Indice pollen

Note méthodologique de calcul des prévisions de comptage à l'échelle nationale

Fédération des associations de surveillance de la qualité de l'air
Atmo France

Version du 2 avril 2025

Rédaction : Mathis Pasquier – Atmo Hauts-de-France
Relecture : Michel Bobbia et Véronique Delmas (Atmo Normandie), Anthony Hezon (Atmo Auvergne-Rhône-Alpes), Jérôme Le Paih – ATMO Grand Est, Stéphanie Leplus (Atmo France)

Table des matières

1 Introduction

Les Associations agréées de surveillance de la qualité de l'air (AASQA) font partie des organismes en charge de la coordination de l'information pollinique selon l'arrêté ministériel du 5 août 2016. C'est la raison pour laquelle elles ont intégré dans leur mission de service public une information pollinique complémentaire à l'information sur la qualité de l'air, étant agréées pour cette dernière depuis la Loi sur l'air et l'utilisation rationnelle de l'énergie de décembre 1996. Ainsi, les AASQA ont décidé de mettre en place, de façon coordonnée avec Atmo France, à partir du 2 avril 2025, un service disponible gratuitement sur toute la France métropolitaine et en Corse. Ce service consiste en la fourniture quotidienne, y compris en open data, de prévisions d'un indice pollinique à l'échelle de la commune pour le jour même et les deux jours suivants.

Ce travail a été réalisé grâce à la collaboration de plusieurs experts des AASQA.

Le présent document a pour objet de présenter la méthode retenue pour l'élaboration de l'indice pollinique, les données utilisées, les limites de cette première version, les résultats obtenus, et les perspectives d'évolution à court terme. Il a été réalisé pour accompagner la diffusion de l'indice pollinique en avril 2025, et est à destination des autorités sanitaires et de tout public intéressé. Il sera réactualisé au fur et à mesure de la production des nouvelles versions du modèle.

2 Stratégie générale

On cherche à calculer des prédictions de comptages journaliers de pollens en grains/m^3 (définis comme la somme des comptages horaires mesurés sur une journée) en tout point du territoire de France métropolitaine, pour le jour-même (j), le lendemain ($j + 1$) et le surlendemain ($j + 2$).

On considère 6 taxons (espèces) à partir desquels est calculé un « indice pollen », en définissant des seuils d'exposition similairement à l'indice ATMO (indice quotidien d'information sur la qualité de l'air basé sur la prévision de présence de plusieurs polluants dans l'air[1]). Les taxons considérés sont les suivants :

- Aulne (ALNUS)
- Armoise (ARTEMISI)
- Ambroisie (AMBROSIA)
- Bouleau (BETULA)
- Graminées
- Olivier (OLEA)

La prévision de l'indice en tout point du territoire est réalisée en trois étapes :

1. Le calcul de prévisions de comptages aux stations de mesure, où l'on dispose de prévisions météorologiques
2. La spatialisation des prévisions, permettant d'obtenir des prévisions sur tout le territoire
3. Le calcul d'un indice pollinique agrégeant les prévisions sur les 6 taxons

Lors de la première étape, on construit pour chaque taxon un modèle statistique de prévision avec une approche d'apprentissage automatique (ou « *machine learning* » en anglais). La seconde étape

[1] https://www.atmo-france.org/sites/federation/files/medias/documents/2022-04/guide_calcul_nouvel_indice_ATMO_VF_version14decembre2020.pdf

est réalisée par krigeage des prévisions aux stations pour corriger les cartes nationales produites par Copernicus (voir section 4 plus bas). Dans la troisième étape on transforme les résultats obtenus pour chaque taxon selon une échelle adaptée en 6 classes et on agrège les résultats obtenus pour chacun des 6 taxons dans l'indice pollinique communal.

L'ensemble du process a nécessité au préalable la collecte et la mise en forme de nombreuses données, afin de permettre la constitution d'une base conséquente de données expertisées et validées, matière première à l'élaboration des modèles statistiques.

