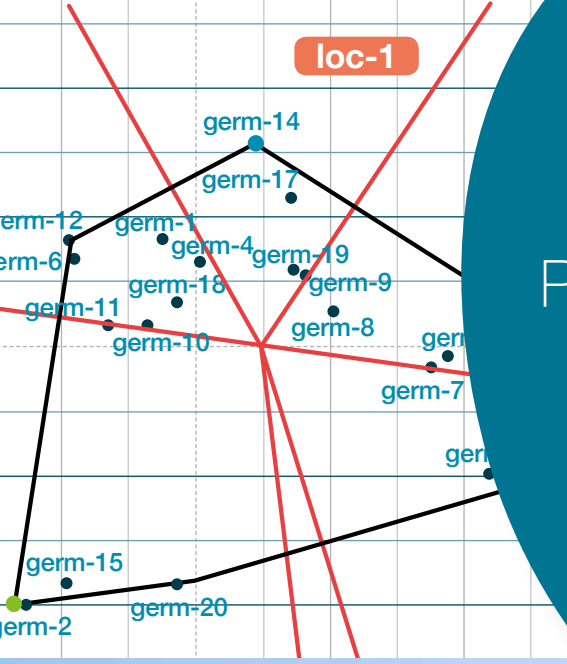
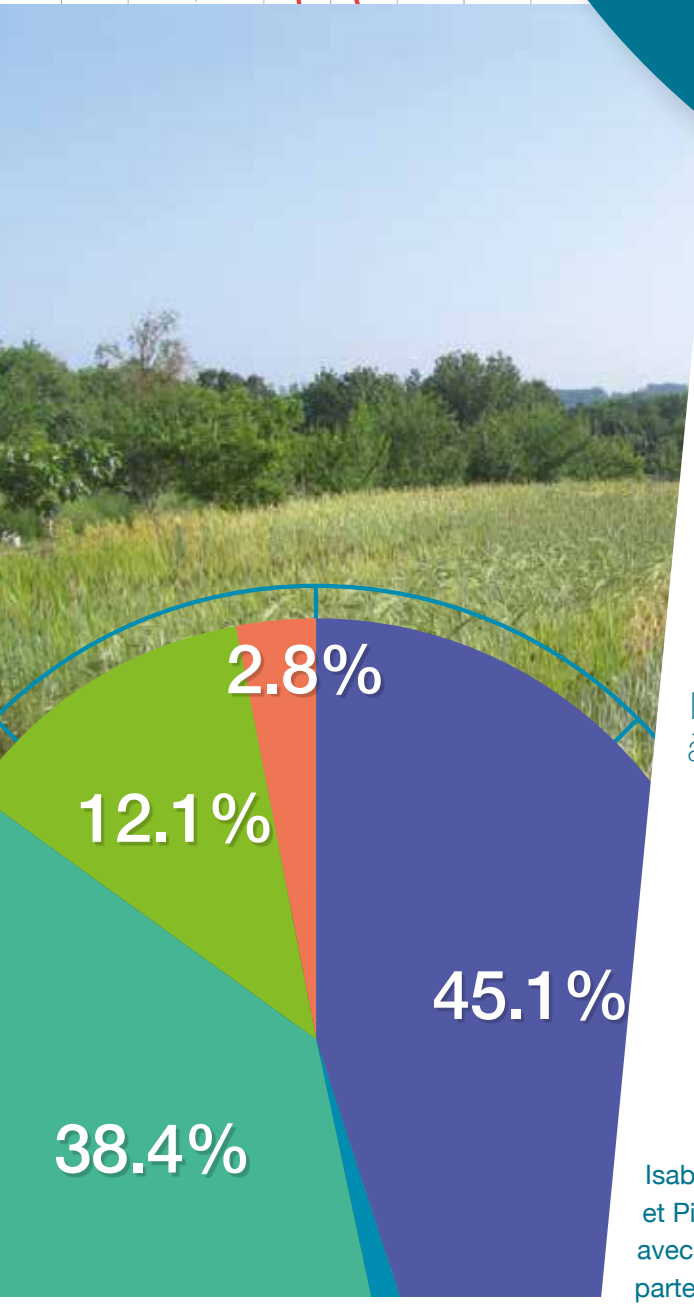


MÉTHODES ET OUTILS POUR LA SÉLECTION DÉCENTRALISÉE À LA FERME



Brochure #3

Ce guide technique présente des dispositifs expérimentaux ainsi que des outils et méthodes statistiques utiles pour la sélection décentralisée à la ferme.

Isabelle Goldringer (INRA) et Pierre Rivière (RSP), avec la contribution des partenaires de Diversifood

DIVERSIFOOD

L'objectif de DIVERSIFOOD est d'ancrer la biodiversité cultivée et de soutenir les réseaux d'acteurs locaux dans des systèmes alimentaires de qualité.

SOMMAIRE

INTRODUCTION

La sélection participative est généralement basée sur une évaluation et une sélection décentralisées à la ferme, qui nécessitent des méthodes et des outils particuliers. Cette brochure décrit les dispositifs expérimentaux ainsi que les méthodes et outils statistiques adaptés pour la sélection décentralisée à la ferme. Bien que la dimension participative soit essentielle dans des programmes de sélection participative pour garantir l'autonomisation de tous les acteurs (agriculteurs, animateurs, transformateurs, jardiniers, consommateurs, etc.) et pour répondre à leurs besoins réels (Sperling et al. 2001), les méthodes participatives ne sont pas présentées ici. Celles-ci sont décrites en détail dans une autre brochure Diversifood intitulé *«Méthodes et cadre méthodologique pour les approches multi-acteurs et la sélection végétale participative»*.

Cette brochure décrit différents dispositifs expérimentaux ainsi que les méthodes statistiques d'analyse pouvant être réalisées, en fonction des objectifs et des contraintes expérimentales du programme de sélection et du groupe d'agriculteurs. La manière d'identifier et de choisir les dispositifs et les méthodes les plus pertinents est basée sur un arbre de décision (Figure 1). La brochure fait également référence à un logiciel qui permet d'appliquer ces méthodes: le package R PPBstats (Rivière, Van Frank, Munoz & David 2018) avec une documentation complète (Rivière, Goldringer & Vindras 2018).

L'INTÉRÊT DE DÉCENTRALISER LA SÉLECTION

On peut modéliser les réponses à la sélection observées dans les fermes à partir de Bernardo (2002) et Gallais (1990), tel que présenté ci-après.

Lors de l'évaluation et de la sélection dans plusieurs environnements, la valeur phénotypique d'un caractère observé sur un individu dans un environnement donné peut être décrite comme la somme de son effet génétique (ou de son potentiel génétique global, G), de l'effet de l'environnement dans lequel il se trouve (E) et de l'interaction entre les deux (G× E), itous ces effets étant modélisés comme des variables aléatoires. Ainsi, le modèle s'écrit comme : P=G+E+G×E+e avec e t l'effet résiduel aléatoire au sein de chaque environnement qui suit une distribution normale N (0, σ^2).

En sélection centralisée classique, l'objectif est de prédire le potentiel génétique global (G) des candidats à la sélection en détectant les plus fortes valeurs. Cette démarche fait l'hypothèse que ce potentiel pourra s'exprimer dans tous les champs des paysans. Ces potentiels génétiques sont prédis à partir des valeurs phénotypiques moyennes sur tous les environnements (qui sont des stations expérimentales en général). Dans ce cadre, l'héritabilité au sens large qui indique la capacité à prédire ce potentiel dans le dispositif (compris entre 0 et 1) est :

$$h^2_{sl} = \frac{\text{var (G)}}{\text{var (G)} + \frac{1}{nE}(\text{var(E)} + \text{var (G}\times\text{ E)}) + \frac{1}{nE\ x nR}\text{var (e)}}$$

avec nE (resp. nR) le nombre d'environnements (resp. le nombre de répétitions dans chaque environnement). Comme les effets environnementaux et les interactions limitent la précision de la prédiction, la stratégie est d'augmenter le nombre d'environnements et d'utiliser des environnements qui sont homogènes et similaires pour minimiser les interactions G×E.

Au contraire, en sélection décentralisée à la ferme, il a été montré que les environnements sont très contrastés en raison de la grande diversité des conditions pédoclimatiques et des pratiques agroécologiques et que les interactions G×E peuvent être fortes (Desclaux et al. 2008). Par conséquent, la prédiction des valeurs génétiques globales (G) n'est pas intéressante et l'objectif est plutôt de prédire la valeur génétique « locale », qui inclue également l'interaction avec l'environnement local, i.e.: Glocij = Gi + (GxE) ij.

La variance génétique dans chaque environnement local peut être décrite comme :

var (Gloc) = var (G) + var (G×E) et l'héritabilité pour prédire la valeur génétique locale basée sur la valeur phénotypique observée dans l'environnement local s'écrit:

$$h^2_{sl} = \frac{\text{var (Gloc)}}{\text{var (Gloc)} + \frac{1}{nR}\text{var(e)}} = \frac{\text{var (G) + var (G×E)}}{\text{var (G) + var (G×E)}\ \frac{1}{nR}\text{var(e)}}$$

On notera que l'interaction G×E se trouve à la fois au dénominateur et au numérateur, ce qui permet de limiter ses effets sur la précision des prédictions. Il s'avère que lorsque qu'une grande diversité d'environnements et de pratiques agroécologique sont rencontrés, alors la sélection décentralisée est un levier clé pour sélectionner des variétés adaptées à des agrosystèmes locaux.


© F. Rey

PLANS EXPÉRIMENTAUX ET MÉTHODES STATISTIQUES EN FONCTION DES OBJECTIFS

L'analyse des données issues des programmes de sélection participative vise cinq principaux objectifs qui structurent l'arbre de décision (Figure 1). Les cinq objectifs ci-dessous s'appliquent à quatre types de caractères **(agronomiques et nutritionnels, sensoriels, données moléculaires, réseau de circulation des semences)**

- **Pour améliorer la prédiction d'une variable cible pour la sélection** à travers l'analyse des caractères agronomiques et nutritionnels.
- **Pour comparer différentes variétés ou populations évaluées pour la sélection dans différents lieux** à travers l'analyse des caractères agronomiques, nutritionnels et sensoriels.
- **Pour étudier les réponses des variétés ou des populations sous sélection sur plusieurs environnements** à travers l'analyses des caractères agronomiques et nutritionnels.
- **Pour étudier la structure de la diversité afin d'identifier des parents à croiser sur la base de leur complémentarité ou de leurs similarité pour certains caractères** à travers l'analyse des caractères agronomiques et nutritionnels et desdonnées moléculaires.
- **Pour étudier les réseau de circulation de semences** à travers l'analyse de la topologie des réseaux.

Pour chaque objectif, plusieurs méthodes sont disponibles. Celles-ci sont basées sur différents dispositifs expérimentaux, en fonction des objectifs et des contraintes expérimentales du programme de sélection et du groupe d'agriculteurs impliqués (Figure 1). Les contraintes à prendre en compte sont en particulier : le nombre de parcelles par site, le nombre de sites, le nombre de variétés répétées dans et entre les sites, qui dépendent tous de la quantité de semences disponible. Une fois qu'un dispositif expérimental et une méthode d'analyse sont sélectionnés, le semis peut être fait !


