Prédiction des niveaux de risque pollinique à partir de données historiques multi-sources

Esso-Ridah Bleza[*,**,***], Valérie Monbet[**]
Pierre-François Marteau[***]

[*]Lify Air
esso-ridah.bleza@lifiair.com
https://www.lifyair.com
[**]IRMAR, Université de Rennes
valerie.monbet@univ-rennes1.fr
https://perso.univ-rennes1.fr/valerie.monbet/
[***]IRISA, Université Bretagne Sud
pierre-francois.marteau@univ-ubs.fr
https://people.irisa.fr/Pierre-Francois.Marteau/

Résumé. Dans la littérature scientifique, de nombreuses études montrent que les conditions météorologiques ont un impact sur l'émission, la dispersion et la suspension des pollens dans l'air. Plusieurs espèces allergisantes menacent la santé des millions de personnes en France. Une information préventive du risque d'exposition pollinique fiable constitue un réel atout pour les allergiques. L'objectif principal de cet article est d'étudier, grâce à des techniques d'apprentissage statistique exploitant des données historiques, et les paramètres météorologiques du jour (J), la capacité à prédire à 3 jours (J+3) à l'avance les niveaux de risques de présence de pollens dans l'air sur un territoire donné (en France Métropolitaine). Nous nous sommes intéressés à la prévision de risque pour 3 familles de pollens qui font partie des espèces les plus allergisantes (ambroisie, cupressacées et graminées). Pour chacun des 4 niveaux de risque considérés, l'agrégation de modèles de régression logistique binaire par un classifieur de type Forêt aléatoire a permis de prédire le niveau du risque pollinique avec des performances de l'ordre de 75% à 90% d'AUC et 70% de précision et de rappel, les confusions concernant principalement les niveaux faible et moyen.

1 Introduction

Plusieurs études récentes montrent que les populations, tout particulièrement en France, souffrent de plus en plus d'allergies à un ou plusieurs pollens. On estime à ce jour que 25 % des français sont concernés et que la population allergique passera la barre des 50 % en 2050 (`https://lejournal.cnrs.fr`). Les seules options pour les allergiques aujourd'hui sont soit le traitement récurrent avec son lot d'inconvénients (effets secondaires, coût, accoutumance, etc.) soit un traitement à l'apparition des

symptômes et les conséquences qui s'en suivent (maladie, absentéisme, etc.). À ce jour, les mesures de concentration des pollens dans l'air sont délivrées au public *a posteriori* avec un décalage temporel important (Cassagne, 2009). En effet, le réseau national de surveillance aérobiologique (RNSA), organisme de référence, diffuse les informations issues des capteurs HIRST dont la technique est basée sur l'analyse et l'identification au microscope des pollens. Ces manipulations induisent un certain délai dans la diffusion des informations au public.Ces constats incitent à développer des modèles prédictifs à courte échéance du risque pollinique pour compléter l'information diffusée par le RNSA et améliorer la prévention des allergies.

Différentes études se sont intéressées au développement de modèles, qui, partant des données météorologiques, prédisent soit la présence d'une espèce de pollen dans un endroit donné, soit le début de saison pollinique d'une espèce (Andersen, 1991; Cassagne, 2009), la variation inter-annuelle des saisons polliniques (Spieksma et al., 1995), ou encore la prévision du niveau du risque ou de concentration pollinique. Par exemple, elles montrent que l'usage des réseaux de neurones permet d'atteindre des résultats satisfaisants pour prédire un risque d'exposition au bouleau (Castellano-Méndez et al., 2005; Muzalyova et al., 2021). Dans (Castellano-Méndez et al., 2005) les auteurs ont ramené le problème à une classification binaire pour chaque niveau de risque en utilisant l'historique de concentration du pollen de bouleau et des données météorologiques pour prédire les jours à haut risque allergénique du bouleau à partir de la (concentration moyenne (grains/m^3), niveau de précipitation et température moyenne journalière). Dans une étude menée à Augsbourg (Allemagne), des modèles prédictifs (régression ARIMA et dynamiques, réseaux neuronaux artificiels, modèles d'autorégression par réseaux neuronaux) ont été déployés pour prévoir la concentration des pollens de bouleau et de poacées, sur la base de mesures automatiques du pollen en temps quasi réel (par pas de 3h) (Muzalyova et al., 2021). Les données météorologiques sont utilisées pour construire les modèles, la température et les précipitations se révélant être les variables les plus significatives. Les réseaux de neurones ont également été utilisés dans d'autres études aérobiologiques, pour construire des modèles supervisés permettant de produire des prévisions de concentration quotidienne de pollen (Sánchez-Mesa et al., 2002; Hidalgo et al., 2002; Ranzi et al., 2003; Iglesias-Otero et al., 2015; Cordero et al., 2021) en ajoutant aux variables météorologiques, des paramètres phénologiques, des données caractéristiques du site et l'historique des concentrations de pollen. De façon plus générale, il existe des études portant sur l'impact des conditions météorologiques sur la diffusion de particules fines dans l'air, i.e. les particules de moins de $1\mu m$ ou $2.5\mu m$ (PM_1, $PM_{2.5}$). Les outils d'apprentissage automatique, notamment les méthodes basées sur les arbres de décision, y sont aussi mis en oeuvre (voir par exemple (Baklanov et al., 2016; Stirnberg et al., 2021) et leurs références).