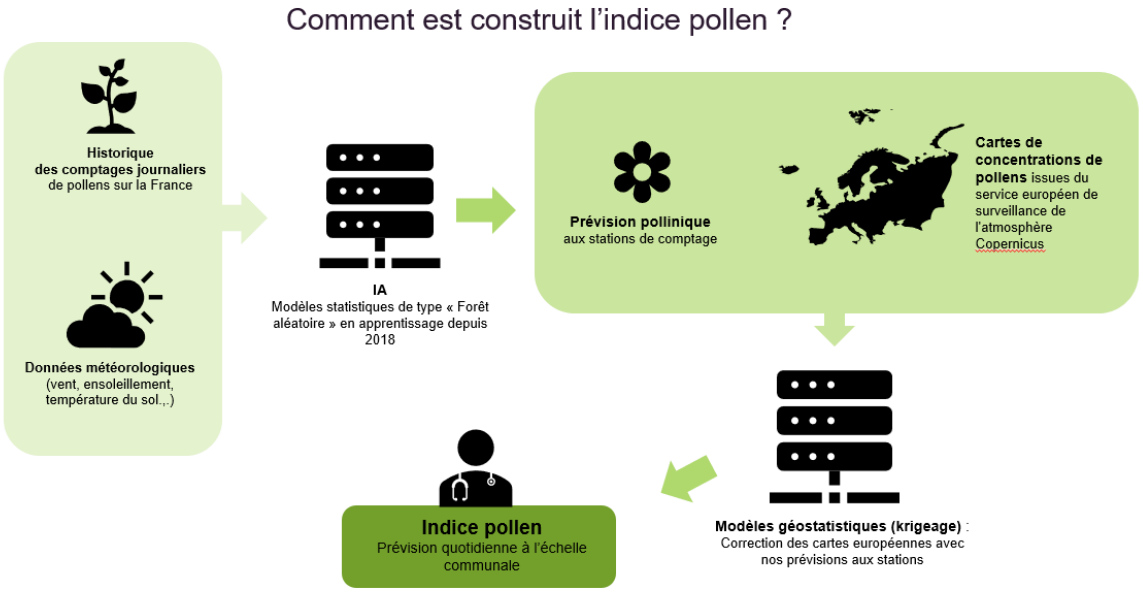

FIGURE 1 – METHODOLOGIE D'ELABORATION DES PREVISIONS D'INDICES POLLINIQUES

3 Calcul des prévisions aux stations

1. Stratégie d'apprentissage automatique

La construction de modèles statistiques pour la prévision des comptages aux stations passe par une phase d'entraînement, aussi dénommée « apprentissage », qui consiste à construire des modèles statistiques capables de capturer le lien entre des « variables explicatives » (ici la date et la météo) et une « variable expliquée » (ici la valeur de comptage de grains par m^3 pour un taxon, une date et une station donnée).

La phase d'apprentissage est menée en utilisant des observations de la variable expliquée (les mesures de comptages aux stations) que l'on croise avec des observations des variables explicatives (les données météo notamment). Les données de mesure utilisées couvrent 56 stations de mesure sur une période étalée du 1er janvier 2015 au 31 décembre 2023. La figure 2 illustre la distribution géographique des 56 stations impliquées dans la phase d'apprentissage des modèles statistiques. 16 de ces stations font partie du parc directement géré par les Associations agréées de surveillance de la qualité de l'air (AASQA), les autres appartiennent à celui administré par le Réseau Nationale de Surveillance Aérobiologique (RNSA). Pour ce dernier réseau, compte tenu de la difficulté d'accès aux données collectées par le RNSA, les séries de mesure sont généralement discontinues. Le nombre total d'observations recueillies pour chaque taxon sur les 16 stations gérées par les AASQA est donné dans le tableau 1.

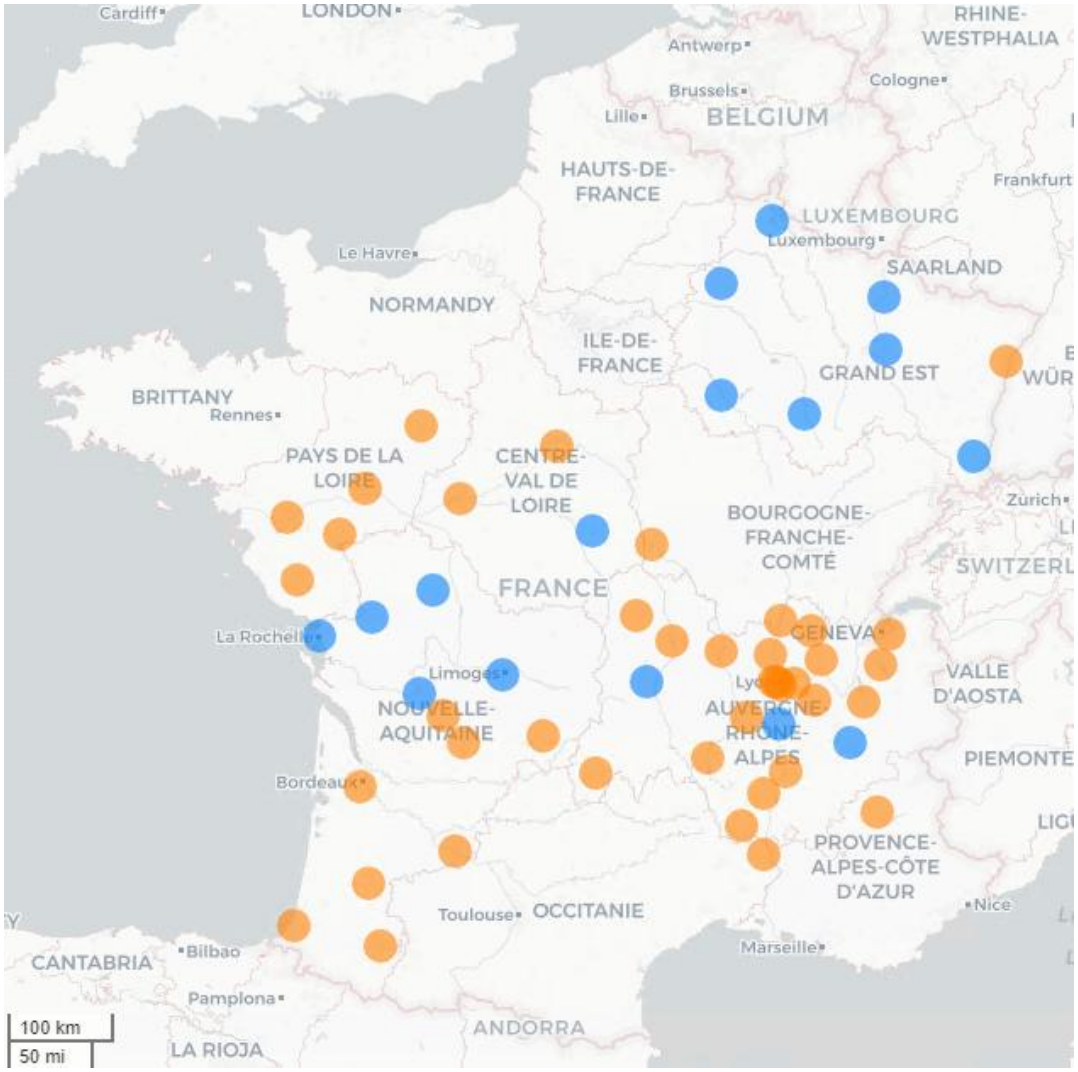

FIGURE 2 – DISTRIBUTION GEOGRAPHIQUE DES 56 STATIONS DE MESURE MOBILISEES POUR L'APPRENTISSAGE DES MODELES STATISTIQUES SUR LES COMPTAGES RELEVES ENTRE 2015 ET 2023 : EN BLEU STATIONS ADMINISTREES PAR LES AASQA, EN ORANGE STATIONS ADMINISTREES PAR LE RNSA. LES DONNEES DES STATIONS DES AASQA RESTENT EN EFFECTIF MAJORITAIRE, LES SERIES DE DONNEES DU RNSA ETANT PARCELLAIRES CAR DIFFICILEMENT ACCESSIBLES.

