

Végéphyll – 12^e CONFÉRENCE INTERNATIONALE SUR LES MALADIES DES PLANTES
TOURS – 11 ET 12 DÉCEMBRE 2018

UNE DEMARCHE GENERIQUE POUR LE DEVELOPPEMENT D'OUTILS D'ANALYSE ET DE
PREDICTION DES DYNAMIQUES EPIDEMIQUES A PARTIR DES DONNEES DES RESEAUX
D'EPIDEMIOSURVEILLANCE DES CULTURES

S. VAN DE KERCKHOVE⁽¹⁾, F. BRUN⁽¹⁾, M. CHEN^(1,4), M. RAYNAL⁽²⁾, C. DEBORD⁽²⁾, F. MAUPAS⁽³⁾,
I. ALEKSOVSKA⁽¹⁾, D. MAKOWSKI⁽⁴⁾.

⁽¹⁾ Acta – les instituts techniques agricoles, Toulouse-Paris, France, francois.brun@acta.asso.fr

⁽²⁾ IFV – Institut Français de la vigne, Blanquefort, France

⁽³⁾ ITB – Institut Technique de la Betterave, Paris, France

⁽⁴⁾ INRA, Paris, France

RÉSUMÉ

Depuis 2010, le Bulletin de Santé du Végétal publie chaque semaine une information sur l'évolution de l'état sanitaire des cultures (plan ECOPHYTO). Il s'appuie sur des réseaux d'observations de parcelles réparties sur le territoire français. Notre approche s'appuie sur l'utilisation de modèles linéaires généralisés à effets mixtes pour analyser et prédire l'état sanitaire. En rendant cette approche générique et opérationnelle, nous cherchons à mettre en œuvre notre méthode sur un maximum de maladies ou ravageurs de différentes cultures. Une démarche de criblage est proposée. Pour chaque maladie ou ravageur, différents modèles sont comparés et en fonction des performances d'analyse et de prédiction, la maladie ou le ravageur sera retenu(e) ou non pour la suite dans les outils opérationnels. Enfin, ces travaux permettent de proposer aux acteurs du BSV des outils opérationnels d'analyse et de prédiction des dynamiques épidémiques et de population basés sur des modèles statistiques dont la qualité de prédiction est mesurable. Pour en faire la démonstration, nous avons aussi construit une application web pour la vigne "Vine Health Watch" basé sur ces approches et développé lors du hackathon API-AGRO 2018.

Mots-clés : épidémiosurveillance, modélisation, Bulletin de Santé du Végétal, data science, modèle mixte.

ABSTRACT

A GENERIC APPROACH FOR THE DEVELOPMENT OF TOOLS TO ANALYZE AND PREDICT EPIDEMIC DYNAMICS USING DATA FROM PLANT HEALTH MONITORING NETWORKS

Since 2010, the Plant Health Bulletin ("Bulletin de Santé du Végétal" in French) publishes weekly information on the evolution of the health status of crops (French ECOPHYTO plan). It relies on observation networks of parcels spread over the French territory. Our approach is based on the use of generalized linear models with mixed effects to analyze and to predict health status. By making this approach generic and operational, we seek to implement our method on a maximum of disease or pest of various crop. A screening process is proposed. For each disease or pest, different models are compared. Depending on the analysis and prediction performances, the pest or the disease will be retained or not for the operational tools. Finally, this work makes it possible to provide BSV actors operational tools for analyzing and predicting epidemic dynamics based on statistical models whose predictive quality is measurable. To demonstrate this, we also built a web application for "Vine Health Watch" based on these approaches and developed during the 2018 API-AGRO hackathon.

Keywords: epidemiological surveillance, modeling, Plant Health Bulletin, data science, mixed model.

INTRODUCTION

Afin de réduire l'utilisation des produits phytosanitaires, le plan Ecophyto a été mis en place au niveau national suite au Grenelle de l'Environnement en 2008 et reconduit en 2015 (MAAF, 2015a), avec pour objectif de réduire de 50% cette utilisation en 10 ans. Avec la mise en place de ce plan, le dispositif de surveillance biologique du territoire s'est généralisé à l'ensemble des productions et des régions afin de mieux quantifier les problèmes sanitaires en cours de saison et ainsi mieux raisonner la protection des cultures et optimiser les interventions phytosanitaires. Ainsi, le Bulletin de Santé du Végétal (BSV) s'est mis en place en 2009. Ce dispositif collectif fournit l'information sur l'état de santé des cultures à l'échelle régionale sous la forme d'une édition périodique (souvent hebdomadaire), à partir d'un réseau de surveillance biologique du territoire fournissant de nombreuses données d'observation, dans certains cas de prédictions issues de modèles épidémiologiques et d'expertises. Ce bulletin, adressé aux agriculteurs et à leurs conseillers, leur permet d'avoir régulièrement un diagnostic quantifié de l'état sanitaire de leurs cultures associé à une analyse de risque afin de mieux cibler leurs interventions phytosanitaires et éviter les traitements systématiques.

Le BSV est maintenant bien établi mais peut néanmoins être amélioré en valorisant mieux les différentes sources d'informations disponibles et en formalisant mieux leur analyse et l'intégration de ces sources d'informations. En effet, l'élaboration du BSV se fait généralement dans un laps de temps réduit entre la collecte des observations et sa publication, souvent moins de 24 heures. Cela oblige les acteurs à faire des choix stratégiques de rédaction et les observations issues du réseau l'épidémiosurveillance sont ainsi sous-utilisées faute de temps et d'outils appropriés. Dans la plupart des cas, un simple résumé avec des statistiques descriptives simples (extrema, moyenne, pourcentage de parcelles atteintes) est proposé.

Nos travaux visent donc à fournir des outils opérationnels aux acteurs du BSV pour appuyer avec des éléments objectifs leur analyse de risque, outils qui seront à la fois des outils d'aide à l'analyse des observations terrains mais aussi des modèles prédictifs à court-terme. Le travail reprend un cadre méthodologique proposée dans une thèse effectuée à l'Acta et l'INRA (Michel, 2016a), dans lequel des modèles linéaires généralisés à effets mixtes étaient utilisés pour analyser et prédire l'état sanitaire pour la septoriose du blé (Michel *et al.*, 2016b) qui a fait l'objet d'une présentation lors du CIMA 2015 (Michel *et al.*, 2015). Ce cadre statistique a été étendu à d'autres maladies et ravageurs d'autres cultures (Michel *et al.*, 2016c). Nous avons repris cette approche, en cherchant à « l'industrialiser » pour la mettre en œuvre sur un maximum de maladies et ravageurs sur différentes cultures et en la rendant plus opérationnelle pour proposer facilement des outils opérationnels aux acteurs du BSV.

