DIVERSIFOOD

Embedding crop diversity and networking for local high quality food systems

Grant agreement n°: 633571

H2020 - Research and Innovation Action

D3.1 : *Smart methods specifically suited for decentralized on farm breeding*

Due date: M28
Actual submission date: April 4th, 2018

Project start date: March 1st, 2015 **Duration:** 48 months
Workpackage concerned: WP3
Concerned workpackage leader: Isabelle Goldringer

Lead Beneficiary: RSP / INRA
Responsible authors: Pierre Rivière / Isabelle Goldringer
Additional author: Salvatore Ceccarelli, João Mendes Moreira, Pedro Mendes Moreira, Gaelle van Frank, Carlota vaz Patto, Camille Vindras

Dissemination level:

✚ **PU:** Public (must be available on the website)

□ **CO:** Confidential, only for members of the consortium (including the Commission Services)

□ **CI:** Classified, as referred to in Commission Decision 2001/844/EC

Abstract

D3.1 describes designs and statistical methods that are relevant for PPB experiments and on-farm decentralized trials. Statistical analyses have been identified, adapted or developed by the partners involved in this activity (RSR, IPC, ITQB, RSP, ITAB, INRA). They are presented following a decision tree organised according to the objectives of the experiments. Most of these methods have been implemented in a R Package, PPBStats, that constitutes the deliverable ***D3.2***: *User-friendly tools incorporating the relevant methods for decentralized on farm breeding*.

Table of content

Abstract.....2

What is Participatory Plant Breeding ?.....4

Decentralize the selection.....4

Involve all actors in the breeding decision process.....5

Designs and statistical methods according to the objectives.....5

Analysis of agronomic traits.....7

Running the analyses.....7

Data format.....8

Experimental designs.....8

Fully-replicated design (D1).....8

Incomplete Block Design (D2).....9

Row-column (D3).....10

Regional and satellite farms (D4).....11

Describe the data.....12

Analysis in order to improve the prediction of a target variable for selection (M1).......12

Classification And Regression Trees (CART).....13

Multivariate Linear Regression.....13

Multivariate Adaptive Regression Splines (MARS).....13

Random Forest.....13

Multivariate analysis to study diversity structure and identify parents to cross based on either good complementarity or similarity for some traits (M2)....................................15

Analyses used to Ccompare different germplasms evaluated for selection in different locations (Family 1: M4a, M4b, M5 & M7a)..15

Spatial analysis (M4b)..16

Incomplete block design and mixed model (M5)..17

Bayesian hierarchical model (M7a)..19

Analyses used to study response of germplasms under selection over several environments (Family 2: M6 & M7b)..21

AMMI (M6)..21

GGE (M6)...23

Bayesian hierarchical model (M7b)..24

Organoleptic analysis...26

Napping Test (M9a)..27

Hedonic Test (M9b)...28

Ranking Tests (M9c)..28

Molecular analysis (M3)..29

Network analysis (M8)...29

Data format...30

Descriptive analysis..31

References...31

Appendix...33

What is Participatory Plant Breeding ?

Decentralize the selection

The following development 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 ($G\times E$), i.e.: $P=G+E+G\times E+e$ with e the random residual effect within each environment following a normal distribution $N(0,\sigma^2)$.

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:

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

with nE (resp. nR) the number of environments (resp. the number of replicates in each environment). As environmental effect and $G\times E$ 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 $G\times E$ 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 $G\times E$ 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 of genotype *i* in environment *j*, $Gloc_{ij}$ which also includes the interaction with the local environment,

i.e.: $Gloc_{ij}=G_i+(G\times E)_{ij}$

Then, the genetic variance in each local environment can be written as: $var(Gloc)=var(G)+var(G\times E)$ and the heritability to predict the local genetic values based on the phenotypic values observed in the local environments is:

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

It can be noted that the $G\times E$ interaction contributes to both denominator and numerator therefore leading to no limiting effect on prediction accuracy. Hence, when facing a wide diversity of agroecological environments and practices, decentralized breeding is a key point to select adapted varieties to local agro-systems.