© iStock


© iStock

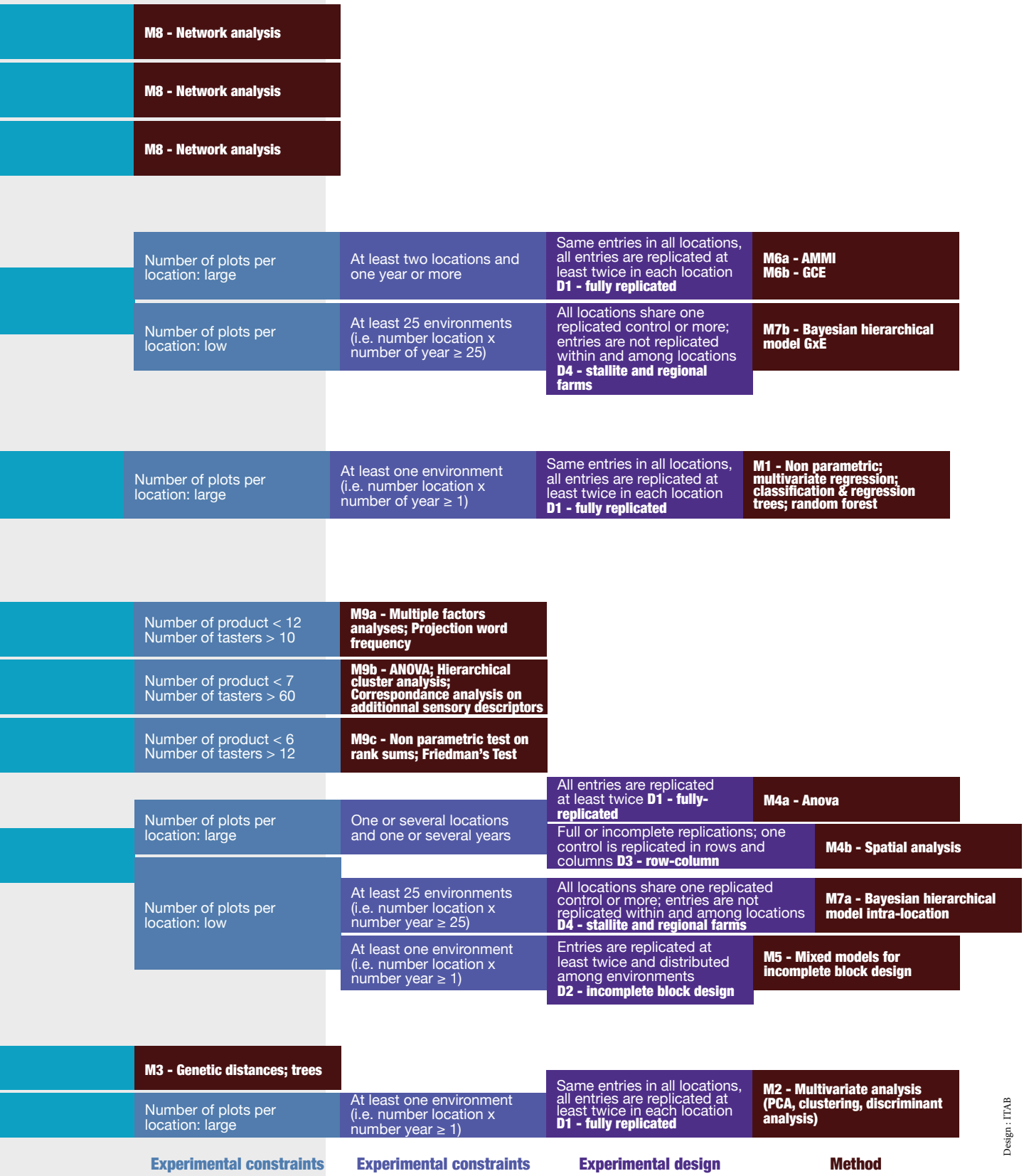
DISPOSITIFS EXPÉRIMENTAUX ET MÉTHODES STATISTIQUES POUR LA SÉLECTION PARTICIPATIVE

ARBRE DE DÉCISION

Dans ce qui suit, chaque branche de l'arbre est expliquée à l'aide d'un exemple de dispositif expérimental et d'analyse dans la section correspondante. Les modèles et les méthodes sont présentés selon les quatre types de caractères **(agronomiques et nutritionnels, sensoriels, données moléculaires, réseau de circulation des semences)** auxquels ils s'appliquent. Ceux-ci constituent les principaux chapitres de la brochure.


© F. Rey

Figure 1 - Arbre de décision


M8 - Network analysis
M8 - Network analysis
M8 - Network analysis
Number of plots per location: large
At least two locations and one year or more
Same entries in all locations, all entries are replicated at least twice in each location D1 - fully replicated
M6a - AMMI
M6b - GCE
Number of plots per location: low
At least 25 environments (i.e. number location x number of year ≥ 25)
All locations share one replicated control or more; entries are not replicated within and among locations D4 - stallite and regional farms
M7b - Bayesian hierarchical model GxE
Number of plots per location: large
At least one environment (i.e. number location x number of year ≥ 1)
Same entries in all locations, all entries are replicated at least twice in each location D1 - fully replicated
M1 - Non parametric; multivariate regression; classification & regression trees; random forest
Number of product < 12
Number of tasters > 10
M9a - Multiple factors analyses; Projection word frequency
Number of product < 7
Number of tasters > 60
M9b - ANOVA; Hierarchical cluster analysis; Correspondance analysis on additionnal sensory descriptors
Number of product < 6
Number of tasters > 12
M9c - Non parametric test on rank sums; Friedman's Test
Number of plots per location: large
One or several locations and one or several years
All entries are replicated at least twice D1 - fully-replicated
M4a - Anova
Full or incomplete replications; one control is replicated in rows and columns D3 - row-column
M4b - Spatial analysis
Number of plots per location: low
At least 25 environments (i.e. number location x number year ≥ 25)
All locations share one replicated control or more; entries are not replicated within and among locations D4 - stallite and regional farms
M7a - Bayesian hierarchical model intra-location
At least one environment (i.e. number location x number year ≥ 1)
Entries are replicated at least twice and distributed among environments D2 - incomplete block design
M5 - Mixed models for incomplete block design
M3 - Genetic distances; trees
Number of plots per location: large
At least one environment (i.e. number location x number year ≥ 1)
Same entries in all locations, all entries are replicated at least twice in each location D1 - fully replicated
M2 - Multivariate analysis (PCA, clustering, discriminant analysis)
Experimental constraints
Experimental constraints
Experimental design
Method
Design : ITAB

ANALYSES DES CARACTÈRES AGRONOMIQUES ET NUTRITIONNELS

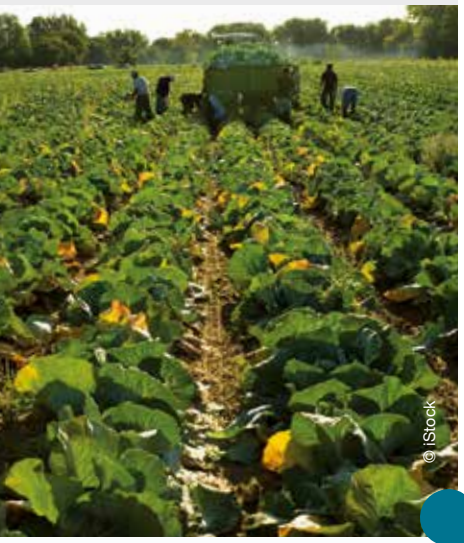

© Rupert Pessl

Les quatre principaux objectifs des analyses agronomiques sont de:

- **Améliorer la prédiction d'un variable cible pour la sélection.** Cela peut être réalisé à travers des méthodes non paramétriques telles que :
  - Des régressions multivariées, des arbres de classification et approches dites random forest **(M1)**, basées sur des dispositifs complets répétés. **(D1)**.
- **Étudier la structure de la diversité afin d'identifier des parents à croiser sur la base de leur complémentarité ou de leurs similarité pour certains caractères .**
  - Cela peut être réalisé à travers des analyse multivariée et de clustering **(M2)**.
  - Et peut être complété par des analyses moléculaires et des arbres de distance génétique **(M3)**.
- **Comparer différentes variétés ou populations évaluées pour la sélection dans chaque lieu.** Cela peut être réalisé à travers différents types d'analyse regroupées dans la famille 1 :
  - Anova classique **(M4a)** basée sur des dispositifs complets répétés **(D1)**,
  - Analyse spatiale **(M4b)** basée sur des dispositifs ligne/colonne **(D3)**,
  - Modèles mixtes **(M5)** basé sur des plans en blocs incomplets **(D2)**,
  - Modèle intra-lieu bayésien hiérarchique **(M7a)** basé sur des dispositifs fermes régionales et satellites **(D4)**.

Ces objectifs peuvent être complétés par des analyses organoleptiques (voir plus loin). Basé sur ces analyses, des objectifs spécifiques comme par exemple l'étude de la réponse à la sélection peuvent être conduits.

- **Étudier les réponses des variétés ou des populations sous sélection sur plusieurs environnements. Cela peut être réalisé à travers différents types d'analyse regroupées dans la famille 2:**
  - AMMI et GGE **(M6)** basées sur des dispositifs complets répétés **(D1)**,
  - Modèle bayésien hiérarchique G×E **(M7b)** basé sur des dispositifs fermes régionales et satellites **(D4)**,


© iStock

FORMAT DES DONNÉES

Selon les logiciels, le format des données sera différent. Néanmoins, les informations importantes nécessaires pour les analyses sont le lieu, l'année, la variété, le bloc, les coordonnées X et Y (ligne et colonne) suivies par les variables et leurs dates de mesure si elles sont disponibles.

EFFETS À ESTIMER ET TYPES D'ANALYSES

Les différents effets qui peuvent être estimés sont:

- **Germplasm:** fait référence à une variété ou à une population
- **Lieu**: fait référence à une ferme ou à une station où les essais sont conduits
- **Environnement:** fait référence à une combinaison d'un lieu et d'une année
- **Entrée:** fait référence à la présence d'un germplasm dans un environnement donné
- **Interaction:** fait référence à l'interaction entre le germplasm et l'environnement année

Deux familles d'analyse sont proposées:

- **Famille 1** rassemble des analyses qui estiment les effets dans chaque environnement (entrées). Cela permet de comparer différentes germplasmes dans chaque lieu et de tester s'il existe des différences significatives entre eux. Une analyse spécifique pour estimer la réponse à la sélection peut également être réalisée.
- **Famille 2** rassemble des analyses qui estiment la valeur globale des germplasmes, les effets environnementaux et les interactions. Cela consiste à analyser la réponse sur un réseau de lieux durant une ou plusieurs années. L'estimation des effets lieu et année séparés est possible en fonction du modèle. Une analyse spécifique visant à tester l'adaptation locale sur la base du modèle migrant versus résidant (Blanquart et al 2013) peut également être réalisée. L'objectif est d'étudier la réponse de différents germplasmes sur plusieurs lieux de sélection.

Les différents modèles et méthodes des familles 1 et 2 sont liés aux dispositifs expérimentaux décrits dans la section suivante et dans l'arbre de décision (Figure 1).

