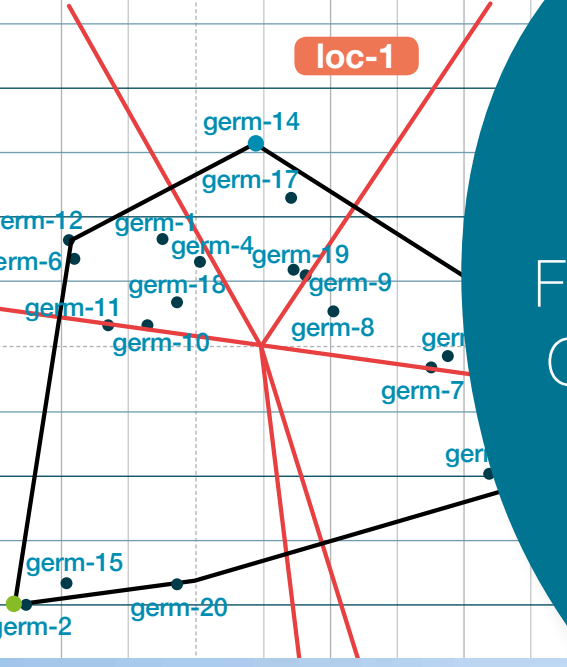
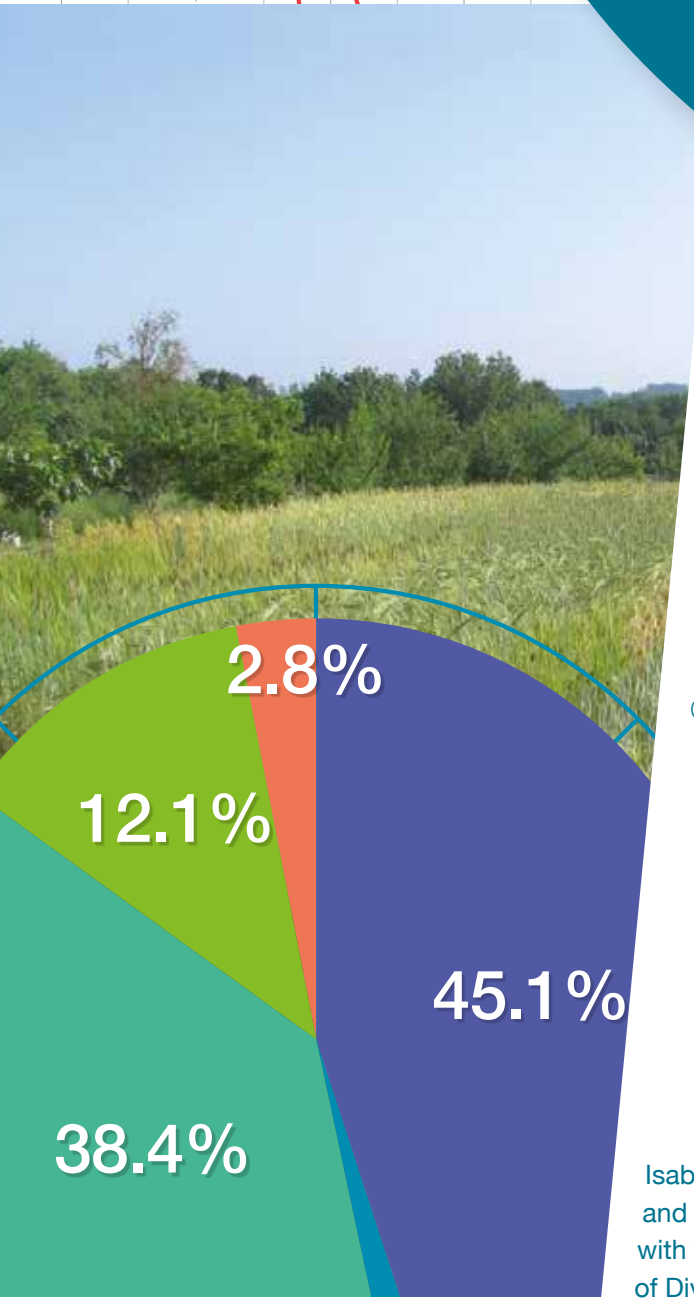


METHODS AND TOOLS FOR DECENTRALIZED ON FARM BREEDING



Booklet #3

This technical booklet describes experimental designs and statistical methods and tools that are relevant for decentralized on-farm breeding.

Isabelle Goldringer (INRA) and Pierre Rivière (RSP), with the contribution of Diversifood partners

DIVERSIFOOD

DIVERSIFOOD is a European H2020 project facing the challenge of promoting a new way of thinking about agriculture. Its ambition is: "*embedding crop diversity and networking for local high quality food systems*".

CONTENTS

INTRODUCTION

Participatory Plant Breeding (PPB) is generally based on decentralized on-farm breeding, which requires particular methods and tools. This technical booklet describes experimental designs and statistical methods and tools that are relevant for decentralized on-farm breeding. Although the participatory dimension is essential in PPB to ensure the empowerment of all actors (farmers, facilitators, processors, gardeners, consumers ...) and to meet their real needs (Sperling et al. 2001), participatory methods are not presented here. However, they are described in details in another Diversifood Technical Booklet entitled *«Methods and methodological framework for multi-actor approaches and participatory plant breeding»*.

This technical booklet describes the possible experimental designs and statistical methods of analysis that can be carried out, according to the objectives and the experimental constraints of the breeding program and the farmers' group. The way to identify and select the most relevant devices and methods is based on a decision tree (Figure 1). The booklet also refers to a user-friendly tool implementing the designs and methods: the R package PPBstats (Rivière, Van Frank, Munoz & David 2018) with full documentation (Rivière, Goldringer & Vindras 2018).

THE INTEREST OF DECENTRALIZING SELECTION

The following is adapted from Bernardo (2002) and Gallais (1990).

When considering multiple environments for evaluation and selection, the phenotypic value of a trait of any individual in a given environment can be written as the sum of its random genetic effect (or overall genetic potential, G), the random environmental effect (E) and the random interaction (G× E), i.e.: P=G+E+G×E+e with e the random residual effect within each environment following a normal distribution N (0, σ^2).

In classical centralized breeding, the objective is to predict the overall genetic potential (G) of the candidates for selection to detect the highest values assuming that this potential would express in all farmers' fields. These genetic potentials are predicted based on the average phenotypic values over all testing environments (usually experimental stations) and therefore the broad sense heritability for prediction is:

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

with nE (resp. nR) the number of environments (resp. the number of replicates in each environment). As environmental effect and G×E interactions limit prediction accuracy,

the option is to increase the number of environments and to use environments that are homogeneous and similar and that minimize G×E interactions.

On the contrary, in decentralized on farm breeding, it has been shown that the environments are very contrasted due to diverse pedo-climatic conditions associated to various agroecological farming practices, and that G×E interactions can be strong (Desclaux et al. 2008). Therefore, the prediction of the overall genotypic value (G) is not interesting and the objective is rather to predict the *«local»* genetic value which also includes the interaction with the local environment, i.e.: Glocij = Gi + (GxE) ij.

Then, the genetic variance in each local environment can be written as:

var (Gloc) = var (G) + var (G×E) and the heritability to predict the local genetic value based on the phenotypic value observed in the local environment is:

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

It can be noted that the G×E interactions contribute to both denominator and numerator therefore leading to no limiting effect on prediction accuracy. Hence, when facing a wide diversity of agroecological environment and practices, decentralized breeding is a key point to select adapted varieties to local agrosystems.


© F. Rey

DESIGN AND STATISTICAL METHODS ACCORDING TO THE OBJECTIVES

The analyses of data from PPB programs aim to address five main objectives that structure the decision tree (Figure 1). The five objectives below apply to four types of information **(agronomic and nutritional traits, sensory traits, molecular data, network of seed circulation)**:

- **To improve the prediction of a target variable for selection** through the analysis of agronomic and nutritional traits.
- **To compare different varieties or populations evaluated for selection** in different locations through the analysis of agronomic and nutritional traits and sensory analysis.
- **To study the response of varieties or populations under selection over several environments** through the analysis of agronomic and nutritional traits.
- **To study diversity structure and identify parents to cross based on either good complementarity or similarity for some traits** through the analysis of agronomic and nutritional traits and molecular data.
- **To study networks of seed circulation** through the analysis of network topology.

For each objective, several methods are available based on different experimental designs according to the objectives and the experimental constraints of the breeding program and the farmers' group (Figure 1). The constraints to be taken into account are the number of plots per location, the number of locations, the number of replicated germplasms within and between locations, etc …, which all depend on the amount of seeds available.