Notre étude se distingue principalement sur deux aspects : premièrement nous avons choisi de prédire le risque pollinique assimilé à une variable discrète qui correspond à des niveaux d'émission (Thibaudon, 2003). Deuxièmement, nous nous intéressons à l'ensemble du territoire français en considérant au total 68 sites (sur 74 disponibles). Finalement, nous étudions avec une même méthodologie plusieurs pollens parmi les plus allergisants (21 espèces étudiées au total dont trois sont présentées dans cet article) répertoriés en France métropolitaine.

Les données sont introduites dans la section 2. Dans la section 3, les algorithmes d'apprentissage sont décrits. La section 4 regroupe les expérimentations : la capacité de généralisation des algorithmes de prédiction proposés est évaluée en temps et en espace. Enfin la section 5 présente quelques conclusions.

2 Données

Les données d'émission de pollen sont produites par le RNSA (`https://www.pollens.fr/reports/database`). En pratique, les grains de pollen se collent sur une lame adhésive qui est relevée puis analysée. On dispose ainsi de données journalières de concentration de pollen pour au moins une vingtaine d'espèces allergisantes et 74 sites pour une période couvrant les années 2000 à 2017. Les capteurs n'ont pas tous été implantés à la même période et les historiques disponibles varient selon les stations de mesure. Certains sites ne sont équipés que depuis 2012.

Comme le montre la figure 1 les concentrations de pollens présentent une forte saisonnalité annuelle avec une variation inter-annuelle qui peut être assez marquée selon les espèces. De plus, la distribution des pollens est dissymétrique. Elle est caractérisée par un nombre important de zéros qui correspondent aux jours sans émission, et une queue lourde liée à quelques jours de l'année présentant de fortes émissions.

Nous présentons dans cette étude trois espèces choisies pour leurs niveaux d'émissions, leurs caractères allergisants et leurs saisonnalités polliniques bien distincts : l'ambroisie, les poacées (graminées) et les cupréssacées avec les seuils de niveaux de risques détaillés dans le tableau 1.

pollen	**seuil faible**	**seuil moyen**	**seuil fort**
ambroisie	2	3	11
cupressacées	70	142	284
poacées	2	5	36

TAB. 1. Seuils de risque d'exposition allergique pour l'ambroisie, les cupréssacées et les poacées, en nombre de grains par m^3/jour. Valeurs issues de (Thibaudon, 2003).

Les données météorologiques sont issues de la base du European Climate Assessment & Dataset project (`https://www.ecad.eu/`). Il s'agit de données enregistrées par des capteurs au sol situés à proximité des aéroports (68 sites). Les stations météorologiques ne se trouvent pas à proximité immédiate des capteurs HIRST souvent situés dans les centres-villes (voir figure 2). Nous avons néanmoins choisi de travailler avec ces données d'observation parce qu'elles sont facilement accessibles en temps réel. En outre, à l'échelle journalière, les champs de variables météorologiques sont assez lisses et la variabilité sur quelques dizaines de kilomètres est faible.