taxon	ALNUS	ARTEMISI	AMBROSIA	BETULA	GRAMINEE	OLEA
N_{obs}	22 204	18 181	17916	21 255	24 853	9 047

TABLE 1 – NOMBRE D'OBSERVATIONS DISPONIBLES ENTRE 2015 ET 2023 SUR LES SEULES 16 STATIONS ADMINISTREES PAR LES AASQA MOBILISEES POUR L'APPRENTISSAGE DES MODELES STATISTIQUES, POUR CHAQUE TAXON CONSIDERE.

2. Variables explicatives identifiées

Deux types de variables explicatives sont utilisées pour les apprentissages des modèles statistiques : des données météorologiques (en grande majorité) et des variables calendaires.

Les variables météorologiques considérées en tant que variables explicatives sont des données issues des réanalyses ERA5[2] pour différentes variables physiques pertinentes (la température du sol et de l'air, l'ensoleillement, la vitesse du vent, la nébulosité, etc.). Les variables explicatives utilisées sont des minima/maxima journaliers et moyennes journalières de ces variables physiques, ainsi que leurs sommes cumulées sur chaque année.

[2] https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5

Pour l'entraînement d'un modèle statistique dédié à la prévision des comptages d'un taxon donné, une sélection préalable des nombreuses variables listées ci-dessus est conduite en utilisant une méthode de sélection sur critère d'Akaike[3]. Cela permet de réduire significativement la quantité de variables utilisées et donc de limiter le temps de calcul nécessaire pour l'entraînement des modèles statistiques.

En plus des variables météorologiques, on emploie deux variables calendaires, correspondant au rang dans l'année du jour de la date pour laquelle on veut calculer une prévision, ainsi qu'au mois de cette date.

3. Modèles statistiques entraînés

On utilise des modèles d'apprentissage automatique classiques, dénommés « forêts aléatoires[4] » (ou *random forests* en anglais, d'où la dénomination « modèle RF » dans la suite de ce document).

L'utilisation de ce type de modèle requiert l'ajustement de paramètres internes dénommés « hyperparamètres ». L'ajustement de ces hyperparamètres permet d'améliorer la performance des modèles entraînés, mais nécessite un temps de calcul important pour trouver la meilleure combinaison de valeurs, et ce d'autant plus qu'il y a d'hyperparamètres à ajuster. Dans notre cas, cet ajustement est réalisé en explorant une grille de valeurs possibles pour les hyperparamètres de façon exhaustive, et en minimisant des scores statistiques en validation croisée comme la racine de l'erreur quadratique moyenne (RMSE) : on parle de stratégie *grid search* en validation croisée.

4. Données de prévisions météorologiques utilisées en phase de production

La « phase de production » arrive après la phase d'apprentissage pendant laquelle les modèles de prévision ont été construits. Dans cette seconde phase, les modèles calculent quotidiennement les prévisions journalières de comptages aux stations métropolitaines.

Les modèles RF ont été entraînés en utilisant les données météorologiques issues des réanalyses ERA5 en tant que variables explicatives, mais par définition ces données ne sont pas disponibles pour réaliser des prévisions en routine. En effet il ne s'agit pas de *prévisions* météorologiques mais de mesures assimilées, elles ne sont donc pas disponibles au jour le jour pour les échéances j à $j + 2$. Ainsi, la production des prévisions quotidiennes devrait idéalement utiliser des données de prévision météorologiques issues de la même source que les données ERA5, à savoir du centre européen pour les prévisions météorologiques à moyen terme (ECMWF[5]). Or nous ne disposons pas pour le moment de ces données de prévisions, mais seulement des données issues du modèle américain GFS[6]. Dans un premier temps, les modèles RF de production n'utilisent donc pas le même type de données que celles avec lesquelles ils ont été entraînés.

En comparant les données ERA5 et les données GFS, on a observé des biais systématiques pour certaines des variables physiques considérées comme des variables explicatives nécessaires au calcul des comptages avec les modèles RF. La figure 3 illustre des exemples de variables météorologiques explicatives pour lesquelles on observe des différences significatives d'une source de données à l'autre. Ces différences se traduisent la plupart du temps sous la forme d'un biais tout en conservant une bonne corrélation linéaire,

[3] https://fr.wikipedia.org/wiki/Critère_d'information_d'Akaike
[4] https://www.ibm.com/fr-fr/topics/random-forest
[5] https://www.ecmwf.int/
[6] https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast

mais pour certaines variables (voir figure 3) la corrélation est quasiment inexistante.

