Etapes clés pour construire un modèle statistique

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

La modélisation des maladies : difficultés rencontrées

Définir le besoin réel de l'utilisateur et le(s) bon(s) critère(s) pour la prise de décision

Définir les modèles adaptés pour chaque pathosystème, selon les objectifs (traiter, observer …) et les jeux de données disponibles

Avoir à disposition des jeux de données adaptés à l'objectif fixé :

- Peu d'essais dédiés (sauf pour validation d'usage, seuils de traitement)
- Vigicultures : besoin important de nettoyage
- Essais variétés : une notation fin de cycle (aucune dynamique, pas de date d'apparition)
- Essais positionnement : peu nombreux
- Essais produits (parcelles NT) : peu nombreux

Choisir les variables agro-climatiques d'intérêt :

- Window pane ? Comment gérer les interactions entre variables climatiques?
- Liste de variables finies (dire d'expert, biblio etc.) puis sélection par méthodes statistiques (random forest, régression PLS etc.). Pb de corrélation entre toutes les variables.
- Utilisation en campagne => des variables agro-climatiques qui ne vont pas jusque des stades trop tardifs, l'idéal = variables quotidiennes mais difficile à mettre en œuvre dans un modèle stat.

Valider les modèles :

- Ajustement de plusieurs modèles : comment faire un choix ?
- Agrégation de modèles ?

Combiner modèles épidémiologiques/nuisibilité et décisionnels :

- Définir des seuils pour déclenchement des traitements, effets de seuil ?

Communiquer sur les sorties :

- Quels critères communiqués, critères agrégés ou indicateurs élémentaires ?
- Prise en compte de l'incertitude ?

Quelle question ?

Question posée ?	Indicateurs / Critères	Maladies	Echelle	Décision
La maladie sera t'elle présente cette année ?	0/1	Toutes (PV, oïdium, rouilles)	Région Parcelle	Stock fongi
A quelle date va-t-elle apparaître ?	Date ou Nb jours avant apparition	Rouilles	Parcelle	Observation
Comment va-t-elle évoluer ?	Symptômes	Septoriose, rouilles	Région Parcelle	Dates intervention
A quelle nuisibilité dois-je m'attendre ?	Nuisibilité Prix blé/fongi	Toutes	Région	intervention

Quelle question ?

- Vérifier la cohérence de l'objectif avec les données disponibles !

- Sinon reformuler….(souvent synonyme de : révision des ambitions à la baisse) ou trouver des données adaptées à l'objectif

Quelles données ?

Existe-t-il des données déjà disponibles ?

Faut-il acquérir de nouvelles données ?

Quelles données ?

Question posée ?	Quelles types de données pour y répondre?	Sources de données actuelles
La maladie sera t'elle présente cette année ?	Notations symptômes, qPCR	Vigicultures Essais Variétés/fongi Enquêtes agriculteurs
A quelle date va-t-elle apparaître ?	Date d'apparition de la maladie, Notations symptômes en dynamique	Vigicultures Essais fongi
Comment va-t-elle évoluer ?	Notations symptômes en dynamique	Vigicultures Essais fongi
A quelle nuisibilité dois-je m'attendre ?	Ecarts T-NT	Base T-NT
Quelle est la date de traitement optimale ?	Seuils de décision à partir d'indicateurs : date apparition, nuisibilité, dynamique etc.	Essais positionnement

Validation et Exploration des données

- Avant toute analyse stat

  => examen initial ou préliminaire des données
- Doit tenir compte de l'objectif de l'étude, de la manière dont les données ont été collectées
- Permet d'orienter le choix d'une méthode d'analyse
- Doit être complété pour vérifier les conditions d'application des méthodes envisagées

Validation et Exploration des données

- Examen visuel des données (ssi "small data")
- Stat descriptive : graphiques, paramètres stat
- Stat inférentielle : ajustement d'une loi, données aberrantes, corrélation, indépendance, etc.
- Parfois : stat multivariée

Réf : Dagnelie, T2, § 2.3

Validation et Exploration des données

• est-ce qu'il y a des données manquantes ? Si oui, quelle est leur origine ? Quelle est leur répartition ?

• les données sont-elles brutes ou calculées ? Si elles sont calculées, comment ?

• comment sont obtenues les données (appareil de mesure, échelle de notation, …) ?