Dans cette étude nous considérons des données journalières, en particulier la température moyenne à 2 mètres, le minimum et le maximum de la température, le cumul de précipitations et l'humidité. Les radiations solaires sont connues pour être un paramètre déterminant pour la floraison et les émissions de pollen. Mais elles ne sont pas

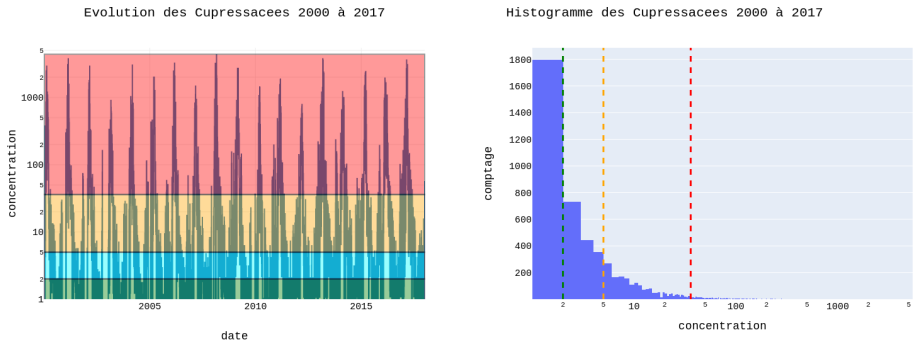

FIG. 1. exemple d'un historique de données de concentration des cupréssacées de 2000 à 2017 (à gauche), avec les niveaux de risque "nul", "faible", "moyen", "fort" matérialisés par les bandes verte, bleue, orange et rouge . Histogramme du même historique de données (à droite) ; Les barres verticales matérialisent les seuils "faible", "moyen" et "fort" de niveaux de risque (de la gauche vers la droite)

disponibles à toutes les stations météorologiques. On pourrait utiliser des sorties de modèle ou des données issues de satellite. Mais nous sommes contraints par un objectif opérationnel qui requiert des données disponibles en temps réel à coût faible. Les radiations ne sont donc pas prises en compte.

3 Modèles

Dans le domaine de l'aérobiologie, plusieurs études ont exploité des algorithmes d'apprentissage automatique, comme la méthode de Gradient Boosting (Cordero et al., 2021), les machines à vecteurs de support (Zewdie et al., 2019), les forêts aléatoires (Zewdie et al., 2019; Nowosad et al., 2018) ou encore les réseaux de neurones artificiels (Cordero et al., 2021; Puc, 2012; Iglesias-Otero et al., 2015; Valencia et al., 2019).

Ici, nous proposons une approche dans laquelle la concentration en pollen est discrétisée. Ainsi, la variable cible $Y_t^{(p)}$ représente le risque allergique lié au pollen. Elle est construite de la façon suivante.

$$\begin{aligned} Y_t^{(p)} &= \text{"}nul\text{"} \quad \text{si } C_t^{(p)} \leq s_{\text{faible}} \\ Y_t^{(p)} &= \text{"}faible\text{"} \quad \text{si } C_t^{(p)} > s_{\text{faible}} \\ Y_t^{(p)} &= \text{"}moyen\text{"} \quad \text{si } C_t^{(p)} > s_{\text{moyen}} \\ Y_t^{(p)} &= \text{"}fort\text{"} \quad \text{si } C_t^{(p)} > s_{\text{fort}} \end{aligned}$$

où $C_t^{(p)}$ est la concentration du pollen p au temps t. Les seuils s_{faible}, s_{moyen}, s_{fort} sont définis a priori et dépendent de l'espèce de pollen (voir table 1). Ainsi, $Y_t^{(p)}$ est une variable ordinale et on peut envisager différent modèles de prédiction.

Nous proposons ici de calibrer des modèles de régression multinomiale. Les paramètres sont estimés en maximisant la vraisemblance. Ces modèles ont a priori de bonnes propriétés de généralisation et leurs sorties sont interprétables. Notons que le modèle ordinal n'est qu'un cas particulier du modèle multinomial avec contraintes sur les paramètres. D'autres algorithmes, qui permettent de conserver une certaine interprétabilité, sont aussi testés mais conduisent à des résultats légèrement moins bons (voir Annexe)

Plus précisément, nous proposons un premier modèle d'agrégation qui consiste à combiner les prédictions de 3 régressions logistiques pourles seuils "faible", "moyen" et "fort" du niveau de risque selon deux approches. Dans la première, on considère la règle de classement suivante. *Si* $\hat{R}_t(moyen) == True$

si $\hat{R}_t(fort) == True$ *alors* $\widehat{Y_t^{(p)}} =$ "$fort$"

sinon $\widehat{Y_t^{(p)}} =$ "$moyen$"

sinon

si $\hat{R}_t(faible) == True$ *alors* $\widehat{Y_t^{(p)}} =$ "$faible$"

$\widehat{Y_t^{(p)}} =$ "nul", avec $\hat{R}_t$ les risques prédits par les modèles binomiaux et $\widehat{Y_t^{(p)}}$ la prédiction.

