Modèle de survie pour prédire l'arrivée de la rouille jaune

ARVALIS
Institut du végétal

le processus de modélisation ARVALIS

1. • Objectif(s) du modèle : **prédiction**, expliquer

2. • Bibliographie + expertise → variables agronomiques / climatiques
   • Données disponibles

Validation du jeu de variables explicatives par les experts thématiques

3. • Validation et Exploration des données
   • Identifier les erreurs, …
   • Corrélation entre variables → variables d'intérêts, variables redondantes, confusions d'effets, …

4. • Choix de la méthode modélisation

5. • ajustement du modèle : estimation, sélection de variables, transformation de variable

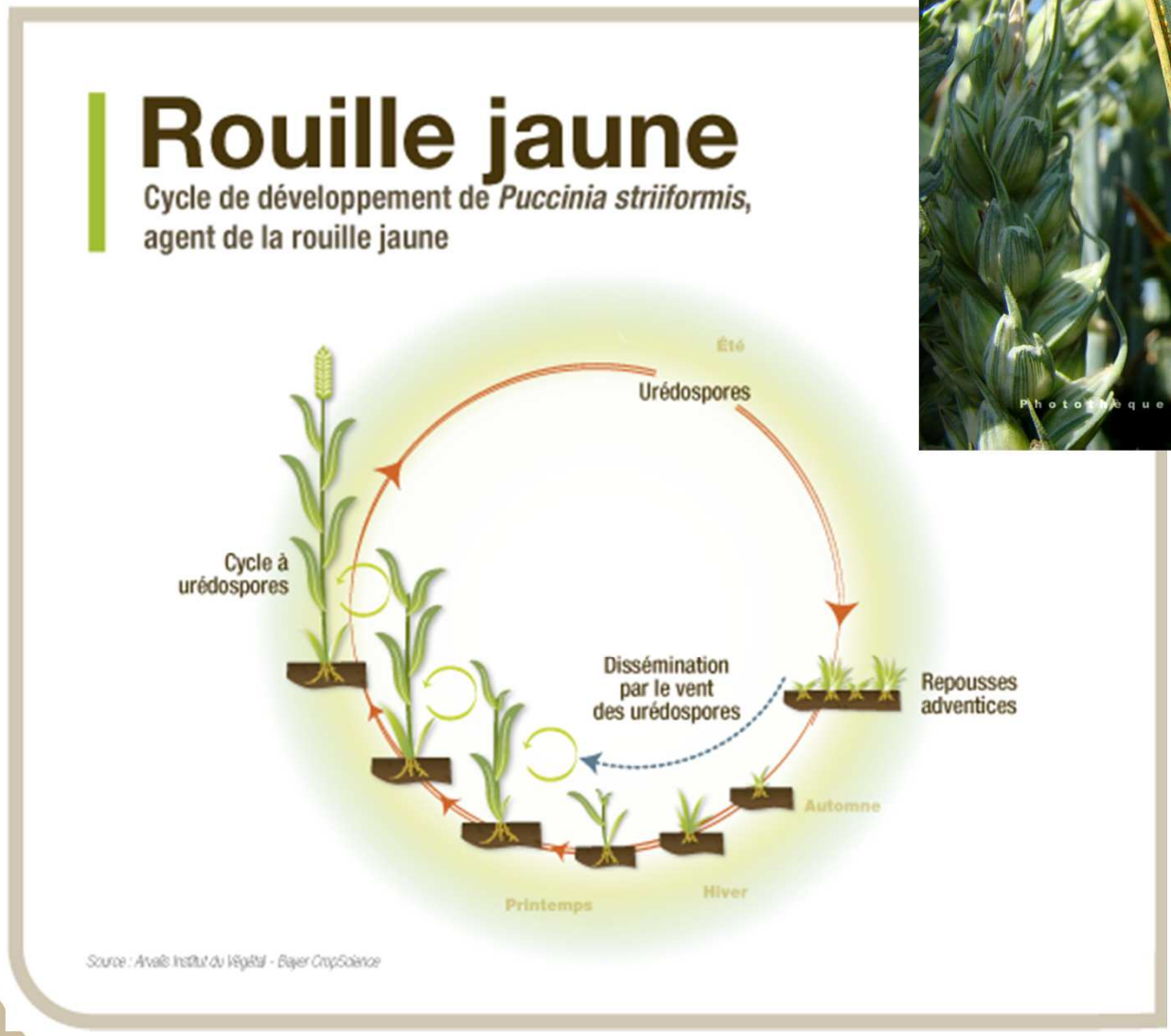
6. • Validation du modèle : cohérence de l'effet des variables
   • Evaluation de la qualité prédictive

7. • Implémentation dans les outils ARVALIS

Quelle question ?

Estimer le délai d'apparition des symptômes de rouille jaune à l'échelle de la parcelle en cours de campagne

Eléments de la littérature



Puccinia striiformis
Biotrophe
Polycyclique
Foyer
Nuisible entre épi 1cm et épiaison

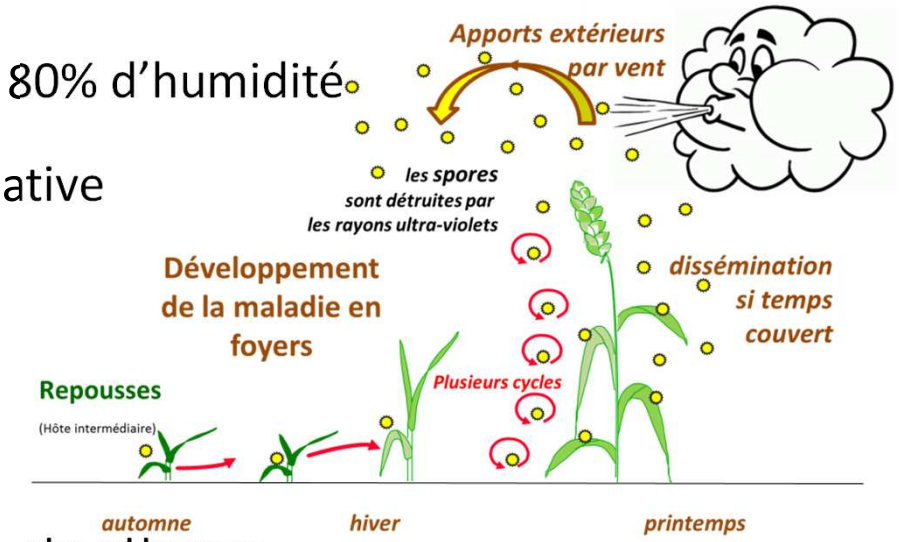
Eléments de la littérature

Nom :
Forme asexuée : *Puccinia striiformis*
Synonyme : *Puccinia glumarum* *(= rouille des glumes)*

Météo favorable :
Printemps doux 4°C<T°<25°C / Optimum 7-10°C + 80% d'humidité

Optimum d'infestation: 8°C à 100% d'humidité relative

Rosée nécessaire pour la germination



Répartition dans la parcelle :
En foyers puis généralisée à la parcelle en cas de forte attaque.

Dégâts :
Jusqu'à 60 % de pertes de rendement pour une forte attaque

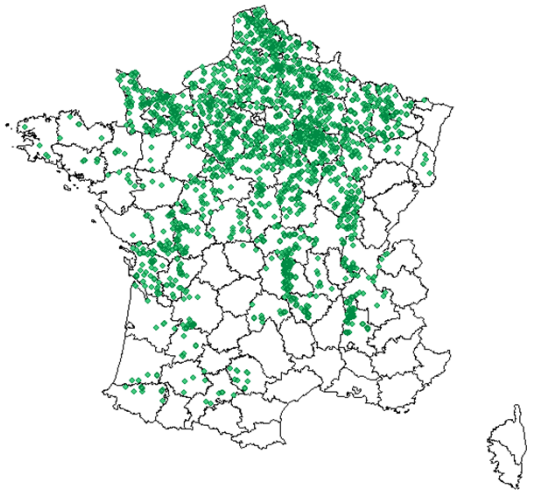
Cultures attaquées :
Blé dur, Blé tendre, Orge, Seigle et Triticale (races différentes)

Données disponibles

Expérimentation : *Pas d'essais dédiés*
Essais fongicides
Parcelles non traitées et essais variétés

Peu de sites / années
2-3 notations campagne

Observations : *Réseau d'épidémio-surveillance*
Parcelles suivies en dynamique
France entière
Depuis 2008



- Vigicultures 2009-2014 parcelles fixes
- 2790 parcelles, moy. 6.4 passages/parcelles/campagne
- 3 protocoles de notations