Involve all actors in the breeding decision process

All actors can be part of the breeding programme: farmers, processors, technicians, researchers, facilitators, consumers ... Such involvement empowers all actors and may better answer the real needs of the actors (Sperling et al. 2001).

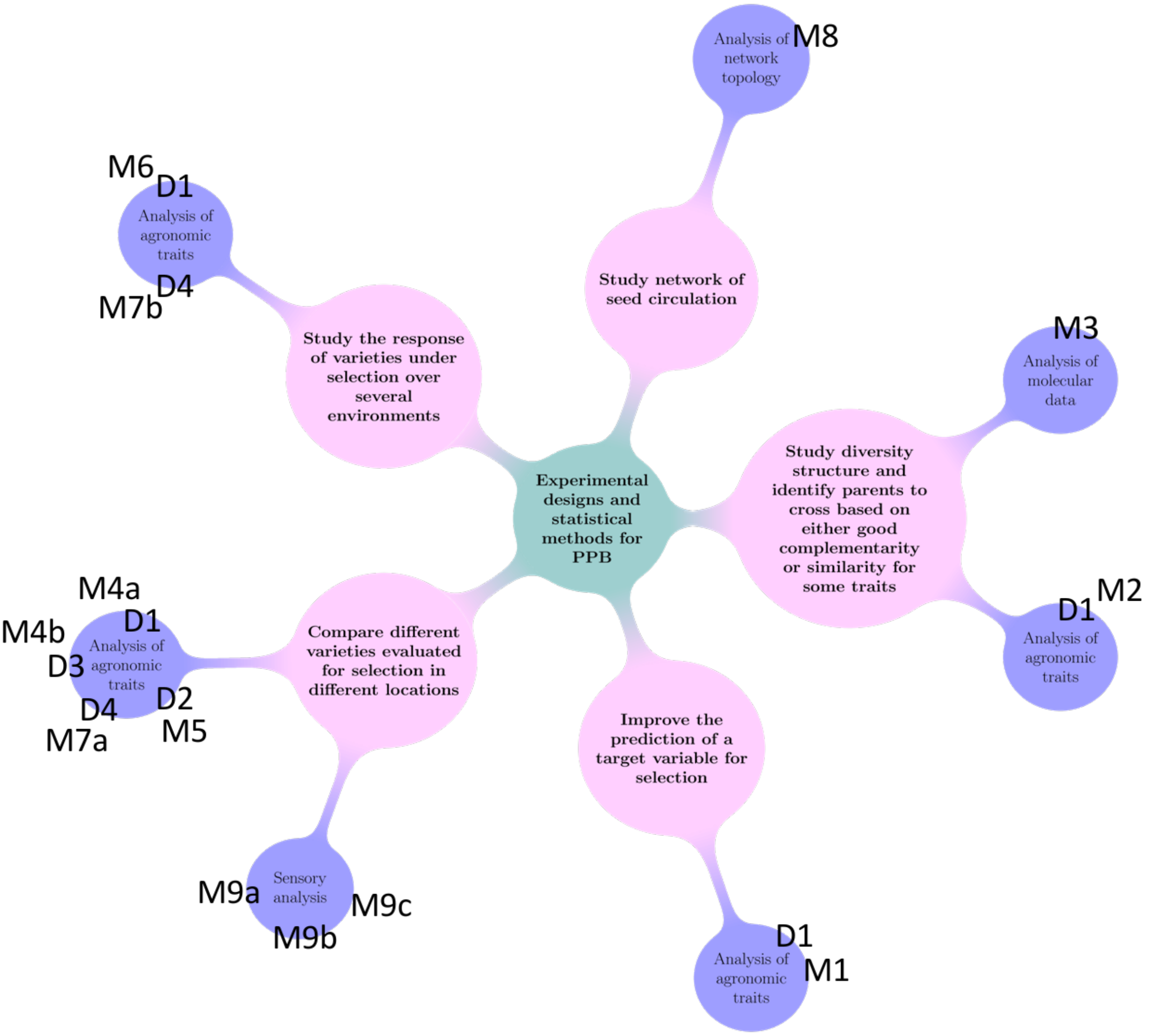
Designs and statistical methods according to the objectives

The analyses of data from PPB programmes aim to address five main objectives:

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

Then, for each objective, there are several methods based on different experimental designs based on number of plots per location, the number of locations, the number of replicated germplasms within and between locations ... all being dependant to the amount of seeds available.