Une démarche de criblage est proposée. Pour chaque maladie ou ravageur, différents modèles sont comparés. En fonction des performances d'analyse et de prédiction, les travaux se poursuivent sur les cas les plus prometteurs. Enfin, les prototypes d'outils ont déjà été proposés aux acteurs du BSV.

MATERIEL ET MÉTHODE

DESCRIPTION DES DONNEES COLLECTEES DANS LE CADRE DU BSV

Description des réseaux d'observation et de leurs données

Un réseau d'observation se compose d'un réseau de parcelles réparties sur le territoire (Figure 1A, pour l'exemple en vigne). Ces parcelles peuvent être des parcelles témoins non traitées ou traitées. Pour le moment, nous avons limité notre étude aux données des parcelles non traitées ou de la période avant le premier traitement.

Ces parcelles font l'objet d'observations périodiques (souvent hebdomadaires) pour caractériser l'état sanitaire à la date de l'observation (Figure 1B). Chaque maladie ou ravageur a son protocole propre, harmonisé au niveau national. Par exemple pour le mildiou de la vigne, cinq observations différentes sont relevées : la fréquence des ceps (MCF, nombre de ceps de vignes touchées par la maladie sur cent ceps), la fréquence des feuilles (MFF, nombre de feuilles touchées sur cent feuilles), l'intensité de la maladie sur les feuilles (MFI, surface moyenne touchée sur les feuilles atteintes), la fréquence des grappes et enfin l'intensité sur les grappes.

Une fois les observations effectuées, elles sont saisies par les observateurs dans un système d'information dédié (Epicure de l'IFV – Institut Français de la Vigne pour la vigne dans la zone ouest) et analysées régionalement par les acteurs du BSV pour produire l'édition du BSV contenant l'analyse de risque pour chaque problème sanitaire.

Chaque réseau d'observation est géré par plusieurs partenaires différents selon les cultures. On y retrouve les chambres d'agriculture, les instituts techniques, les coopératives et les négociants notamment.

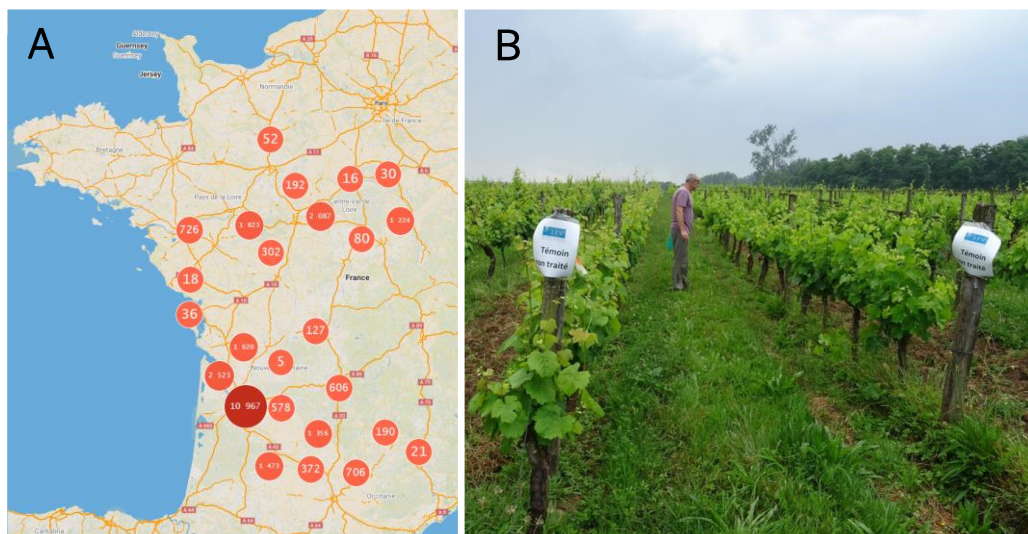


Figure 1: Collecte d'observations en vigne. A: Réseau d'observations de la zone ouest collectées sous l'outil EPICURE (IFV) : les chiffres des bulles représentent le nombre d'observations des 4 maladies collectées sur 2009-2017. B: Observations dans une parcelle non traitée.

Collecting observations in the vineyard. A: Western Zone Observations Network collected under the IFV EPICURE tool. B: Observations in an untreated plot.

Des exemples de jeux de données de tailles différentes sont présentés dans la Table 1 : il y a beaucoup plus de données pour la vigne que pour la betterave. En vigne, les données sont mieux réparties sur les sites et sur plus d'années. Ces cultures ont été choisies pour la disponibilité des données. Les données d'autres grandes cultures (colza et pois notamment) ont également été disponible, mais trop tardivement pour les inclure dans ce travail.

	Vignes - maladies	Vignes - ravageurs	Betteraves - maladies
Maladies/ravageurs observés	Mildiou, Oïdium, Black-rot, Pourriture grise	Cochylis, Eudémis	Oïdium, Cercosporiose, Rouille, Ramulariose
Années d'observation	8	6 à 7	3
Parcelles par an*	107 (82 à 156)	132 (28 à 227)	25 (20 à 32)
Observations par Parcelle par an*	14 (7 à 23)	31 (10 à 183)	5 (3 à 10)

Table 1: Caractéristiques des données selon les réseaux d'observations. * : valeur médiane (minimum à maximum).

*Characteristics of the data according to the networks of observations. *: median value (minimum to maximum).*

Exemple des séries de données considérées

Deux exemples de séries temporelles sont donnés en Figure 2. La première série (Figure 2.A) représente les observations de fréquence de mildiou sur les feuilles de vigne (MFF) effectuées sur une parcelle donnée et sur deux années. Dans ce cas, les dynamiques épidémiques présentent une forme de sigmoïde, la dynamique de la maladie n'étant pas stoppée (parcelle non traitée) et présentant un démarrage plus tardif en 2017 qu'en 2016.