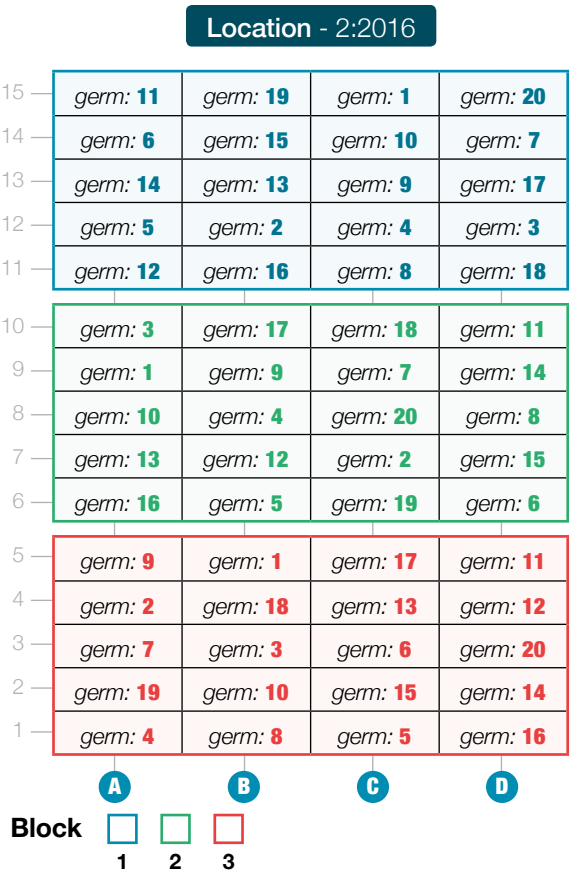
DISPOSITIFS EXPÉRIMENTAUX

Le dispositif expérimental est décrit par le nombre de parcelles par site, le nombre de lieux, le nombre de répétitions des différents germplasmes dans et entre les lieux. Vous trouverez ci-dessous des exemples de plusieurs dispositifs expérimentaux. Chaque dispositif expérimental est suivi d'une analyse spécifique, telle que décrite dans l'arbre de décision (Figure 1).

DISPOSITIFS COMPLETS RÉPÉTÉS (D1)

Dans les dispositifs complets répétés (Figure 2), tous les germplasmes (entrées) sont répétés et répartis aléatoirement dans les différents blocs.

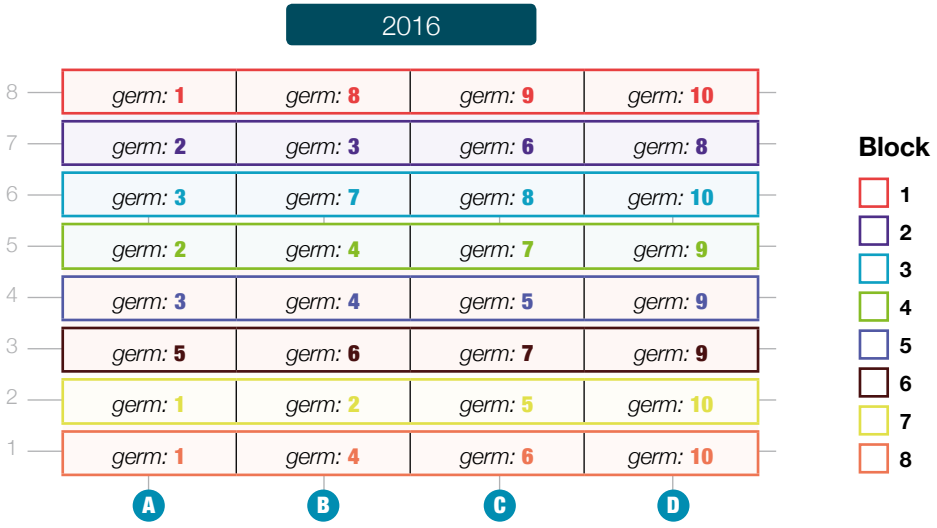
Figure 2: Dispositif complet répété où tous les germplasmes sont répétés trois fois dans des blocs complets.



BLOCS INCOMPLETS (D2)

Dans le dispositif en blocs incomplets (Figure 3), les germplasmes (entrées) ne sont pas répétés systématiquement au sein de chaque lieu. Chaque germplasme est retrouvé dans plusieurs lieux. Chaque bloc est une unité indépendante qui peut être allouée à n'importe quel lieu. Un agriculteur peut choisir de mettre en place un ou plusieurs blocs. Par conséquent, l'expérimentation peut être effectuée sur plusieurs lieux qui ne reçoivent pas le même nombre de blocs.

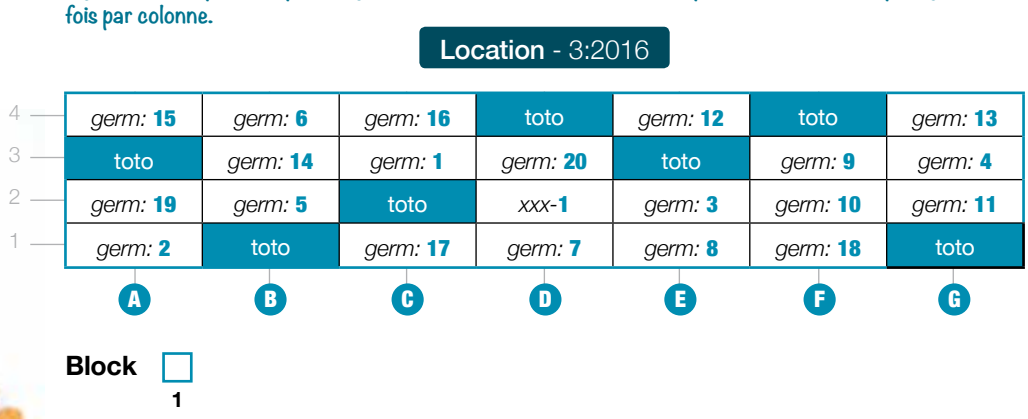
Figure 3: Exemple de dispositif en bloc incomplet où différents germplasmes sont répétés dans différents blocs. Les blocs représentent les lignes horizontales.



LIGNE-COLONNE (D3)

Dans un dispositif ligne colonne (Figure 4), un contrôle est répété une ou plusieurs fois dans chaque ligne et colonne pour capturer la variation environnementale au mieux.

Figure 4: Exemple de dispositif ligne colonne où le contrôle (toto) est répété une ou deux fois par ligne et une fois par colonne.



© iStock

FERMES RÉGIONALES ET SATELLITES (D4)

Dans ce dispositif, les essais à la ferme sont divisés en deux types: les fermes régionales et les fermes satellites (Figures 5 & 6). Les fermes régionales reçoivent plusieurs germplasmes dans deux blocs ou plus dont certains (les contrôles) sont répétés dans chaque bloc. Les fermes satellites ont un seul bloc et un seul germplasme (le contrôle) répété deux fois. Les agriculteurs choisissent pour leur ferme tous les germplasmes non répétés. Le nombre d'entrées peut varier d'une ferme à l'autre. A noter qu'au moins 25 environnements (lieu x année) sont nécessaires pour obtenir des résultats robustes.

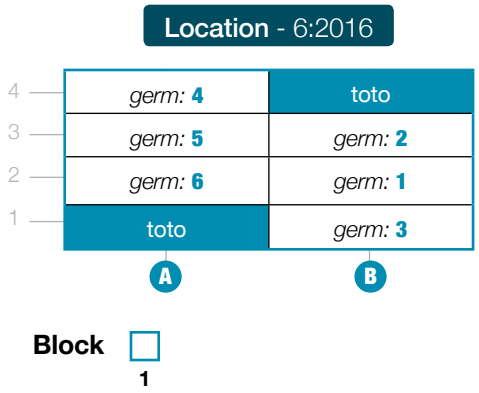
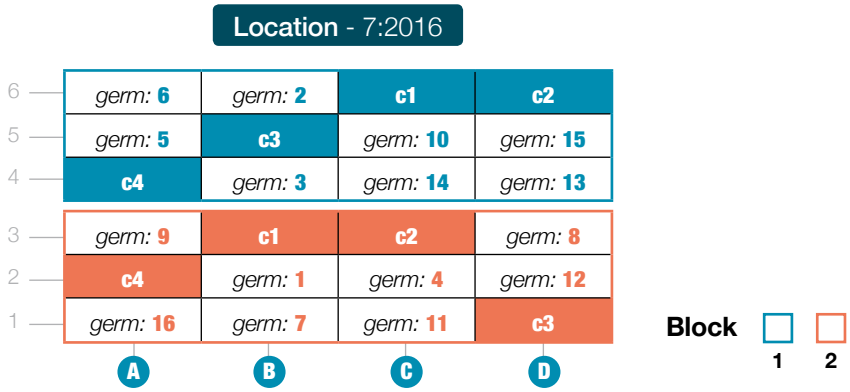


Figure 5: Exemple de dispositif ferme satellite
Figure 6: Exemple de dispositif ferme régionale




© iStock

DESCRIPTION DES DONNÉES

Une fois que les données ont été collectées, une première étape des analyses consiste à les décrire à l'aide de statistiques descriptives et de graphiques tels que des histogrammes, des graphiques en barres, où les écarts types sont affichés, des graphiques en boîtes à moustache, des normes de réactions, des biplots ou des radars.

ANALYSE

AFIN D'AMÉLIORER LA PRÉDICTION D'UNE VARIABLE CIBLE POUR LA SÉLECTION (M1)

Le problème est, sachant un ensemble de p variables prédictives X1, X2,..., Xp, d'estimer la valeur d'une variable cible y. En notant les estimateurs de y ^y nous avons ^y=^f (X1 , X2 ,... , Xp). Un exemple pourrait être d'estimer le rendement produit en utilisant le caractère poids d'un épi de maïs comme variable prédictive. La fonction ^f peut être obtenue par n'importe quel algorithme prédictif, mais seuls les algorithmes qui sont capables de prédire des variables cibles quantitatives sont présentés dans la suite. De plus, nous nous concentrons sur des algorithmes interprétables, c'est à dire des algorithmes qui peuvent expliquer comment la valeur de ^y a été prédite sachant les valeurs X1 , X2 ,. . . ,X p. Quatre algorithmes peuvent être utilisés:|

- Les arbres de classification et de régression (CART)
- Les régressions linéaires multivariées (MLR)
- Les régressions adaptatives multivariées splines (MARS)
- La classification par Random Forest

Chacune de ces méthodes est décrite ci-dessous.


© F. Rey

ARBRES DE CLASSIFICATION ET DE RÉGRESSION (CART)

CART (Breiman et al. 1984) divise à chaque itération les exemples en deux sous-ensembles. Le fractionnement est effectué en choisissant la variable et une valeur qui minimise la somme des carrés de l'erreur moyenne des deux sous-ensembles résultants. Le résultat de cette procédure est une structure arborescente dans laquelle chaque division est définie par une règle. L'interprétation de chaque nœud-feuille est obtenue par l'ensemble des règles des nœuds qui définissent ce nœud-feuille.

RÉGRESSION LINÉAIRE MULTIVARIÉE

La régression linéaire multivariée est une méthode bien établie qui utilise le modèle d'optimisation des moindres carrés ordinaires afin d'ajuster un modèle linéaire aux données d'apprentissage.