The figure in the Appendix presents a decision tree with all objectives, experimental constraints in order to choose and optimize experimental designs and methods of analysis. Each branch is explained through an example for each experimental design and analysis in the corresponding section. Below is a simplified version showing only the objectives and traits to be analysed. Once an experimental design and a methods is used, sowing can be done!


Analysis of network topology
M8
Study network of seed circulation
M6
D1
Analysis of agronomic traits
D4
M7b
Study the response of varieties under selection over several environments
M3
Analysis of molecular data
Experimental designs and statistical methods for PPB
Study diversity structure and identify parents to cross based on either good complementarity or similarity for some traits
M2
D1
Analysis of agronomic traits
M4a
M4b
D1
D3
Analysis of agronomic traits
D2
D4
M7a
M5
Compare different varieties evaluated for selection in different locations
Improve the prediction of a target variable for selection
M9a
Sensory analysis
M9c
M9b
D1
Analysis of agronomic traits
M1

Analysis of agronomic traits

The four main objectives in PPB 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 analysis of molecular data and genetic distance trees (**M3**).
- **Compare different varieties evaluated for selection in different locations**. This can be done through family 1 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**).

It can be completed by organoleptic analysis. Using these designs and analyses, particular comparisons of populations (i.e. estimating response to selection) can also be carried out.

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

Running the analyses

After describing the data, you can run statistical analysis. The various effects that can be estimated are:

- **germplasm**: a variety or a population,
- **location**: a farm or a station where a trial is carried out,
- **environment**: a combination of location by year,

- **entry**: the occurrence of a germplasm in a given environment or location,
- **interaction**: interaction between germplasm and location or germplasm and environment or germplasm and year.

Regarding agronomic analyses, two main families are proposed:

- **Family 1** gathers analyses that estimate entry effects. It allows to compare different entries in each location and test for significant differences among entries. Specific analysis including response to selection can also be carried out. The objective is to compare different germplasms in each location in order to apply selection.
- **Family 2** gathers analyses that estimate germplasm, location and interaction effects. This is to analyse the response over a network of locations. Estimation of location and year effects is possible depending on the model. Specific analysis including migrant and resident can also be done. It allows studying the response of germplasm over several locations or environments. The objective 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 listed in the next section and in the decision tree.

Data format

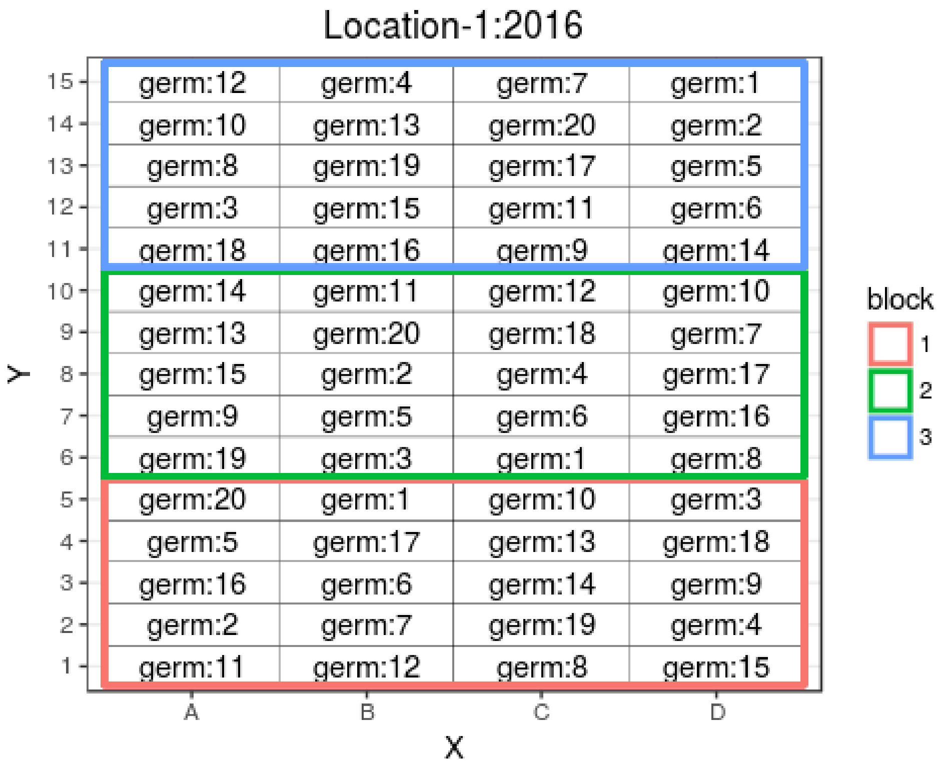
Depending on the software you use, data format may be different. Anyhow, the important information needed for analysis is location, year, germplasm, bloc, rows and columns followed by the variables and their corresponding dates if available.