La seconde série (Figure 2.B) représente les comptages de papillons d'eudémis, dont les larves produisent des dégâts sur les fruits (vers de la grappe), capturés grâce à un piège à phéromone. Les dynamiques épidémiques forment des pics de présence cycliques (vers mi-juin, mi-juillet et mi-août), représentant les vols de papillons eudémis des différentes générations.

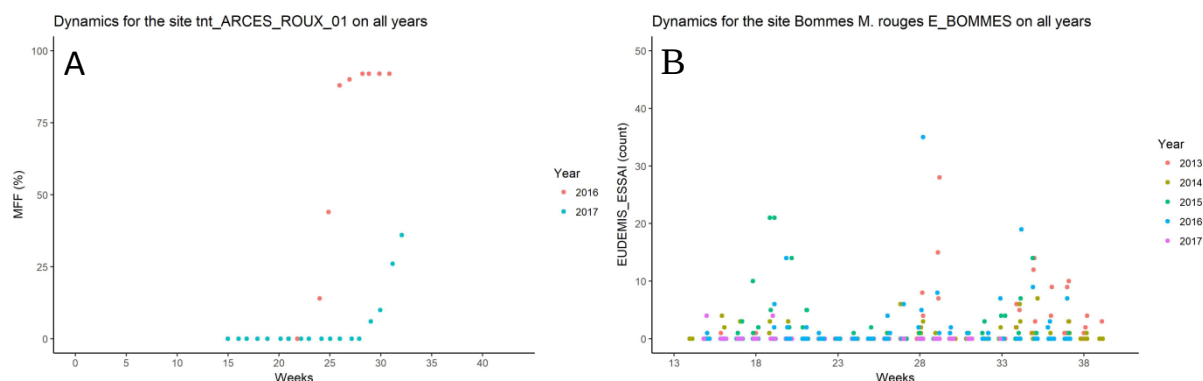


Figure 2 : Exemple des séries d'observation en vigne. A. Fréquence de mildiou sur les feuilles de vigne (MFF) sur une parcelle non traitée sur deux années. B. Comptage de papillons d'eudémis sur 1 site et 5 années.

Example of observation series in vines. A. Downy mildew (MFF) frequency on one untreated plot over two years. B. Counting butterflies of eudemis on 1 site and 5 years.

MODELES LINEAIRES GENERALISES A EFFETS MIXTES POUR L'ETUDE DES DYNAMIQUES EPIDEMIQUES

Présentation du cadre de modélisation et des variantes

Le modèle linéaire à effets mixtes tient son nom du fait qu'il contient à la fois des effets fixes et des effets aléatoires. Ainsi, plusieurs sources de variabilité peuvent être considérées (Laird and Ware, 1982). Les effets fixes s'interprètent de la même façon qu'une régression linéaire : à une variable caractéristique de l'individu (autrement appelé prédicteur) correspond un coefficient multiplicatif associé unique pour tous les individus. Les effets aléatoires sont une généralisation de ce concept où l'on considère un effet aléatoire pour le prédicteur.

Le modèle linéaire généralisé à effets mixtes permet de prendre en considération des variables réponses de distributions différentes et dépendantes non linéairement des prédicteurs. Pour cela, les effets fixes et aléatoires sont combinés pour former un prédicteur linéaire. La variable réponse est ensuite reliée à ce prédicteur linéaire par une fonction de lien choisie, et la distribution conditionnellement aux effets aléatoires est également choisie par l'utilisateur. Des propositions de distributions et fonctions de liens en fonction de grands types de situations rencontrées dans les données d'épidémiologie est expliqué par Michel *et al.* (2016c). Ainsi, on utilisera une distribution binomiale avec une fonction de lien logit pour des incidences de maladie (proportion de plantes atteintes par exemple) et une distribution de Poisson avec une fonction de lien log pour des données de comptage (nombre d'insectes piégés par exemple).

Ces modèles ont l'avantage d'être assez souples pour traiter aussi bien des données d'incidence de maladie que de comptage que l'on rencontre fréquemment dans les réseaux d'épidémiologie.

Cela permet également de traiter des données longitudinales, comme ici avec des observations sur chaque site répétées à plusieurs dates dans l'année. Enfin, ils peuvent facilement être ajustés à l'arrivée de nouvelles données d'observation au fur et à mesure de la campagne et, ainsi, on peut profiter des premières observations en début de campagne pour affiner les prédictions. Ces modèles ont été implémentés et ajustés dans le langage statistique R en utilisant la librairie lme4 (Bates *et al.*, 2015).

DEMARCHES DE CRIBLAGE UTILISANT LE MODELE GENERALISE A EFFETS MIXTES

Présentation de la démarche de criblage

L'objectif est de faire le lien entre les besoins d'analyse et de prédiction des rédacteurs du BSV pour renforcer leur analyse de risque, les données disponibles sur plusieurs années, cultures et maladies ou ravageurs grâce à leur réseau d'observations et enfin les modèles linéaires généralisés à effets mixtes dans le but de fournir des outils opérationnels au BSV. Dans cette optique, une démarche générique de criblage est proposée pour évaluer rapidement l'intérêt de cette approche de modélisation sur un maximum de maladies et ravageurs des différentes cultures. Celle-ci permet d'identifier les maladies ou ravageurs pour lesquels un modèle avec des bonnes qualités d'analyse et de prédiction peut être ajusté, afin de proposer finalement aux acteurs du BSV des outils opérationnels utilisant ces modèles.

La démarche de criblage utilisée est schématisée dans la Figure 3.