© iStock

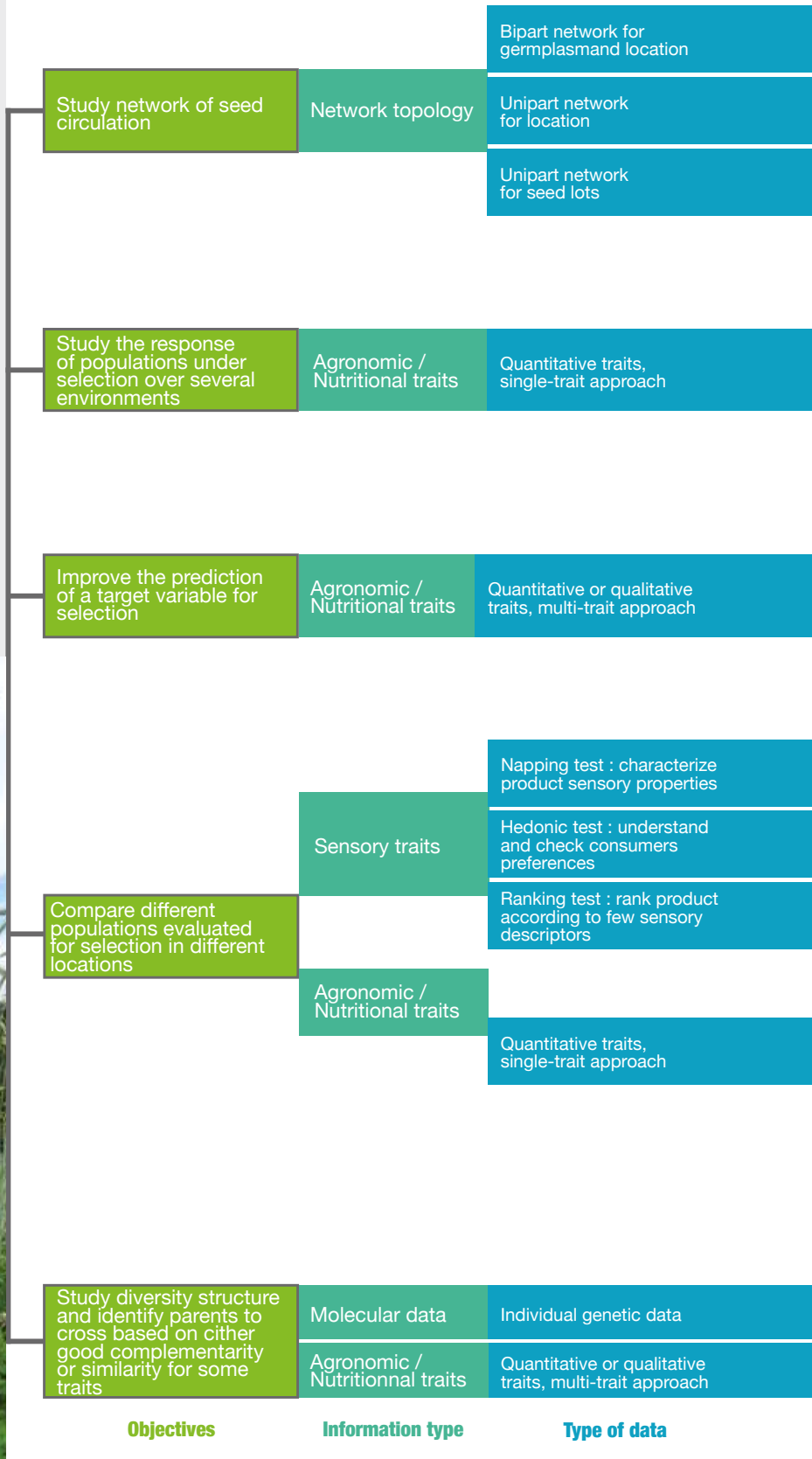

© iStock

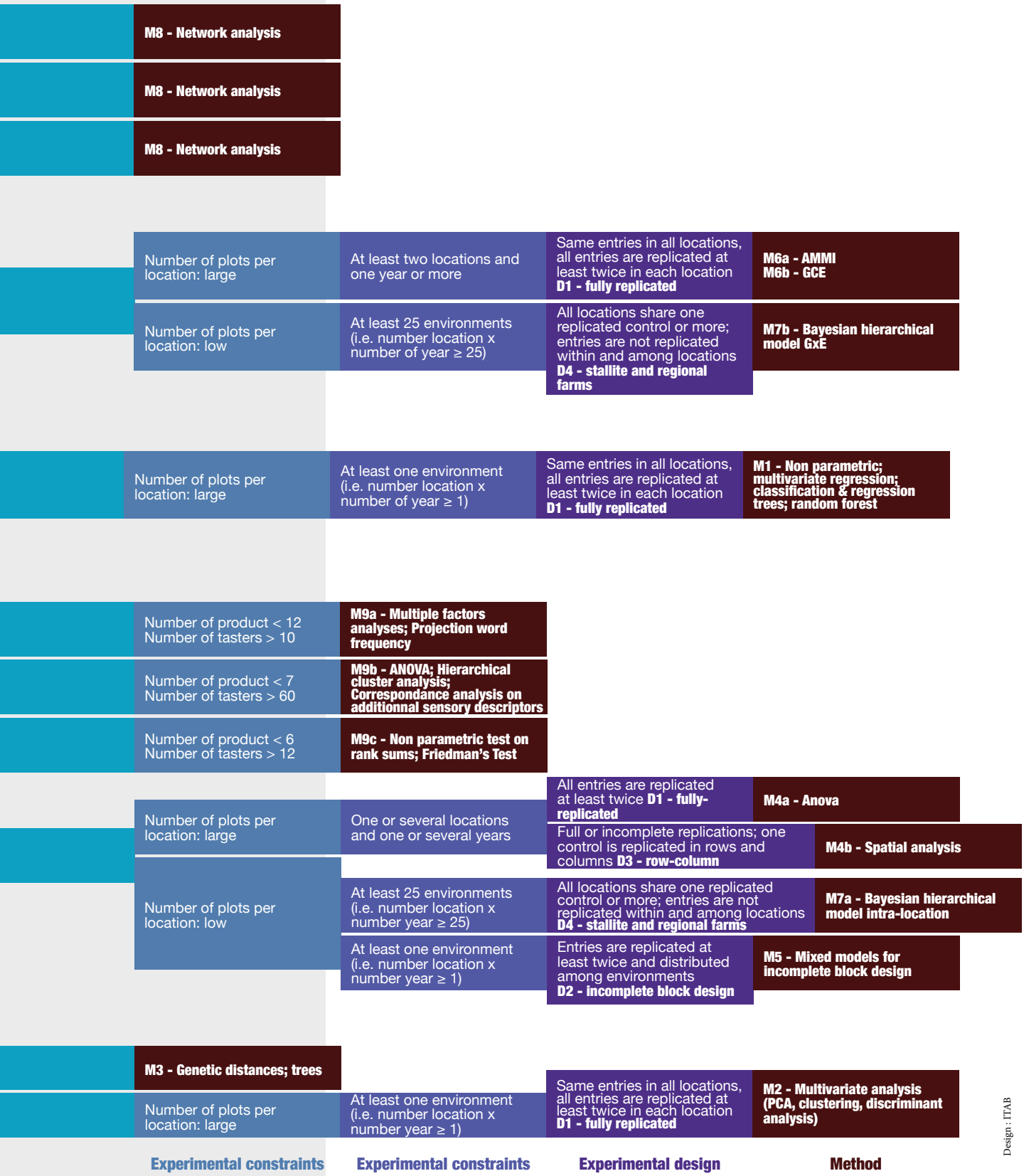
EXPERIMENTAL DESIGNS AND STATISTICALS METHODS FOR PPB

DECISION TREE

In the following, each branch is explained using an example for each experimental design and analysis in the corresponding section. The designs and methods are presented according to the four types of traits **(agronomic and nutritional traits, sensory traits, molecular data, network of seed circulation)** to which they apply. This constitutes the main chapters of the booklet.