En pratique, les classes sont déséquilibrées et les prédictions sont réalisées en comparant la probabilité prédite à un seuil. Pour chaque modèle le seuil est déterminé de façon à maximiser le F1-score sur l'ensemble d'apprentissage.

Dans une seconde version du modèle d'agrégation, les probabilités $\hat{\pi_t}^{R_(r)}$ des modèles binomiaux associées au risque ${R_t}^r$, $r \in \{faible, moyen, fort\}$ sont utilisées en entrée d'un algorithme de forêt aléatoire pour obtenir un classement final à quatre modalités de risque : "nul", "faible", "moyen" ou "fort". Les variables météorologiques sont elles aussi transformées conformément aux pratiques communément admises en aérobiologie (voir références citées plus haut). Les variables introduites dans le modèle sont alors :

- la température du jour t et ses écarts entre deux jours successifs pour les jours t et $t-1$;
- les écarts entre les températures maximum et minimum aux jours t, $t-1$ et $t-2$;
- les degrés jour $D_t = \sum_{j=t-30}^{t} \frac{Tmax_j - Tmin_j}{2} \mathbf{1}_{Tmin_j > 10}$ où $\mathbf{1}$ est la fonction indicatrice et T_{min} et T_{max} les min et max journaliers de la température ;
- l'humidité du jour t et ses écarts entre deux jours successifs pour les jours t et $t-1$;
- le cumul de précipitations.
- La saisonnalité annuelle est prise en compte en introduisant une variable qualitative qui est le numéro de la semaine et une variable quantitative $\cos(\frac{2\pi t}{365})$.

Finalement, l'information spatiale est introduite via la latitude et la longitude des capteurs HIRST ainsi qu'une variable qualitative de région qui prend les valeurs Nord-Est, Nord-Ouest, Sud-Est et Sud-Ouest.

4 Résultats

Nous avons entraîné des modèles dits "locaux" lorsque les données d'apprentissage utilisées sont restreintes à un site donné et "spatiaux" lorsqu'elles intègrent les données de 62 sites avec des variables géographiques (longitude, latitude et régions) ajoutées dans les descripteurs. Les modèles sont entraînés sur 80 % de données et évalués sur les 20 % restant qui correspondent aux quatre dernières années d'historique. Six (6) sites ont été retirés de la base d'apprentissage pour évaluer la capacité de généralisation en espace. Nous présentons une étude comparative des différents modèles sur la base de critères calculés sur les échantillons tests, en particulier l'aire sous la courbe ROC (AUC), la précision et le rappel.Nous avons testé les modèles : arbres de décision, forêt aléatoire, régression logistique multinomial, et ordinal et une combinaison de modèles de régression logistique binaire. Les résultats donnent un avantage à méthode de combinaison de modèles de régression logistique (voir le détail en **Annexe**)

L'interprétation des paramètres estimés dans les modèles de régression binaire (non montrés) révèle que les variables les plus importantes sont les numéros du jour et dans quelques cas ceux de la semaine, permettant de reproduire la forte saisonnalité. Les variables météorologiques, en particulier la température, aident à caractériser la variation autour de la saison. Selon les types de pollens les autres variables météorologique sont d'importance variable bien que souvent faible.

4.1 Performances globales des modèles binaires

Les performances globales des modèles binaires sont bonnes comme le montre le tableau 2. On observe que les résultats obtenus pour l'ambroisie sont légèrement meilleurs que ceux des autres espèces. Ceci est dû au fait que l'ambroisie émet très peu dans les sites du Nord de la France. La prédiction est alors facile car le risque nul est largement dominant. En outre, pour l'ensemble des pollens, une analyse plus fine basée sur les matrices de confusions (non présentées) indique que les erreurs viennent principalement des classes de risque "faible et "moyen" qui sont difficiles à séparer contrairement aux classes de risque "nul" et "fort". Ceci s'explique par le fait que dans la plupart des cas les seuils de risques faible et moyen sont très proches (voir tableau 1).

4.2 Analyse par site géographique

Une cartographie des AUC (figure 2) montre que les sites du Nord-Est et du Centre sont légèrement mieux prédits que les sites de l'Ouest. On remarque aussi que les sites voisins ont des AUC similaires. Par ailleurs, on observe en général que les prédictions sont meilleures quand les stations météorologiques sont proches des sites des capteurs HIRST (voir Rennes et Dinan par exemple).