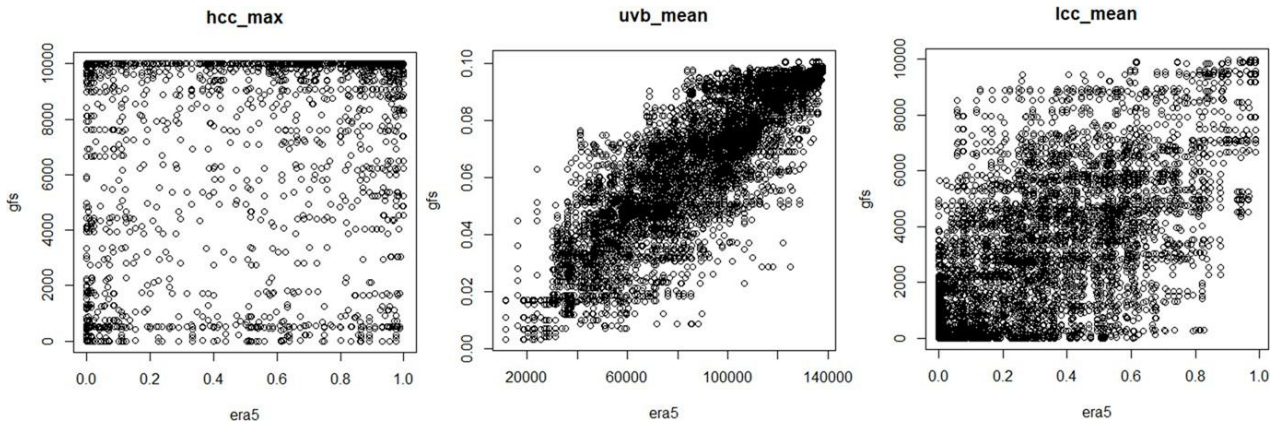

FIGURE 3 – EXEMPLES DE VARIABLES METEOROLOGIQUES EXPLICATIVES POUR LESQUELLES ON CONSTATE DES DIFFERENCES SIGNIFICATIVES ENTRE LES DONNEES ERA5 ET LES DONNEES GFS : NEBULOSITE EN HAUTE TROPOSPHERE MAXIMALE JOURNALIERE (GAUCHE, ABSENCE DE CORRELATION ET BIAIS SIGNIFICATIF), RADIATION SOLAIRE DESCENDANTE CAPTEE A LA SURFACE MOYENNE JOURNALIERE (CENTRE, BIAIS SIGNIFICATIF) ET NEBULOSITE EN BASSE TROPOSPHERE MOYENNE JOURNALIERE (DROITE, BIAIS SIGNIFICATIF).

Ainsi, afin de minimiser le risque de créer des disparités significatives entre les prévisions calculées en production (avec les données GFS) et les prévisions calculées lors de l'apprentissage en validation croisée (avec les données ERA5), la correction de biais suivante est proposée pour corriger les données GFS utilisées en production. Si on considère une variable météorologique explicative quelconque non nulle pour laquelle on dispose de N_{met} jours de données (par exemple une année entière), en notant :

- $x = \{x_1, \dots, x_{N_{met}}\}$ un échantillon de données de prévisions GFS pour cette variable
- $\tilde{x} = \{\tilde{x}_1, \dots, \tilde{x}_{N_{met}}\}$ un échantillon de données de prévisions ERA5 de référence pour cette même variable aux mêmes dates

alors on calcule un échantillon de données GFS corrigé, noté $\hat{x} = \{\hat{x}_1, \dots, \hat{x}_{N_{met}}\}$, de la façon suivante :

$$\forall i \in \{1, \dots, N_{met}\}, \qquad \hat{x}_i = \frac{\mu(\tilde{x})}{\mu(x)} \times x_i$$

avec $\mu(x)$ définie comme la moyenne de l'échantillon x. Cette opération garantit que le biais entre les deux échantillons $\hat{x}$ et $\tilde{x}$ est nul. Cela ne corrige pas les défauts de corrélation linéaire entre les jeux de données mais ramène les données GFS au même ordre de grandeur que les données ERA5.

Pour nos calculs de prévisions quotidiennes, les coefficients de correction ont été calculés en utilisant les données ERA5 et GFS sur l'année 2023. La figure 4 et le tableau 2 illustrent un exemple de prévisions de comptages de graminées obtenues pour la saison pollinique de 2023 en corrigeant les données météorologiques GFS et en utilisant un modèle RF entraîné sur la période 2015-2022.
On constate une nette amélioration après correction, même si les prévisions sont de moins bonne qualité qu'en utilisant les données ERA5.

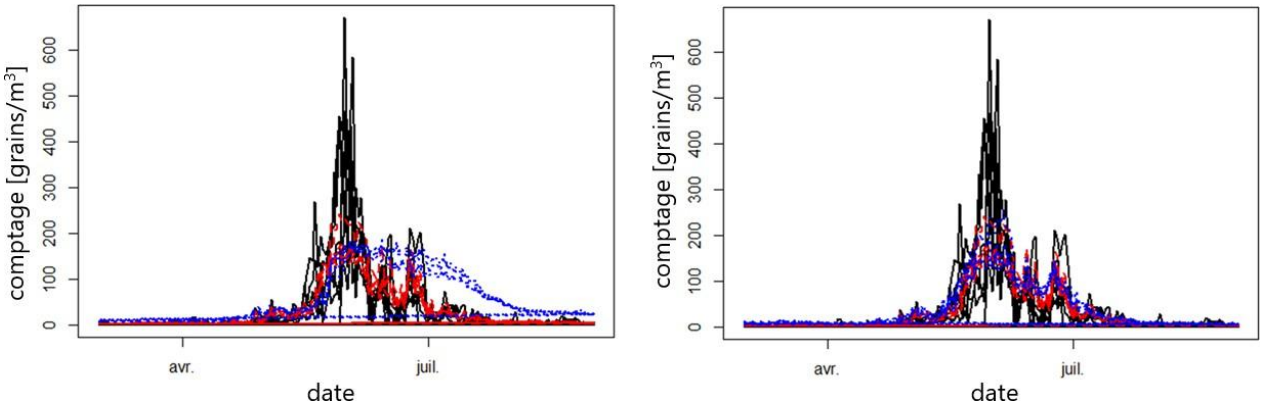

FIGURE 4 – PREVISIONS DE COMPTAGES POUR LES GRAMINEES POUR LA SAISON POLLINIQUE DE 2023 OBTENUES AVEC LE MODELE RF DE PRODUCTION ENTRAINE SUR 2015-2022 EN UTILISANT LES DONNEES METEOROLOGIQUES ERA5 (COURBES ROUGES) ET GFS (COURBES BLEUES), SANS CORRECTION DE BIAIS (GAUCHE) ET APRES CORRECTION (DROITE). LES COURBES NOIRES REPRESENTENT LES MESURES DE COMPTAGES AUX 56 STATIONS CONSIDEREES.