© F. Rey



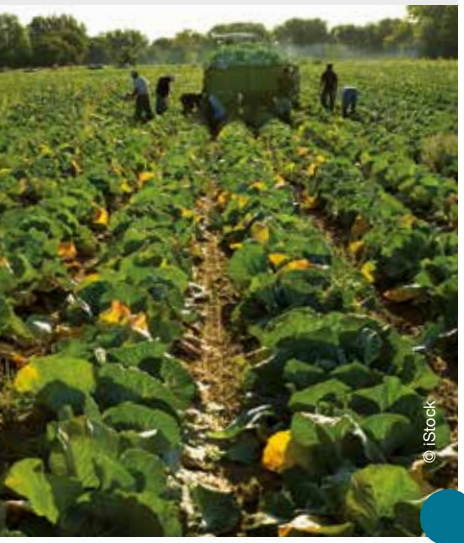

M8 - Network analysis
M8 - Network analysis
M8 - Network analysis
Number of plots per location: large
At least two locations and one year or more
Same entries in all locations, all entries are replicated at least twice in each location D1 - fully replicated
M6a - AMMI
M6b - GCE
Number of plots per location: 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

ANALYSIS OF AGRONOMIC AND NUTRITIONAL TRAITS


© Rupert Pessl

The four main objectives for agronomic analyses are to:

- **Improve the prediction of a target variable for selection.** This can be done through non parametric methods such as:
  - Multivariate regression and classification trees, random forest **(M1)**, based on fully-replicated design **(D1)**.
- **Study diversity structure and identify parents to cross based on either good complementarity or similarity for some traits.**
  - This can be done through multivariate analysis and clustering **(M2)**.
  - It can be completed by the analysis of molecular data and genetic distance trees **(M3)**.
- **Compare different varieties or populations evaluated for selection in different locations.** This can be done through family 1 type of analyses
  - Classic anova (M4a) based on on fully replicated designs **(D1)**,
  - Spatial analysis **(M4b)** based on row-column designs **(D3)**,
  - Mixed models **(M5)** for incomplete blocks designs **(D2)**,
  - Bayesian hierarchical model intra-location **(M7a)** based on satellite-regional farms designs **(D4)**.


© iStock

It can be completed by organoleptic analysis (see below). Based on these analysis, specific objective including studying the response to selection can also be done.

- **Study the response of varieties or populations under selection over several environments.** This can be done through family 2 type of analyses:
  - AMMI and GGE **(M6)** based on on fully replicated designs **(D1)**,
  - Bayesian hierarchical model G× E **(M7b)** based on satellite-regional farms designs **(D4)**.

DATA FORMAT

Depending on the software, data format may be different. Anyhow, the important information needed for the analyses are the location, year, germplasm, bloc, X and Y (row and column) followed by the variables and their corresponding dates if available.

EFFECTS TO BE ESTIMATED
AND TYPES OF ANALYSES

The various effects that can be estimated are:

- **Germplasm:** refers to a variety or population
- **Location:** refers to a farm or a station where a trial is carried out
- **Year**
- **Environment:** refers to a combination of a location by a year
- **Entry:** refers to the occurence of a germplasm in a given environment or location
- **Interaction:** refers to the interaction between germplasm and location or germplasm and environment

Regarding agronomic analyses, two main families are proposed:

- **Family 1** gathers analyses that estimate entry effects. It allows to compare different entries on each location and test for significant differences among entries. A specific analysis to estimate the response to selection can also be conducted. The objective is to compare different germplasms on each location for selection.

- **Family 2** gathers analyses that estimate germplasm and location and interaction effects. This is to analyse the response over a network of locations Estimation of environment and year effects is possible depending of the model. A specific analysis to test for local adaptation based on the local vs foreign model (Blanquart et al 2013) can also be conducted. The objectives is to study response of different germplasms over several locations for selection.

The different models and methods in Family 1 and 2 correspond to experimental designs that are described in the next section and in the decision tree (Figure 1).

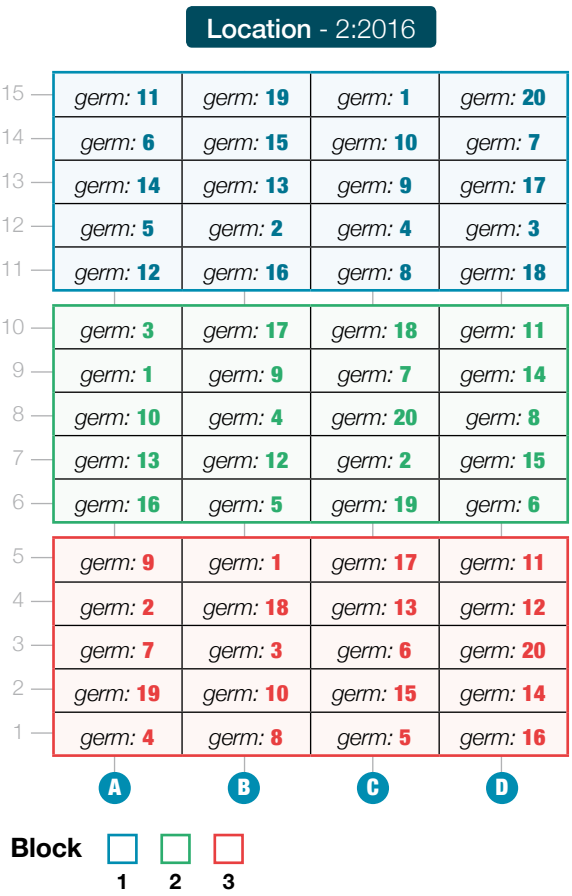
EXPERIMENTAL DESIGNS

The experimental design is described by the number of plots per location, the number of locations, the number of replications of the different germplasms within and between locations. Below are examples of several experimental designs. Each experimental design is followed by a specific analysis as described in the decision tree (Figure 1).

FULLY-REPLICATED DESIGN (D1)

In fully replicated design (Figure 2), all entries are replicated with a random order into different blocks.

Figure 2: Fully replicated design where all germplasms are replicated three time in complete blocks.

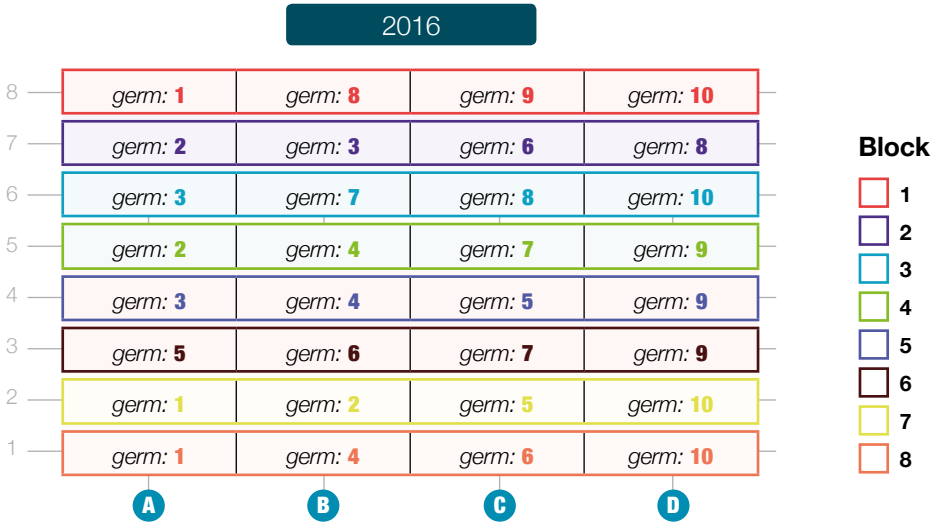


INCOMPLETE BLOCK DESIGN (D2)

In the incomplete block design (Figure 3), entries are not replicated in each location. Some entries are common to some locations. Each block is an independent unit and can be allocated to any location.

Each farmer has to choose one or several predesigned blocks. Therefore, the experiment can be handled by several locations that cannot each receive many plots.

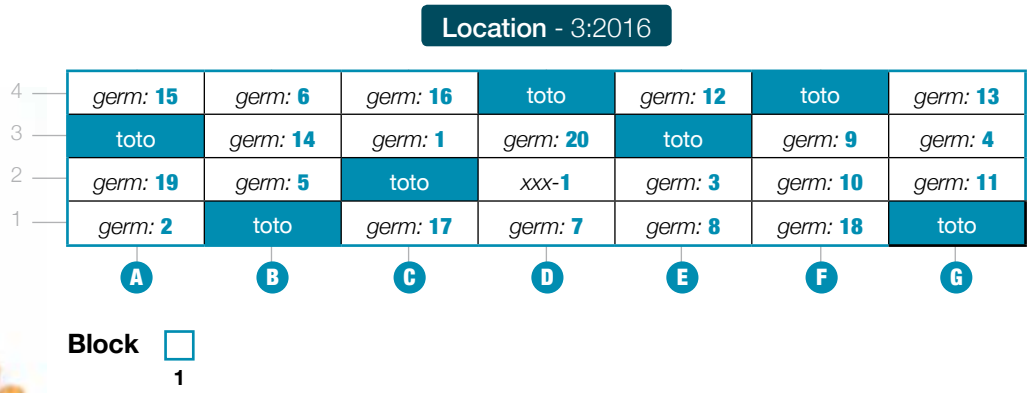
Figure 3: Example of incomplete block design where different germplasms are replicated over some blocks. Blocks are displayed as horizontal lines.