Données disponibles : Vigicultures

Identifiant champ	Description du champ	Information
region	réseau d'observation	Origine observation
id_plot	Identifiant de la parcelle	
nom_parcelle	texte	
organisme_oad	texte	
observateur_oad	texte	
latitude	calcul automatique à partir CP commune et éventuellement repositionnement sur carte	Localisation parcelle
longitude	calcul automatique à partir CP commune et éventuellement repositionnement sur carte	
altitude	saisie	
cp	Code postal	
lib_commune	menu déroulant à partir du CP	
code_insee	calcul automatique à partir CP commune et éventuellement repositionnement sur carte	
culture	liste de saisie	Information parcelle et observation
code_culture	agroEDI ?	
type_sol	liste finie dans la saisie transformé en code	
variete	menu déroulant ET saisie manuelle	
code_variete	agroEDI ?	
date_semis	calendrier	
dico_id_elt_observe	Identifiant élément noté (feuille ou inflorescence)	
observation	Maladie + organe + unité	
obs_val_num	Dépend des protocoles	
dateobs	Date de l'observation	
stade	Menu déroulant zadock	
annee	campagne agricole	
semaine	semaine d'observation	
labour	oui/non	Actions anthropiques
precedent	liste	
traitements	champ avec dates, type, produits	
date_crea	date du serveur au moment de l'enregistrement de la donnée	Information liée au fichier
valide	1 si l'observation est validée	

Données disponibles : protocoles

Dates protocoles	Parcelles fixes	Parcelles flottantes
Avril 2008 – Février 2010	0=absence / 1= pustules isolées sur une des 20 plantes (sur F1, F2,F3) / 2= foyer de rouille jaune ou plusieurs pustules par feuille **-> Note 0|1|2**	globale : 0=absence / 1= pustules isolées / 2= foyer de rouille jaune ou plusieurs pustules par feuille **-> Note 0|1|2**
Depuis le 05/10/2010	Echelle 0 à 10 par feuille f1/f2/f3 a partir d'une fréquence de feuille touchées sur 20 plantes **-> Note 0 : 10**	globale : 0=absence / 1= pustules isolées / 2= foyer de rouille jaune ou plusieurs pustules par feuille **-> Note 0|1|2**
Depuis le 05/10/2010 (note Expert)	Echelle 0 à 100 par feuille f1/f2/f3 a partir d'une intensité (surface infestée) de feuille touchées sur 20 plantes **-> Note 0 : 100**	

Données disponibles : agronomie / environnement

- **Variables liées à la plante :**

variete, Sensibilite_variete, note_variete, date_epi1cm_recalcule, date_Grainpateux

- **Variables liées aux pratiques culturales :**

Precedent, groupe_precedent (6), precedent_hote (oui/non), labour (oui/non), jour_semis, mois_semis, annee

- **Variables liées à la position géographique de la parcelle :**

latitude, longitude, altitude, Region_Arvalis, (type_sol/Code_postal ?)

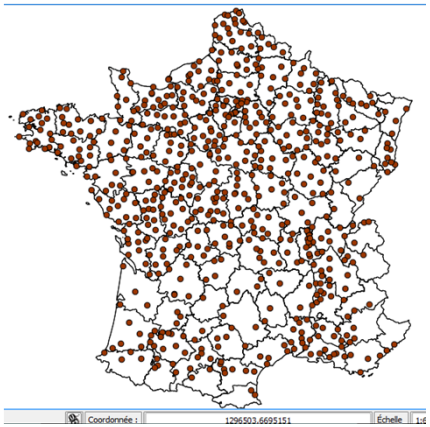
- **Variables Yello (modèle SRPV):**

date_traitement_Yello, traitement_yello_oui_non

Données disponibles : climat

- Variables climatiques de la campagne précédente (hiver précédent, sortie hiver précédent, printemps précédent, été précédent)
- Variables climatiques de la campagne en cours (entre levée, épi 1 cm et 2 nœuds)

=> 200 variables climatiques générées à partir des stades et du réseau de station météo ARVALIS



Validation des données

- Suppression des parcelles flottantes
- Notes de résistance variétale en fonction de l'année
- Données/feuille => Données/plante (somme)

- 3 lignes par date d'observation => 1 ligne par observation

- Regroupe les données/parcelles => plus petite date d'observation de maladie

- X lignes par parcelles (nb d'obs) => 1 ligne par parcelle

- Récupération de la dernière date d'observation

- Permet retirer les parcelles peu observées

Parcelle	Date_obs1	Date_obs2	Date_obs3	…	Dernière_ date_obs
Parcelle_1	0	1	1	…	1
Parcelle_2	0	0	0	…	0
Parcelle_2	1	1	1	…	1

Validation des données

8% des données supprimées, 4% des données corrigées

Principaux types d'erreur rencontrés	Donnée
Fautes orthographe nom de la variété et code gnis référencé dans Varcom	CORRIGEE
Fautes orthographe nom de la variété et code gnis inconnu du référentiel Varcom	EXCLUE
Mois semis incohérent (hors septembre-avril)	EXCLUE
Date de semis/Date observation incohérente (séparées de plus d'un an, ou observation avant semis)	EXCLUE
Année campagne	CORRIGEE
Localisation géographique hors France métropole	EXCLUE
Altitude	CORRIGEE
Stade Zadock incohérent	CORRIGEE
Traitement fongicide + absence notation maladie	EXCLUE

Exploration des données

- Etapes

  1) Visualisation brute des variables, graphiques univariés

  2) Visualisation des données manquantes

  3) Analyse en composante principale (ACP)

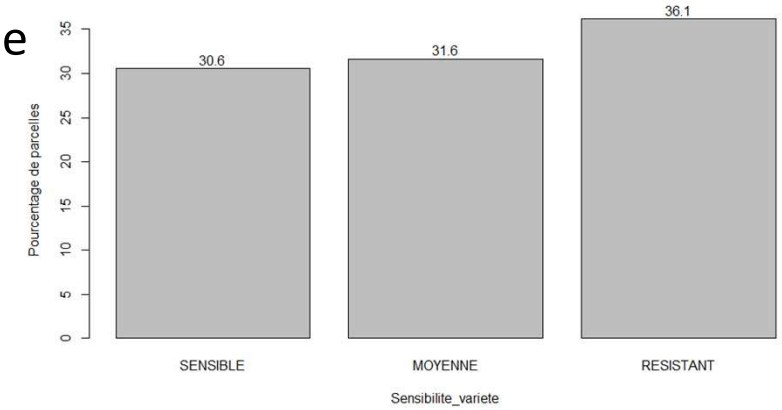
  4) Test Khi-deux et Visualisation graphique de toutes les variables en fonction de la variable à expliquer

  5) Analyse des correspondances Multiples (ACM)

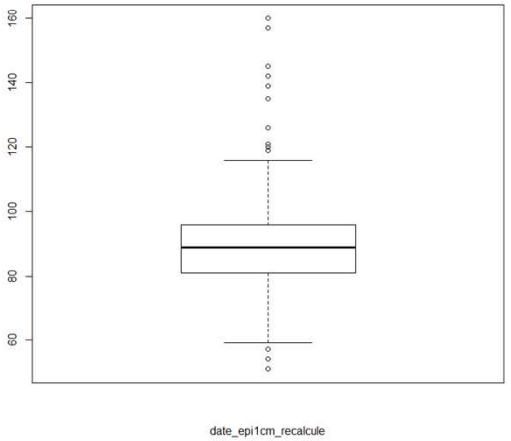
  6) Analyse de la variance (ANOVA)