données météorologiques	biais	RMSE	corrélation
ERA5	−0.16	43.37	0.79
GFS	−25.19	61.96	0.59
GFS corrigées	−6.05	45.33	0.75

TABLE 2 – SCORES STATISTIQUES POUR LES PREVISIONS 2023 DONNEES SUR LA FIGURE 4, EN UTILISANT LES DONNEES METEOROLOGIQUES ERA5, GFS ET GFS CORRIGEES.

4 Spatialisation des prévisions utilisant Copernicus

Le Service Copernicus de surveillance de l'atmosphère (CAMS, atmosphere.copernicus.eu/) est un service environnemental atmosphérique mondial et régional, exploité par le Centre européen pour les prévisions météorologiques à moyen terme (CEPMMT) dans le cadre du programme Copernicus d'observation de la Terre de l'Union européenne. Le service régional CAMS produit quotidiennement des prévisions, des analyses et des réanalyses de la qualité de l'air en Europe. Ce service s'appuie sur une modélisation distribuée produite par onze équipes dans dix pays européens. Chaque modèle produit quotidiennement des analyses sur 24 h pour la veille et des prévisions sur 97 h pour 19 espèces chimiques et six espèces de pollen. Les onze modèles individuels sont ensuite combinés pour former une médiane ENSEMBLE, exploitée par INTERpollens tel que décrit ci-dessous.

Pour chaque taxon, et pour chaque échéance de prévision ($j+0$, $j+1$ et $j+2$), la différence entre les prévisions et le modèle d'ensemble de Copernicus est interpolée sur l'ensemble du territoire avec une méthode de krigeage (interpolation spatiale). La carte des différences ainsi obtenue est ajoutée à la carte du modèle Copernicus, permettant de corriger ce modèle par nos prévisions. Le résultat respecte les valeurs prévues aux sites retenus – on parle de « krigeage en dérive externe ».

La carte est alors projetée sur une grille plus fine (mailles kilométriques) afin de pouvoir être croisée avec les contours des communes.

5 Calcul de l'indice pollen

Les modèles statistiques sont destinés à fournir des prévisions de comptages (en grains/m^3) pour les 6 taxons mentionnés plus haut, et ces comptages seront utilisés pour calculer un « indice pollen » mesurant le risque d'exposition aux pollens dans l'air ambiant pour les populations. Pour plus de détails sur cet

indice, on renvoie le lecteur à la note dédiée sur le site d'Atmo France ou par mail à l'adresse pollens@atmo-france.org.

L'indice pollen est défini de façon similaire à l'indice ATMO et varie entre 1 (exposition faible) et 6 (exposition extrêmement élevée), voir figure 5. Il est défini comme le maximum des sous-indices calculés pour chacun des 6 taxons impliqués, les sous-indices étant eux-mêmes définis par rapport à des seuils propres à chaque taxon. Pour chaque classe de concentration, la borne inférieure est incluse dans le calcul de l'indice.