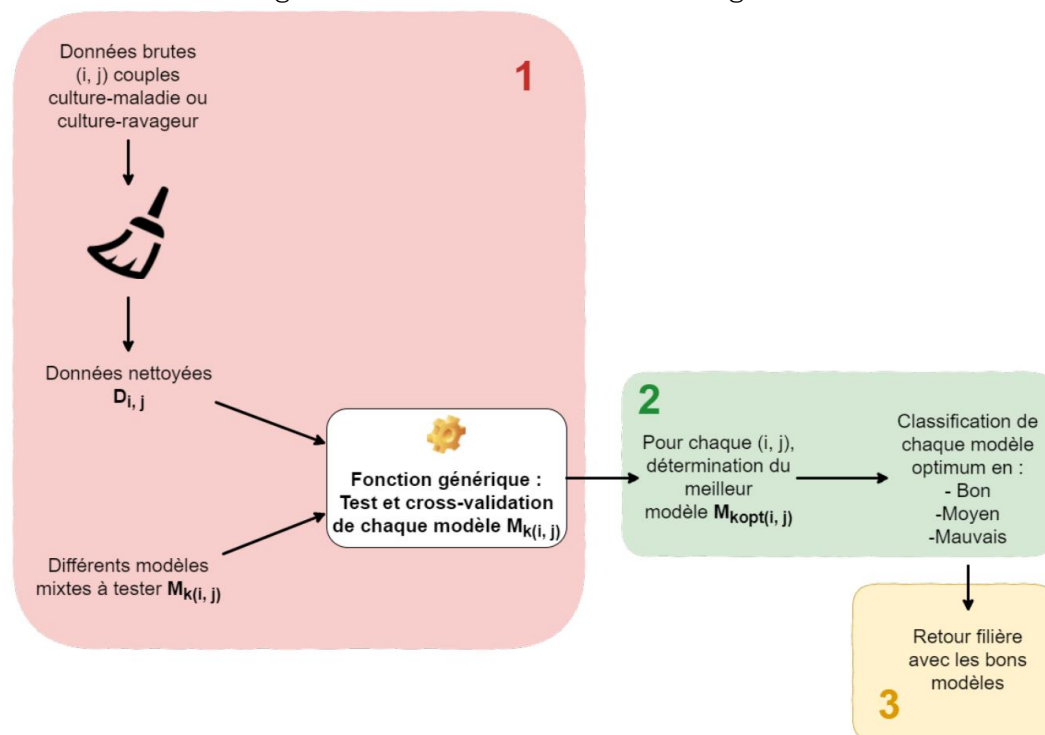


Figure 3 : Démarche de criblage proposée.
Screening process proposed.

Ce travail de criblage consiste en 3 étapes :

Etape 1. Pour chaque bio-agresseur (maladie ou ravageur), le nettoyage des données consiste uniquement en une élimination des données aberrantes ou outliers, notamment l'élimination d'années pour lesquelles il y a très peu de données ou alors de sites avec trop peu d'observations. Ces règles ont été précisées pour chaque cas en cherchant à limiter l'élimination de données.

A chaque couple est ensuite associé ses informations descriptives (nom de la culture, de la variable d'intérêt, nom de certaines variables comme la variable année, semaine ou département ...) et les modèles linéaires généralisés à effets mixtes à tester (nom du modèle, distribution d'erreur, fonction de lien et effet fixes et aléatoires).

Les performances de chaque modèle sont alors calculées.

Etape 2. Pour chaque couple, le modèle le plus performant est identifié en sortie de cette fonction. Les performances des meilleures modèles de tous les couples peuvent enfin être comparées et les couples sont ainsi classifiés comme ayant un bon, moyen ou mauvais modèle associé. Cette qualification est un premier indicateur automatique basé sur des seuils permettant de prioriser les efforts par la suite.

Les performances sont variables entre couples selon bien sûr à quel point le modèle est adéquat pour décrire la dynamique de la maladie ou du ravageur, mais également selon le nombre de données d'observation pour un couple donné. En effet, a priori, plus un couple a de données sur plusieurs années et bien réparties sur les sites, plus son meilleur modèle peut être ajusté de la meilleure façon.

Etape 3. Cette classification permet finalement de sélectionner les couples culture-maladie ou ravageur et leur meilleur modèle associé qui intéresseraient le plus les acteurs du BSV, de par leurs bonnes qualités d'analyse et de prédiction. Cela ouvre aussi des possibilités d'approfondissement (ajout de co-variables par exemple).

Évaluation des performances d'analyse et de prédiction des modèles

Pour évaluer les qualités d'analyse et de prédiction des différents modèles testés, différents indicateurs de performance sont calculés pour évaluer les performances de chacun des modèles. Seront présentés d'une part les indicateurs de performance d'analyse et d'autre part les indicateurs de performance de prédiction ainsi que la validation croisée utilisée pour estimer ceux-

ci. Ils sont calculés pour comparer les valeurs des observations aux prédictions du modèle étudié correspondantes.

Pour calculer les indicateurs de performance d'analyse (Table 2), le modèle linéaire généralisé à effets mixtes considéré est ajusté sur l'ensemble des données nettoyées disponibles pour le couple culture-maladie ou culture-ravageur étudié. Il n'y a donc pas de notion d'ensemble d'apprentissage ou de validation ici. Ces indicateurs visent uniquement à décrire à quel point le modèle peut s'ajuster aux données afin d'en analyser les dynamiques. Un modèle avec de bons indicateurs de performance d'analyse sera capable par exemple de bien comparer les incidences médianes de maladie (présenté dans la Figure 5).

AIC	critère d'information d'Akaike	Combine la qualité de l'ajustement et la complexité du modèle en pénalisant les modèles avec un grand nombre de paramètres
BIC	critère d'information bayésien	
RMSE	racine carrée de l'erreur quadratique moyenne	qualité d'ajustement du modèle sans pénalisation de la complexité

Table 2 : Indicateurs de performance d'analyse
Performance indicators of analysis

Les indicateurs de prédiction sont calculés par une procédure de validation croisée. L'ensemble des données doit être divisé en un ensemble d'apprentissage pour ajuster le modèle et un ensemble de validation sur lequel la prédiction et la comparaison avec les observations est effectuée. Cette division de l'ensemble des données est ici faite plusieurs fois en suivant le principe d'une validation croisée.