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 and possibly the controls and the way they are replicated within and among trials. Below are examples of several experimental designs. Each experimental design is followed by a specific analysis as described in the decision tree.

Fully-replicated design (D1)

In a fully replicated design, all entries are replicated with independent and randomly chosen orders in the different blocks.

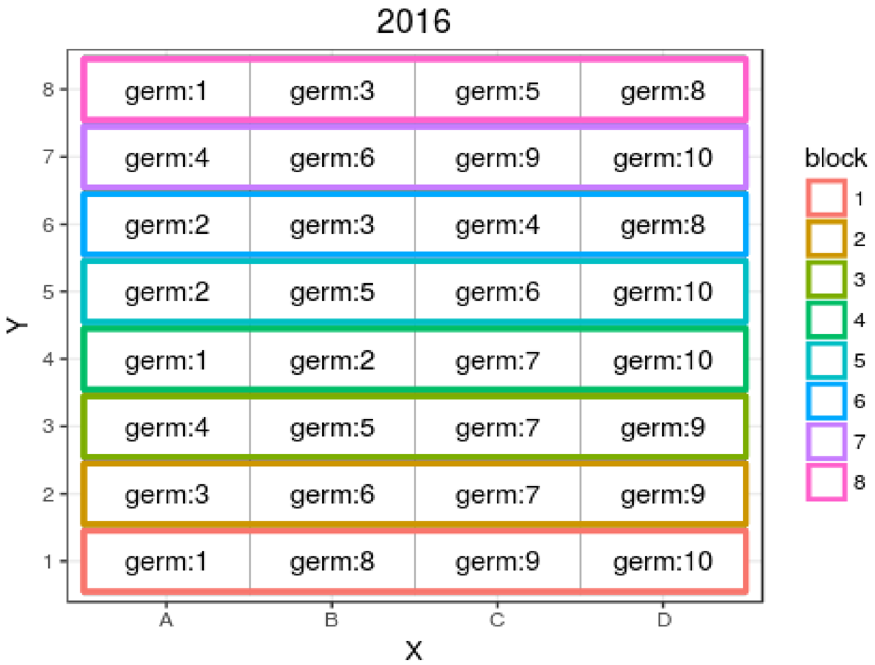


Y	A	B	C	D
15	germ:12	germ:4	germ:7	germ:1
14	germ:10	germ:13	germ:20	germ:2
13	germ:8	germ:19	germ:17	germ:5
12	germ:3	germ:15	germ:11	germ:6
11	germ:18	germ:16	germ:9	germ:14
10	germ:14	germ:11	germ:12	germ:10
9	germ:13	germ:20	germ:18	germ:7
8	germ:15	germ:2	germ:4	germ:17
7	germ:9	germ:5	germ:6	germ:16
6	germ:19	germ:3	germ:1	germ:8
5	germ:20	germ:1	germ:10	germ:3
4	germ:5	germ:17	germ:13	germ:18
3	germ:16	germ:6	germ:14	germ:9
2	germ:2	germ:7	germ:19	germ:4
1	germ:11	germ:12	germ:8	germ:15

Fully replicated design where all germplasms are replicated three times in blocks.

Incomplete Block Design (D2)

Entries are not replicated in a 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 pre-designed blocks. Therefore, the experiment can be handled by several farmers (in several locations) who cannot receive a large number of plots.