RÉGRESSIONS ADAPTATIVES MULTIVARIÉES EN SPLINES (MARS)

La méthode MARS (Friedman 1991) a été choisie car elle ne repose sur aucune hypothèse et est interprétable (Hasties, Tibshirani, and Friedman 2001). Elle ressemble à une régression par étapes, mais les relations entre chaque variable dépendante et la variable indépendante ne doivent pas nécessairement être linéaires, car chaque relation est définie par un ensemble de segments linéaires connectés, au lieu d'un seul. A l'instar de la régression linéaire, le résultat de MARS est exprimé sous la forme d'une équation un peu plus complexe que la régression linéaire mais toujours interprétable. MARS est utilisé autant de fois que le nombre de variables indépendantes non normales. A chaque fois, une seule variable est utilisée.

RANDOM FOREST

Random Forest (RF) (Breiman 2001) est une approche CART, qui utilise un ensemble de méthodes, au lieu d'une seule, pour accomplir sa tâche. RF génère plusieurs CART. Chaque CART généré est différent car l'arbre est entraîné sur un sous-ensemble de l'ensemble original obtenu à l'aide de l'ensachage (Breiman 1996) et d'un sous-ensemble aléatoire du sous-ensemble original d'entités sur chaque nœud. Les résultats RF peuvent être interprétés en utilisant deux métriques différentes (adaptées à la régression à partir de (Kuhn J. 2008) : **(I)** la précision de la réduction moyenne (% IncMSE), qui est construite en permutant les valeurs de chaque variable de l'ensemble test, en enregistrant les prédictions et en les comparant avec les prédictions des variables de l'ensemble test non permuté. Plus la valeur en % de IncMSE est élevée, plus la variable a de l'importance; **(II)** la diminution moyenne MSE (IncNodePurity), qui mesure la qualité d'une scission pour chaque variable d'un arbre. Chaque fois qu'une division d'un nœud est fait sur une variable, la somme de l'erreur quadratique moyenne (MSE) des deux sous-ensembles descendants est inférieure à l'erreur MSE du sous-ensemble parent. Additionner la diminution de MSE pour chaque variable individuelle sur toutes les arborescences générées fournit un bon indicateur. Plus la valeur IncNodePurity est élevée, plus l'importance de la variable est grande.

© Rupert Pessl

ANALYSE MULTIVARIÉE POUR ÉTUDIER

LA STRUCTURE DE LA DIVERSITÉ AFIN D'IDENTIFIER DES PARENTS À CROISER SUR LA BASE DE LEUR COMPLÉMENTARITÉ OU DE LEURS SIMILARITÉ POUR CERTAINS CARACTÈRES M2)

A partir de dispositifs complets répétés, des analyses en composante principale, un regroupement (clustering) et des analyses discriminantes peuvent être réalisés pour identifier les germplames qui pourraient être intéressants pour de prochains croisements.

ANALYSE POUR COMPARER

DIFFÉRENTES VARIÉTÉS OU POPULATIONS ÉVALUÉES POUR LA SÉLECTION DANS CHAQUE LIEU (FAMILLE 1: M4A, M4B, M5 & M7A)

Quatre analyses sont proposées: anova classique, analyse spatiale, modèle mixte pour les dispositifs en blocs incomplets et un modèle bayésien hiérarchique. Les anovas classiques **(M4a)** ne sont pas expliquées ici car ce sont des analyses très classique. Seules les analyses spatiales, le modèle mixte et le modèle bayésien hiérarchique sont détaillés.

ANALYSE SPATIALE (M4B)

L'analyse spatiale s'appuie sur un dispositif en lignes colonnes **(D3)**. Le modèle est basé sur les statistiques fréquentistes. Il tient compte des variation environnementales au sein des blocs avec quelques témoins répétés sur les ligne et sur les colonnes. L'analyse est basée sur un modèle SpATS (Analyse Spatial d'essais au champ avec des splines) proposé par Rodríguez-Álvarez et al. (2016). La variation environnementale est prise en compte en incluant des effets lignes et colonnes ainsi qu'une fonction bivariée lisse qui prend en compte simultanément la tendance spatiale dans les deux directions du champ (ligne et colonne). Plus d'informations sur le modèle ainsi que des exemples du package R SpATS peuvent être trouvés dans Rodríguez-Álvarez et al. (2016). Cette analyse peut être effectuée par PPBstats en différentes étapes : lancer le modèle, vérifier les sorties du modèle et les visualiser, calculer les comparaisons de moyenne sur les germplasmes (Figure 7) (https://priviere.github.io/PPBstats_book/family-1.html#spatial-analysis).


© Pro Specie Rara

MODÈLE MIXTE POUR DES DISPOSITIFS EN BLOCS INCOMPLETS (M5)

Les dispositif expérimentaux utilisés sont les blocs incomplets **(D2)**.

L'objectif des dispositifs en blocs incomplets est de contrôler la variation entre microparcelles et idéalement de comparer les germplasmes deux à deux (Mead 1997), mais cela est rarement possible avec un nombre d'entrée élevé et un faible nombre de répétitions. Les dispositifs que l'on peut résoudre sont ceux avec des blocs complets répétés, chaque répétition étant divisée en petits blocs incomplets. Les dispositifs Lattice sont des dispositifs particuliers de blocs incomplets que l'on peut résoudre où le nombre d'entrées g est le carré d'un entier et la taille du bloc est $\sqrt{g}$. L'introduction des dispositifs alpha (Patterson and Williams 1976) supprime la restriction en terme de nombre d'entrées. L'avantage du dispositif en blocs incomplets est que chaque bloc incomplet (une séquence de 4 microparcelles dans l'exemple montré en Figure 3), est une unité indépendante qui peut par conséquent être allouée à un champ différent. Le nombre de blocs incomplets qui peut être semé dans chaque ferme dépend seulement de la taille de la ferme. Il est également possible qu'une répétition complète soit semée dans une ferme plus grande et que les 10 blocs incomplets des autres répétitions soit semés dans 10 fermes différentes. Les inconvénients de ce dispositif sont i) la restriction que le nombre total d'entrée (g) soit un multiple de la taille du bloc (k) de telle sorte que $g = sk$ avec s = le nombre de blocs incomplets par répétition; ii) la perte du dispositif ligne colonne qui permet d'augmenter la précision avec une analyse spatiale. Plus d'informations sont disponibles dans (Singh and El-Shama'a 2015) (Patterson and Williams 1976)(Mead 1997). Le modèle est expliqué dans (Sarker and Singh 2015).

MODÈLE BAYÉSIEN HIÉRARCHIQUE (M7A)

Ce modèle est basé sur les dispositifs expérimentaux fermes régionales et satellites **(D4)**.

Au niveau de la ferme, la résiduelle a peu de degrés de liberté, ce qui conduit à une estimation médiocre et instable de la variance résiduelle et à un manque de puissance pour comparer les populations. Le modèle bayésien hiérarchique **(M7a)** a été mis en œuvre pour améliorer l'efficacité de la comparaison des moyennes sur chaque ferme. Il est efficace si l'on dispose d'au moins 20 environnements (i.e. lieu année) (Rivière et al. 2015). Le modèle est basé sur des statistiques bayésiennes.

Le modèle est décrit dans Rivière et al (2015). La spécificité du modèle est que le terme résiduel dans chaque environnement suit une distribution normale centrée sur zéro avec une variance propre à l'environnement mais qui est supposée provenir d'une distribution des variances résiduelles commune à tous les essais du réseau. Cette hypothèse est raisonnable en raison de la structure similaire des essais dans tous les environnements du réseau. Une approche hiérarchique est utilisée et des distributions préalables vagues sont placées sur les hyper-paramètres de la distribution. En d'autres termes, la variance résiduelle d'un essai dans un environnement donné est estimée en utilisant toute l'information disponible sur le réseau plutôt qu'en utilisant uniquement les données de cet essai particulier. Des distributions a priori, non informatives, sont également supposées pour les paramètres des germplasmes et des blocs.

D'un point de vue agronomique, l'hypothèse selon laquelle les variances résiduelles des essais sont hétérogènes (elles sont modélisées avec une distribution gamma inverse) est cohérente avec l'agriculture biologique : il existe autant d'environnements que de fermes et d'agriculteurs (des pratiques telles que la date de semis, la densité des semis, le labour, etc différentes ainsi que les conditions pédo-climatiques, stress biotique et abiotique, etc.) conduisant à une grande hétérogénéité. De plus, les variances résiduelles suivent une distribution gamma inverse montrant des propriétés de conjugaison qui facilitent la convergence MCMC.

La variance résiduelle estimée à partir des témoins est supposée être représentative de la variance résiduelle des autres entrées. Les blocs ne sont inclus dans le modèle que si l'essai comporte des blocs. L'analyse peut se faire avec PPBstats en plusieurs étapes : exécuter le modèle, vérifier les sorties du modèle et les visualiser, obtenir des comparaisons moyennes pour les germplasmes dans chaque lieu (https://priviere.github.io/PPBstats_book/family-1.html#model-1).

© Arche Noah

ANALYSES POUR ÉTUDIER LES RÉPONSES DES VARIÉTÉS OU DES POPULATIONS SOUS SÉLECTION SUR PLUSIEURS ENVIRONNEMENTS (FAMILLE 2: M6 & M7B)

Trois analyses sont proposées: AMMI et GGE **(M6)** et le modèle bayésien hiérarchique **(M7b)**.


© Rupert Pessl

AMMI (M6)

Le dispositif expérimental complet répété est utilisé ici **(D1)** sur plusieurs environnements. Le modèle AMMI (Additive Main effects and Multiplicative Interaction) est basé sur des statistiques fréquentistes. L'analyse peut être décomposée en deux étapes décrites dans Gauch 2006.

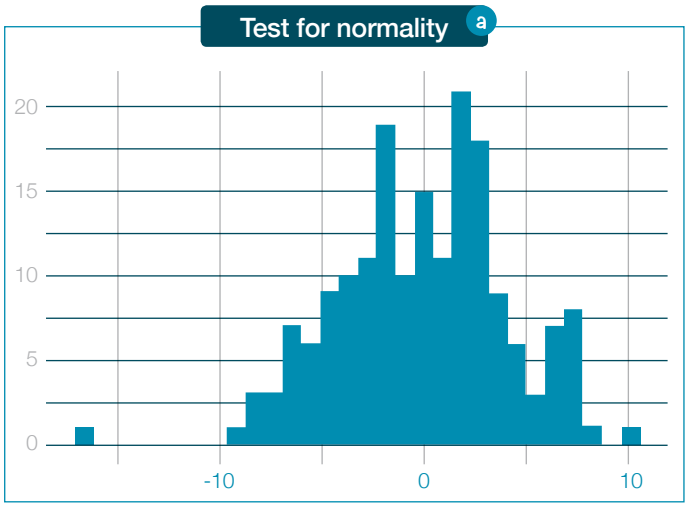
La première étape est une ANOVA avec des effets germplasme, environnement et (germplasme x environnement). Tous les autres effets nécessaires, tels que les blocs dans l'environnement ou la décomposition de l'environnement en effets lieu et année, peuvent être inclus.