4.3 Résultats après agrégation de modèles

La table 3 montre que l'agrégation des modèles binomiaux par forêt aléatoire permet d'améliorer les performances en prédiction en comparaison de l'approche basée sur

Pollen	Risque	AUC			Précision			Rappel		
		global	quantile		global	quantile		global	quantile	
			5 %	95 %		5 %	95 %		5 %	95 %
Ambroisie	**faible**	89.2	85.5	93.6	85.6	77.9	99.6	88.5	70.5	97.1
	moyen	89.4	83.7	94.7	86	77.8	98.1	89	67.4	95.8
	fort	89.7	85.3	92.1	88.7	77.2	98.8	91.9	66.3	98.7
Cuprés-sacées	**faible**	75.8	79.1	84.6	70.4	68	81.6	70	54.8	80.1
	moyen	78.7	74.5	86.3	72.6	70	81.1	71.7	55.5	81.1
	fort	83.5	78.5	89.6	80.8	72.1	94.7	83.9	70.9	93.4
Poacées	**faible**	84.7	78.3	89.6	77.6	74.6	82.6	76.6	67	81.6
	moyen	82.7	78.2	87.5	75.3	72.3	80.3	74.2	64.7	80.5
	fort	79	73.3	85.3	72.2	67.9	79.9	74.5	58.8	83.3

TAB. 2. Valeurs globales ("global") d'AUC, précision et rappel obtenues en considérant l'ensemble des sites et quantiles à 5 % et 95 % des valeurs par site.

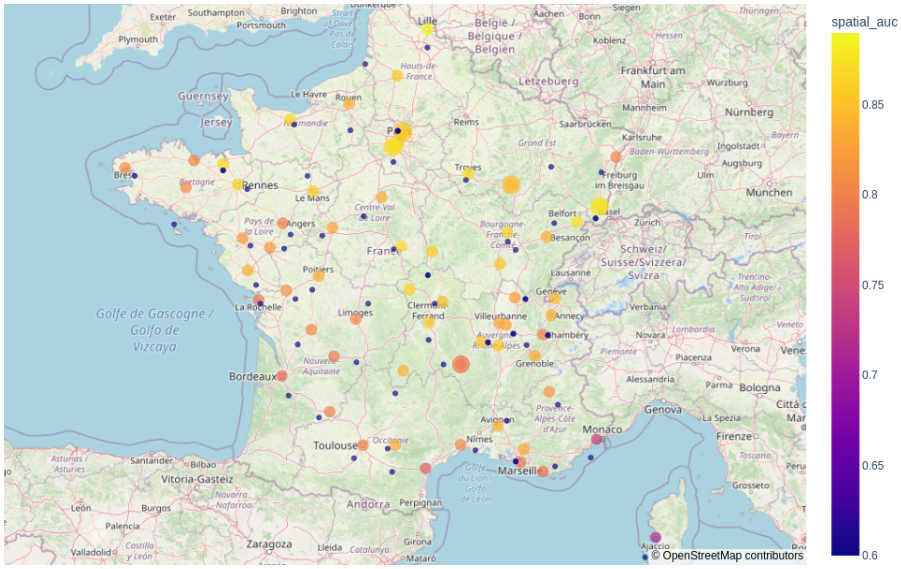

FIG. 2. Carte des AUC des prédictions du risque moyen des Poacees des données tests (points colorés suivant l'échelle $spaltial_auc$ et indiquant la position des capteurs HIRST). Les points bleus matérialisent la position des capteurs météos utilisés

des règles *ad hoc*. Nous montrons les résultats de deux sites en particulier mais des performances similaires sont constatées pour les autres sites.

La figure 3 montre les prédictions en fonction de la date. Les niveaux de concentrations en échelle log sont séparés par des lignes horizontales et les points matérialisent les différentes prédictions de risque (les prédictions du risque "nul" non affichées pour plus de lisibilité de la figure). Les bandes horizontales démarquent les zones de niveau de risque allant graduellement de "nul", "faible", "moyen", à "fort". La figure montre que les prédictions suivent bien les niveaux de concentrations. Les fausses alertes principa-

sites	méthode	précision	rappel	F1-score
Rennes	RD	0.71	0.59	0.62
	RF	0.72	0.66	0.68
Ajaccio	RD	0.75	0.56	0.61
	RF	0.79	0.68	0.72

TAB. 3. Comparaison du bagging par forêt aléatoire (RF) et des règles de décision ad hoc (RD) définies en 3

lement sur les risques faibles arrivent surtout en fin de saison pollinique. Ce phénomène est observé de façon générale.