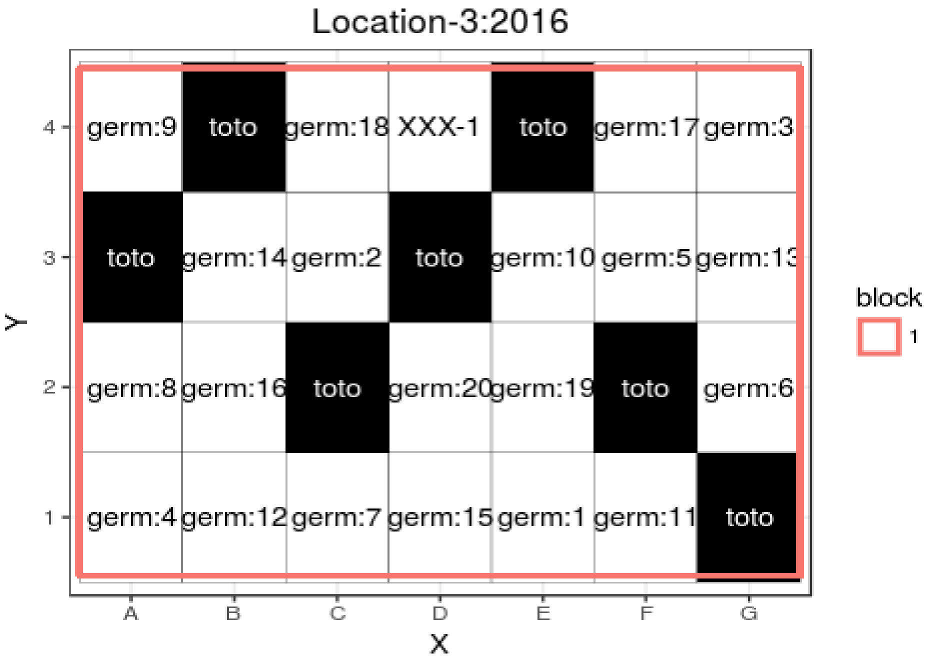


Y	A	B	C	D
8	germ:1	germ:3	germ:5	germ:8
7	germ:4	germ:6	germ:9	germ:10
6	germ:2	germ:3	germ:4	germ:8
5	germ:2	germ:5	germ:6	germ:10
4	germ:1	germ:2	germ:7	germ:10
3	germ:4	germ:5	germ:7	germ:9
2	germ:3	germ:6	germ:7	germ:9
1	germ:1	germ:8	germ:9	germ:10

Example of incomplete block design where different germplasms are replicated over some blocks.

Row-column (D3)

In a Row-column design, a control is replicated in rows and columns to catch as much as possible of the variation.