Exploration des données

Histogramme pour les variables qualitatives

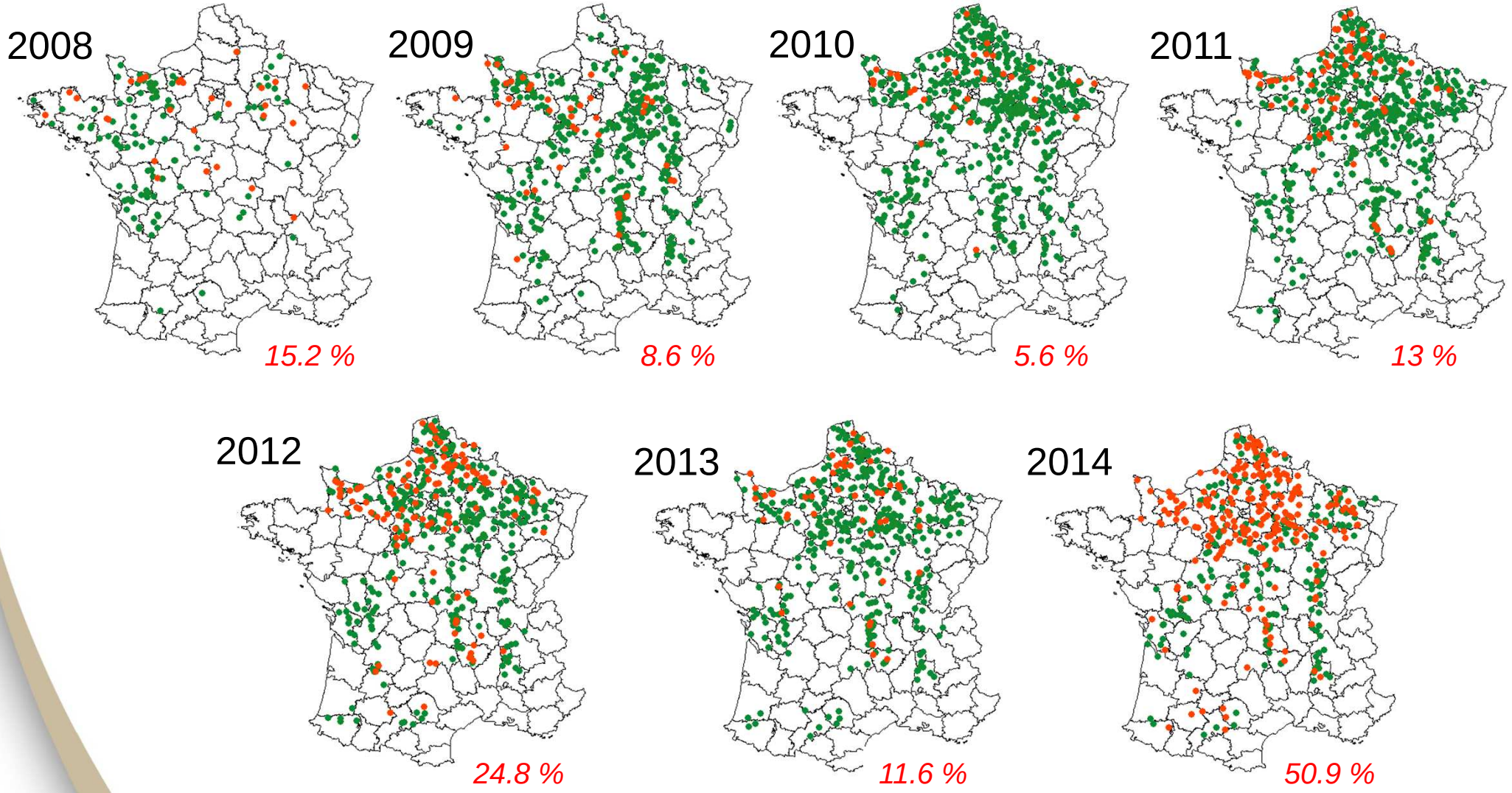


Boxplot ou Violinplot pour les variables quantitatives



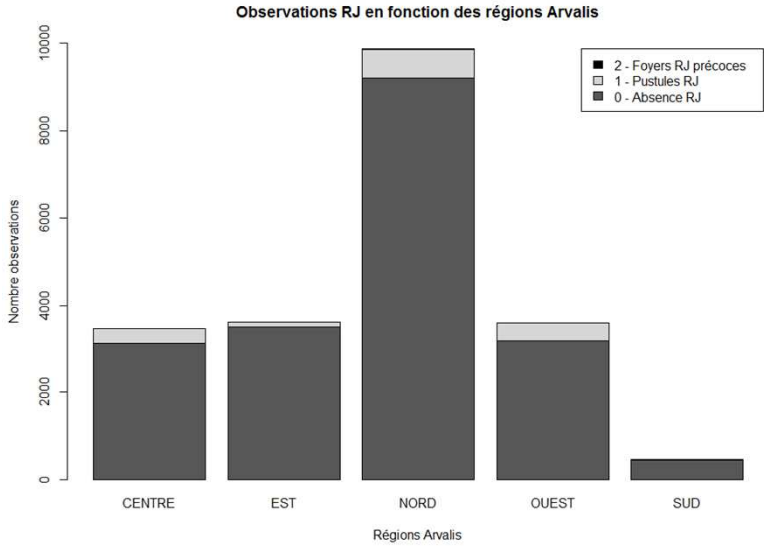
- 2008 présente moitié moins de données que les autres années
- On compte 3 à 9 fois moins de données sur la région SUD que sur les autres régions (région NORD est la plus représentée avec 40% des notations)
- 80% des parcelles n'ont pas été infestées dans le JDD
- 7% parcelles avec précédents culturaux hôtes de la maladie, 56% précédent non hôte et 35% précédent cultural inconnu

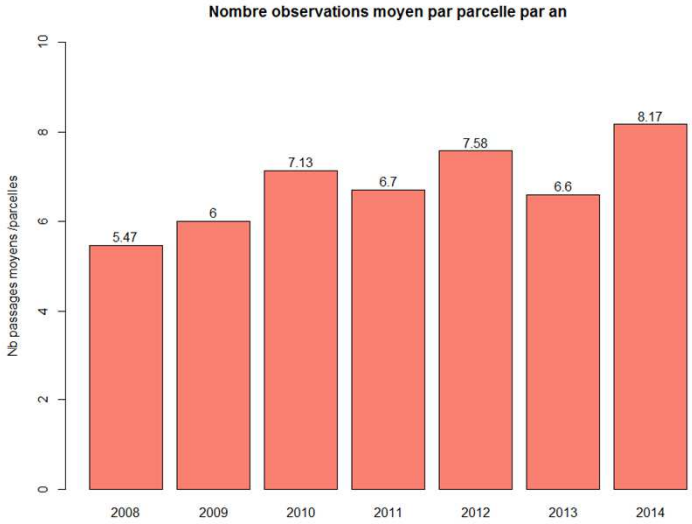
Exploration des données

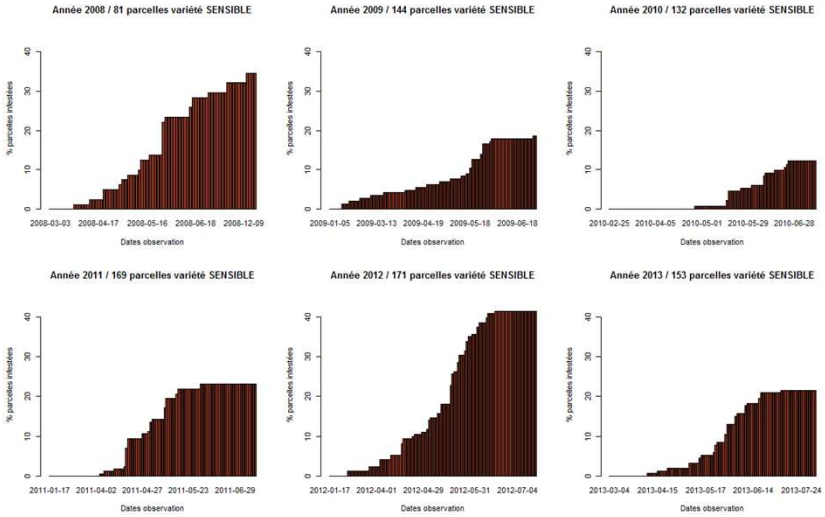


- Parcelles BTH saines
- Parcelles BTH ayant subit une infestation RJ durant la campagne
- % Pourcentage de parcelles BTH infestées par la RJ durant la campagne