ROW-COLUMN (D3)

In a Row-Column design (Figure 4), a control is replicated in rows and columns to catch the environmental variation as much as possible.

Figure 4: Example of a Row-Column design where a control (toto) is replicated in rows and columns.



© iStock

REGIONAL AND SATELLITE FARMS (D4)

In this device, on farm trials are divided into two types: satellite farms and regional farms (Figures 5 & 6). Regional farms receive several entries (i.e. a germplasm in an environment) in two or more blocks with some entries (i.e. controls) replicated in each block. Satellite farms have a single block and only one entry (i.e. the control) is replicated twice. Farmers choose all entries that are not replicated. The number of entries may vary between farms. Note that at least 25 environments (location x year) are needed in order to get robust results.

Figure 5: Example of a satellite farm design.

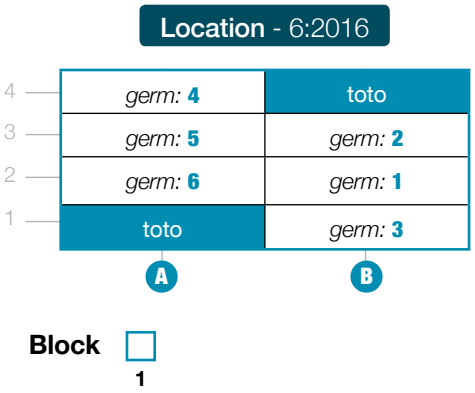
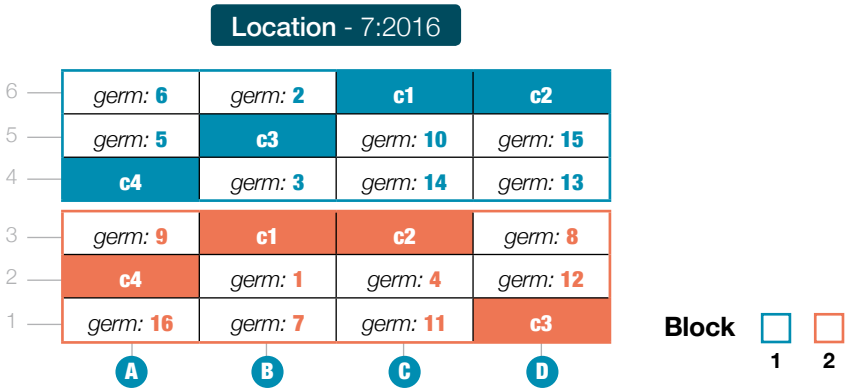


Figure 6: Example of a regional farm design.



© iStock

DATA DESCRIPTION

Once the data have been collected, a first step is to describe them with descriptive statistics and plots such as histograms, barplots, where standard errors are displayed, boxplots, interactions, biplots or radars.

ANALYSIS IN ORDER

TO IMPROVE THE PREDICTION OF A TARGET VARIABLE FOR SELECTION (M1)

The problem is, given a set of p predictive attributes X1 , X2 ,. . . ,X p , to estimate the value of a target attribute y . Denoting the estimator of y by ^y , we have ^y=^f (X1 , X2 ,... , Xp) . An example would be to estimate the yield produced using the maize ear traits as predictive attributes. The ^f function can be obtained by any predictive algorithm, but only algorithms that are able to predict quantitative target attributes have been used. Moreover, we focused on interpretable algorithms, i.e., algorithms that can explain somehow how the value of ^y was predicted given the values of X1 , X2 ,. . . ,X p . Four different algorithms were used:

- Classification And Regression Trees (CART)
- Multivariate Linear Regression (MLR)
- Multivariate Adaptive Regression Splines (MARS)
- Random Forest

Each of these four methods is described below.

CLASSIFICATION AND REGRESSION TREES (CART)

The CART (Breiman et al. 1984) splits, at each iteration, the examples in two subsets. The split is done by choosing the variable and a value that minimizes the sum of the mean squared error of the two resulting subsets. The result of this procedure is a tree like structure where each split is defined by a rule. The interpretation of each leaf-node is obtained by the set of rules in the nodes that defines that leaf-node.

MULTIVARIATE LINEAR REGRESSION

Multivariate linear regression is a well established method that uses the ordinary least squares optimization model in order to adjust a linear model to the training data.

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

MARS (Friedman 1991) was chosen because it has no assumptions and has good interpretability (T., R., and Friedman 2001). It is similar to a stepwise regression but the relations between each dependent variable and the independent one do not need to be linear, because each relation is defined by a set of connected linear segments, instead of a single one. Like linear regression, MARS result is expressed as an equation a bit more complex than linear regression but still interpretable. MARS is used as many times as the number of non-normal independent variables. At each time just one variable is used.

RANDOM FOREST

Random Forest (RF) (Breiman 2001) is a CART based approach, that uses of a set of methods, instead of just one, in order to accomplish its task. RF generates several CART. Each generated CART is different because the tree is trained on a subset of the original set obtained using bagging (Breiman 1996) and using a random subset of the original subset of features at each node. RF results can be interpreted using two different metrics (adapted for regression from (Kuhn J. 2008): (i) the Mean Decrease Accuracy (% IncMSE), which is built by permuting the values of each variable of the test set, recording the prediction and comparing it with the unpermuted test set prediction of the variable. The higher % IncMSE value, the higher the variable importance; (ii) the Mean Decrease MSE (IncNodePurity), which measures the quality of a split for every variable of a tree. Every time a split of a node is made on a variable, the sum of the mean squared error (MSE) for the two descendent subsets is less than the MSE for the parent subset. Adding up the MSE decrease for each individual variable over all the generated trees provides a good indicator. The higher the IncNodePurity value the higher the variable importance.

© Rupert Pessl

MULTIVARIATE ANALYSIS TO STUDY

DIVERSITY STRUCTURE AND IDENTIFY PARENTS TO CROSS BASED ON EITHER GOOD COMPLEMENTARITY OR SIMILARITY FOR SOME TRAITS (M2)

Based on fully replicated design, Principal Component Analysis, clustering and Discriminant analysis can be carried out to identify germplasms that may be used for further crosses.

ANALYSES USED TO COMPARE

DIFFERENT GERMPLASMS EVALUATED FOR SELECTION IN DIFFERENT LOCATIONS (FAMILY 1: M4A, M4B, M5 & M7A)

Four analyses are proposed: classic anova, spatial analysis, mixed models for incomplete block design and a Bayesian hierarchical model. Classic anova **(M4a)** is not explained here as it is a very classic analysis. Only spatial analysis and the Bayesian hierarchical model are detailed.

SPATIAL ANALYSIS (M4B)

The experimental design used is the row-column design **(D3)**. The model is based on frequentist statistics. The model allows taking into account environmental variation within a block with few controls replicated in rows and columns.

It is based on a SpATS (Spatial Analysis of Field Trials with Splines) model proposed by Rodríguez-Álvarez et al. (2016). Environmental variation is accounted for by including row's and column's effects as well as a smooth bivariate function that simultaneously accounts for the spatial trend across both directions of the field (rows and columns).

More information regarding the model as well as example of R package SpATS can be found in Rodríguez-Álvarez et al. (2016). The analysis can be done with PPBstats in several steps : run the model, check model outputs and visualize it, get mean comparisons for germplasms (Figure 7) (https://priviere.github.io/PPBstats_book/family-1.html#spatial-analysis).


© Pro Specie Rara

INCOMPLETE BLOCK DESIGN AND MIXED MODEL (M5)

The experimental design used is the incomplete block design **(D2)**.

The objective of Incomplete block designs is to control the plot-to-plot variation and ideally they should allow the comparisons for all pairs of genotypes (Mead 1997), but this is rarely achievable with large numbers of genotypes and small numbers of replications. Resolvable designs are designs in complete replicated blocks with each replicate split into small incomplete blocks. Lattice designs are a special type of resolvable incomplete blocks where the number of genotypes g is the square of an integer and the block size is $\sqrt{g}$. The introduction of the alpha-designs (Patterson and Williams 1976) removed the restrictions in term of number of genotypes. The advantage of an incomplete block design is that each incomplete block (a sequence of 4 plots in the example shown in Figure 3), is an independent unit and therefore can be allocated to a different field from each of the other incomplete blocks within the same location. The number of incomplete blocks which can be planted on each farm depends only on the farm size. It is also possible that one full replication is planted by a larger farm and the 10 incomplete blocks of another replication in 10 different farms. The disadvantages of this layout are i) the restriction that the total number of entries (g) is a multiple of the block size (k) so that $g = sk$ where s = number of incomplete block per replication; ii) the loss of the row and column design which allows a further increase in precision with spatial analysis. More information can be fond in (Singh and El-Shama'a 2015) (Patterson and Williams 1976) (Mead 1997). The model can be found in (Sarker and Singh 2015).