4.4 Comparaison des modèles spatiaux et locaux

Une étude comparative entre modèle spatial et local est proposée sur les sites de Rennes et d'Ajaccio. Par comparaison des AUC obtenus pour chaque type de modèle, un classement est présenté dans le tableau 4. Dans 87.5 % des cas le modèle local est meilleur en AUC que le modèle spatial à Ajaccio contrairement au site de Rennes pour lequel 85.71 % des modèles spatiaux sont meilleurs que les modèles locaux. Ceci confirme les premiers résultats montrés sur la figure 2, le score du modèle spatial à Ajaccio correspond à l'un des plus mauvais scores.

sites	modèle	nb meilleur (auc)	taux meilleur (auc)
Ajaccio	Local	7	87.5 %
	Spatial	1	12.5 %
Rennes	Local	1	12.29 %
	Spatial	6	85.71 %

TAB. 4. Comparaison des modèles spatiaux et locaux sur Rennes et d'Ajaccio

Pour le cas des sites de Rennes de d'Ajaccio nous avons comparé la classification finale des poacées, en effectuant une combinaison par forêts aléatoires des modèles locaux, d'une part et des modèles spatiaux d'autre part (tableau comparatif 5 (voir tableau 5) . La classification par combinaison des modèles spatiaux permet d'obtenir de meilleurs scores avec environ 2 % d'écart.

sites	modèle	précision	rappel	F1-score
Rennes	spatial	0.74	0.68	0.70
	local	0.72	0.64	0.67
Ajaccio	spatial	0.79	0.68	0.72
	local	0.77	0.65	0.69

TAB. 5. Comparaison des classifications finales par combinaison des modèles (de régression logistique binaire) spatiaux et locaux

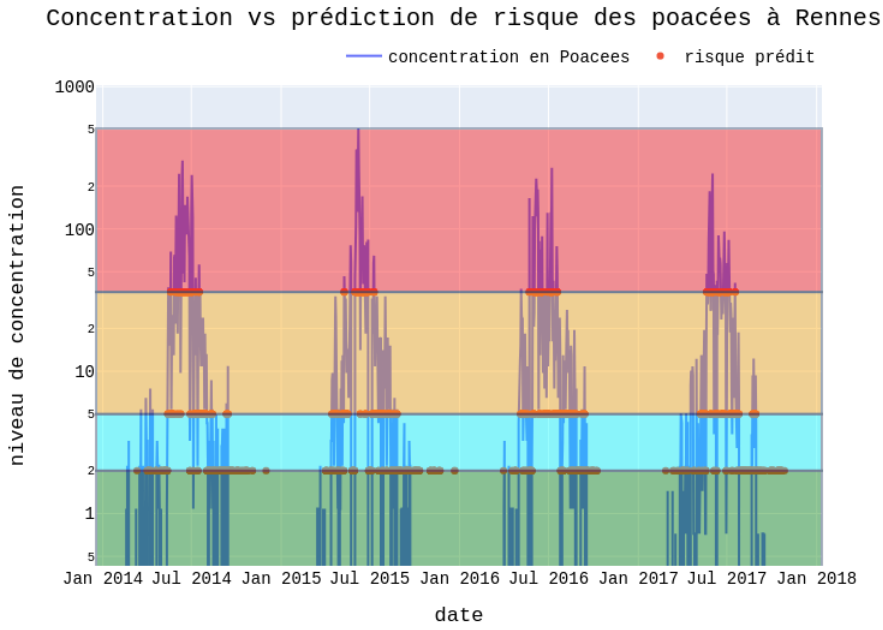

FIG. 3. Prédiction de risque des poacées et évolution de la concentration de l'échantillon test sur le site de Rennes : 2014-2017

La comparaison des modèles spatiaux et locaux a montré de façon générale que les écarts entre les AUC des deux types de modèles sont inférieurs à 10 %. Dans environ 70 % des cas, les modèles spatiaux conduisent à de meilleures performances. Ces derniers présentent l'avantage de réduire le nombre de modèles à entraîner d'un facteur 62 et de permettre l'extrapolation spatiale.