• vérifier la nature des variables après importation dans un logiciel

• vérifier la cohérence des données :

- Date de récolte postérieure à la date de semis
- Précipitations >= 0
- Saint-Hilaire-en-Woëvre, Saint-Hilaire, Saint Hilaire, St-Hilaire, SAINT-HILAIRE, SAINTHILAIRE
- Nombre de décimales
- Présence de caractères indésirables (#VALEUR!, NA, ?, …)
- Codage des données manquantes ("", ?, NA, 0, 999, …)
- Codage des modalités (1=OUI, 2=NON, 3=?)
- …

Validation et Exploration des données

- graphiques :
  - Univariés : boxplot, histogramme, diagr. en bâtons, tsplot
  - Bivariés : nuage de points, graph univarié par catégorie
  - Multivariés : suite à une AF
- calcul de paramètres statistiques :
  - min, max, moyenne, variance, …
  - Corrélation, khi2, distance
  - Centrage - réduction
- Vérification des conditions d'application de la méthode stat (normalité, …)

Choix de la méthode modélisation

METHODES DESCRIPTIVES

- n variables quantitatives → **ACP**
- 2 variables qualitatives → **AFC**
- n variables qualitatives → **ACM**
- 2 ou plusieurs groupes de variables qualitatives → **AFM**
- n variables qualitatives et/ou quantitatives → **CAH**
- Matrice de distances entre individus → **CAH**, **MDS**

Variable(s) Y « à expliquer » MODELISATION

- Une variable Y quanti
  - X quanti et/ou quali → **Modèle linéaire et non linéaire fixe et mixte**; **Modèle linéaire généralisé fixe et mixte**; **Régression PLS**
  - X quanti → **Régression pénalisée (ridge, lasso, …)**
- Une variable Y quali
  - X quanti → **Analyse discriminante**
  - X quanti et/ou quali → **Régression logistique**; **Régression PLS2**
- Une variable Y quali ou quanti
  - X quanti et/ou quali → **Segmentation**; **Forêts aléatoires**
- Plusieurs variables Y quanti → **Régression PLS2**

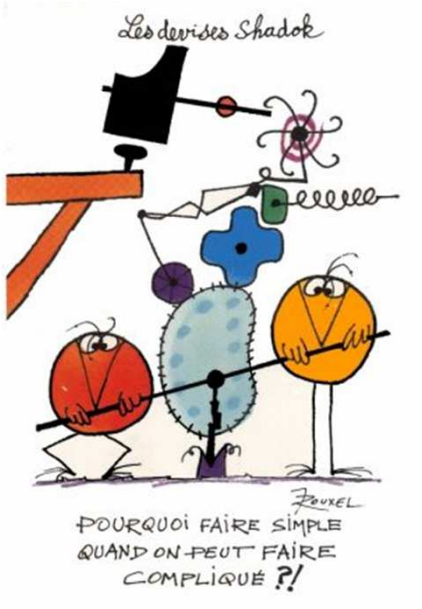
Éléments supplémentaires à prendre en compte :

- Dimension du jeu de données (nb individus/nb variables)
- Colinéarité
- Distribution de Y et des X
- …

Choix de la méthode modélisation

- Sur-ajustement

- **Pour éviter** un éventuel sur-ajustement, deux principes doivent gouverner le choix d'un modèle :
  - principe de parcimonie (ne pas utiliser un modèle plus complexe que nécessaire).



  - principe de causalité (il faut pouvoir justifier le modèle utilisé par la connaissance du phénomène étudié).

L'ACP

Données

Variables quantitatives
Possibilité d'avoir plus de variables que d'individus

But

Etudier **les liaisons linéaires entre les variables** :

- Identifier les variables corrélées à une variable d'intérêt
- Identifier des variables redondantes

Distinguer **des profils d'individus** qui se « ressemblent »:

- Identifier des individus aberrants
- Identifier des groupes d'individus

Principe

Projection des observations qui sont dans un espace à n dimensions (n = nombre de variables) dans un espace de dimension réduite (souvent un plan).

Représentation graphique en 2D de l'ensemble du jeu de données.	Ne prend en compte que les liaisons linéaires

La classification (CAH, k-means…)

Données

Variables quantitatives ou qualitatives

But

Faire des groupes d'individus et les caractériser

Principe

Utiliser des algorithme pour faire des groupes d'individus . Au sein d'un groupe les individus sont les plus homogènes possibles ET les groupes diffèrent le plus possibles les uns des autres au regard des variables du jeu de données.