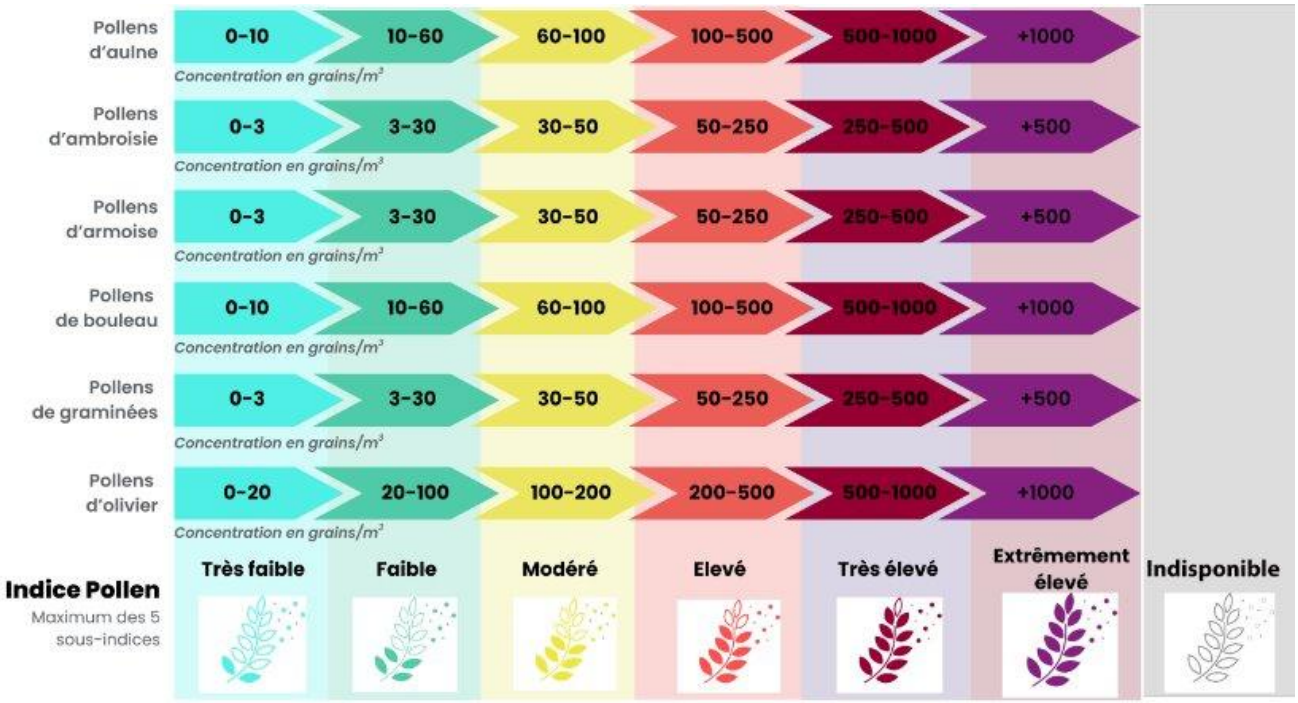

FIGURE 5 – SEUILS DEFINISSANT LES SOUS-INDICES POLLEN POUR CHAQUE TAXON CONSIDERE DANS LE CALCUL DE L'INDICE POLLEN TOTAL. LES COULEURS INDIQUENT LES DIFFERENTES VALEURS D'INDICE CORRESPONDANT AUX NIVEAUX D'EXPOSITION. DE GAUCHE A DROITE : 1 (TRES FAIBLE), 2 (FAIBLE), 3 (MODERE), 4 (ELEVE), 5 (TRES ELEVE) ET 6 (EXTREMEMENT ELEVE)

6 Limites

Plusieurs limitations principales sont à garder à l'esprit concernant l'entraînement des modèles statistiques et le calcul des prévisions de comptages en phase de production.

Une première limitation est liée à l'absence d'une base de données publique permettant l'accès aux données historiques des comptages de grain de pollens coordonnées par le RNSA. Cette situation a conduit les AASQA à créer une base de données qui s'avère parcellaire sur les stations directement gérées par le RNSA. Il s'agit d'une première limite dans la mesure où les résultats des modèles statistiques s'avèrent généralement plus précis lorsqu'ils sont créés à partir de séries de données continues. En outre, avoir accès à la base de données du RNSA permettrait une meilleure couverture du territoire.

Ensuite, le choix de modèles de type « forêt aléatoire » a été fait de façon à construire rapidement des modèles de prévision fonctionnels, et ce sans mobiliser trop de ressources de calcul. En effet, les modèles de type « forêt aléatoire » disposent de peu d'hyperparamètres à ajuster en validation croisée contrairement aux modèles comme l' « amplification de gradient » (*gradient boosting* avec la librairie R

XGboost). Cela constitue un avantage (les modèles sont plus robustes et facile à entraîner) mais aussi une limite (on perd en précision).

Une autre limitation des prévisions calculées est liée à l'emploi de données de prévisions météorologiques issues du modèle américain GFS alors que les modèles ont été entraînés sur des données issues du modèle ECMWF – cf section 2.1. Même avec une correction empirique permettant d'ajuster les données météorologiques GFS pour les rendre exploitables par les modèles (voir section 3.4 plus haut), cela n'exploite pas pleinement le potentiel des modèles entraînés, et des erreurs peuvent se propager et endommager les prévisions.

Ensuite, il faut noter que les modèles de prévision n'ont pas pu être entraînés sur le maximum de données de comptages disponibles. Cela est dû au fait que pour certaines stations et certaines régions, les historiques de mesures de comptages sur les années passées ont été renseignés a posteriori dans la base de données, sans récupérer conjointement les données météo ERA5 associées. Cela a donc empêché d'utiliser ces historiques de mesures pour entraîner les modèles statistiques de production, mais une récupération ultérieure des données météo ERA5 manquantes permettra un nouvel apprentissage plus précis.

Enfin, il faut mentionner que la méthodologie présentée dans cette note ne permet pas de calculer de prévisions de comptages pour les départements et régions d'outre-mer (DROM) : cela est principalement lié à l'absence de cartes de prévisions à méso-échelle comme celles fournies par Copernicus, qui sont à la base de notre méthodologie car corrigées localement par krigeage en utilisant les prévisions calculées aux stations de comptage. En outre, la flore diffère significativement entre la France métropolitaine et les DROM, ce qui nécessiterait d'entraîner des modèles statistiques spécifiques pour capturer les différentes dynamiques de pollinisation.

7 Evaluation de la performance des prévisions

La performance de notre stratégie de construction des modèles statistiques est évaluée en « validation croisée ». Cette approche consiste à construire des modèles statistiques en les entraînant sur une partie des données d'apprentissage, et en les testant sur le reste des données non utilisées. Ici on utilise une approche de type *leave-one-out* sur les années, consistant à entraîner des modèles sur toutes les années sauf une à chaque fois, l'année restante servant pour les tests.

En pratique, on construit un premier modèle en utilisant les données de comptage et les données météorologiques pour les années 2016 à 2023, et on calcule des prévisions pour l'année 2015, que l'on peut comparer aux données de comptages non prises en compte pour l'apprentissage. On répète cette opération en entraînant cette fois un modèle sur les données de l'année 2015 et des années 2017 à 2023 et en l'évaluant sur l'année 2016, et ainsi de suite.

Formellement, si on dispose de N années $a_1, \ldots, a_N$ de données de comptages, on construit N modèles $M_1, \ldots, M_N$ tels que pour tout i entre 1 et N, M_i est entraîné sur $a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_{N-1}, a_N$.