BAYESIAN HIERARCHICAL MODEL (M7A)

The experimental design used is satellite and regional farms **(D4)**.

At the farm level, the residual has few degrees of freedom, leading to a poor and unstable estimation of the residual variance and to a lack of power for comparing populations. **M7a** was implemented to improve efficiency of the comparison of means. It is efficient with more than 20 environment (i.e. location × year) (Rivière et al. 2015). The model is based on Bayesian statistics.

The model is described in Rivière et al. (2015). The specificity of the model is that the residual term in each environment follows a normal distribution centered on zero with a variance specific to the environment but that is assumed to come from a common distribution of residual variances in all trials of the network. This is reasonable because of the similar structure of the trials in all environments of the network. A hierarchical approach is used and vague prior distributions are placed on the hyperparameters of the distribution. In other words, the residual variance of a trial in a given environment is estimated using all the information available on the network rather than using the data from that particular trial only. Vague prior distributions are also assumed for the germplasm and block parameters.

From an agronomical point of view, the assumption that trial residual variances are heterogeneous (i.e. follow an inverse gamma distribution) is consistent with organic farming: there are as many environments as farms and farmers (practices such as sowing date, sowing density, tilling, etc... pedo climatic conditions, biotic and abiotic stress, ...) leading to a high heterogeneity. Moreover, the residual variances follow an inverse gamma distribution showing conjugate properties that facilitate MCMC convergence.

The residual variance estimated from the controls is assumed to be representative of the residual variance of the other entries. Blocks are included in the model only if the trial has blocks. The analysis can be done with PPBstats in several steps : run the model, check model outputs and visualize it, get mean comparisons for germplasms in each location (https://priviere.github.io/PPBstats_book/family-1.html#model-1).

© Arche Noah

ANALYSES USED TO STUDY

RESPONSE OF GERMPLASMS UNDER SELECTION OVER SEVERAL ENVIRONMENTS (FAMILY 2: M6 & M7B)

Three analyses are proposed: AMMI and GGE **(M6)** and a Bayesian hierarchical model **(M7b)**.

AMMI (M6)

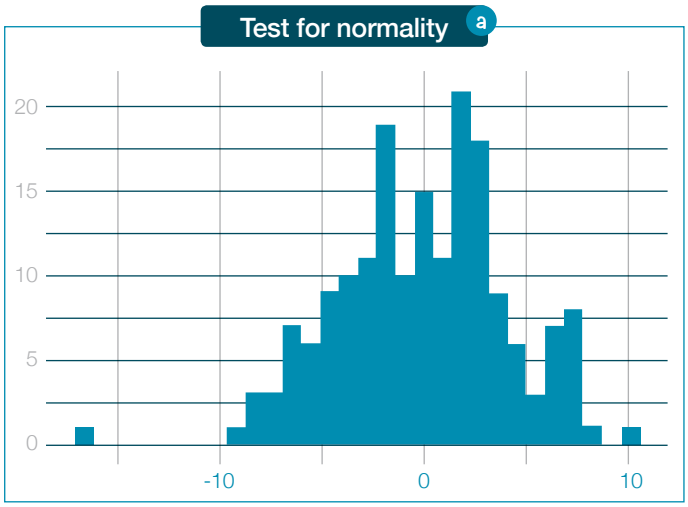
The experimental design used is fully replicated **(D1)**. The Additive Main effects and Multiplicative Interaction (AMMI) model is based on frequentist statistics. The analysis can be broken down in two steps described in Gauch 2006.

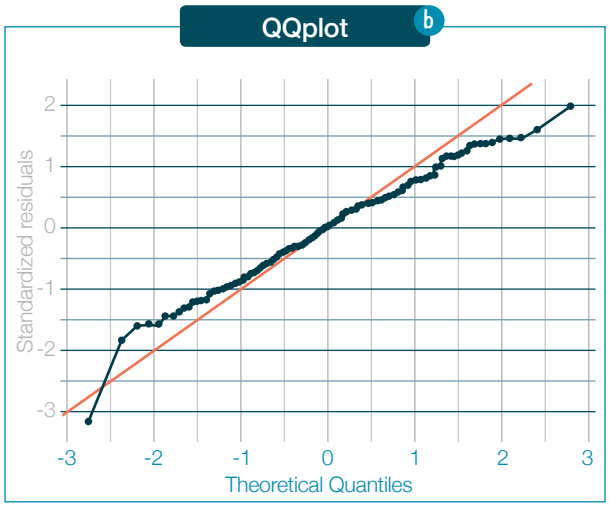
The first step is an ANOVA with germplasm, environment and (germplasm x environment) effects. All other necessary effects such as block in environment, or decomposing environment in location and year effects must be included. Then a Principal Component Analysis (PCA) is run on the interaction terms (matrix of dimensions g x e). The data are double centered on environments and germplasms. The PCA studies the structure of the interaction matrix. The locations are the variables and the germplasms are the individuals. It allows to detect:

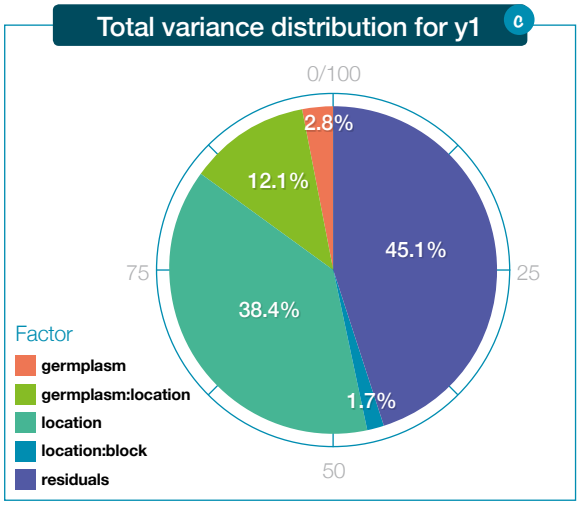
- The germplasms that are stable (i.e. that contribute less to the interaction),
- The germplasms that contribute the most to the interaction and with which environment,
- The locations that have the same profiles regarding interaction.

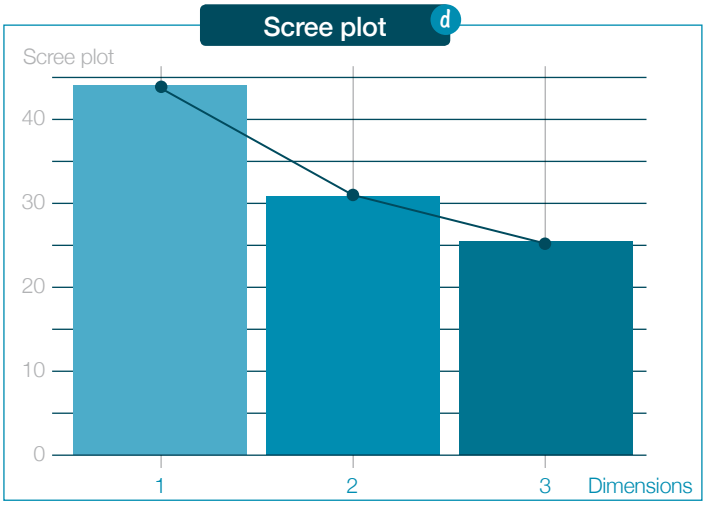
The analysis can be done with PPBstats in several steps : run the model, check model outputs and visualize it, get mean comparisons for each factor from the ANOVA, get biplot from the PCA (https://priviere.github.io/PPBstats_book/family-2.html#ammi).