Le principe de cette validation croisée (Figure 4) a été défini pour évaluer la qualité de prédiction de manière réaliste par rapport à l'utilisation opérationnelle visée par la suite. Une étape de validation croisée est déterminée par le choix d'une année et d'une semaine de transition. Étant donné ces deux paramètres (l'année 2013 et la semaine représentée par la ligne bleue sur le schéma), l'ensemble d'apprentissage est l'ensemble des données sur les années autres que l'année choisie complété par les données sur l'année choisie observées jusqu'à la semaine de transition choisie. L'ensemble de validation représente le reste des données, soit les observations de l'année choisie après la semaine de transition. Ce découpage de l'ensemble des données permet de se placer en condition réelle, où l'on a les données de plusieurs années précédant la campagne en cours et les données de la campagne en cours jusqu'au jour de l'étude, et où l'on cherche à prédire les dynamiques épidémiques pour le reste de la campagne.

Ce découpage est ensuite itéré sur l'ensemble des années d'étude et sur un ensemble de semaines de transition raisonnable. Cet ensemble raisonnable de semaines est trouvé automatiquement, en cherchant l'ensemble des semaines telles que tous les sites ou presque aient deux observations avant cette semaine et deux observations après. L'utilisation de 2 observations avant la semaine de prédiction n'est pas strictement nécessaire, mais ainsi, on se met en situation où ces deux premières observations ont permis de capturer le début de l'épidémie. Sans cela, le modèle renverrait une prédiction moyenne quelle que soit l'année, ce qui limite l'intérêt de l'approche.

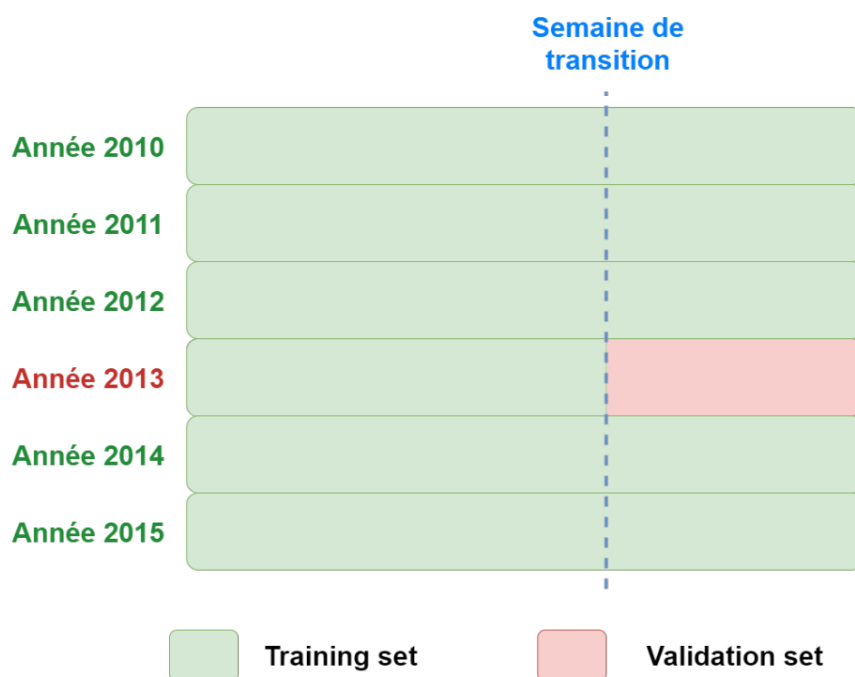


Figure 4 : Principe de la validation croisée utilisée
Principle of cross-validation used

Chaque indicateur de performance (Table 2) de prédiction est alors calculé à 3 échéances de prédiction, soit à une, deux et trois semaines. Cela signifie que les indicateurs sont calculés sur les données observées et prédites à une et à deux semaines (nous avons également évalué à trois semaines pour voir le comportement) de la semaine de transition considérée. Les résultats qui seront présentés dans le tableau final seront également divisés en ces trois échéances de prédiction, afin de montrer l'évolution de la qualité de prédiction avec le temps.

RMSEP	racine carrée de l'erreur quadratique moyenne de prédiction	Calculé sur des données de validation lors de la validation croisée.
RRMSEP	valeur relative de la racine carrée de l'erreur quadratique moyenne de prédiction	Relativisation en fonction de la valeur de l'observation.
EF	efficience	Quantifie la variance expliquée par le modèle par rapport à la variance totale. Sa valeur est inférieure à 1, 1 correspondant à un modèle parfait.
Kappa	Kappa de Cohen	Cet indicateur représente la capacité de classement en fonction d'un seuil d'intérêt défini (pour chaque maladie ou ravageur). Sa valeur est inférieure à 1. Un modèle classifiant parfaitement les données aura un Kappa de 1.

Tableau 2 : Indicateurs de qualité de prédiction
Prediction quality criteria

RESULTATS

L'objectif est de fournir aux acteurs du BSV des outils opérationnels d'analyse objective et de prédiction des dynamiques épidémiques.

DEUX EXEMPLES DE VALORISATION : OUTIL ANALYSE ET OUTIL DE PREDICTION.

Un exemple d'outil d'analyse est la comparaison des dynamiques épidémiques en ajustement sur différentes années. Après ajustement du modèle, les médianes des réponses sur les différents sites d'une région peuvent être calculées pour les différentes années et comparées (Figure 5). En cours de campagne, cela permet notamment au BSV d'indiquer si la dynamique épidémique de la campagne en cours est en avance ou retardée par rapport à celles des campagnes précédentes. C'est intéressant aussi en bilan de campagne.

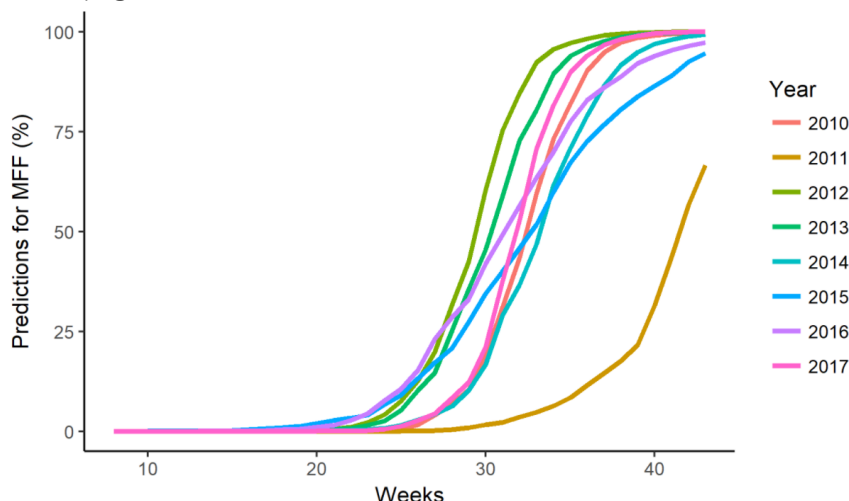


Figure 5: Comparaison des médianes d'incidence de mildiou de la vigne sur différentes campagnes en Aquitaine en utilisant un modèle linéaire généralisé à effets mixtes.
Comparison of the median incidence of downy mildew of vine for different years in Aquitaine using a generalized linear model with mixed effects.

