M. Lovett, P. Levett, R. Gilman, M. Willig, E. Gotuzzo, et J. Vinetz (2003). Leptospirosis : a zoonotic disease of global importance. *The Lancet Infectious Diseases 3*(12), 757–771.

Bonvallot, J., J.-C. Gay, et E. Habert (2012). *Atlas de la Nouvelle Calédonie*. Marseille (FRA) ; Nouméa : IRD ; Congrès de la Nouvelle-Calédonie.

Bénichou, P. et O. Le Breton (1987). Aurelhy : une méthode d'analyse utilisant le relief pour les besoins de l'hydrométéorologie. In *Deuxièmes journées hydrologiques de l'ORSTOM à Montpellier*, Colloques et Séminaires, pp. 299–304. Paris : ORSTOM.

Dhewantara, P., M. Riandi, et T. Wahono (2022). Effect of climate change on the geographical distribution of leptospirosis risk in western java, indonesia. *IOP Conference Series : Earth and Environmental Science 1089*(1), 012074.

Fernández, A., S. del Río, N. V. Chawla, et F. Herrera (2017). An insight into imbalanced big data classification : outcomes and challenges. *Complex & Intelligent Systems 3*(2), 105–120.

Goarant, C., M. Picardeau, S. Morand, et K. M. McIntyre (2018). Leptospirosis under the bibliometrics radar : evidence for a vicious circle of neglect. *J Glob Health 9*(1), 010302.

Guernier, V., C. Goarant, J. Benschop, et C. Lau (2018). A systematic review of human and animal leptospirosis in the pacific islands reveals pathogen and reservoir diversity. *PLOS NTD 12*(5), e0006503.

Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, et G. Bing (2017). Learning from class-imbalanced data : Review of methods and applications. *Expert Systems with Applications 73*, 220–239.

Jian, C., J. Gao, et Y. Ao (2016). A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing 193*, 115–122.

Lefèvre, J., P. Marchesiello, N. Jourdain, C. Menkes, et A. Leroy (2010). Weather regimes and orographic circulation around new caledonia. *Marine Pollution Bulletin 61*(7-12), 413–431.

Longadge, R. et S. Dongre (2013). Class imbalance problem in data mining review.

Mayfield, H., C. Smith, J. Lowry, C. Watson, M. Baker, M. Kama, E. Nilles, et C. Lau (2018). Predictive risk mapping of an environmentally-driven infectious disease using spatial bayesian networks : A case study of leptospirosis in fiji. *PLOS NTD 12*(10), e0006857.

Rivoilan, P. (2020). La croissance démographique fléchit nettement en nouvelle-calédonie entre 2014 et 2019. Technical Report 45, ISEE.

Teles, A. J., B. C. Bohm, S. C. M. Silva, et F. R. P. Bruhn (2023). Socio-geographical factors and vulnerability to leptospirosis in south brazil. *BMC Public Health 23*(1), 1311.

Thibeaux, R., P. Genthon, R. Govan, N. Selmaoui-Folcher, C. Tramier, M. Kainiu, M.-E. Soupé-Gilbert, K. Wijesuriya, et C. Goarant (2024). Rainfall-driven resuspension of pathogenic leptospira in a leptospirosis hotspot. *Science of The Total Environment 911*, 168700.

WRB, I. W. G. (2007). World reference base for soil resources 2006. TechRep 103, FAO.

Summary

The advent of big data collection has contributed to the development of knowledge extraction methods, but it has also introduced new challenges. One of the main issues is dealing with highly imbalanced datasets, particularly in class labels for categorical classification tasks. This article presents a comprehensive strategy developed to address the issue of imbalanced data in a spatio-temporal epidemiological study of leptospirosis. The approach was evaluated using real data for a binary classification task, predicting the presence of contamination risk with the bacteria associated with leptospirosis, where the majority class represents 95% of the labels. By applying under-sampling, training 200 machine learning models, and using weighted predictions, our strategy achieved a balanced accuracy of 76.19%, an AUC-ROC of 83.29%, and a recall of 83.93%.