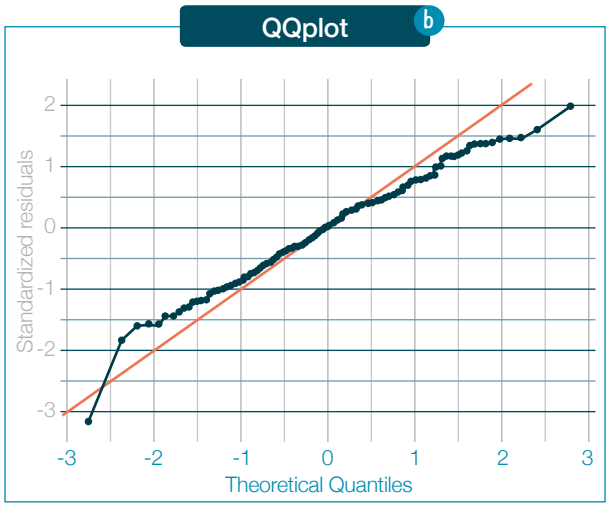
Ensuite, une Analyse en Composante Principale (ACP) est exécutée sur les termes d'interaction (matrice des dimensions g x e). Les données sont doublement centrées sur les environnements et les germplasmes. L'ACP étudie la structure de la matrice d'interaction. Les environnements sont les variables et les germplasmes sont les individus. L'ACP permet de détecter :

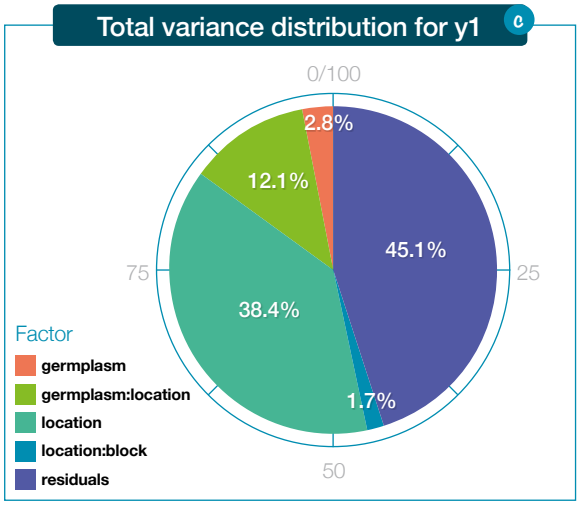
- Les germplasmes qui sont stables (c'est-à-dire qui contribuent moins à l'interaction, ils répondent comme la moyenne des germplasmes aux différents environnements),
- Les germplasmes qui contribuent le plus à l'interaction (qui s'écartent de la moyenne) et avec quel environnement,
- Les lieux qui ont les mêmes profils d'interaction.

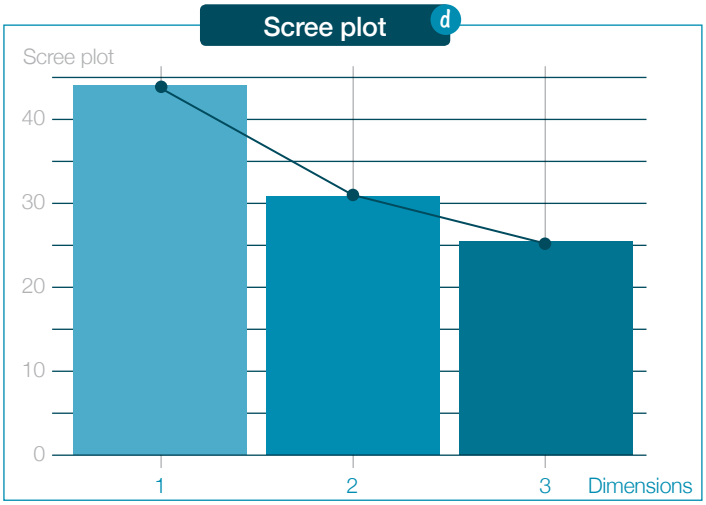
L'analyse peut se faire avec PPBstats en plusieurs étapes : exécuter le modèle, vérifier les sorties du modèle et les visualiser, obtenir des comparaisons de moyennes pour chaque facteur à partir de l'ANOVA, obtenir des biplot d' l'ACP (https://priviere.github.io/PPBstats_book/family-2.html#ammi).

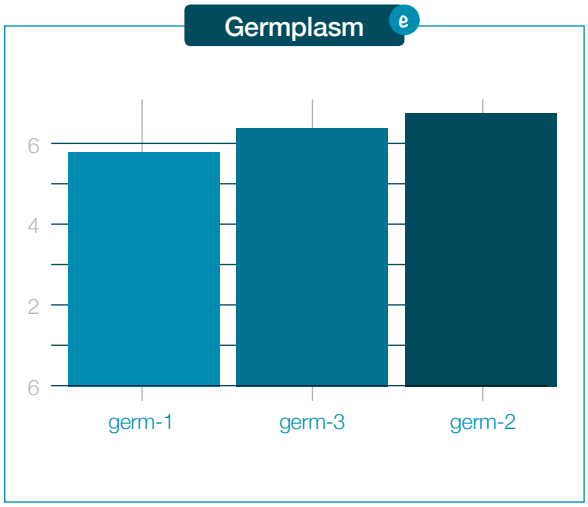
Figure 7. Exemples de sorties de PPBstats pour vérifier les sorties de l'anova (a.b.c.) et de l'ACP (d,); comparaisons de moyennes à partir de l'ANOVA sur les germplasmes (e.) et biplot à partir de l'ACP (f.).

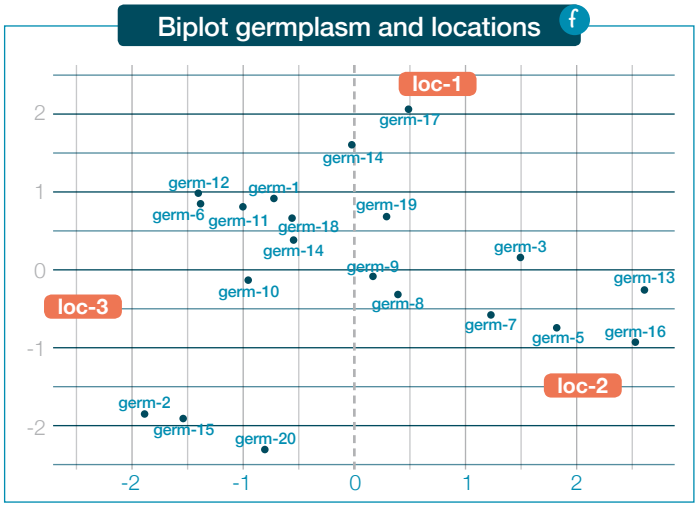

Test for normality
a
20
15
10
5
0
-10
0
10


QQplot
b
Standardized residuals
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Total variance distribution for y1
c
0/100
25
50
75
2.8%
12.1%
45.1%
38.4%
1.7%
Factor
germplasm
germplasm:location
location
location:block
residuals


Scree plot
d
Scree plot
40
30
20
10
0
1
2
3
Dimensions


Germplasm
e
6
4
2
6
germ-1
germ-3
germ-2


Biplot germplasm and locations
f
loc-1
germ-17
germ-14
germ-12
germ-1
germ-6
germ-11
germ-18
germ-19
germ-14
germ-3
germ-9
germ-13
germ-10
germ-8
loc-3
germ-7
germ-5
germ-16
loc-2
germ-2
germ-15
germ-20
2
1
0
-1
-2
-2
-1
0
1
2

GGE (M6)

Le dispositif expérimental complet répété est utilisé ici **(D1)** sur plusieurs environnements. Le modèle GGE est le même que le modèle AMMI sauf que l'ACP est faite sur une matrice centrée sur les lieux : les effets germplames et interactions sont confondus. Le modèle est basé sur les statistiques fréquentistes. Comme pour l'AMMI, les analyses peuvent être réalisées avec PPBstats (https://priviere.github.io/PPBstats_book/family-2.html#gge). Certains graphiques spécifiques à l'analyse GGE peuvent être réalisés (Figure 8).

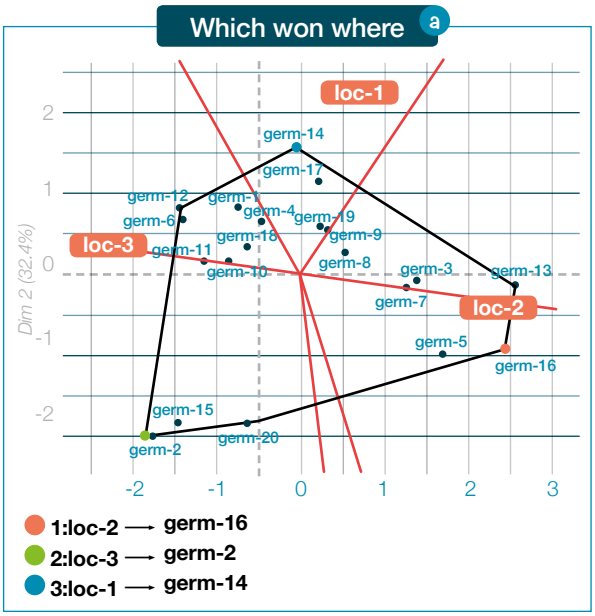
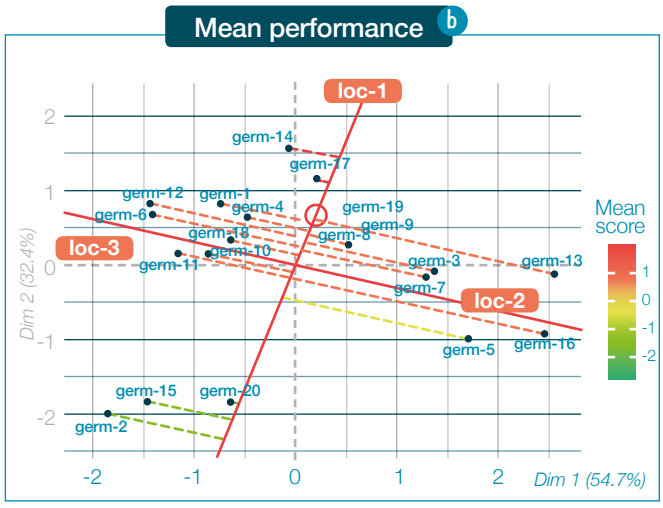
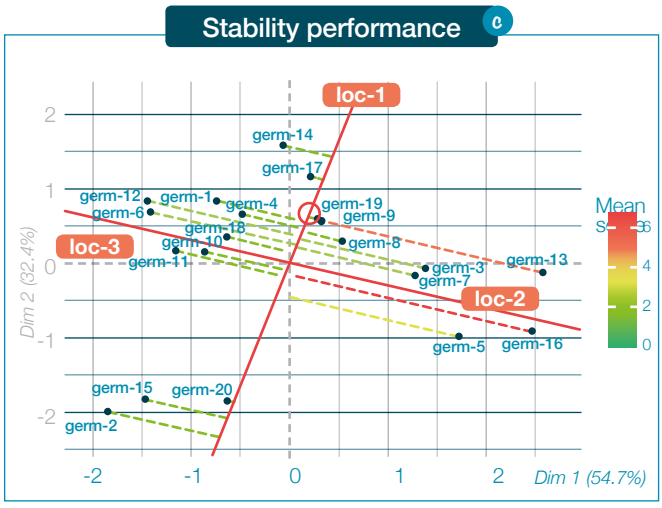
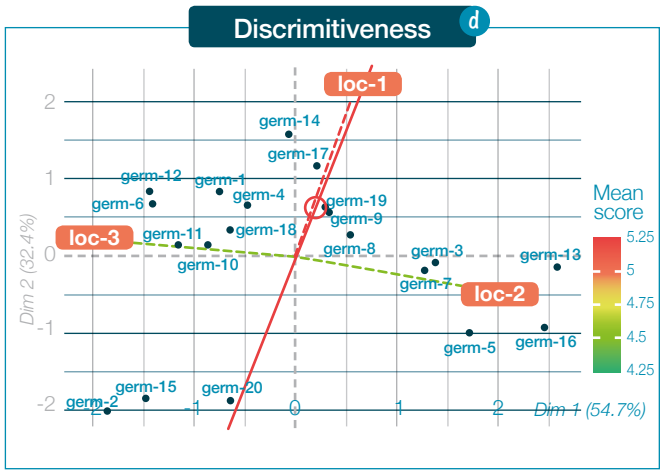
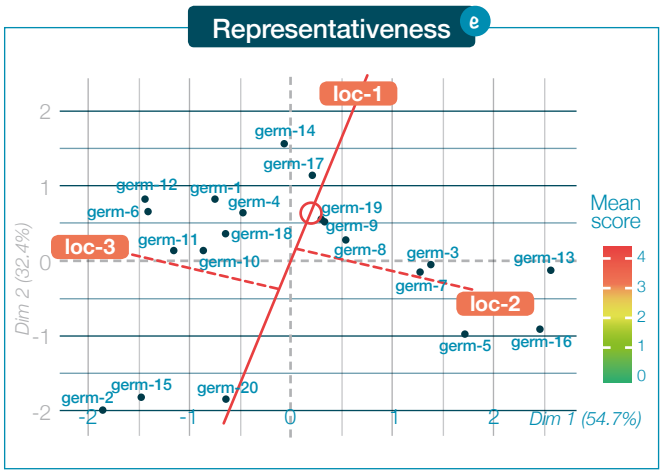