	
Typologie automatique	Difficulté de déterminer les groupes à retenir
	Les groupes peuvent être différents selon l'algorithme utilisé

La segmentation

Données

Variables quantitatives ou qualitatives

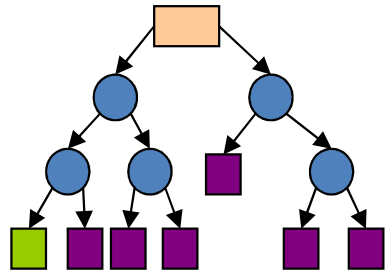
But

Modéliser une variable Y à partir d'un ensemble de variables « explicatives » Xi sous forme d'un arbre de décision :

- prédiction
- description de la relation entre Y et X

Principe

A chaque nœud, les individus sont divisés en 2 sous-ensembles selon la valeur d'une variable explicative Xi. La segmentation s'arrête lorsque le nombre d'individus restant est trop faible, ou si ça n'explique plus « assez ».



Prise en compte des relations non linéaires entre les X et le Y	Nécessite un nombre important d'individus (d'autant qu'on a beaucoup de variables) sinon risques d'instabilité
Prise en compte des interactions entre les Xi	
Sortie facile à communiquer (arbre de décision)	

Les forêts aléatoires (Random Forest)

Données

Variables quantitatives ou qualitatives

But

Modéliser une variable Y à partir d'un ensemble de variables « explicatives » Xi :

- prédiction
- quantifier importance des variables explicatives
- obtenir une matrice de distances entre individus

Principe

Partir du principe de la segmentation pour faire plusieurs arbres et les agréger pour former une forêt.

	
Prise en compte des relations non linéaires entre les X et le Y	Un peu « boîte noire » : impossible de visualiser le « modèle final » (pas d'arbre unique, pas d'équation…)
Prise en compte des interactions entre les Xi	
+ robuste que la segmentation (moins sensible au sur-ajustement)	

Le modèle mixte

Données

Variables quantitatives ou qualitatives

But

Pouvoir prendre en compte dans le modèle:
- une structure particulière des données (variances hétérogènes, autocorrélation)
- des effets aléatoires (split-splot, criss-cross, alpha-plan, réseau d'essais, …)

Autocorrélation : lorsque 2 observations proches dans le temps (ou l'espace) se ressemblent plus que 2 observations éloignées (ex : mesures répétées sur le même individu)

Facteur à effet fixe : lorsqu'on ne s'intéresse qu'aux modalités effectivement observées dans l'expérimentation (ex: traitement)

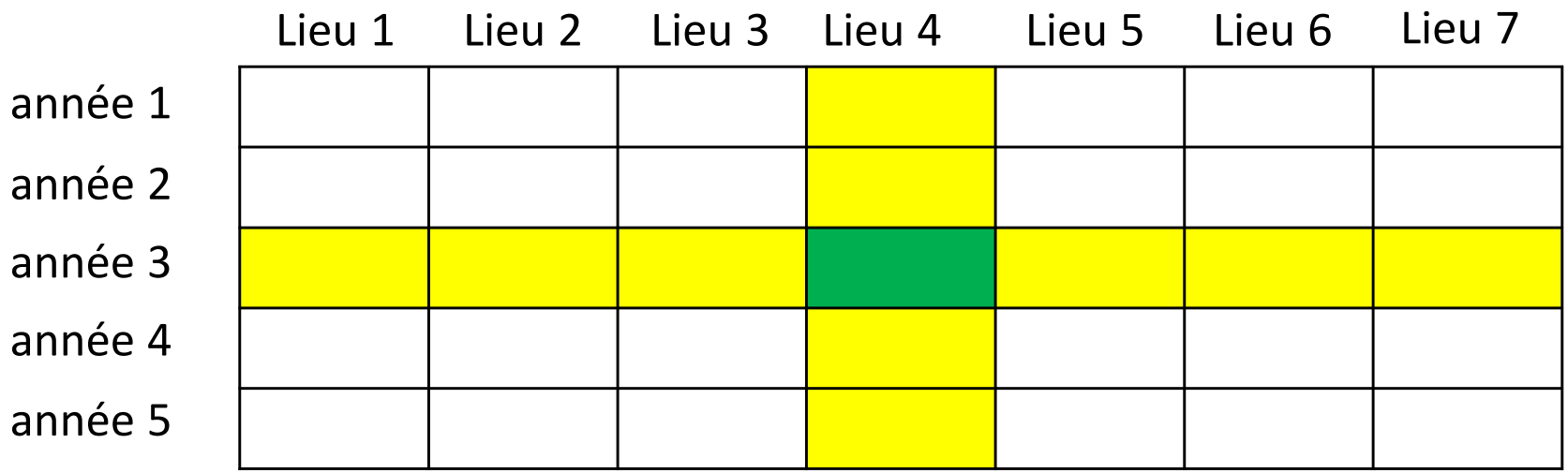
Facteur à effet aléatoire : lorsque les modalités observées dans l'expérimentation sont issues d'une population de modalités et que l'on souhaite généraliser les conclusions à l'ensemble de la population (ex : plante, microparcelle,…)