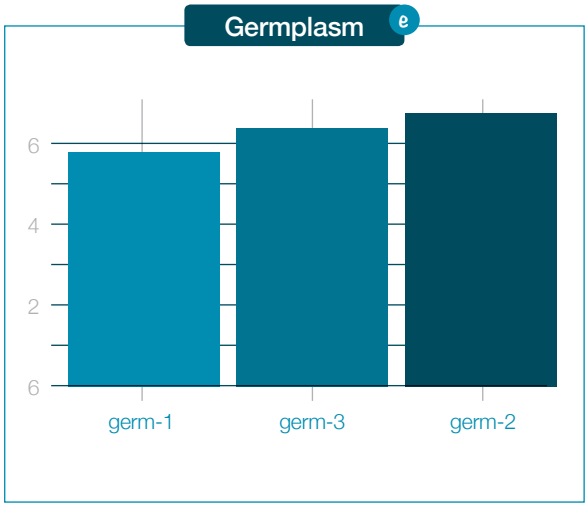
Figure 7.
Example of outputs from PPBstats regarding check of the anova (a.b.c.) and the PCA (d,) ; mean comparisons resulting from the ANOVA for germplasm (e.) and biplot resulting from PCA (f.).

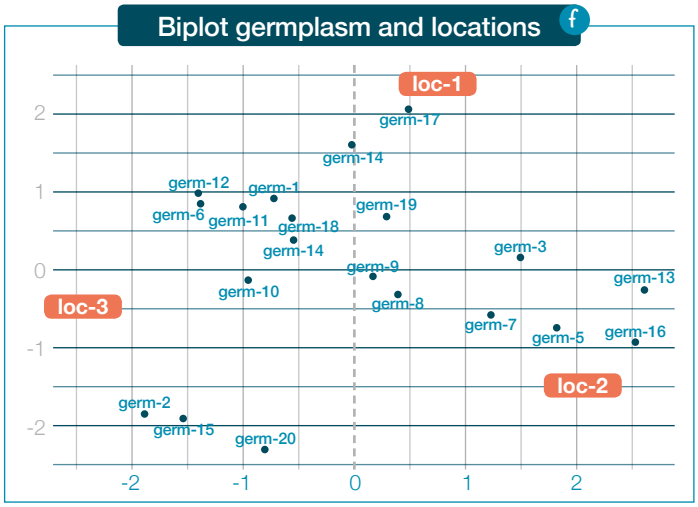

Test for normality
a
20
15
10
5
0
-10
0
10


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)

The experimental design used is fully replicated **(D1)**. The GGE model is the same as the AMMI model except that the PCA is done on a matrix centered on the locations: germplasm and interaction effects are merged. The model is based on frequentist statistics. As for AMMI, the analysis can be done with PPBstats (https://priviere.github.io/PPBstats_book/family-2.html#gge). In addition to AMMI, severals plots can be done that are GGE specifics (Figure 8).

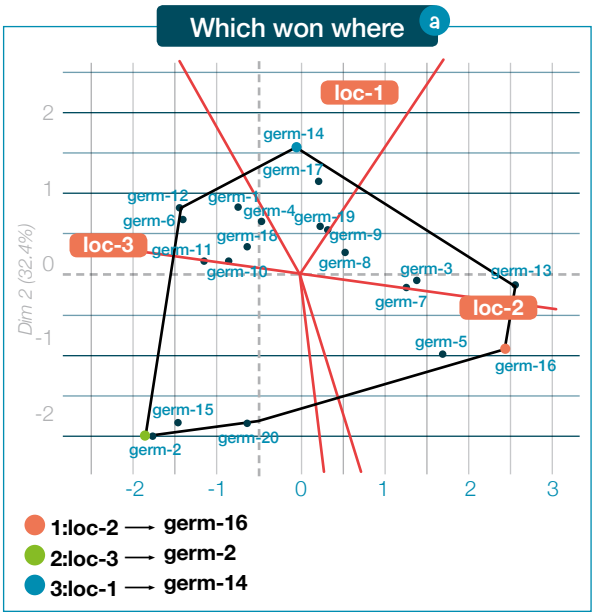
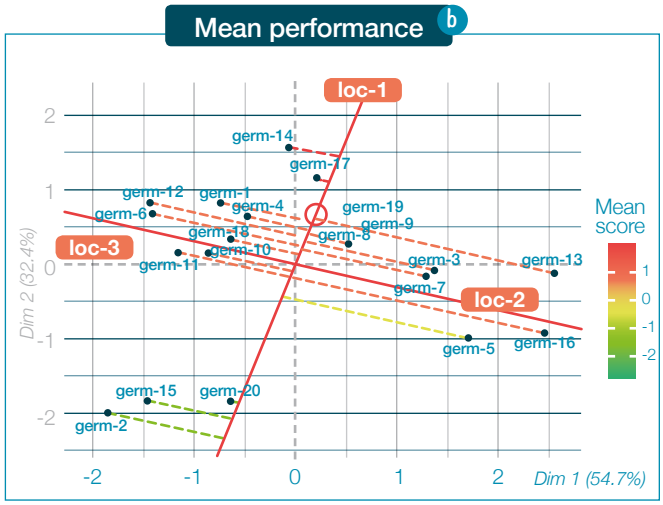
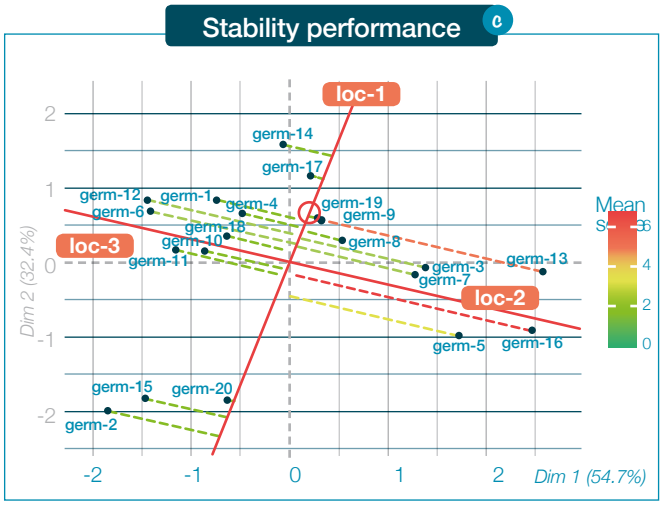
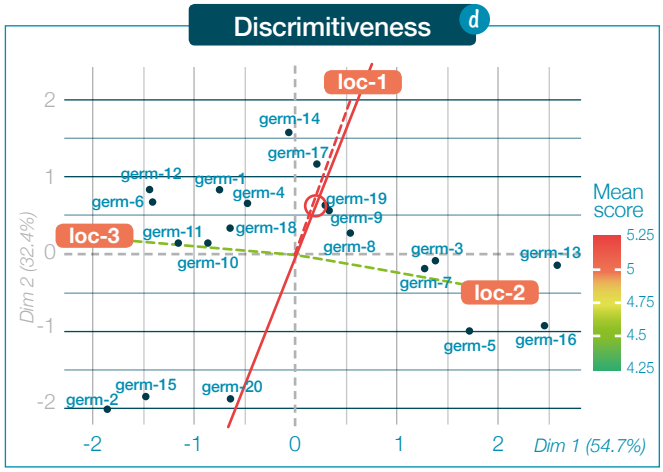
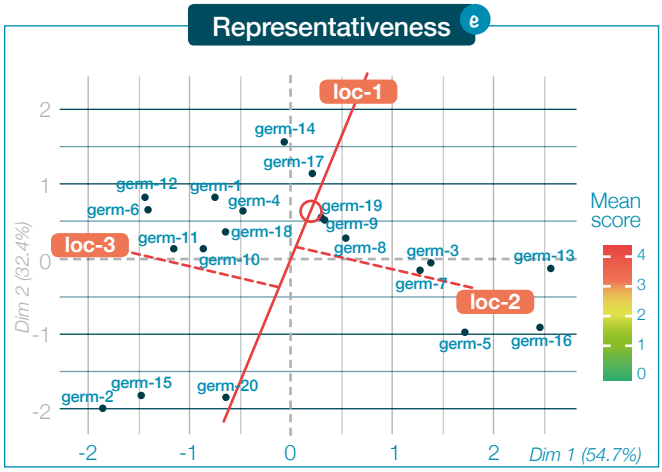


Figure 8: GGE biplot with PPBstats: which won where (a.), mean performance (b.), Stability performance (c.), discrimitivness (d.) and representativeness (e.)









BAYESIAN HIERARCHICAL MODEL (M7B)

The experimental design used is the satellite and regional farms **(D4)**.

M7b is of particular relevance when, at the **network level**, there is a large number of germplasm × environment combinations that are missing, leading to a poor estimation of germplasm, environment and interaction effects. Implementing the **M7b** method requires that the data includes at least around 75 environments with 120 germplasms present in at least two environments (95% of missing G× E combinations). It is based on Bayesian statistics.