Ensuite, pour tout i entre 1 et N, le modèle M_i est utilisé pour calculer des prévisions sur l'année a_i, que l'on peut comparer aux données de comptage.
Cette analyse en validation croisée permet de générer des prévisions sur la totalité de la période courant

de 2015 à 2023, en générant pour chaque année i des prévisions avec le modèle M_i. Ensuite, des scores statistiques et des mesures de taux de détection de niveaux de comptages élevés peuvent être réalisés. La figure 6 illustre par exemple les prévisions calculées pour les graminées avec des modèles RF en validation croisée entre 2015 et 2023, comparées aux comptages correspondants.

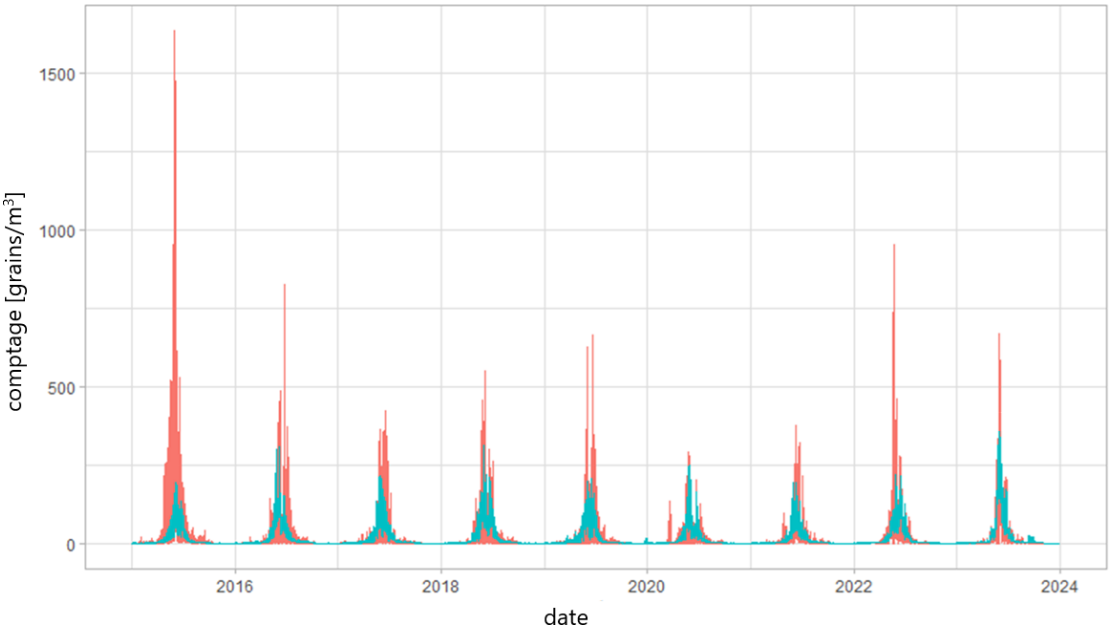

FIGURE 6 – PREVISIONS DES COMPTAGES DE GRAMINEES CALCULES AVEC DES MODELES RF EN VALIDATION CROISEE SUR LES ANNEES 2015 A 2023 SUR L'ENSEMBLE DES 56 STATIONS DE MESURE (COURBES BLEUES SUPERPOSEES), COMPAREES AUX COMPTAGES RELEVES AUX STATIONS (COURBES ORANGES).

Afin d'analyser la plus-value par rapport à l'existant, on compare les mesures avec les prévisions calculées en validation croisée et les prévisions correspondantes fournies par Copernicus pour les échéances j. Les scores statistiques sont donnés dans le tableau 3.

Les résultats conduisent aux observations suivantes :

1. Les prévisions de Copernicus tendent le plus souvent à sous-estimer les mesures (biais positif) alors que le modèle RF les surestime (biais négatif)
2. Certains taxons sont plus difficiles à prévoir avec précision que d'autres. Par exemple les comptages de graminées semblent mieux prédits par les modèles (corrélation supérieure à 60% et à l'inverse l'armoise semble moins bien prédite (corrélation inférieure à 40%). Cela peut s'expliquer par la quantité de données disponible, qui est variable selon le taxon considéré (on a beaucoup plus de données pour les graminées que pour les autres taxons, cf tableau 1).
3. Le modèle RF est sensiblement plus performant globalement que Copernicus sur cette analyse en validation croisée. En effet :
   - La RMSE est plus faible pour tous les taxons sauf l'aulne où elle est légèrement supérieure mais reste comparable
   - Le biais absolu est plus faible sauf pour l'olivier (très proche) et l'ambroisie (augmentation significative)
   - On observe un gain de corrélation de 10 points pour l'aulne, l'armoise et le bouleau, et des corrélations comparables pour les autres taxons (légère baisse pour l'ambroisie bien que toujours supérieure à 50%).

taxon	score	Copernicus	modèle RF

ALNUS	biais	8.82	−2.91
	corr	0.29	0.42
	RMSE	38.68	41.39
AMBROSIA	biais	0.08	−1.86
	corr	0.58	0.54
	RMSE	13.59	11.79
ARTEMISI	biais	−0.39	−0.32
	corr	0.25	0.33
	RMSE	3.79	2
BETULA	biais	7.7	−3.97
	corr	0.38	0.55
	RMSE	63.75	56.11
GRAMINEE	biais	6.05	−4.06
	corr	0.66	0.68
	RMSE	36.26	31.68
OLEA	biais	0.21	−0.24
	corr	0.31	0.36
	RMSE	1.31	1.28

TABLE 3 – SCORES CALCULES POUR LES PREVISIONS COPERNICUS ET LES PREVISIONS DU MODELE RF POUR CHAQUE TAXON. LES SCORES SONT CALCULES SUR LA PERIODE 2018-2023 CAR LES PREVISIONS COPERNICUS NE SONT DISPONIBLES DANS LA BASE DE DONNEES INTERPOLLENS QUE DEPUIS 2018.