L'un des enjeux de cette étude est de développer des modèles capables de fournir une prédiction pour un site non observé. Pour évaluer la généralisation spatiale, six sites ont été retirés de l'échantillon d'entraînement. Les résultats du tableau 6 montrent les valeurs minimales des critères de performance de prédiction pour ces sites. L'AUC et la précision sont bonnes. Mais le rappel peut être faible comme pour les cupréssacées par exemple. Les résultats d'AUC sont confirmés sur la carte de la figure 2 où les points les plus gros représentent ces mêmes sites. Les prédictions des sites extrapolés montrent les mêmes niveau de performance que les sites voisins entraînés (voir par exemple Saclay). On note une exception pour Le-Puy-en-Velay (comparativement à Saint-Etienne ou Lyon) qui s'explique par des conditions météorologiques très locales. Dans ce type de situation, une substitution par un modèle local est effectuée.

Pollen	Risque	AUC min	Précision min	Rappel min
Ambroisie	**faible**	85.23	76.54	66.18
	moyen	87.59	80.46	73.24
	fort	81.84	87.61	85.51
Cupréssacees	**faible**	69.30	70.10	48.27
	moyen	72.60	72.85	52.28
	fort	84.48	76.75	63.49
Poacées	**faible**	80.01	75.78	76.14
	moyen	79.23	73.65	75.59
	fort	89.65	82.28	78.22

TAB. 6. Valeurs minimales d'AUC, précision et rappel pour l'ensemble des sites retirés de l'échantillon d'entraînement.

5 Conclusion

Nous avons présenté une étude de faisabilité concernant la prédiction de niveaux de risque d'émission de pollens à trois jours en partant de données aérobiologiques historiques, de données météorologiques et de géo-localisation. L'originalité de l'approche se situe dans la prise en compte d'informations spatiales et temporelles. Il ressort de notre étude qu'il est possible en général de prédire de manière acceptable les épisodes polliniques en tenant compte de niveaux de risques. Toutefois, nous avons montré que si les modèles "spatiaux" (entraînés sur l'ensemble des sites) surpassent en général les modèles "locaux" (entraînés site par site), l'agrégation de modèles spatiaux et locaux s'avère nécessaire pour atteindre de bons niveaux de prédiction. Les perspectives à court terme de cette étude concernent : i) modèles auto-régressifs en réinjectant les prévisions à J+1 dans le modèle J+2 et J+1, J+2 dans le modèle J+3, ii) le développement de méta modèles pour l'agrégation de modèles locaux et globaux, iii) l'évaluation de modèles à base d'approches neuronales, notamment constitués de réseaux convolutionnels spatio-temporels, iv) le couplage des données historiques HIRST et météorologiques avec des données "temps réel" issues de nouveaux capteurs optiques dédiés à la détection des émissions polliniques.

Références

Andersen, T. B. (1991). A model to predict the beginning of the pollen season. *Grana 30*(1), 269–275.

Baklanov, A., L. T. Molina, et M. Gauss (2016). Megacities, air quality and climate. *Atmospheric Environment 126*, 235–249.

Cassagne, E. (2009). Revue bibliographique des principaux seuils de détermination et méthodes de prévision de la date de début de pollinisation (ddp). *Revue Francaise d'Allergologie 49*(8), 571–576.

Castellano-Méndez, M., M. Aira, I. Iglesias, V. Jato, et W. González-Manteiga (2005). Artificial neural networks as a useful tool to predict the risk level of betula pollen in the air. *International Journal of Biometeorology 49*(5), 310–316.

Cordero, J. M., J. Rojo, A. M. Gutiérrez-Bustillo, A. Narros, et R. Borge (2021). Predicting the olea pollen concentration with a machine learning algorithm ensemble. *International Journal of Biometeorology 65*(4), 541–554.

Hidalgo, P. J., A. Mangin, C. Galán, O. Hembise, L. M. Vázquez, et O. Sanchez (2002). An automated system for surveying and forecasting olea pollen dispersion. *Aerobiologia 18*(1), 23–31.

Iglesias-Otero, M., M. Fernández-González, D. Rodríguez-Caride, G. Astray, J. Mejuto, et F. Rodríguez-Rajo (2015). A model to forecast the risk periods of plantago pollen allergy by using the ann methodology. *Aerobiologia 31*(2), 201–211.

Muzalyova, A., J. O. Brunner, C. Traidl-Hoffmann, et A. Damialis (2021). Forecasting betula and poaceae airborne pollen concentrations on a 3-hourly resolution in augsburg, germany : toward automatically generated, real-time predictions. *Aerobiologia*, 1–22.