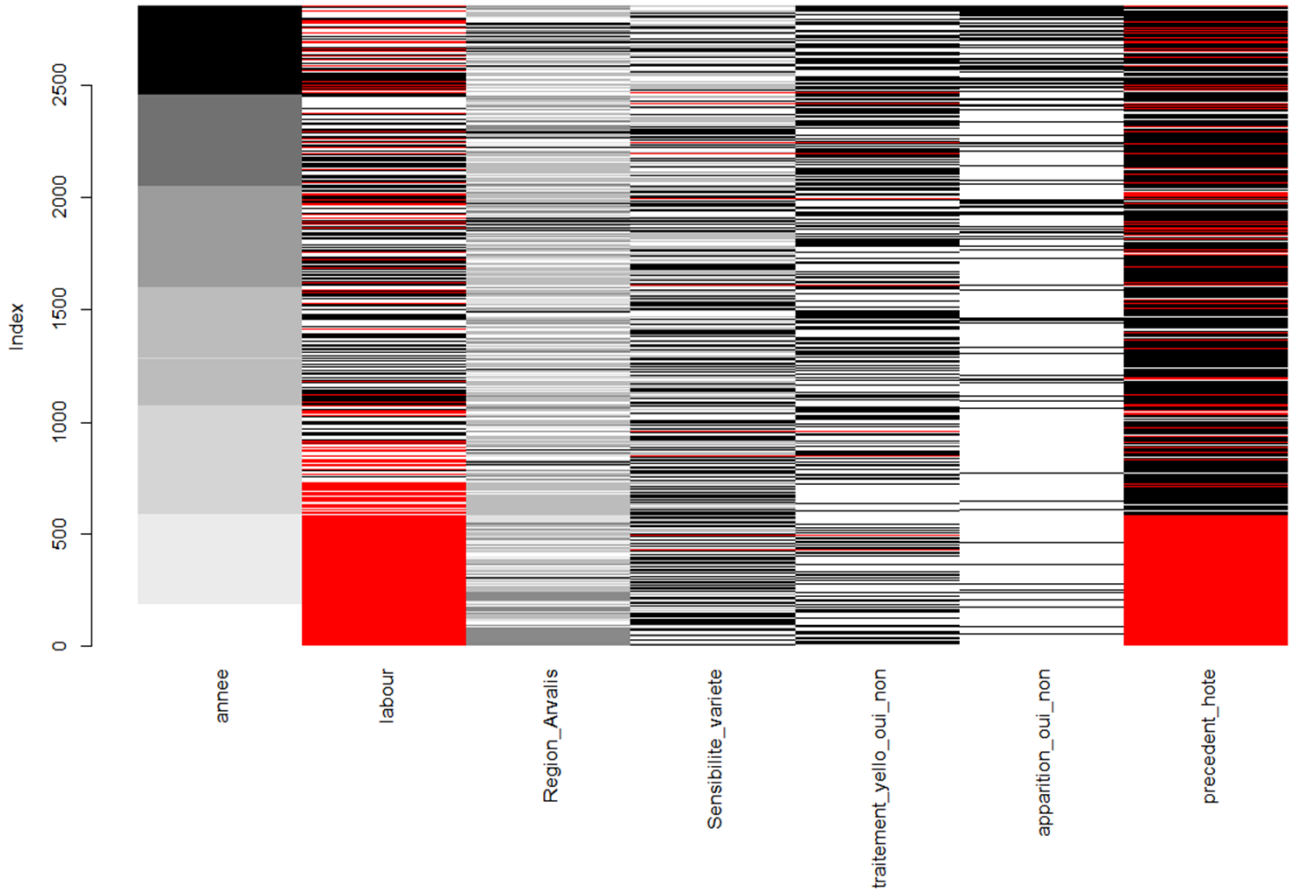
Exploration des données







Exploration des données

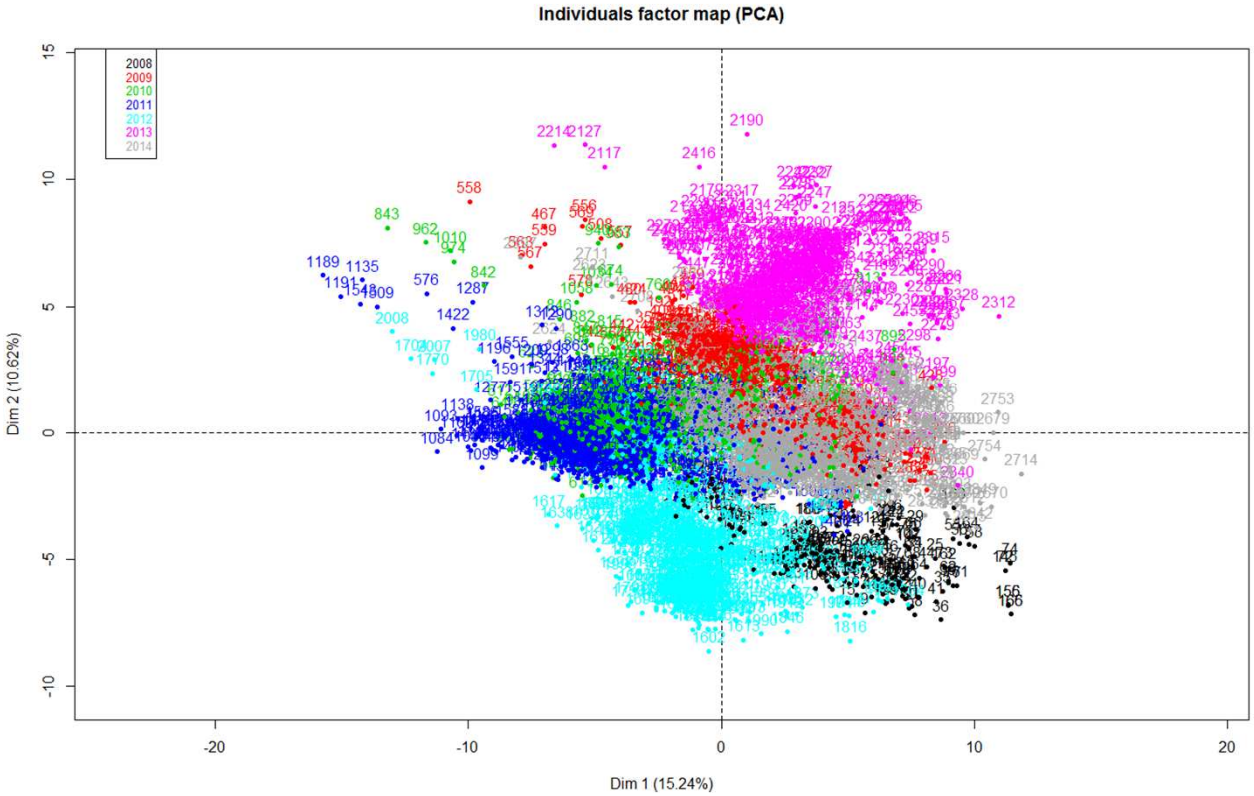


En rouge, données manquantes

Bilan : • Variables labour et précédent cultural inventoriées qu'à partir de 2010

Exploration des données

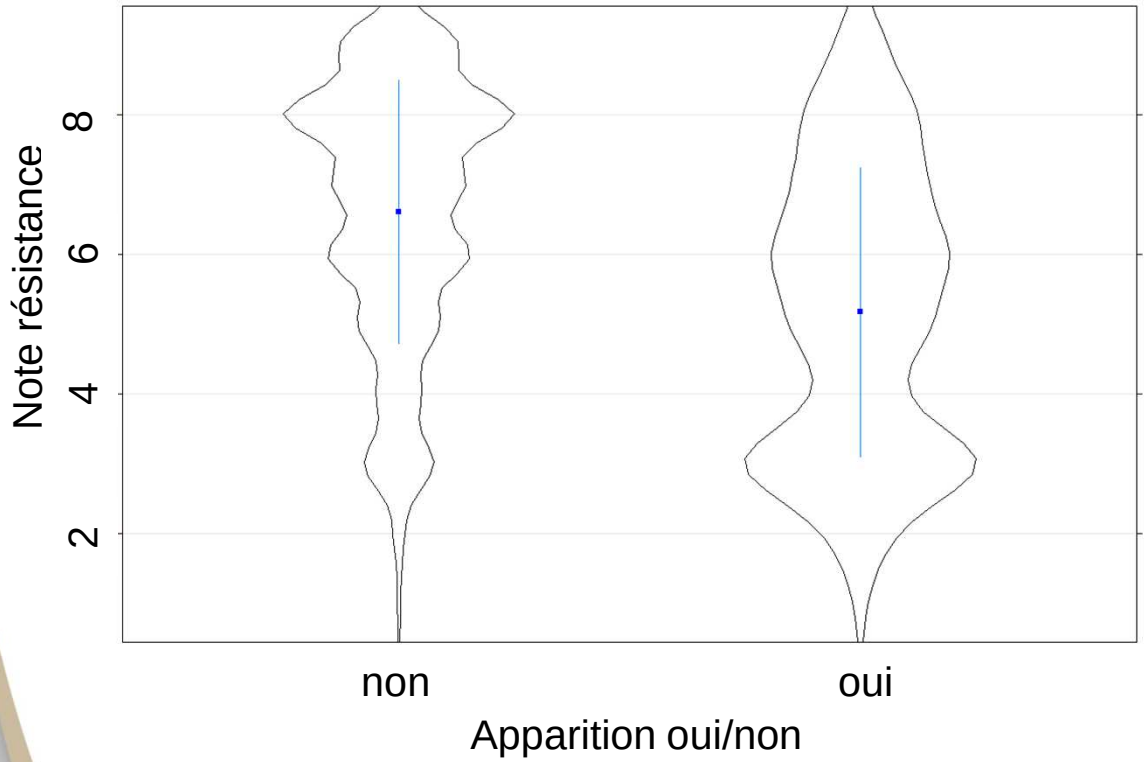
Graphique des individus illustrés par année



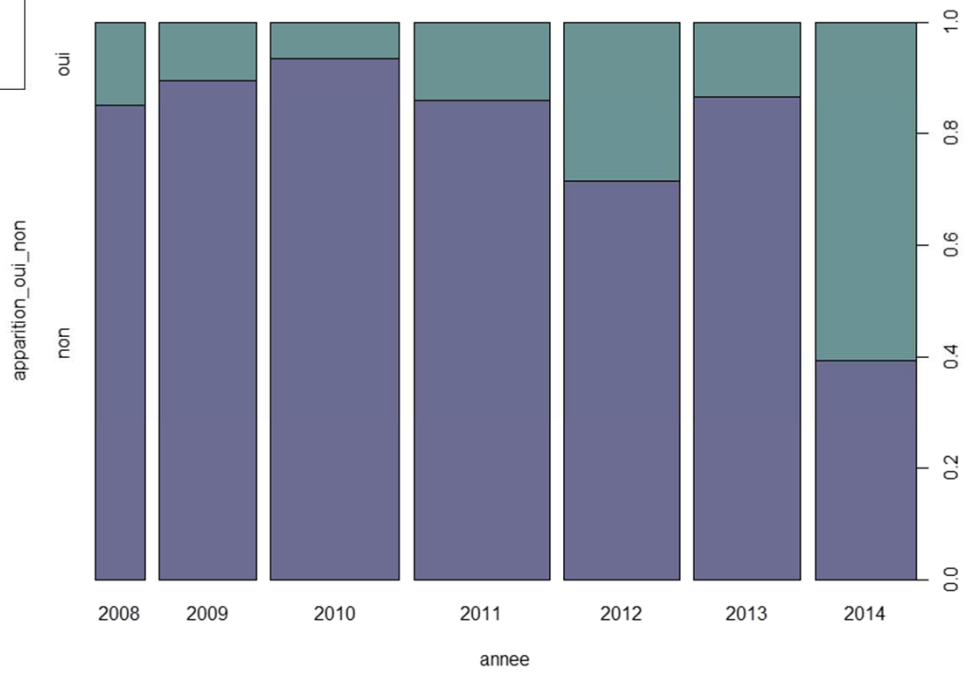
Bilan :