The model includes germplasm and environment effects and the interaction is expressed as a multiplicative term consisting in the environment effect times a regression coefficient that depends on the germplasm. The remaining part of the interaction goes into the residual. The model can further be reduced in a germplasm effect plus a multiplicative term consisting in the environment effect times a coefficient that represents the **sensitivity** of each germplasm to the environments. This model is known as the Finlay Wilkinson model or as the joint regression (1963).

Germplasms' sensitivity quantifies the stability of germplasms' performances over environments. The average sensitivity is equal to 1 so that a gemplasm with a value greater (resp. lower) than 1 is more (resp. less) sensitive to environments than a germplasm with the average sensitivity (Nabugoomu, Kempton, and Talbot 1999).

Given the high disequilibrium of the data and the large amount of data, this model is implemented with a hierarchical Bayesian approach. Hierarchical priors are used for the germplasm, environment and sensitivity effects while a vague prior is used for the residual variance.

The analysis can be done with PPBstats in several steps : run the model, check model outputs and visualize it, perform cross validations studies, get mean comparisons for each factor, predict the past (Figure 9) (https://priviere.github.io/PPBstats_book/family-2.html#model-2).

Figure 9. Example of plots to predict the past with PPBstats: there are two values: one estimated by the model (i.e. the combinaison germplasm x location exists in the data set) and one predicted by the model for germplasm that are not present in a given location.

SENSORY ANALYSIS


© iStock

In the following, the products that are tasted through organoleptic analysis should be seen as extensions of the varieties or populations and the sensory analyses as phenotypic evaluations that require particular statistical analyses. All the analysis can be done with PPBstats (https://priviere.github.io/PPBstats_book/organoleptic.html).

NAPPING TEST (M9A)

The Napping allows to look for sensory differences between products. Differences are on global sensory characteristics and should be complemented with a verbalisation task to ease the understanding of the differences. It offers greater flexibility, as no trained panel is needed.

Two tasks are done in a Napping:

- The **sorting task**: each taster is asked to position the whole set of products on a sheet of blank paper (a tablecloth) according to their similarities/dissimilarities. Thus, two products are close if they are perceived as similar or, on the contrary, distant from each other if they are perceived as different. Each taster uses his/her own criteria.
- The **verbalisation task**: After performing the napping task, the panelists are asked to describe the products by writing one or two sensory descriptors that characterize each group of products on the map.

Panels should be composed of 12 to 25 tasters according to the judge's experience with the product and to the objective of the experiment. For example ten farmersbakers should be enough to have reliable results as they are used to eat and taste bread. In case of consumers, a panel of twenty could be more adapted.

No more than ten products should be evaluate simultaneously. A random, three-digit code should be assigned to each sample. Samples are presented simultaneously and the assessors can taste as much as they need. Napping data lead to a quantitative table. The rows are the products. Two columns per panelist give the spatial coordinates (x, y) of each product.


© iStock

Sensory descriptors are coded through a *«products x words»* frequency table. First a contingency table counting the number that each descriptor has been used to describe each product is created. Then this contingency table is transformed in frequencies so that the *«word frequency»* becomes a qualitative variable with the number of words cited as modalities.

A Multiple Factor Analysis (MFA) is performed. Each subject constitute a group of two un-standardized variables. The MFA leads to a synthesis of the panelists' tablecloth. Two products are close if all judges consider them close on the napping.

The more the first two components of MFA explain the initial variability, the more judges agreet.

The frequency table crossing products and word frequency is considered as a set of supplementary variables: they do not intervene in the axes construction but their correlations with the factors of MFA are calculated and represented as in usual PCA (Figure 10).

Figure 10. Exemple of MFA with PPBstats.

HEDONIC TEST (M9B)

The hedonic evaluation test involves asking consumers to rate their preference from 1 (I dislike extremely) to 9 (I like very much) for 3 to 4 sensory attributes specific to the test product. The overall preference is ascertained at the beginning of the questionnaire in order not to influence the consumer and be closer to typical conditions of consumption. Additional information concerning sex, age and organic consumption frequency are asked at the end of the test in order to characterize the population sample. Additional sensory descriptors to describe products are asked after evaluation of each product. One of the main objectives of hedonic tests is to determine differences of appreciation for a given attribute between a set of samples. The data distribution determines the type of tests that should be used to analyze the data set.

- If the distribution is Normal, one-way analysis of variance (ANOVA) can be performed, the source of variance being the sample, followed by multiple comparison of mean data values from each assessor. The aim is to obtain a final ranking based on consumers' preferences.
- If the data set doesn't follow a Normal distribution, a Friedman test on the rank should be used to indicate if the varieties are perceived differently by assessors.

Finally a Hierarchical Cluster Analysis can be implement to identify groups of preferences.


© iStock

RANKING TESTS (M9C)

A panel of assessors compares several products simultaneously and ranks them according to the perceived magnitude of a given sensory characteristic (e.g. acidity, fibrousness). This method has the advantage of being easy to implement. The jury ideally includes 12 semi-naive assessors (consumers initiated to sensory analyses) according to the ISO 8587 standard[1], although it is possible to highlight significant differences with a smaller number of assessors. Key characteristics:

- Products are presented simultaneously This requires that the whole set of samples to be tested is available at the same time. Some vegetable species show marked differences in precocity (e.g. broccoli), and therefore care should be taken to ensure that samples of the same precocity are compared.
- The assessors can taste as much as they need.
- When they answer, assessors cannot put any two products at the same rank, i.e. all ranks assigned must be unique.

It is advised not to exceed 6 samples per session. Null hypothesis (H0): all varieties have exactly the same responses (rank means are equal) Friedman's test (non parametric test on k independent samples) leads to the rejection or acceptance of this hypothesis, based on α value (<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.

MOLECULAR ANALYSIS (M2)

© iStock

Molecular analysis can be used to study diversity structure and identify complementary or similar parents for cross through genetic distances and trees. They are based on individual genetic data.

NETWORK ANALYSIS (M8)


© iStock

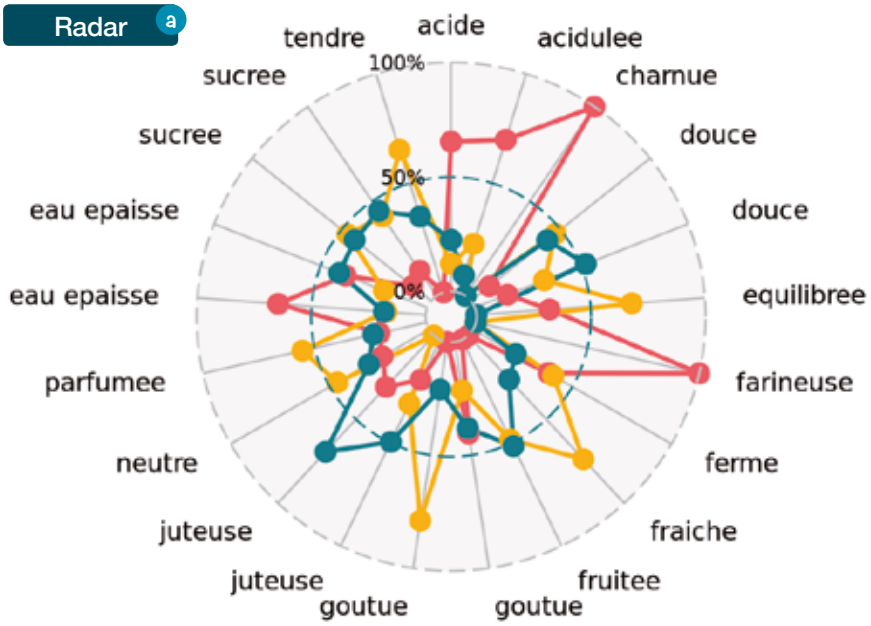
Describing the topology of networks of seed circulation is interesting since it gives insight on how exchanges are organized within a PPB programme or a Community Seed Bank (Vernooy, Shrestha, and Sthapit 2015) (Pautasso et al. 2013). Analysis can be done at several geographical or organizing scales, for example local, regional or national. Two types of network can be studied: (I) unipart networks : - where a node can be a seed lot (i.e. a combinaison of a germplasm in a given location a given year) and edges are relationships such as diffusion, mixture, reproduction, crosses or selection for example ; - where a node can be a location and edges are diffusion events between locations; (II) bipart networks where a node can be a location or a germplasm.