D'autre part, ces modèles peuvent être utilisés pour prédire en cours de campagne l'évolution la plus probable de la dynamique épidémique (Figure 6). Ceci est particulièrement utile aux agriculteurs et à leurs conseillers pour orienter leur stratégie d'intervention.

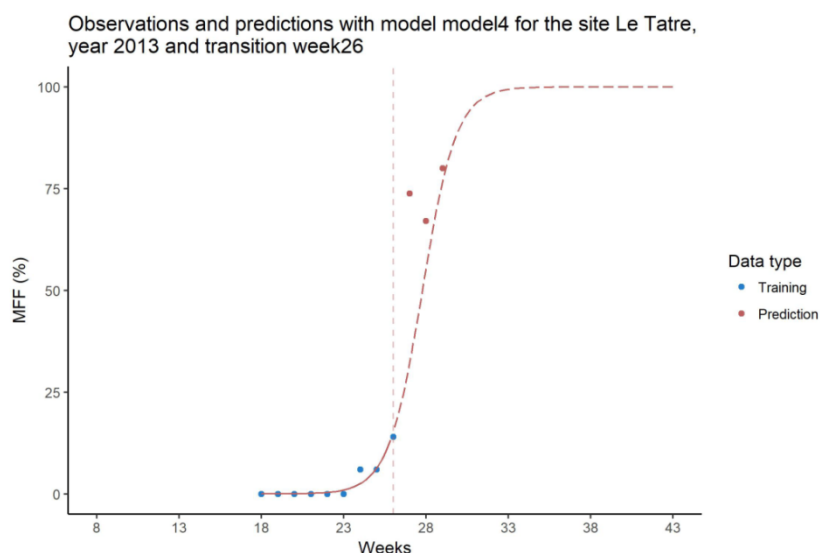


Figure 6: Prédiction de l'incidence de mildiou de la vigne après la semaine 26 sur un site observé en utilisant un modèle linéaire généralisé à effets mixtes.
Prediction of the incidence of downy mildew of Vine after week 26 on an observed site using a generalized linear mixed effects model.

PERFORMANCE DES MODELES SUR LES DIFFERENTS MALADIES ET RAVAGEURS

Comparaison des différents modèles pour une maladie ou un ravageur

Un exemple de tableau résultat pour le mildiou de la vigne est donné dans la figure 7. Celui-ci montre le résultat de la comparaison de 5 modèles intégrant différents effets fixes et aléatoires :

tous ont une constante et un coefficient pour la semaine comme effet fixe sauf le modèle model4 qui intègre également le cépage, et les effets aléatoires (entre parenthèses dans l'écriture des modèles) sont tous pour une constante et un coefficient et la semaine, mais regroupés par différents groupes (années, sites, et/ou sites-années selon les modèles). On observe que trois modèles se distinguent (model4, model3 et modelLM), tous ayant un effet aléatoire par sites-années, et ont les mêmes caractéristiques avec de bonnes qualités d'analyse et prédiction. Ces trois modèles présentent des qualités de prédiction moyennement satisfaisantes (efficacité à 1 semaine de l'ordre de 0.215 à 0.277). Le critère du Kappa montre aussi une bonne capacité de classification, mais il reste à échanger sur la pertinence du seuil avec les acteurs. Ainsi, ces modèles peuvent constituer un bon choix de départ pour fournir un premier outil opérationnel aux acteurs du BSV ou pour les améliorer en rajoutant d'autres co-variables par la suite.

	formula	family	link	AIC (16)	BIC (8)	RMSE (12)	RMSEP W1 (8)	RMSEP W2 (4)	RMSEP W3 (2)	RRMSEP W1 (4)	RRMSEP W2 (2)	RRMSEP W3 (1)	EF W1 (8)	EF W2 (4)	EF W3 (2)	Kappa W1 (4)	Kappa W2 (2)	Kappa W3 (1)
model4 (67.3)	cbind(MFF, 100 - MFF) ~ SEMAINE_OBS + CEPAGE + (1 + SEMAINE_OBS ANNEE) + (1 + SEMAINE_OBS NOM_PARCELLE) + (1 + SEMAINE_OBS ANNEE.NOM_PARCELLE)	binomial	logit	68061.4	68314.3	6.337	10.702	18.080	25.554	170.450	177.455	159.191	0.277	-0.115	-0.434	0.659	0.548	0.373
model3 (65.5)	cbind(MFF, 100 - MFF) ~ SEMAINE_OBS + (1 + SEMAINE_OBS ANNEE) + (1 + SEMAINE_OBS NOM_PARCELLE) + (1 + SEMAINE_OBS ANNEE.NOM_PARCELLE)	binomial	logit	68056.2	68138.0	6.334	10.938	18.621	26.860	172.407	180.924	163.492	0.245	-0.183	-0.585	0.665	0.538	0.391
modelLM (64.1)	cbind(MFF, 100 - MFF) ~ SEMAINE_OBS + (1 + SEMAINE_OBS ANNEE) + (1 + SEMAINE_OBS ANNEE.NOM_PARCELLE)	binomial	logit	68153.6	68213.2	6.333	11.150	18.896	27.502	175.487	183.725	167.138	0.215	-0.218	-0.661	0.650	0.513	0.351
model2 (36.5)	cbind(MFF, 100 - MFF) ~ SEMAINE_OBS + (1 + SEMAINE_OBS ANNEE) + (1 + SEMAINE_OBS NOM_PARCELLE)	binomial	logit	164848.8	164908.3	12.766	12.552	21.640	33.830	208.411	205.258	191.291	0.006	-0.597	-1.514	0.379	0.266	0.227
model1 (13.7)	cbind(MFF, 100 - MFF) ~ SEMAINE_OBS + (1 + SEMAINE_OBS ANNEE)	binomial	logit	281656.1	281693.3	17.626	13.322	23.269	37.837	209.222	201.342	195.721	-0.120	-0.847	-2.144	0.180	0.079	0.072

Figure 7 : Comparaison des performances de différents modèles pour le mildiou de la vigne (MFF). Le gradient de couleur représente les performances des modèles : vert (bonnes performances) à rouge (mauvaises performances) (W1, W2, W3 : échéances semaines 1, 2, 3).