Permet de modéliser une large gamme de situations (voir But)	Théorie pas complètement aboutie (ex :test des effets aléatoires)
Possibilité d'avoir un dispositif incomplet	Interprétation des sorties parfois difficile

Validation et évaluation du modèle

L'évaluation d'un modèle peut se faire sur :

- Validation croisée : schéma de validation croisée



Validation croisée : pour chaque lieu:année, on retire la croix jaune et on prédit le lieu:année vert

Validation et évaluation du modèle

L'évaluation d'un modèle peut se faire sur :

- Données indépendantes :

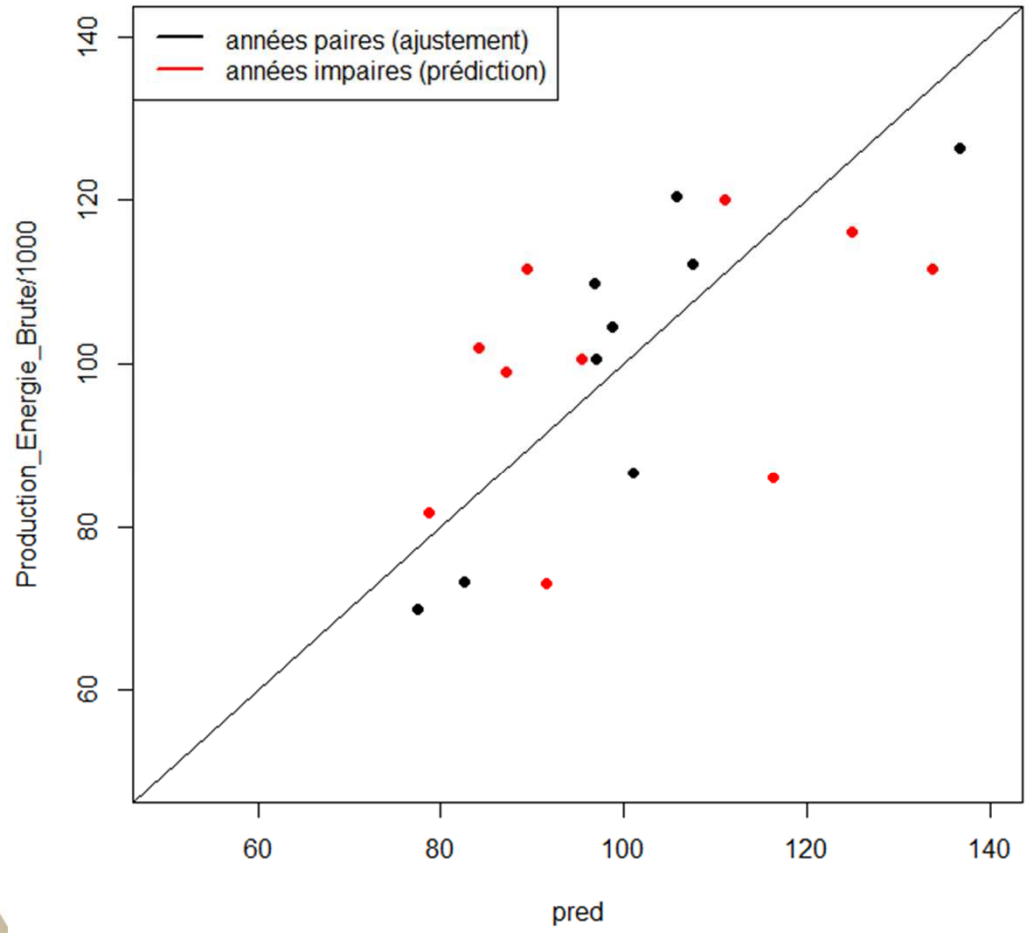
Exemple : rdt du système « monoculture de blé »

Modèle retenu après sélection de variables :

```
Production_Energie_Brute ~ IFT_Insecticide +  Ecart_Epandage_20mm

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         110.9265     9.3100  11.915 2.12e-05 ***
IFT_Insecticide      14.9449     5.3072   2.816   0.0305 *
Ecart_Epandage_20mm  -1.1089     0.3149  -3.521   0.0125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Validation et évaluation du modèle



R^2 ajustement = 0.71
R^2 prédiction = -0.32