- Individus de 2009 (2010) et 2014 semblent se superposer (= même profil dans les variables climatiques choisies)
- **ATTENTION**, l'axe 1 explique seulement 15% des résultats on ne peut pas interpréter précisément et avec assurance les résultats de l'ACP

Exploration des données



Violinplot
pour les variables
quantitatives



Histogramme
pour les
variables
qualitatives

Exploration des données

- **Année – Région** : il y a plus de données dans l'Ouest et moins dans le nord en 2008
  → *A partir 2009 VégéObs*

- **Année – Sensibilité variété** : + variété R et - de variétés M et S en 2009
  → *Après 2009 Warrior apparait => contournement donc – de R après*

- **Région – Sensibilité variété** : Il est noté plus de variété R et moins de variété M dans le Sud ; il est noté plus de variété S dans le Centre
  → *Caractéristique du JDD*

- **Région – Traitement Yello** : Yello prédit plus d'infestation dans le Centre et moins dans le Sud
  → *Caractéristique du JDD*

…

Mise en évidence d'éventuelles confusions d'effets

Exploration des données

200 variables climatiques

- ✓ Expertise maladie
- ✓ Expertise climatique
- ✓ Tests de corrélation

100 variables

Choix de la méthode de modélisation

Choisir le(s) type(s) de modèle(s) désiré(s)

- Modèle qualitatif « oui-non »

  => Prédit une infestation ou non

- Modèle quantitatif

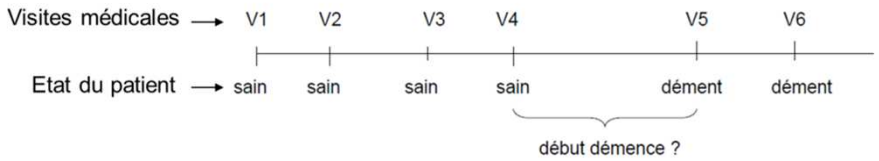
  => Prédit une date d'infestation

Choix de la méthode de modélisation

- Random forest / Extra-trees
- Modèle mixte
- Régression logistique
- Analyse discriminante linéaire
- Analyse factorielle discriminante
- Modèle de survie

Modèle de survie

- Durée de survie ou temps de survie (T) : temps qui s'écoule depuis un instant initial (début du traitement, diagnostic,...), jusqu'à la survenue d'un événement d'intérêt final (décès du patient, rechute, rémission, guérison...)



- Les données de survie consistent en une variable réponse (durée de vie) et éventuellement une ou plusieurs variables potentiellement explicatives de la durée de vie. Ces variables explicatives, appelées covariables ou facteurs pronostiques (concomitant variables, covariates, or prognostic factors), peuvent être qualitatives, comme le sexe ou la l'espèce, ou continue, comme l'âge ou la température

- ***L'objectif d'une analyse de survie est de modéliser la distribution de la durée de survie et d'estimer l'effet des variables explicatives sur cette distribution.***

Modèle de survie

Le temps de survie T est considéré comme une variable aléatoire. La distribution du temps de survie T (v.a. positive et continue) est décrite par 5 fonctions :

- **la fonction de survie** $S(t) = P(T > t)$

→ Elle désigne la probabilité de survivre au moins jusqu'a la date t

- **la fonction de répartition** $F(t) = P(T \leq t)$

→ Elle désigne la probabilité de décéder avant t

- **la densité** $f(t)$ avec $f(t) = \lim_{dt \to 0} \frac{1}{dt} P(t < T \leq t + dt)$

- **le risque instantané de décès** $h(t)$ avec $h(t) = \lim_{dt \to 0} \frac{1}{dt} P(t < T \leq t + dt | T > t)$

→ probabilité que l'évènement survienne dans un petit intervalle de temps après t , sachant qu'il n'a pas eu lieu jusqu'à l'instant t . Si t est mesuré en jours, h(t) approxime la probabilité d'un individu vivant le jour t , de mourir le jour suivant

- **le risque cumulé de décès** $\mathrm{H}(t)$ avec $\mathrm{H}(t) = \int_0^t h(s)ds$

Modèle de survie type COX

- Méthode **descriptive** et **prédictive (≠ Kaplan-Meier)**
- Hypothèses de fonctionnement
  - Hypothèse : la fonction de survie dépend du **temps** plus une ou plusieurs **covariables qualitatives ou quantitatives**
  - Hypothèse 2 : vérifier que les risques soient proportionnels

    (test effet covariable indépendant du temps)
- Possibilité de faire une sélection de variables quantitatives et qualitatives (fonction stepwise() dans R |cv.glmnet() -> approche ridge et Lasso)
- Cette approche permet aussi de prendre en compte un effet aléatoire (année en général)

Modèle de survie type COX

```
Library (survival)
na.omit(JDD) #ne prend pas les données manquantes
## Création d'un objet de type survival
Objet.surv <- Surv(JDD$evenement, JDD$statut)
#Evenement = visite : 'date dernière observation'(numéro de semaine ici)
#Statut = évènement produit ou non (donnée censurée ou non) : '+ ou rien'
```

```
> RJtot.surv
  [1] 23+ 20+ 27+ 23+ 21+ 21+ 24+ 25+ 15+ 26+ 15  20+ 20+ 24  21+ 22+  9+ 23+ 18  24  23+ 26+ 26+ 24+ 24+
 [32] 25+ 23  19+ 19  17+ 25+ 23+ 24+ 22+ 24+ 19   9  20+ 24+ 23+ 23+ 22+ 24+ 18+ 19  20+ 20+ 12+ 22+ 19
 [63] 22+ 25+ 21+ 18  24+ 19+ 22+ 26+ 24+ 19+ 20+ 19+ 24+ 25  23+ 20  15+ 14  23  16+ 21+ 15  21+ 25+ 12+
 [94] 27  23+ 24+ 17  19+ 20+ 23+ 25+ 16  18+ 22+ 16+ 26+ 14+ 22+ 26+ 20+ 23+ 18  24+ 15  24  21  19  19+
[125] 22+ 27+ 16  12+ 16+ 18  19+ 15+ 27+ 25+ 20  13+ 22+ 15+ 22+ 14+ 17+ 24+ 26+ 21  23+ 24+ 21+ 20+ 22+
[156] 25+ 24+ 22+ 14+ 26+ 25+ 24+ 22+ 20+ 17+ 27+ 17+ 26+ 21  18+ 23+ 24+ 20+ 18+ 22+ 21+ 26+ 22+ 25+ 22+
```

```
## Modèle de Cox
modele.cox <- coxph(Objet.surv ~ cluster(annee) + var1 + var2 + var3 +… , data=JDD)
## Prédiction
survfit(modele.cox , newdata= newdata) #plot possible
```

Ajustement du modèle

- Sélection de variables sur Jeu de donnée Vigicultures 2009-2014 (méthode Lasso puis Ridge):
  - résistance variétale (note) et date épi 1cm
  - climat année en cours jusqu'à 2 Nœuds
  - climat année précédente
  - variables incubation/contamination Yello campagne précédente et en cours