Nowosad, J., A. Stach, I. Kasprzyk, K. Chłopek, K. Dąbrowska-Zapart, Ł. Grewling, M. Latałowa, A. Pędziszewska, B. Majkowska-Wojciechowska, D. Myszkowska, et al. (2018). Statistical techniques for modeling of corylus, alnus, and betula pollen concentration in the air. *Aerobiologia 34*(3), 301–313.

Puc, M. (2012). Artificial neural network model of the relationship between betula pollen and meteorological factors in szczecin (poland). *International journal of biometeorology 56*(2), 395–401.

Ranzi, A., P. Lauriola, V. Marletto, et F. Zinoni (2003). Forecasting airborne pollen concentrations : Development of local models. *Aerobiologia 19*(1), 39–45.

Sánchez-Mesa, J., C. Galán, J. Martínez-Heras, et C. Hervás-Martínez (2002). The use of a neural network to forecast daily grass pollen concentration in a mediterranean region : the southern part of the iberian peninsula. *Clinical & Experimental Allergy 32*(11), 1606–1612.

Spieksma, F. T. M., J. Emberlin, M. Hjelmroos, S. Jäger, et R. Leuschner (1995). Atmospheric birch (betula) pollen in europe : Trends and fluctuations in annual quantities and the starting dates of the seasons. *Grana 34*(1), 51–57.

Stirnberg, R., J. Cermak, S. Kotthaus, M. Haeffelin, H. Andersen, J. Fuchs, M. Kim, J.-E. Petit, et O. Favez (2021). Meteorology-driven variability of air pollution (pm_1) revealed with explainable machine learning. *Atmospheric Chemistry and Physics 21*(5), 3919–3948.

Thibaudon, M. (2003). The pollen-associated allergic risk in france. *European annals of allergy and clinical immunology 35*(5), 170–172.

Valencia, J., G. Astray, M. Fernández-González, M. Aira, et F. Rodríguez-Rajo (2019). Assessment of neural networks and time series analysis to forecast airborne parietaria pollen presence in the atlantic coastal regions. *International journal of biometeorology 63*(6), 735–745.

Zewdie, G. K., X. Liu, D. Wu, D. J. Lary, et E. Levetin (2019). Applying machine learning to forecast daily ambrosia pollen using environmental and nexrad parameters. *Environmental monitoring and assessment 191*(2), 1–11.

Annexe

Comparaison de la combinaison des modèles binomiaux et autres approches de classification

Les modèles sont notés comme suivant : Combinaison de modèles binaires (MetaModel), modèle de régression logistique multinominal (MNL) et ordinal (OL), forêt aléatoire (RF) , arbre de décision (DT).

pollen	DT	RF	MNL	OL	MetaModel
Ambroisie	89.52	90.89	86.00	90.73	91.81
Cupressacées	67.55	76.27	72.51	73.89	77.50
Poacées	77.66	80.57	78.32	81.33	80.91

Tab. 7. Comparaison des AUC en % des différents modèles spatiaux

Le tableau 7 montre que notre approche de combinaison de modèles régression logistique binaire à un léger avantage en performance des AUC sur les données test. Le modèle proposé est facilement interprétable : il nous permet d'observer aussi que les modèles de classification ont tendance à sélectionner différemment les variables importantes en fonction du type de pollen. La variable de saisonnalité (numéro du jour de l'année), les variables géographiques (région Nord Ouest et Est, Sud Ouest, longitude, latitude) et météorologique (degré jour, température) sont les plus importantes pour les classifications.

Summary

In the scientific literature, numerous studies show that meteorological conditions have an impact on the emission, dispersion and suspension of pollens in the air. Several allergenic species permanently threaten the health of millions of people in France. Preventive information on the risk of pollen exposure would become a real asset for allergy sufferers. The main objective of this article is to study, thanks to statistical learning techniques using historical data and meteorological parameters of the day (D), the ability to predict 3 days (D+3) in advance the risk levels of pollen presence in the air on a given territory (in Metropolitan France). We were interested in the prediction -in 4 levels- of risk for 3 families of pollens which are among the most allergenic species (Ambrosia, Cupressaceae and grasses). The aggregation of binary logistic regression models for each level of risk by a random forest classifier allowed us to predict the level of pollen risk with performances in the range of 75% to 90% of AUC and 70% of precision and recall, confusions concerning mainly low and medium levels.