Le constat selon lequel les prévisions Copernicus tendent à sous-estimer les mesures de comptages et les prévisions des modèles RF à les surestimer peut être en outre établi en calculant les taux de bonne détection, de sous-estimation et de surestimation pour des seuils donnés pour les 6 taxons. Le tableau 4 présente des mesures de taux de bonne prévision/surestimation/sous-estimation pour chaque taxon en utilisant à chaque fois le seuil défini par le passage à un indice pollen égal à 4 (exposition forte).

taxon	seuil exposition élevée [grains/m^3]	taux de bonnes détections [%]	taux de surestimation [%]	taux de sous-estimation [%]
ALNUS	100	*1.2* / **18.7**	*0.6* / **58.2**	*98.3* / **23**
AMBROSIA	50	*23.6* / **22.7**	*47.8* / **47.6**	*28.6* / **29.6**
ARTEMISI	50	*0* / **0**	*0* / **0**	*100* / **100**
BETULA	100	*16* / **34.7**	*14.7* / **46.7**	*69.3* / **18.6**
GRAMINEE	50	*43.8* / **44.8**	*20.8* / **46.8**	*35.4* / **8.4**
OLEA	200	*0* / **0**	*20* / **15.6**	*80* / **84.4**

TABLE 4 – TAUX DE BONNE DETECTION, DE SURESTIMATION ET DE SOUS-ESTIMATION DES DEPASSEMENTS DE SEUIL D'EXPOSITION FORTE POUR CHAQUE TAXON, POUR LES PREVISIONS DE COPERNICUS (EN ITALIQUE) ET DES MODELES RF (EN GRAS). LES PERFORMANCES SONT CALCULEES SUR LA PERIODE 2018-2023.

On retrouve bien les tendances à la surestimation des modèles RF et la tendance à la sous-estimation de Copernicus, notamment pour l'aulne, le bouleau et les graminées. On ne constate pas d'amélioration pour les autres taxons en termes de détection de dépassement du seuil d'exposition élevée.
La mesure de surestimation des modèles RF s'explique par le fait que les modèles ont tendance à ne pas

percevoir les valeurs extrêmes des comptages (cf figure 6 pour les graminées) lorsqu'ils surviennent, mais ont tendance à détecter des comptages non nuls en dehors des saisons polliniques. Cette observation a conduit dans la phase de production à forcer le modèle à zéro en dehors des périodes de pollinisation, dans l'objectif de limiter la surestimation des comptages. Globalement, on observe que les modèles de prévision capturent correctement le profil annuel des comptages – c'est-à-dire leur tendance, les dates de début et de fin des périodes de pollinisation, le nombre de pics observés, etc.

En résumé, on constate que les scores obtenus sont les meilleurs pour les graminées (taxon pour lequel on a le plus de données) et les moins bons pour l'olivier et l'armoise pour lesquels le nombre de données est plus faible (de l'ordre de 10 000 pour l'olivier contre environ 25 000 pour les graminées par exemple).

8 Conclusion et perspectives

L'approche de prévision mise en œuvre par les AASQA et Atmo France, à savoir le couplage de deux méthodes statistiques (forêts aléatoires puis krigeage) permet de redresser les cartes produites par COPERNICUS par la prise en compte des données de comptage historiques des pollens. Cette approche intègre une réévaluation régulière des modèles, qu'ils soient statistiques ou déterministes (Copernicus), afin de privilégier pour un taxon donné le meilleur modèle disponible, en laissant la possibilité d'utiliser directement un des modèles de Copernicus s'il s'avère plus précis que le modèle RF/krigeage pour un taxon donné.

Parmi les résultats du projet, il faut citer la constitution d'une base de données à l'échelle de la France métropolitaine et de la Corse, base de données qu'il a fallu constituer entièrement, puisque les AASQA n'ont pas eu accès à la base de données du RNSA.
Un accès à la base du RNSA permettrait de compléter les séries de données disponibles et utilisées pour le projet, et contribuerait à améliorer les performances des modèles et faciliter leur évaluation.

Une fois que les données de mesure des stations de comptages sur 2024 seront disponibles sur l'ensemble du territoire, les modèles statistiques pourront être réentraînés sur davantage d'historique de mesures, en complétant l'ajustement des hyperparamètres des modèles RF, voire d'amplification de gradient. En outre, l'identification de nouvelles variables explicatives pourra être envisagée afin d'améliorer la précision des modèles et de permettre de mieux prévoir les dates de début et de fin des saisons polliniques.

En date de la fin du mois de mars 2025, des démarches ont été engagées pour obtenir les données de prévision météorologiques issues de l'ECMWF, conformes aux données d'entraînement des modèles, via la signature d'une convention avec Météo-France. L'utilisation de ces données permettra de remplacer les données du modèle américain GFS et d'améliorer la précision des prévisions des modèles entraînés sur des données de réanalyses ERA5 issues des modèles européens.

Dans une optique d'amélioration continue, une nouvelle évaluation de modèles de prévision entraînés entre 2015 et 2023 sera réalisée afin de contrôler quantitativement la fiabilité des prévisions. Cette évaluation a vocation à être effectuée a minima chaque année.

Enfin, les 6 taxons présentés ici et pris en compte dans le calcul de l'indice pollen couvrent les essences dont les pollens sont les plus répandus et allergisants en France métropolitaine. L'indice pollen sera progressivement enrichi avec l'intégration d'autres taxons comme le noisetier et le cyprès grâce aux données élargies de la plateforme Copernicus.