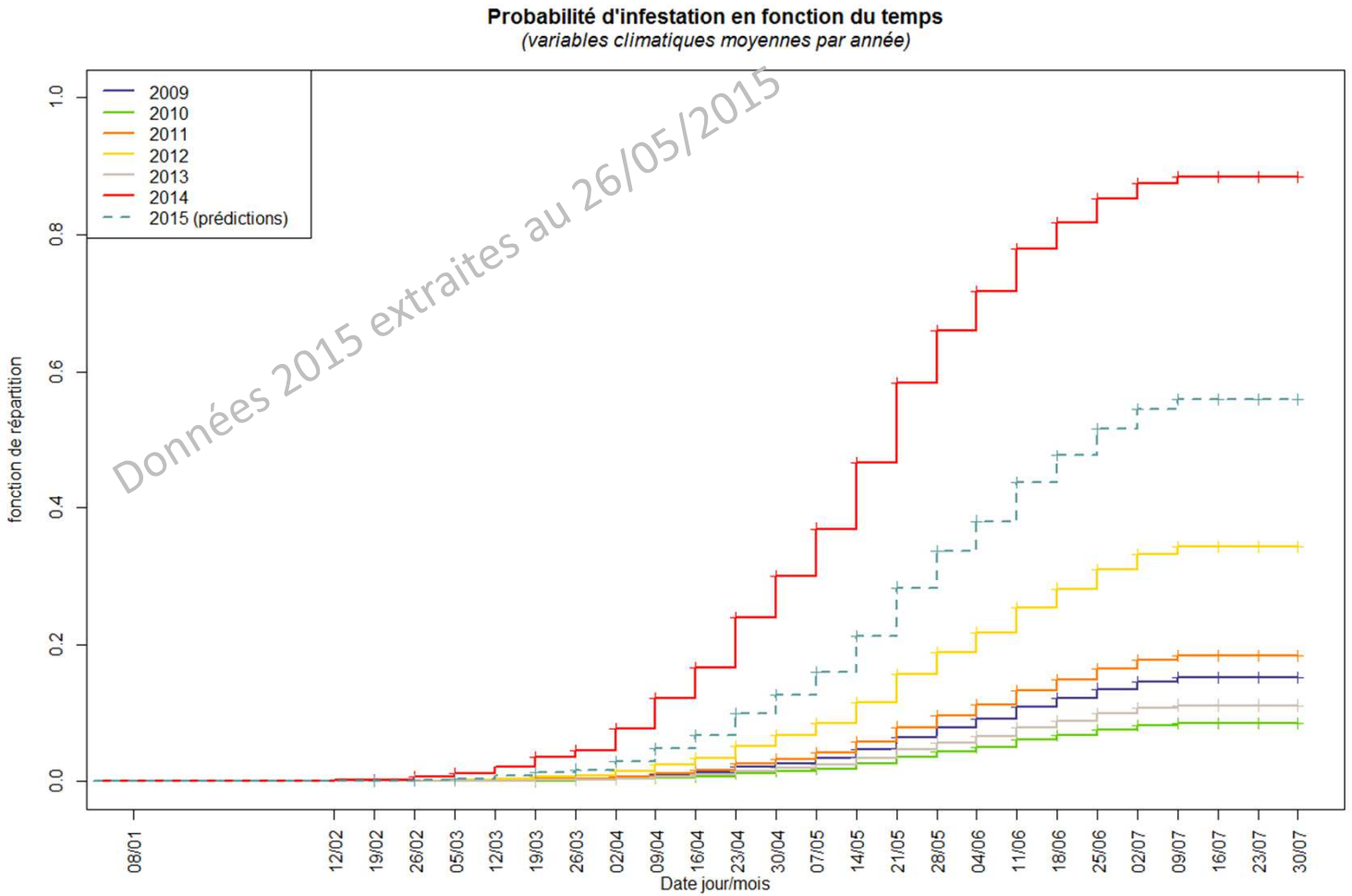
} Inoculum primaire

Bilan : 2500 variables élémentaires et en interaction double
=> 15 interactions retenues

Ajustement du modèle

Variables	Définition
note_variete	note de résistance variétale entre 1 et 9
date_epi1cm_recalcule	date épi 1 cm en nombre de jours depuis le 1er janvier
sTMiniinf0_leveeepi1cm	Somme des T°min > 0°C Levée - Epi 1cm Campagne N
sTMoynTMoysup7_leveeepi1cm	Somme des T°moyenne >7°C /Nombre de jours où T°moyenne >7°C Levée - Epi 1cm Campagne N
mTMini_leveeepi1cm	Moyenne des T°minimales Levée - Epi 1cm Campagne N
sIncubation_Yello_leveeepi1cm	Somme incubation (indice Yello entre 0 et 10) Levée - Epi 1cm Campagne N
sContamination_Yello_leveeepi1cm	Somme contamination (indice Yello entre 0 et 100) Levée - Epi 1cm Campagne N
mTMini_epi1cm2N	Moyenne des T°minimales Epi 1cm - 2 noeuds Campagne N
sContamination_Yello_hiver_precedent	Somme contamination (indice Yello entre 0 et 100) Hiver Campagne N-1
sTMoynTMoysup0_hiver_precedent	Somme des T°moyenne >0°C /Nombre de jours où T°moyenne >0°C Hiver Campagne N-1
sPETPsup0_sortiehiver_precedent	Somme pluie - ETP >0°C Sortie hiver Campagne N-1
nPluieinf02_sortiehiver_precedent	Nombre de jours de sécheresse Sortie hiver Campagne N-1
c2Pluieinf02_printemps_precedent	Nombre de séquence de 2 jours de sécheresse Printemps campagne N-1
sPluienPluiesup0_printemps_precedent	Somme des pluies/nombre de jours de pluie Printemps campagne N-1
sIncubation_Yello_printemps_precedent	Somme incubation (indice Yello entre 0 et 10) Printemps campagne N-1
sTMaxi_ete_precedent	Somme des T°maximales Été Campagne N-1

Ajustement du modèle



- Probabilité de mourir avant l'instant t
- Ici = probabilité que l'infestation survienne avant chaque date d'observation selon les années