Figure 8: biplot GGE avec PPBstats: qui gagne où (a.), performance des moyennes (b.), performance de stabilité (c.), discrimitivité (d.) et représentativité (e.)









MODÈLE BAYÉSIEN HIÉRARCHIQUE (M7B)

Ce modèle est basé sur les dispositifs expérimentaux fermes régionales et satellites **(D4)**.

M7b est particulièrement pertinent lorsque, au **niveau du réseau,** il manque un grand nombre de combinaisons germplasme environnement, ce qui conduit à une mauvaise estimation des effets germplasme, environnement et de l'interaction. La mise en œuvre de la méthode **M7b** requiert que les données comprennent au moins environ 75 environnements avec 120 germplasmes présents dans au moins deux environnements (95% des combinaisons manquantes). Lorsque le déséquilibre diminue (% de combinaisons manquantes plus faible), le nombre d'environnements et de germplasmes nécessaire diminue également. Le modèle est basé sur des statistiques bayésiennes.

Le modèle inclut les effets germplasme et environnement. L'interaction est exprimée sous la forme d'un terme multiplicatif constitué de l'effet environnement multiplié par un coefficient de régression qui dépend du germplasmes. La partie restante de l'interaction va dans la résiduelle. Le modèle peut être réduit en un effet germplasme plus un terme multiplicatif consistant en l'effet environnement multiplié par un coefficient qui représente la **sensibilité** de chaque germplasme aux environnements.

Ce modèle est connu sous le nom de modèle de Finlay Wilkinson ou de régression conjointe (1963).

La sensibilité des germplasmes quantifie la stabilité de leurs performances sur l'ensemble des environnements. La sensibilité moyenne est égale à 1, de sorte qu'un germplasme ayant une valeur supérieure (resp. inférieure) à 1 est plus (resp. moins) sensible aux environnements que la moyenne des germplasmes observés (Nabugoomu, Kempton, et Talbot 1999).

Compte tenu du déséquilibre élevé des données et de la grande quantité de données, ce modèle est mis en œuvre selon une approche bayésienne hiérarchique. Des distributions a priori, hiérarchiques sont utilisées pour les effets germplasme, environnement et la sensibilité à l'interaction, tandis qu'un prior non informatif est utilisé pour la variance résiduelle.

L'analyse peut se faire avec PPBstats en plusieurs étapes : exécuter le modèle, vérifier les sorties du modèle et les visualiser, effectuer des études de validation croisée, obtenir les comparaisons des moyennes pour chaque facteur, « *prédire le passé* » (Figure 9) (https://priviere.github.io/PPBstats_book/family-2.html#model-2).

Figure 9. Exemple de graphiques pour « prédire le passé » avec PPBstats: il y a deux valeurs: une estimée par le modèle (i.e. la combinaison germplasme x lieu existe dans le jeu de données) et une prédite par le modèle pour les germplasms qui n'étaient pas présents dans les lieux.

ANALYSE ORGANOLEPTIQUE


© iStock

Dans ce qui suit, les produits qui sont dégustés à travers l'analyse organoleptique doivent être considérés comme des extensions des variétés ou des populations et les analyses sensorielles comme des évaluations phénotypiques qui nécessitent des analyses statistiques particulières. Toute les analyses peuvent être faites avec PPBstats (https://priviere.github.io/PPBstats_book/organoleptic.html).

TEST NAPPING (M9A)

Le napping permet de rechercher des différences sensorielles entre les produits. Les différences portent sur les caractéristiques sensorielles globales et doivent être complétées par une tâche de verbalisation pour faciliter la compréhension des différences. Il offre une grande flexibilité, car il n'est pas nécessaire de former.un panel.


© iStock

Deux étapes sont effectuées lors du napping :

- **L'étape de tri :** chaque dégustateur est invité à positionner l'ensemble des produits sur une feuille de papier blanc (une nappe) en fonction de leurs similitudes/disparités. Ainsi, deux produits sont proches s'ils sont perçus comme similaires ou, au contraire, éloignés l'un de l'autre s'ils sont perçus comme différents. Chaque dégustateur utilise ses propres critères.
- **L'étape de verbalisation :** Après avoir effectué le napping, on de mande aux dégustateurs de décrire les produits en écrivant un ou deux descripteurs sensoriels qui caractérisent chaque groupe de produits sur la carte.

Le groupe doit être composés de 12 à 25 dégustateurs en fonction de l'expérience des juges avec le produit et de l'objectif de l'expérience. Par exemple, dix paysans-boulangers suffiront pour obtenir des résultats fiables pour une analyse de pains car ils sont habitués à en manger et en goûter. Dans le cas de consommateurs, un groupe de vingt personnes serait mieux adapté.

Un maximum de dix produits peut être évalué simultanément. Un code aléatoire à trois chiffres doit être attribué à chaque échantillon. Les échantillons sont présentés simultanément et les évaluateurs peuvent goûter autant

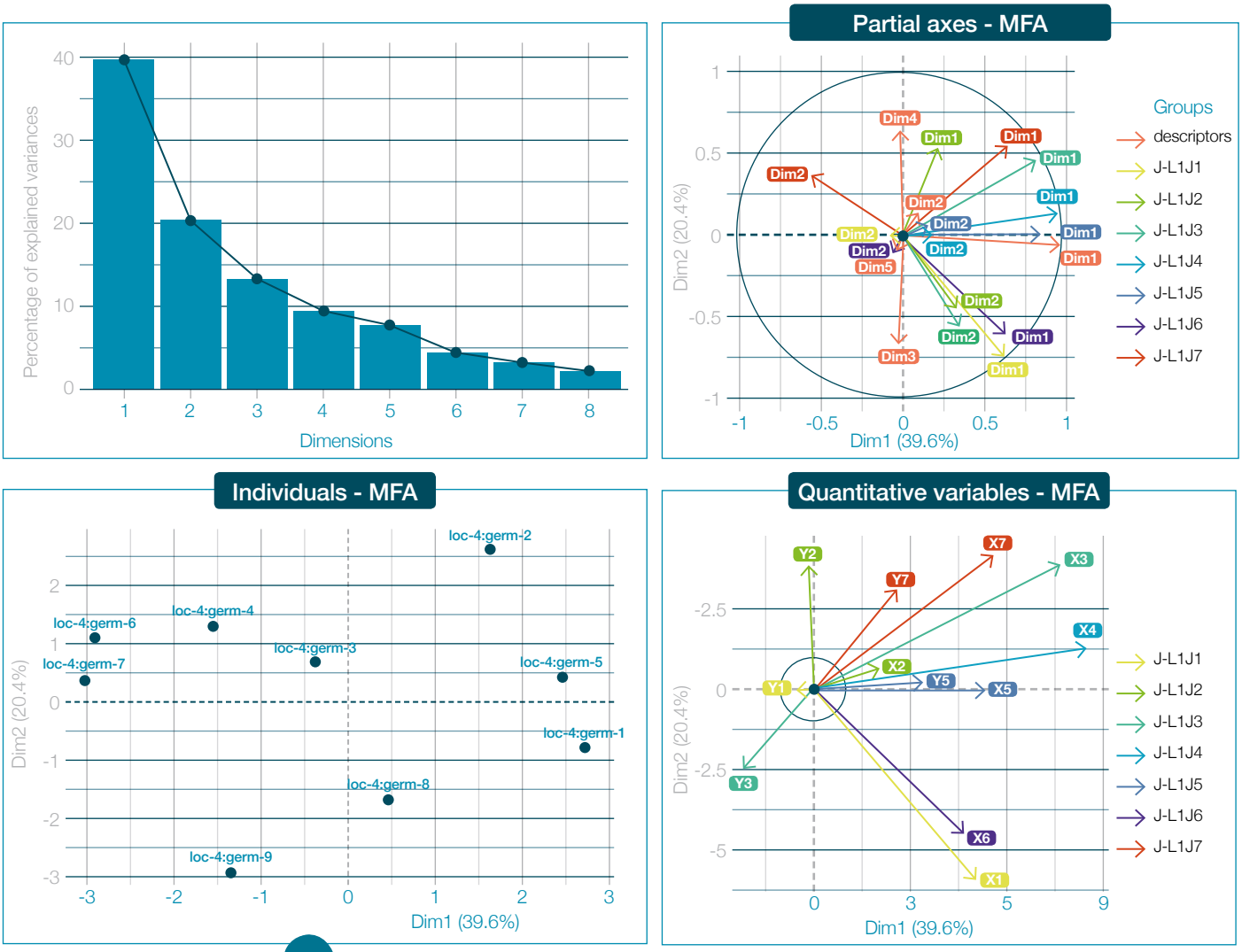
qu'ils le souhaitent. Les données de napping conduisent à un tableau quantitatif. Les lignes sont les produits. Deux colonnes par juge donnent les coordonnées spatiales (x, y) de chaque produit sur la nappe.

Les descripteurs sensoriels sont codés à l'aide d'un tableau de fréquences « *produits x mots* ». Tout d'abord, un tableau de contingence comptant le nombre de fois que chaque descripteur a été utilisé pour décrire chaque produit est créé. Cette table de contingence est ensuite transformée en fréquences pour que la « *fréquence des mots* » devienne une variable qualitative avec le nombre de mots cités comme modalités.

Une analyse factorielle multiple (AFM) est effectuée. Chaque sujet constitue un groupe de deux variables non normalisées. L'AFM conduit à une synthèse de la nappe des dégustateurs. Deux produits sont proches si tous les juges les considèrent proches sur la nappe. Plus les deux premières composantes de l'AFM expliquent la variabilité initiale, plus les juges sont d'accord.