Comparison of the performances of different models for downy mildew of vine (MFF)

Résultats synthétiques du criblage sur les différents couples culture-maladie ou culture-ravageur

Une fois les meilleurs modèles choisis en fonction des performances pour chaque couple, les couples peuvent être ordonnés et classifiés comme ayant un bon, moyen ou mauvais modèle associé. Ceci permettra de sélectionner les couples qui ont un modèle avec de bonnes qualités d'analyse et de prédiction et de se rapprocher des acteurs du BSV pour leur proposer des outils opérationnels utilisant ces modèles.

Cette classification est effectuée grâce aux valeurs de certains des indicateurs de performances pour les meilleurs modèles. Parmi eux, seuls les efficacités et les Kappa peuvent être comparés entre couples puisqu'ils présentent la même échelle. En effet, les AIC et BIC permettent de comparer des modèles utilisant le même jeu de données, les RMSE et RMSEP n'auront pas les mêmes unités selon que l'on s'intéresse à des valeurs d'incidence (en pourcentage) ou des données de comptage (positives non majorées) et ne sont donc pas comparables. Les RRMSEP sont également difficilement comparables.

Un score est donc établi en fonction des valeurs de l'efficacité et du Kappa de chaque meilleur modèle de chaque couple et la classification est établie selon ce score. Les résultats de la classification sur les dix jeux de données étudiés jusqu'à présent sont donnés dans la figure 8.

On peut voir que les modèles sur vigne sont globalement meilleurs, notamment sur BCF (fréquence sur ceps de Black rot) ou sur plusieurs variables de l'oïdium (OGI, OGF, OFF). Sur la betterave, la cercosporiose est la mieux prédite avec une efficacité faible (0.111). Mais comme il y a très peu de données sur la betterave en comparaison de la vigne, il serait intéressant d'approfondir. Les modèles sur les ravageurs de la vigne (non présentés dans la figure 8) présentent encore des problèmes de formulation et de convergence à régler.

Quality	Database	VarInterest	BestModel	EF_W1	EF_W2	EF_W3	Kappa_W1	Kappa_W2	Kappa_W3
	EPICURE2	BCF	model4	0.644	0.263	-0.088	0.808	0.740	0.685
	EPICURE2	OGI	model4	0.555	0.199	-0.105	0.743	0.450	0.355
	EPICURE2	OGF	model4	0.685	0.026	-0.297	0.691	0.526	0.440
	EPICURE2	OFF	model4	0.499	0.115	-0.324	0.694	0.543	0.483
	EPICURE2	MGI	model4	0.265	0.067	-0.031	0.627	0.475	0.435
	EPICURE2	PGI	model3	0.392	0.011	-0.517	0.800	0.666	0.419
	EPICURE2	PCF	modelLM	0.513	-0.065	-0.718	0.762	0.630	0.601
	EPICURE2	MGF	modelLM	0.589	0.200	-0.490	0.476	0.389	0.209
	EPICURE2	MFF	model4	0.277	-0.115	-0.434	0.659	0.548	0.373
	ITB2	CER	model5	0.111	0.114	0.386	0.284	0.135	0.202
	EPICURE2	BGF	model4	-0.126	-0.553	-0.182	0.560	0.353	0.328
	EPICURE2	PGF	model4	-0.059	-0.258	-0.170	0.499	0.233	0.125
	EPICURE2	BGI	model4	-0.132	-0.549	-0.258	0.434	0.215	0.138
	EPICURE2	BFI	model4	0.165	-0.601	-1.762	0.728	0.565	0.460
	EPICURE2	PFI	model3	0.300	-0.581	-2.065	0.626	0.561	0.502
	EPICURE2	MFI	model4	0.027	-0.618	-1.589	0.612	0.481	0.426
	EPICURE2	BFF	modelLM	0.174	-1.007	-2.124	0.731	0.615	0.504
	ITB2	OID	model5	0.361	-0.723	-2.131	0.556	-0.041	-0.076
	EPICURE2	PFF	model4	-0.353	-1.508	-3.278	0.567	0.488	0.462
	ITB2	RAM	model5	0.177	-3.583	-1.295	0.280	0.150	-0.042
	ITB2	ROU	model5	-0.812	-1.967	-4.192	0.487	0.509	0.310

Figure 8 : Classification des meilleurs modèles de chaque couple (performances bonnes : vert, moyennes : orange ou mauvaises : rouge). Exemple de variable d'intérêt : BCF : Fréquence sur ceps de botrytis; OFF : fréquence sur feuille d'oidium; CER : cercosporiose;...)
Classification of the best models of each couple (good : green, medium : orange or bad performances : red)

INTEGRATION DANS DES OUTILS OPERATIONNELS

Lors du hackathon Api-Agro 2018 (concours de codage d'application web en 48h), nous avons entrepris le développement d'un prototype opérationnel utilisant ces méthodes à destination du BSV (MAAF, 2018). Ce prototype ajuste les meilleurs modèles identifiés dès l'entrée de nouvelles données d'observation en cours de campagne, puis permet des prédictions représentées sous la forme de graphiques de tendances régionales et des cartes de dynamique (figure 9). Ce premier prototype permet aux rédacteurs du BSV de se familiariser dès à présent avec ces méthodes statistiques et de commencer à les intégrer pour leur analyse de risque.