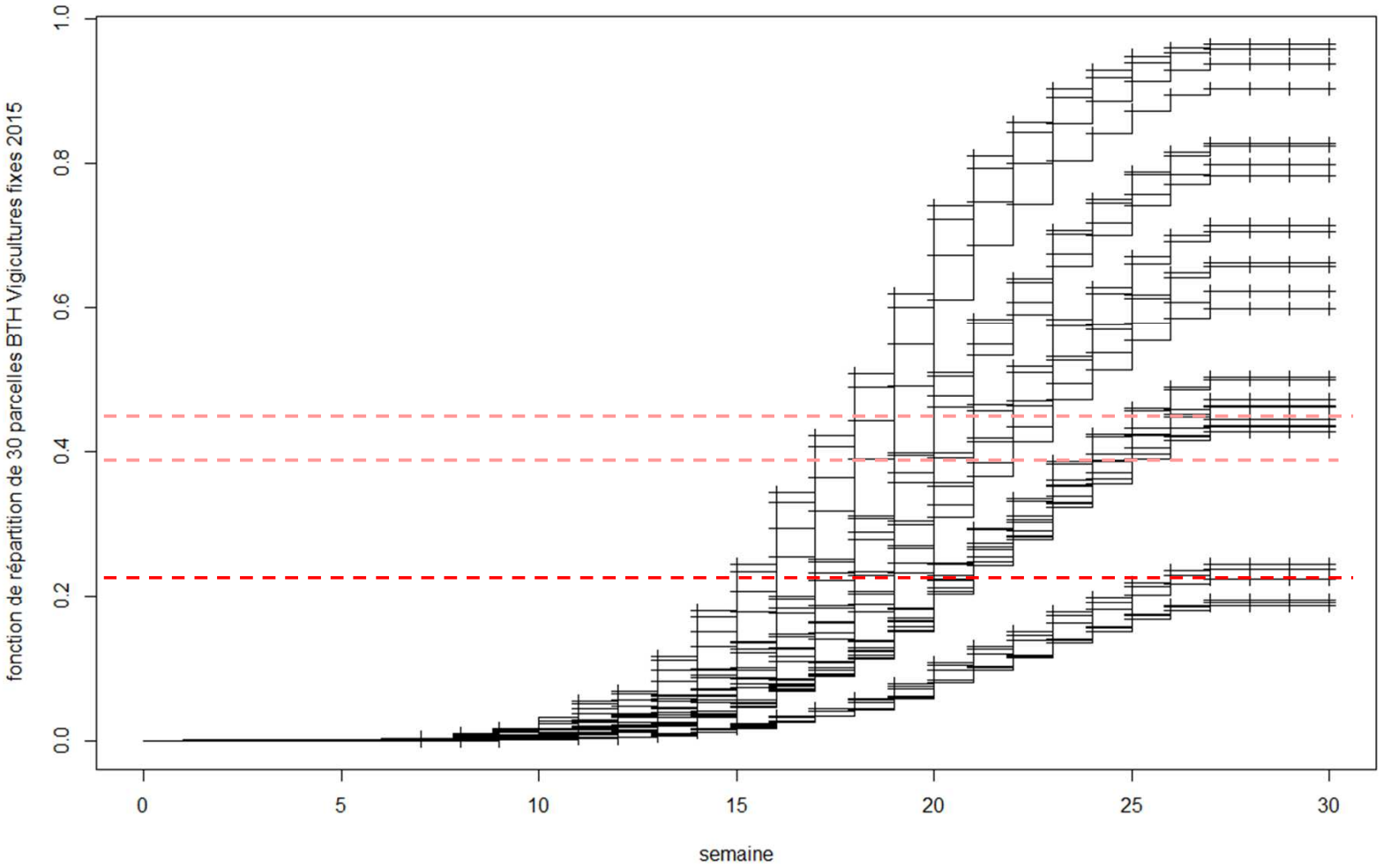
Validation et évaluation du modèle

…

Seuil 3

Seuil 2

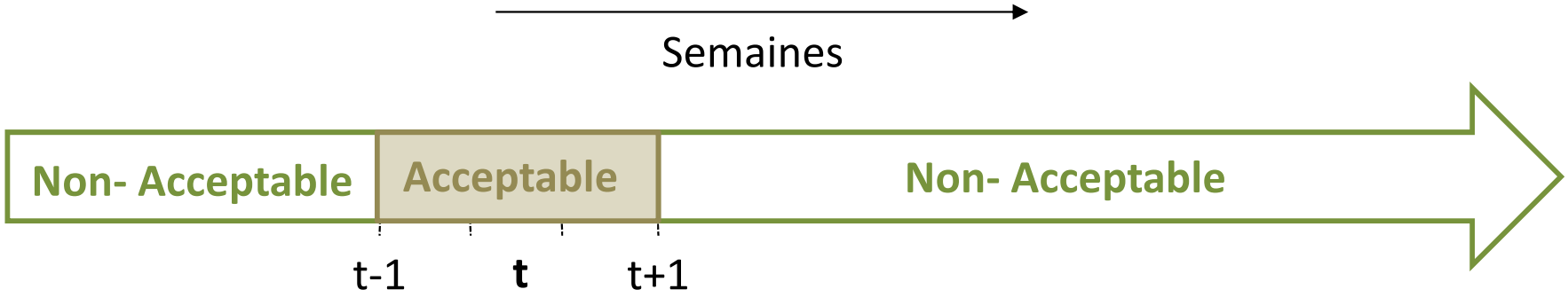
Seuil 1



- Seuil 1 -> Date de visite 1 Seuil 2 -> Date visite 2 …
- Optimiser les seuils de décision = probabilité à partir de laquelle l'évènement infestation a le plus de chance de se produire

=> critère d'acceptabilité des résultats

Validation et évaluation du modèle



Nature de la prédiction

t : semaine d'apparition réelle

**Acceptable lorsque modèle prédit de une semaine avant jusqu'à deux semaines après l'apparition réelle de la maladie*

Validation et évaluation du modèle

Apparition_prédite	Apparition_réelle	
	NON	OUI
Acceptable	Vrai Négatif (VN)	Vrai Positif (VP)
Non Acceptable	Faux Positif (FP)	'Faux Négatif' (FN)

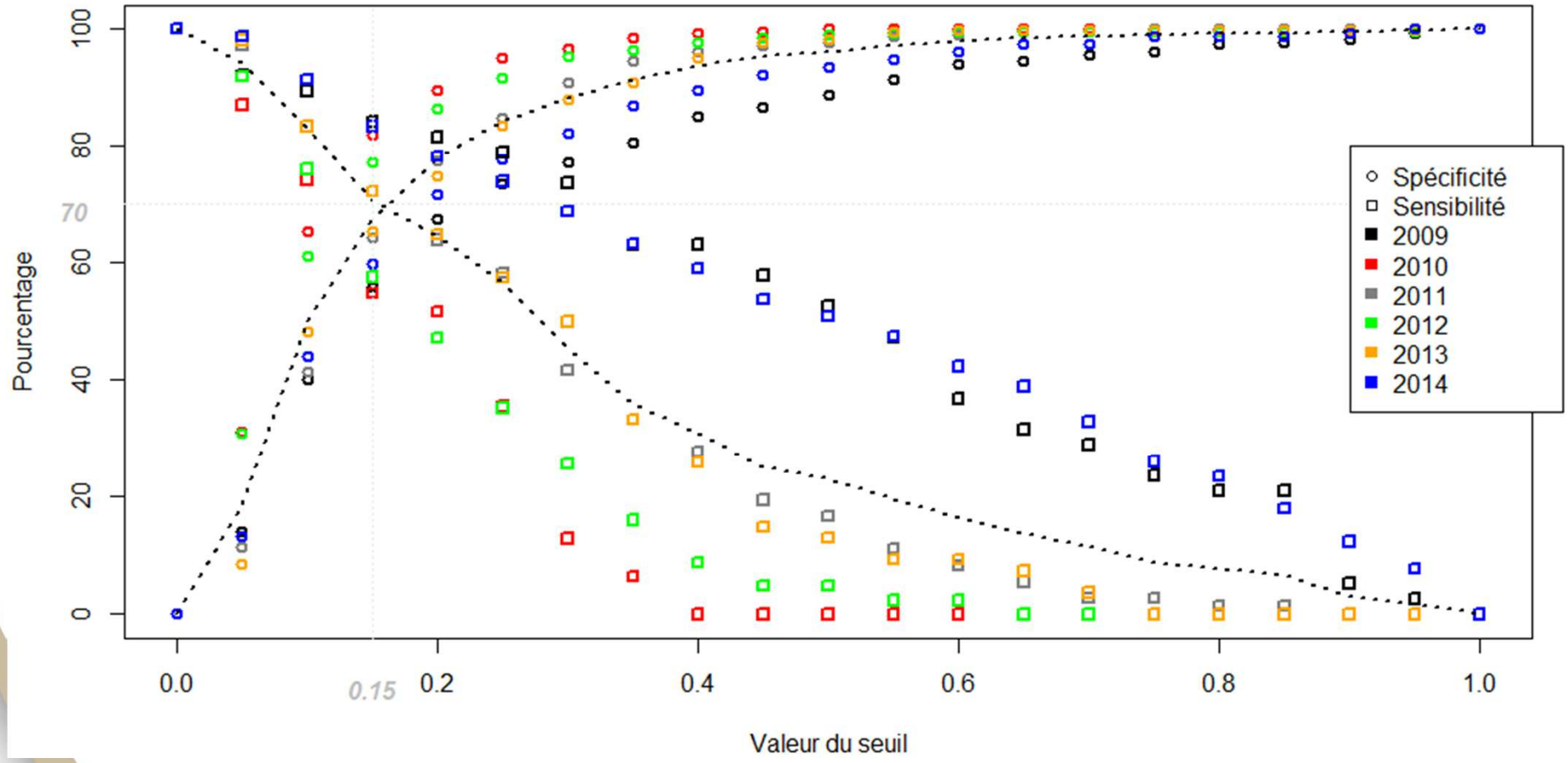
« Sensibilité » acceptable *(= capacité à dire 'oui' dans le bon créneau)*
= proportion de vrais positifs (PVP) parmi les cas R=1
= VP/(FN+VP)
Taux de FN = 1- « Sensibilité » = FN /(FN+VP)

« Spécificité » acceptable *(= capacité à dire 'non')*
= proportion de vrais négatifs (PVN) parmi les cas R=0
= VN/(FP+VN)
Taux de FP = 1- « Spécificité » = FP /(FP + VN)

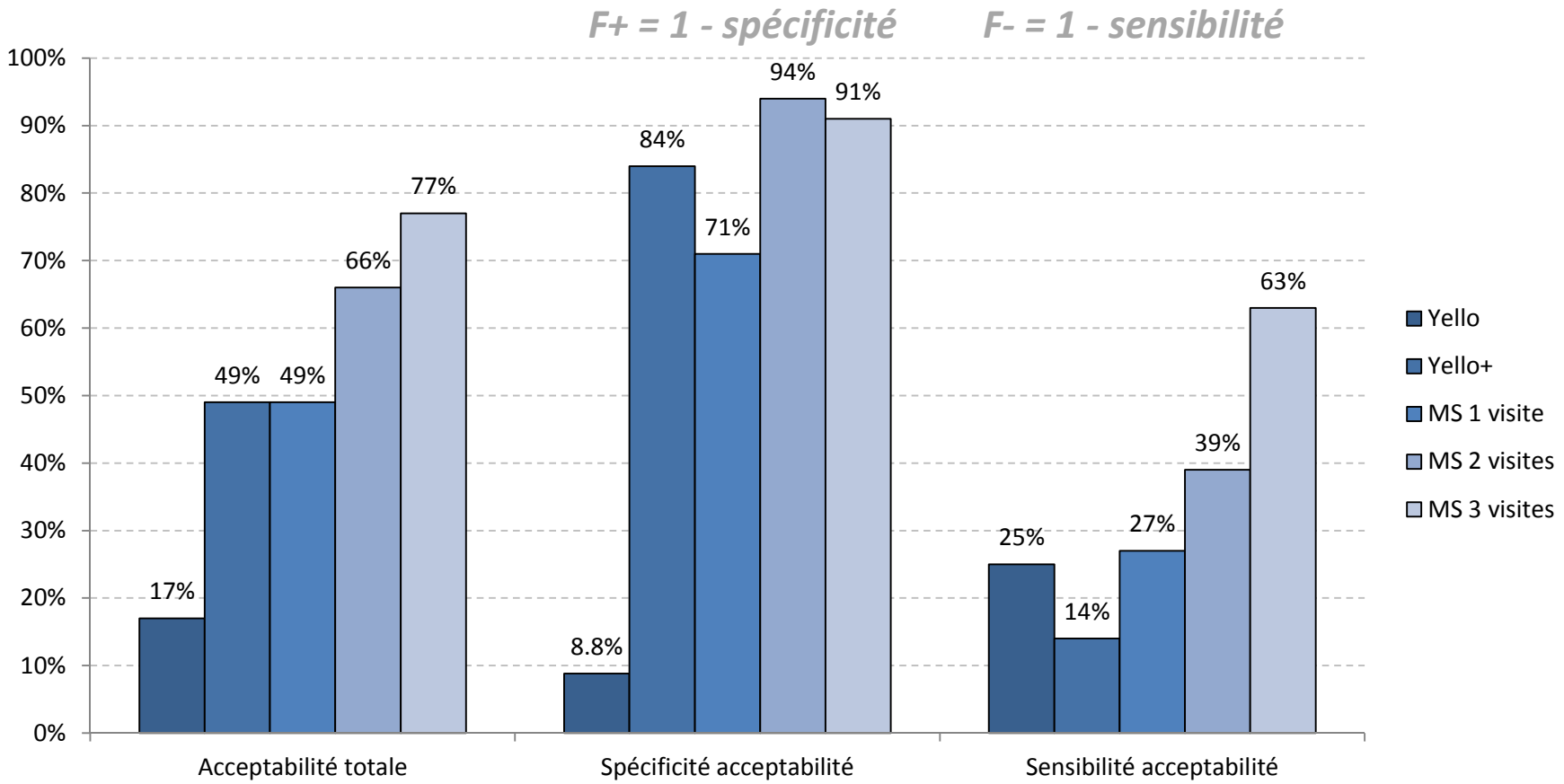
« Spécificité » acceptable + « Sensibilité » acceptable
= Acceptabilité totale