Les tableaux de fréquences qui croisent les produits et la fréquence des mots sont considérées comme un ensemble de variables supplémentaires : ils n'interviennent pas dans la construction des axes mais leurs corrélations avec les facteurs d'AFM sont calculées et représentées comme dans une ACP classique (Figure 10).

Figure 10. Exemple d'AFM avec PPBstats.



TEST HÉDONIQUE (M9B)

Le test d'évaluation hédonique consiste à demander aux consommateurs d'évaluer leur préférence de 1 (je n'aime pas beaucoup) à 9 (j'aime beaucoup) pour 3 ou 4 attributs sensoriels spécifiques au produit testé. La préférence globale est déterminée au début du questionnaire afin de ne pas influencer le consommateur et d'être plus proche des conditions typiques de consommation. Des informations supplémentaires concernant le sexe, l'âge et la fréquence de consommation de produits biologiques sont demandées à la fin du test afin de caractériser l'échantillon de population. Des descripteurs sensoriels supplémentaires pour décrire les produits sont demandés après l'évaluation de chaque produit. L'un des principaux objectifs des tests hédoniques est de déterminer les différences d'appréciation pour un attribut donné entre un ensemble d'échantillons testés. La distribution des données détermine le type de tests à utiliser pour analyser l'ensemble de données.


© iStock

- Si la distribution est normale, une analyse de variance (ANOVA) peut être effectuée, la source de variance étant l'échantillon, suivie d'une comparai son multiple des valeurs moyennes des données issues de chaque évaluateur. L'objectif est d'obtenir un classement final basé sur les préférences des consommateurs.
- Si l'ensemble de données ne suit pas une distribution normale, un test de Friedman sur les rangs devra être utilisé pour indiquer si les variétés sont perçues différemment par les évaluateurs.

Enfin, une analyse hiérarchique par cluster peut être mise en œuvre pour identifier les groupes de préférences.

TESTS SUR LES RANGS (M9C)

Un panel d'évaluateurs compare plusieurs produits simultanément et les classe en fonction de l'ampleur perçue d'une caractéristique sensorielle donnée (p. ex. acidité, fibrosité). Cette méthode a l'avantage d'être facile à mettre en œuvre. Le jury comprend idéalement 12 évaluateurs semi-naïfs (consommateurs initiés aux analyses sensorielles) selon la norme ISO 8587, bien qu'il soit possible de mettre en évidence des différences significatives avec un plus petit nombre d'évaluateurs. Caractéristiques clés :

- Les produits sont présentés simultanément, Ceci nécessite que l'ensemble des échantillons à tester soit disponible en même temps. Certaines espèces végétales présentent des différences marquées de précocité (par exemple, le brocoli), et il faut donc veiller à ce que les échantillons de la même précocité soient comparés.
- Les évaluateurs peuvent goûter autant que nécessaire.
- Lorsqu'ils répondent, les évaluateurs ne peuvent pas classer deux produits au même rang, c'est-à-dire que tous les rangs attribués doivent être uniques.

Il est conseillé de ne pas dépasser 6 échantillons par séance. Hypothèse nulle (H0) : tous les produits ont exactement les mêmes réponses (les moyennes de classement sont égales). Le test de Friedman (test non paramétrique sur k échantillons indépendants) conduit au rejet ou à l'acceptation de cette hypothèse, basée sur la valeur α (<0.05).

1 - ISO 8587:2006 is a standard from International Organisation for Standardisation which describes a method for sensory evaluation with the aim of placing a series of test samples in rank order.

ANALYSE MOLÉCULAIRE (M2)

© iStock

L'analyse moléculaire peut être utilisée pour étudier la structure de la diversité et identifier des parents complémentaires ou similaires pour faire des croisements, à travers des distances génétiques et des arbres. Ces analyses sont basées sur des données génétiques individuelles.

ANALYSE RÉSEAU (M8)


© iStock

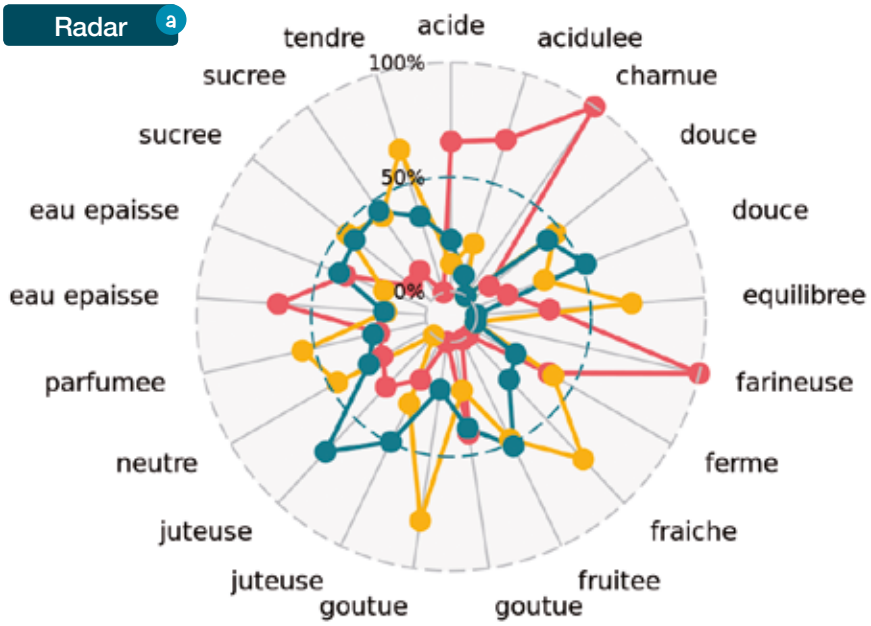
La description de la topologie des réseaux de circulation des semences est intéressante car elle donne un aperçu de l'organisation des échanges au sein d'un programme de sélection participative ou d'une maison des semences paysannes (Vernooy, Shrestha, and Sthapit 2015) (Pautasso et al. 2013). L'analyse peut se faire à plusieurs échelles géographiques ou organisationnelles, par exemple locale, régionale ou nationale. Deux types de réseaux peuvent être étudiés : **(I)** les réseaux unipartites : - où un nœud peut être un lot de semences (c'est-à-dire une combinaison d'un germplasme dans un lieu donné une année donnée) et les liens entre nœuds sont des relations telles que la diffusion, le mélange, la reproduction, les croisements ou la sélection par exemple ; - où un nœud peut être un lieu et les liens sont des événements de diffusion entre lieux ; **(II)** les réseaux biparties où un nœud peut être un lieu ou un germplasme.

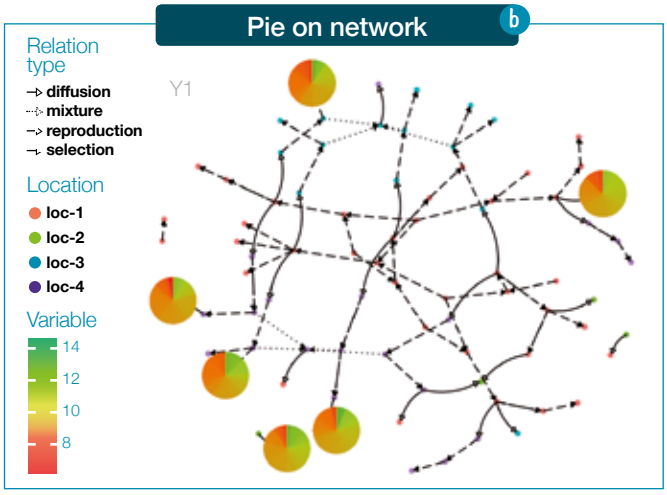
ANALYSE DESCRIPTIVE

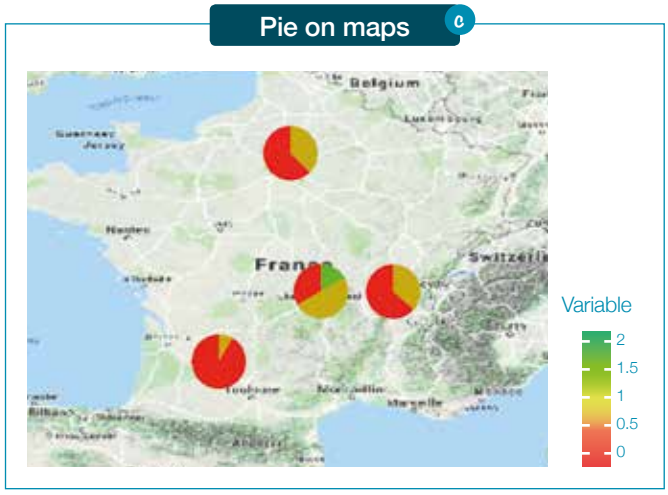
Des analyses descriptives peuvent être effectuées pour mieux comprendre comment les échanges sont organisés au sein d'une maison des semences paysannes ou d'un programme de sélection participative (Figure 11). Les réseaux unipartites de lots de semences peuvent être affichés dans l'ordre chronologique. Des diagrammes peuvent être utilisés pour montrer la répartition des germplasmes par lieu ou par année. Dans les réseaux unipartites de lieux, les événements de diffusion entre les lieux et leurs fréquences peuvent être affichés. Les réseaux bipartites de germplasmes et de lieux montrent les relations entre les germplasmes et les lieux (c.-à-d. quel germplasme dans quel lieu).

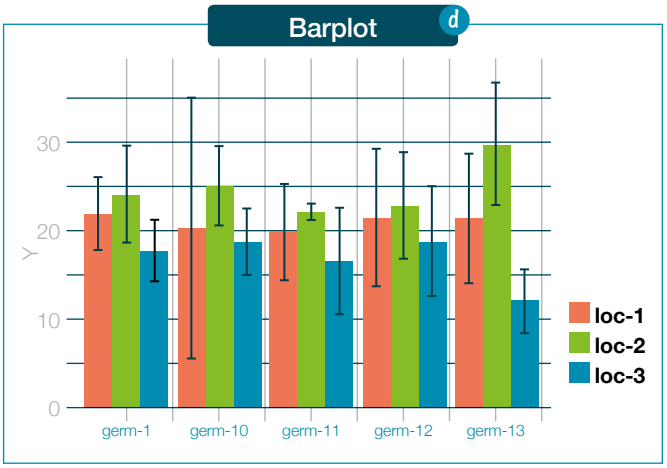

© iStock

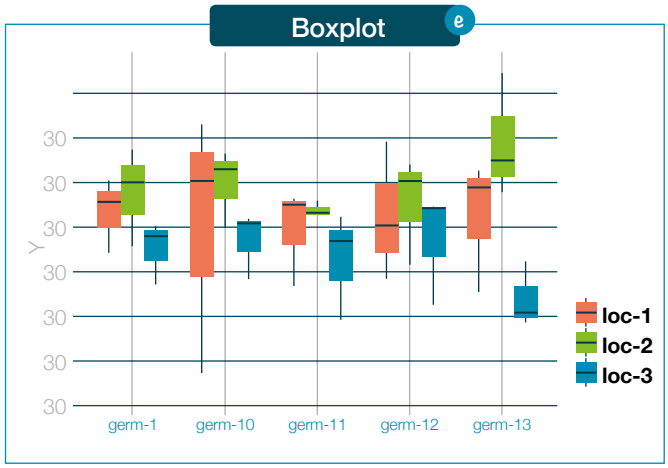
Figure II.
Exemples de graphiques descriptifs avec PPBstats : radar (a.), camembert sur un réseau (b.), camembert sur une carte (c.), diagramme (d.), boîte à moustache (e.).