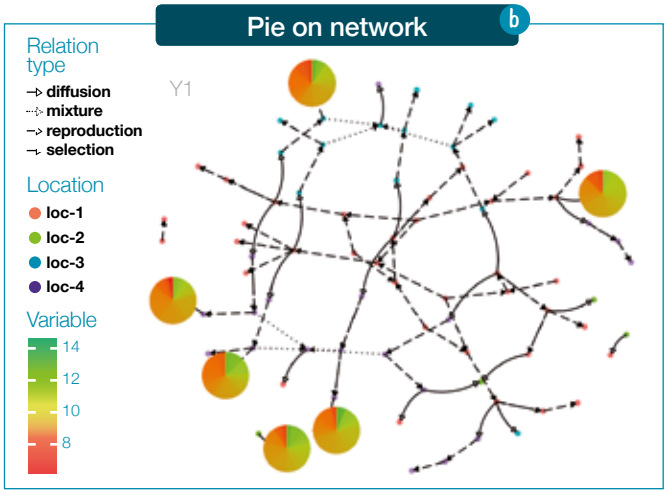
DESCRIPTIVE ANALYSIS

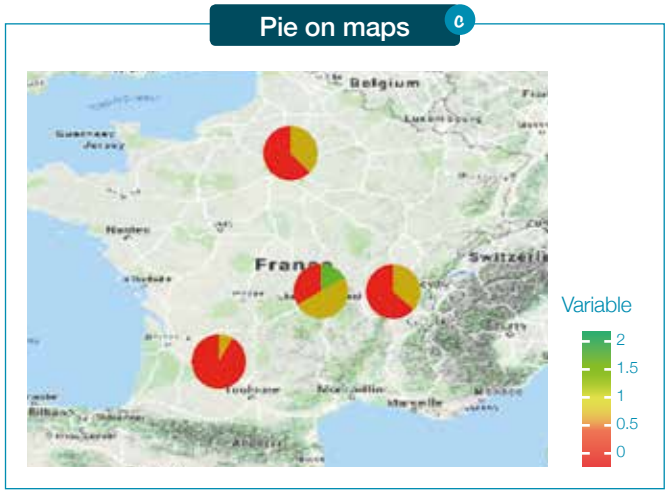
Descriptive analyses can be carried out to better understand how exchanges are organized within a CSB or a breeding programme (Figure 11). Unipart networks of seed-lots can be displayed in the chronological order. Barplots can be used to show the repartition of germplasms per location or per year. In unipart networks of locations, diffusion events between locations and their frequencies can be displayed. Bipart networks of germplasms and locations display the relationships between germplasms and locations (i.e. which germplasm in which location).

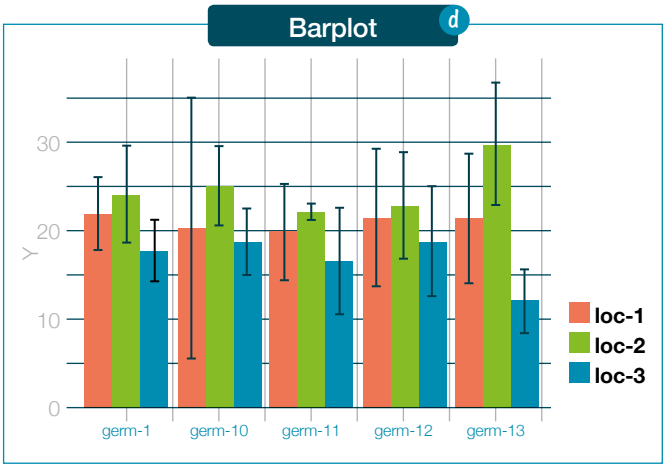

© iStock

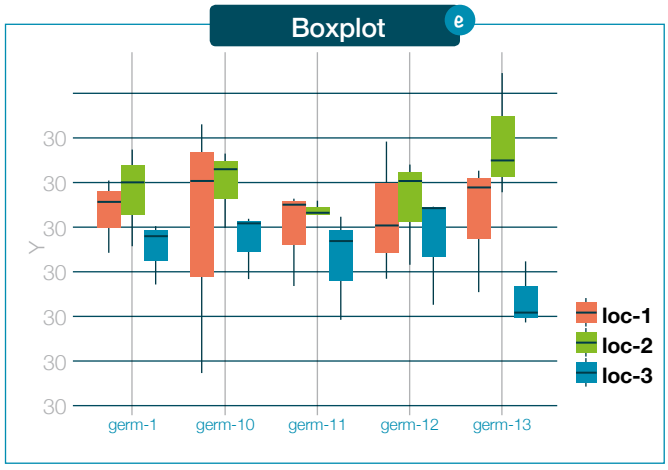
Figure 11.
Examples of descriptive plots in PPBstats : radar (a.), pie on network (b.), pie on maps (c.), barplot (d.), boxplot (e.).











TOOLS TO IMPLEMENT THE METHODS : THE R PACKAGE PPBSTATS


© iStock

A first set of these methods has been consistently implemented in a new software: the R package PPBstats.

This package is based on R software that is open source and widely used in breeding and agronomy community. PPBstats aims at performing the analyses found within PPB programmes at four levels:

- Agronomic trials
- Organoleptic tests
- Molecular experiments
- Network of seeds circulation

PPBstats is still under development and the code is hosted on Github to facilitate collaboration: https://github.com/priviere/PPBstats

The following experimental designs can be used :

- **D1:** fully-repicated block design
- **D2:** incomplete block design
- **D3:** row-colum design
- **D4:** satellite-farms & regional-farms

The following methods have been implemented (but further tests are welcome):

- **M2:** Multivariate analyses (PCA)
- **M4a:** Classical ANOVA
- **M4b:** Spatial analysis
- **M6:** AMMI and GGE
- **M7a:** Bayesian hierarchical model intra-location
- **M7b:** Bayesian hierarchical model GxE
- **M8:** Network analyses

The following methods are not yet implemented and can be done through other software:

- **M1:** Non parametric; multivariate regression; classification & regression trees; random forest:
  - Classification And Regression Trees (CART): rpart, the recursive partitioning algorithm, is the function used to train Classification And Regression Trees (CART). The function rpart is available in the R package rpart.

- Multivariate Linear Regression (MLR): lm, the linear model function is available in the base package of R. I.e., you don't need to open any specific R package.
- Multivariate Adaptive Regression Splines (MARS): function earth from the R-project.
- Random Forest: the function random Forest from the R package randomForest.

- **M2:** Multivariate analyses (clustering, discriminant analysis):
  - R package FactoMineR, *http://factominer.free.fr/index.html*
- **M3:** Genetic distances; trees: R package adegenet. Diversity analysis can be done through:
  - 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:** Mixed models for incomplete block designs: Genstats module.
- **M9a:** Multiple Factors Analyses; Projection Word Frequency:
  - R packages FactoMineR, *http://factominer.free.fr/index.html*
  - R package SensoMineR, *http://sensominer.free.fr/*
- **M9b:** ANOVA; Hierarchical Cluster Analysis; Correspondance Analysis on additionnal sensory descriptors:
  - R packages FactoMineR, *http://factominer.free.fr/index.html*
  - R package SensoMineR, *http://sensominer.free.fr/*
- **M9c:** Non parametric Test on Rank Sums; Friedman's Test: bascis R functions.

Methods **M2, M5, M9a, M9b** and **M9c** will be implemented in PPBstats by the end of Diversifood project.

A website dedicated to PPBstats and the exhaustive tutorial to collaborate and use the package can be found here : https://priviere.github.io/PPBstats_web_site.


© F. Rey

REFERENCES


© 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.

Booklet #3

Authors: Isabelle Goldringer (INRA) and Pierre Rivière (RSP)

Editors: Frédéric Rey (ITAB), Isabelle Goldringer (INRA) and Pierre Rivière (RSP)

Acknowledgements: We thank 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) for their contribution to some methods' description.

How to cite this document: Goldringer I., Rivière P., 2018. Methods and Tools for decentralized on farm breeding. Booklet #3. Diversifood Project.

December 2018

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

Contact: isabelle.goldringer@inra.fr

www.diversifood.eu

21 partners
DIVERSIFOOD CONSORTIUM

France

INRA • Institut National de la Recherche Agronomique
ITAB • Institut Technique de l'Agriculture Biologique
RSP • Réseau Semences Paysannes
IT • INRA Transfert

UK

ORC • Organic Research Centre

Switzerland

FiBL • Forschungsinstitut für biologischen Landbau
PSR • ProSpecieRara

The Netherlands

LBI • Louis Bolk Instituut

Portugal

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

Italy

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

Cyprus

ARI • Agricultural Research Institute

Finland

LUKE • Luonnonvarakeskus

Spain

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

Hungary

ÖMKI • Ökológiai Mezőgazdasági Kutatóintézet

Austria

ARCHE NOAH • Arche Noah Schaugarten GMBH

Norway

FNI • Fridtjof Nansen Institute

DIVERSIFOOD

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 633571.