Validation et évaluation du modèle

Spécificité et sensibilité du modèle de Cox en fonction des seuils de décision considérés
(travaux sur la fonction de répartition obtenus lors de la cross validation par année)



Validation et évaluation du modèle

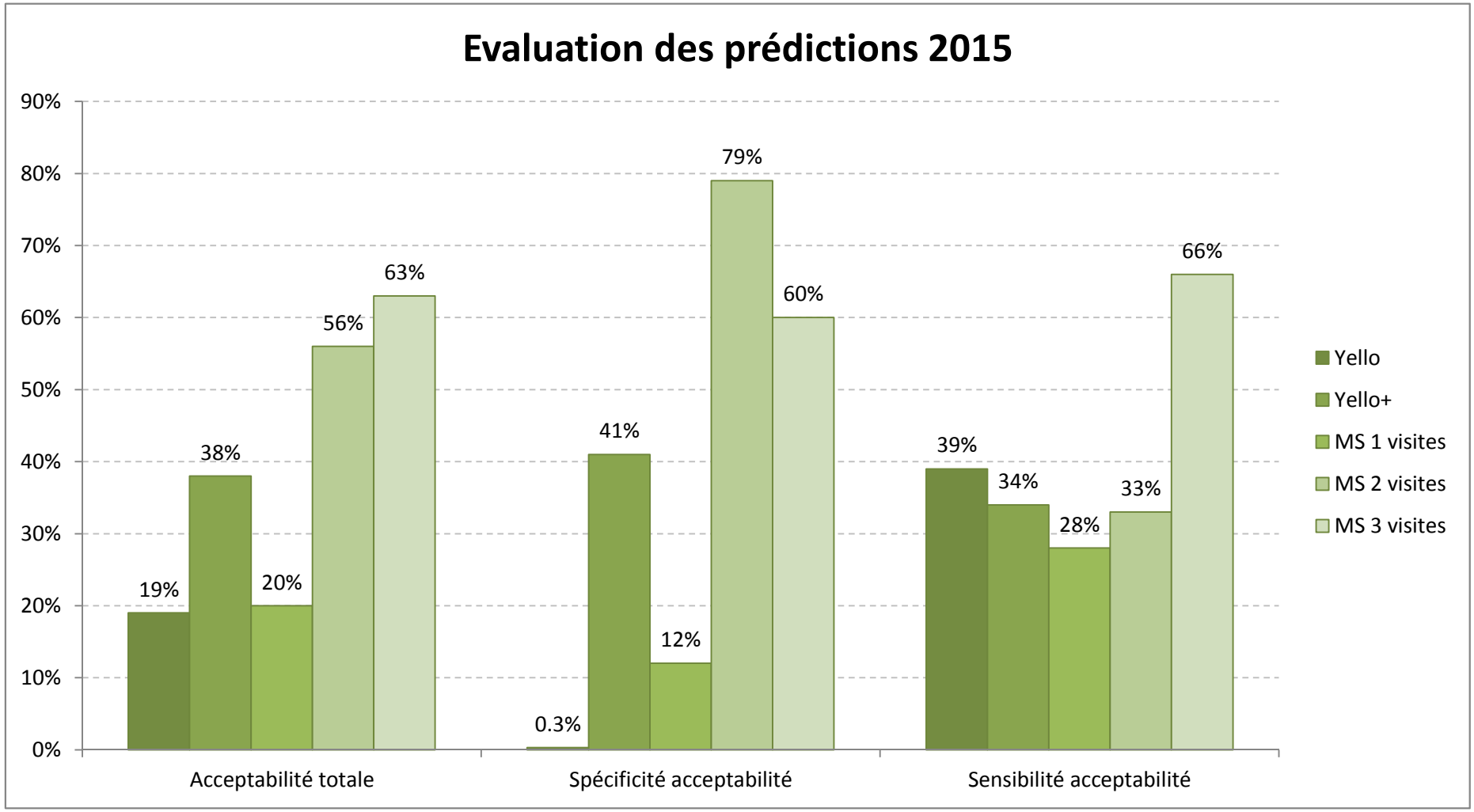


Capacité à dire non

Capacité à dire oui dans le bon créneau

Validité des résultats de Yello et Plus concernant la date d'alerte ET de traitement

Validation et évaluation du modèle



Capacité à dire non

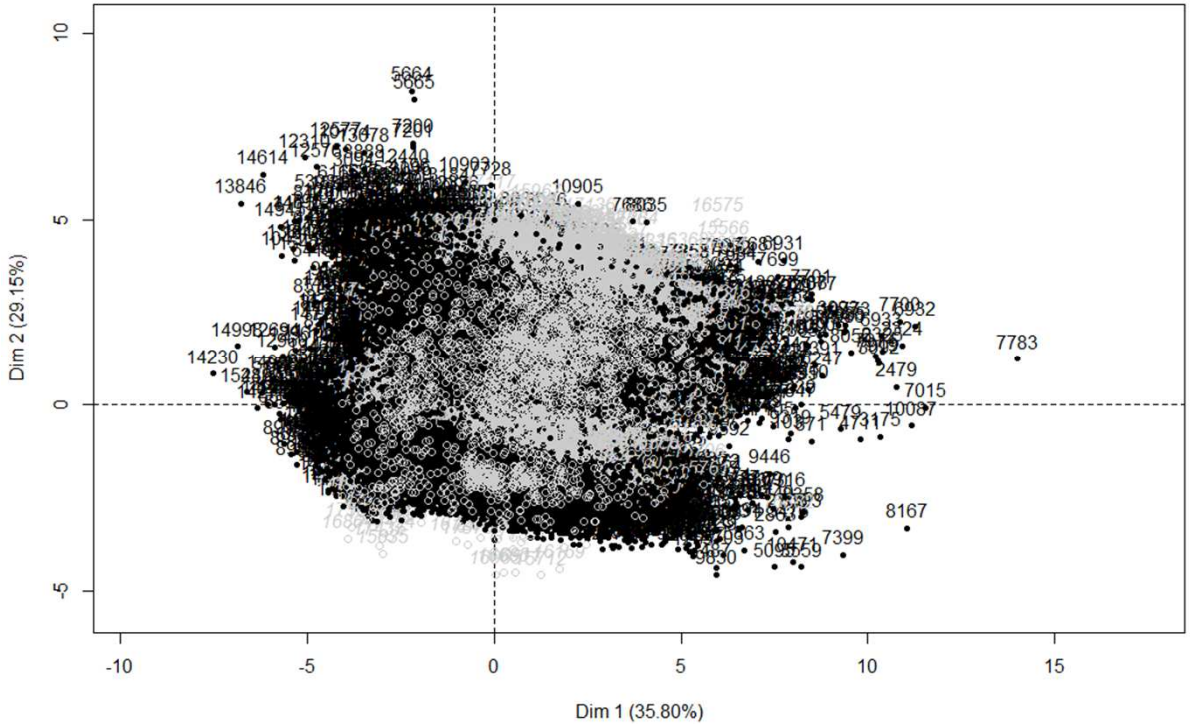
Capacité à dire oui dans le bon créneau

Validité des résultats de Yello et Plus concernant la date d'alerte ET de traitement

Validation et évaluation du modèle

- Comparer répartition des variables du JDD ajusté avec répartition des variables d'un JDD toutes stations depuis 1995 (380 stations) ; variété Apache et Alixan semis 20/10

ACP des variables CrustY du JDD ajustement (en gris) et JDD stations météo France entière depuis 1995 (en noir)



- Les variables du JDD ajusté semblent représentatives des variables France entière
  ⇒ **CrustY semble encore viable pour les années à venir**

Perspectives

Méthode intéressante qui nécessite d'être explorer :

- **Prise en compte de variables dynamiques**
- **Autres approches de survie**
- **Application à plusieurs évènements**
- **Ajustement d'un modèle par groupe de sensibilité variétale**