OUTILS POUR METTRE ŒUVRE LES MÉTHODES: LE PACKAGE R PPBSTATS


© iStock

Un premier ensemble de ces méthodes a été implémenté de manière cohérente dans un nouveau logiciel : le package R PPBstats.

Ce progiciel est basé sur un logiciel R qui est open source et largement utilisé dans la communauté de la sélection et de l'agronomie. PPBstats vise à effectuer les analyses rencontrées dans les programmes PPB à quatre niveaux :

- essais agronomiques (évaluation agronomique et/ou nutritionnelle)
- tests organoleptiques
- expérimentation moléculaires
- réseau de circulation de semences

PPBstats PPBstats est en cours de développement et le code est hébergé sur Github pour faciliter la collaboration: https://github.com/priviere/PPBstats

Les dispositifs expérimentaux suivant peuvent être réalisés :

- **D1:** dispositifs complets répétés
- **D2:** dispositifs en blocs incomplets
- **D3:** dispositifs ligne/colonne
- **D4:** dispositifs fermes régionales et satellites

Les méthodes suivantes ont été implémentées (mais des tests approfondis sont les bienvenus !):

- **M2:** Analyse multivariée (ACP)
- **M4a:** ANOVA classique
- **M4b:** Analyse spatiale
- **M6:** AMMI et GGE
- **M7a:** Modèle intra-lieu bayésien hiérarchique
- **M7b:** modèle bayésien hiérarchique GxE
- **M8:** Analyse réseau

Les méthodes suivantes ne sont pas encore intégré dans le logiciel mais peuvent être réalisée avec d'autres logiciels:

- **M1:** Méthode non paramétrique, régression multivariée, classification et arbre de régression, random forest:
  - Arbres de régression et de classification (CART): rpart, the recursive partitioning algorithm, est une fonction utilisée pour réaliser les analsyes CART. La fonction rpart est disponible dans le package R rpart.

- Régression linéaire multivariée (MLR): lm, le modèle linéaire est disponible dans R par défaut. i.e., aucun package R n'est nécessaire.
- Régression adaptative multivariée Splines (MARS): fonction earth du projet R.
- Random Forest: la fonction random Forest du package R randomForest.

- **M2:** Analyse multivariée (clustering, analyse discriminante):
  - R package R FactoMineR, *http://factominer.free.fr/index.html*
- **M3:** Distance génétique; arbres
  - R package adegenet
  - PowerMarker, *http://statgen.ncsu.edu/powermarker/downloads.htm* - V3.23 (Liu, 2002)
  - GENEPOP, *http://kimura.univ-montp2.fr/~rousset/Genepop.htm* - 4.0 (Raymond and Rousset 1995)
  - FSTAT, *http://www2.unil.ch/popgen/softwares/fstat.htm* - FSTAT v. 2.9.3.2, program package (Goudet 2002)
  - ARLEQUIN, *http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html* - ARLEQUIN ver. 3.0 (Excoffier et al., 2005)
  - PHYLIP, *http://evolution.genetics.washington.edu/phylip/getme.html*
  - PHYLIP ver. 3.6b software package (Felsenstein 1993) - STRUCTURE, *http://pritchardlab.stanford.edu/structure.html* - STRUCTURE ver. 2.3.3 (Pritchard et al., 2000)
  - STRUCTURE HARVESTER, *http://taylor0.biology.ucla.edu/structureHarvester/*
  - STRUCTURE HARVESTER v0.6.92 (Earl and van Holdt, 2012)
- **M5:** Modèle mixte pour les blocs incomplets: module Genstats.
- **M9a:** Analyse factorielle multiple; Projection des fréquence de mots:
  - R packages FactoMineR, *http://factominer.free.fr/index.html*
  - R package SensoMineR, *http://sensominer.free.fr/*
- **M9b:** ANOVA; Analyse de cluster hiérarchique; analyse sur correspondance sur des descripteurs sensoriels supplémentaires:
  - R packages FactoMineR, *http://factominer.free.fr/index.html*
  - R package SensoMineR, *http://sensominer.free.fr/*
- **M9c:** Méthodes non paramétrique, test sur la somme des rangs; test de Friedman: fonction basique dans R.

Les méthodes **M2, M5, M9a, M9b** et **M9c** seront insérés dans PPBstats.

Un site web dédié à PPBstats et un tutoriel exhaustif pour collaborer et utiliser le logiciel peuvent être consultés ici : https://priviere.github.io/PPBstats_web_site.


© F. Rey

RÉFÉRENCES


© Pro Specie Rara

- **Bernardo, R. 2002.** Breeding for quantitative traits in plants. Stemma Press, Woodbury, Minnesota.
- **F. Blanquart, O. Kaltz, S. L. Nuismer, et S. Gandon. 2013.** A practical guide to measuring local adaptation. Ecology Letters, 16(9) :1195–1205.
- **Breiman, L. 1996.** «Bagging Predictors.» Machine Learning 26: 123–40.
- **Breiman, L. 2001.** «Random Forests.» Machine Learning 45: 5–32.
- **Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984.** Classification and Regression Tree. Edited by Chapman and Hall/CRC.
- **Desclaux, D., J. M. Nolot, Y. Chiffoleau, C. Leclerc, and E. Gozé. 2008.** «Changes in the Concept of Genotype X Environment Interactions to Fit Agriculture Diversification and Decentralized Participatory Plant Breeding: Pluridisciplinary Point of View.» Euphytica 163: 533–46.
- **Friedman, J.H. 1991.** «Multivariate Adaptive Regression Splines.» Journal of Ann. Stat. 19: 1–141.
- **Gallais, A. 1990.** Théorie de la sélection en amélioration des plantes. Masson. Sciences Agronomiques.
- **Gauch, H.G. 2006.** «Statistical Analysis of Yield Trials by AMMI and GGE.»Crop Sci 46 (4): 1488–1500.
- **Kuhn J., S. Neumann, B. Egert. 2008.** «Building Blocks for Automated Elucidation of Metabolites: Machine Learning Methods for Nmr Prediction.» BMC Bioinformatics 9: 400.
- **Mead, R. 1997.** Design of Plant Breeding Trials. Edited by London Kempton RA Fox PN (eds) Statistical Methods for Plant Variety Evaluation pp 40-67. Chapman & Hall.
- **Nabugoomu, F., R.A. Kempton, and M. Talbot. 1999.** «Analysis of Series of Trials Where Varieties Differ in Sensitivity to Locations.» Journal of Agricultural, Biological and Environnmental Statistics 4 (3): 310–25.
- **Patterson, H.D., and E.R. Williams. 1976.** «A New Class of Resolvable Incomplete Block Designs.» Biometrika 63: 83–90.
- **Pautasso, M., G. Aistara, A. Barnaud, S. Caillon, P. Clouvel, O. Coomes, M. Delêtre, et al. 2013.** «Seed exchange networks for agrobiodiversity conservation. A review.» Agronomy for Sustainable Development 33.
- **Rivière, P., J.C. Dawson, I. Goldringer, and O. David. 2015.** «Hierarchical Bayesian Modeling for Flexible Experiments in Decentralized Participatory Plant Breeding.» Crop Science 55 (3).
- **Rivière, P., Goldringer, I. and Vindras C. 2018.** Analysis of Participatory Plant Breeding programme with the R package PPBstats. (version 0.23). https://priviere. github.io/PPBstats_book/.
- **Rivière, P., G. Van Frank, F. Munoz and O. David, 2018,** PPBstats: An R package to perform analysis found within PPB programmes regarding network of seeds circulation, agronomic trials, organoleptic tests and molecular experiments. Version 0.24, URL: https://github.com/priviere/PPBstats_web_site.
- **Rodríguez-Álvarez, M.X., M. P. Boer, F. A. van Eeuwijk, and P. H. C. Eilers. 2016.** «Spatial Models for Field Trials.» ArXiv E-Prints, July.
- **Sarker, A., and M. Singh. 2015.** «Improving Breeding Efficiency Through Application of Appropriate Experimental Designs and Analysis Models: A Case of Lentil (Lens Culinaris Medikus Subsp. Culinaris) Yield Trials.» Field Crops Research 179: 26–34.
- **Singh, M., and K. El-Shama'a. 2015.** Experimental Designs for Precision in Phenotyping.
- **Sperling, L., J.A. Ashby, M.E. Smith, E. Weltzien, and S. McGuire. 2001.** «A Framework for Analyzing Participatory Plant Breeding Approaches and Results.» Euphytica 122 (3): 439–50.
- **T., Hastie, Tibshirani R., and J.H. Friedman. 2001.** The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Edited by Springer.
- **Vernooy, R., P. Shrestha, and B. Sthapit. 2015.** Community Seed Banks: Origins, Evolution and Prospects. Issues in Agricultural Biodiversity. Earthscan for Routledge.

Ce **guide technique #3** présente des dispositifs expérimentaux ainsi que des outils et méthodes statistiques utiles pour la sélection décentralisée à la ferme.

Brochure #3

© iStock

© iStock

Auteurs : Isabelle Goldringer (INRA) et Pierre Rivière (RSP)

Éditeur : Frédéric Rey (ITAB)

Remerciements : nous remercions Salvatore Ceccarelli (RSR), João Mendes Moreira (Univ. of Porto), Pedro Mendes Moreira (IPC), Moreira, Gaelle van Frank (INRA), Carlota Vaz Patto (ITBQ NOVA), Camille Vindras (ITAB) pour leur contribution à la description de certaines méthodes.

Comment citer ce document : Goldringer I., Rivière P., 2019. Méthodes et outils pour la sélection décentralisée à la ferme. Brochure #3. Projet Diversifood

Février 2019

Graphisme: Galerie du Champ de Mars, *floredelataille.grafic@gmail.com*

Contact: isabelle.goldringer@inra.fr

www.diversifood.eu

21 partenaires du CONSORTIUM DE DIVERSIFOOD

France

INRA • Institut National de la Recherche Agronomique
ITAB • L'institut de l'agriculture et de l'alimentation biologiques
RSP • Réseau Semences Paysannes
IT • INRA Transfert

Angleterre

ORC • Organic Research Centre

Suisse

FiBL • Forschungsinstitut für biologischen Landbau
PSR • ProSpecieRara

Pays Bas

LBI • Louis Bolk Instituut

Portugal

IPC • Instituto Politécnico de Coimbra
ITQB NOVA • Instituto de Tecnologia Quimica e Biologica-Universidade Nova de Lisboa

Italie

UNIBO • Alma Mater Studiorum Università di Bologna
UNIPI • Università di Pisa
RSR • Rete Semi Rurali
FORMICABLU • Science communication agency

Chypre

ARI • Agricultural Research Institute

Finlande

LUKE • Natural Resources Institute Finland

Espagne

CSIC • Agencia Estatal Consejo Superior de Investigaciones Cientificas
RAS • Asociacion Red Andaluza de Semillas Cultivando Biodiversidad

Hongrie

ÖMKI • Ökológiai Mezőgazdasági Kutatóintézet

Autriche

ARCHE NOAH • ARCHE NOAH - Vielfalt erleben GmbH

Norvège

FNI • Fridtjof Nansen Institute

DIVERSIFOOD

Ce projet a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne, au titre de la convention de subvention n ° 633571.