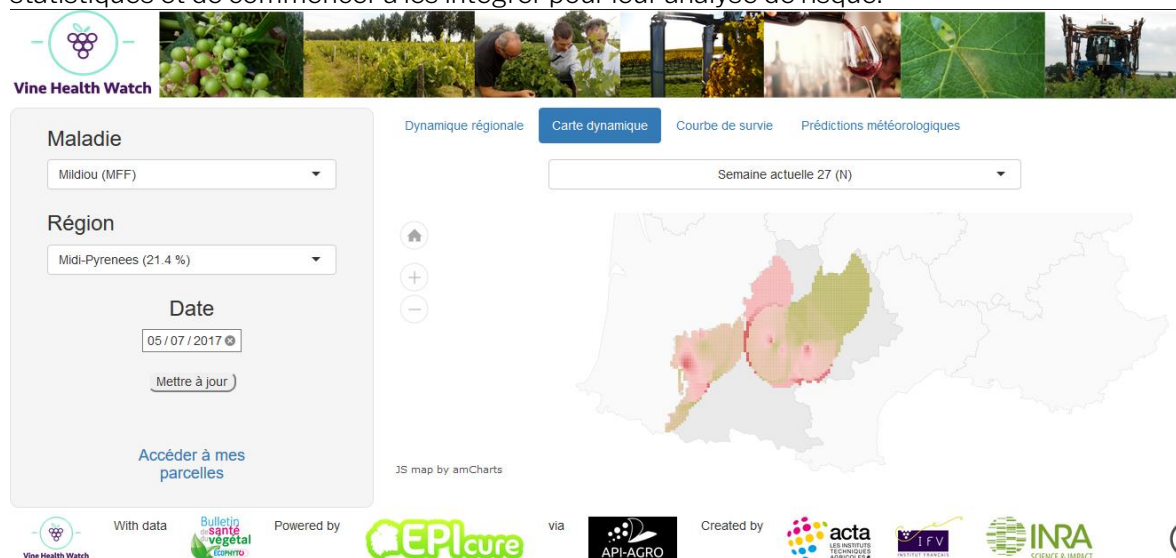


Figure 9: Prototype Vine Health Watch développé lors du Hackathon Api-Agro 2018, lauréat du 2^{ème} prix.
Prototype Vine Health Watch developed during the Api-Agro Hackathon 2018, 2nd prize winner.

DISCUSSION

INTERETS ET LIMITES DE L'APPROCHE

Les principaux écueils statistiques rencontrés ont été liés à la fonction du package lme4 utilisée pour implémenter les modèles linéaires généralisés à effets mixtes. On peut en effet rencontrer des problèmes de convergence, notamment avec l'ajout de paramètres aux modèles, ce qui ne facilite pas le travail d'automatisation.

La démarche générique de criblage va se continuer sur de nouveaux jeux de données du BSV dans le cadre du projet Boîte à Outils pour le BSV. Il s'agira notamment d'approfondir la question de ce genre de modèle adapté aux ravageurs qui présentent un profil assez différent (comptage et pic). L'intégration de co-variables météorologiques doit également être testée plus largement. Certains seuils utilisés pour calculer le Kappa doivent aussi être affinés après échanges avec les experts et utilisateurs.

Sur les cas les plus prometteurs, d'autres approches statistiques ou de science des données pourront aussi être explorées et comparées, toujours dans l'optique de proposer des outils performants et opérationnels (utilisable en cours de saison) d'analyse ou de prédiction de l'état sanitaire.

REMERCIEMENTS

Ces travaux ont été menés dans le cadre du projet SMART-PIC - Concevoir des outils pour l'aide à la décision en protection intégrée des cultures intégrant l'ensemble des leviers par la mobilisation de nouvelles données et de formalismes innovants (2015-2019, financement CASDAR, AAP RT). Ils valorisent également des acquis du projet SYNOEM - Mieux profiter de la synergie entre réseaux d'observations, expertise et modélisation pour l'élaboration du Bulletin de Santé du Végétal (2013-2016, financement ONEMA, AAP PSPE).

Merci à Nicolas Lenne (Direction Générale de l'Alimentation, Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt) et Danièle Simoneau (Arvalis, l'Institut du Végétal) pour leur aide pour faire le lien avec le Comité National d'Epidémiosurveillance.

Merci à Barbara Cichosz (CRA Occitanie) et Thierry Bordin (CRA Centre) pour leurs échanges réguliers sur ces travaux et leurs soutiens à la démarche.

Merci à Corentin Barbu (INRA, UMR Agronomie) pour avoir fourni ses routines de pré-traitements des données de Vigiculture sur le colza.

BIBLIOGRAPHIE

Akaike H., 2011 - Akaike's information criterion. In *International encyclopedia of statistical science*, pages 25–25. Springer.

Bates D., Mächler M., Bolker B., Walker S., 2015 - Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1) :1–48.

Laird N.M., Ware J.H., 1982 - Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

MAAF, 2015a - Le plan Ecophyto, pour réduire l'utilisation des produits phytosanitaires en France. <http://agriculture.gouv.fr/le-plan-ecophyto-pour-reduire-lutilisation-des-produits-phytosanitaires-en-france>. Accédé le : 2017/11/27.

MAAF, 2018 - Vine Health Watch: la data science au service de la santé de la vigne. <http://agriculture.gouv.fr/vine-health-watch-la-data-science-au-service-de-la-sante-de-la-vigne> consulté le 02/05/2018

Michel L., Decarrier A., Franche M., Bochu V., Benredjem K., Hugerot G., Simonneau D., Piraux F., Gourdain E., Veslot J., Makowski D., Brun F., 2015 - Un outil d'analyse des dynamiques épidémiologiques pour le Bulletin de Santé du Végétal: application à la septoriose du blé. *AFPP – Onzième Conférence Internationale sur les Maladies des Plantes*. Tours – 7 au 9 décembre 2015.

Michel L., 2016a - Mieux valoriser les réseaux d'épidémiosurveillance lors de l'élaboration du Bulletin de Santé du Végétal. Mémoire de thèse.

Michel L., Brun F., Piraux F., Makowski D., 2016b - Estimating the incidence of Septoria leaf blotch in wheat crops from in-season field measurements. *European Journal of Plant Pathology*.

Michel L., Brun F., Piraux F., Makowski D., 2016c - A framework based on generalised linear mixed models for analysing pest and disease surveys. *Crop Protection* DOI:10.1016/j.cropro.2016.12.013