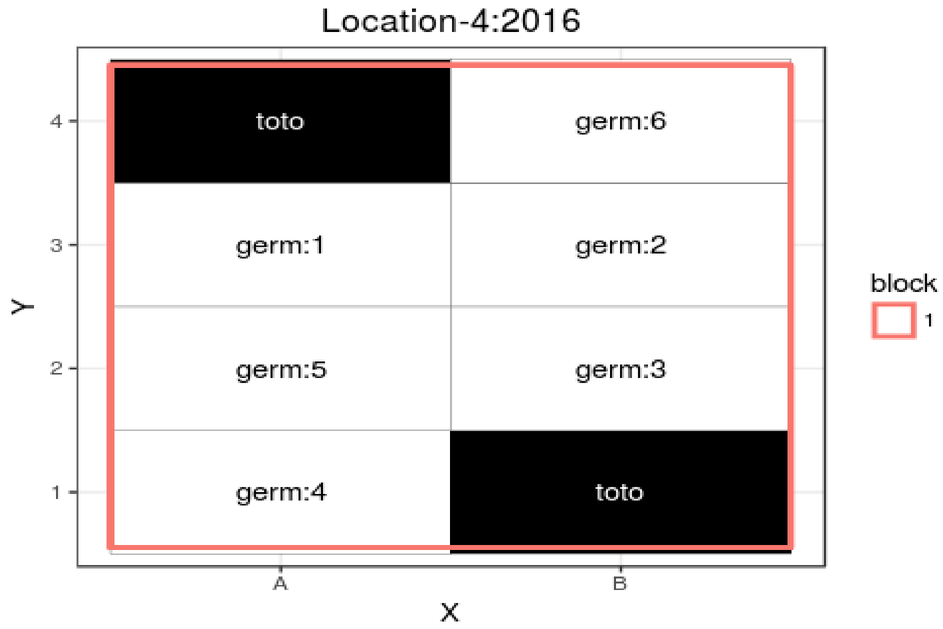


Y	A	B	C	D	E	F	G
4	germ:9	toto	germ:18	XXX-1	toto	germ:17	germ:3
3	toto	germ:14	germ:2	toto	germ:10	germ:5	germ:13
2	germ:8	germ:16	toto	germ:20	germ:19	toto	germ:6
1	germ:4	germ:12	germ:7	germ:15	germ:1	germ:11	toto

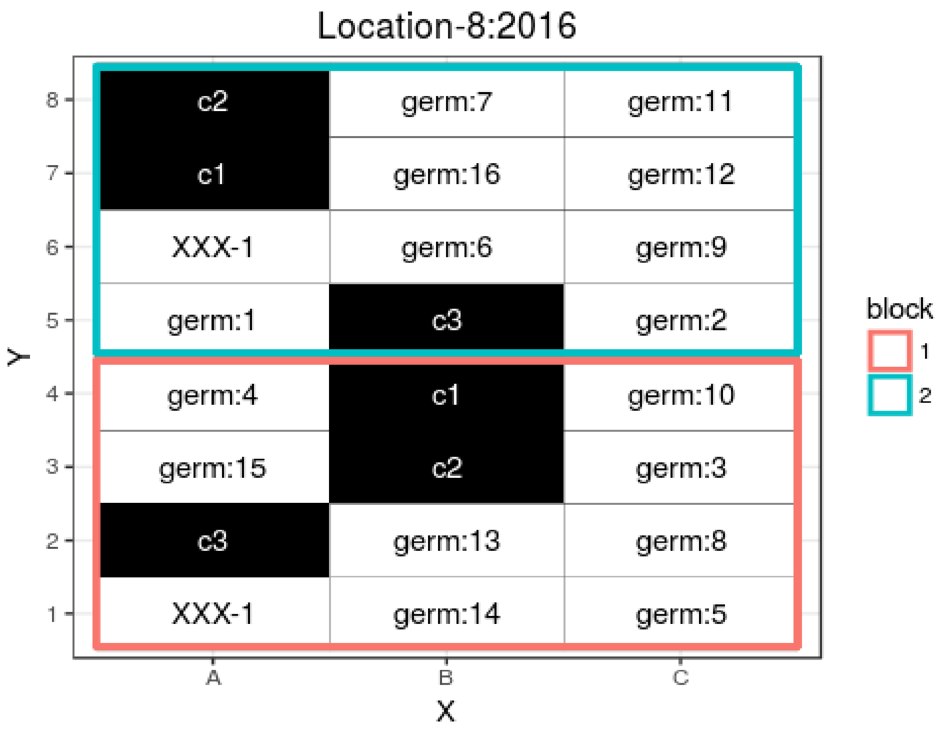
Row column design where a control (toto) is replicated in rows and columns.

Regional and satellite farms (D4)

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 except the controls. 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.



Example of a satellite farm design.



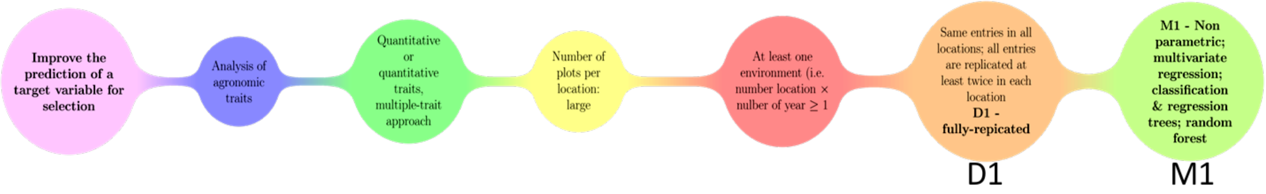
Y	A	B	C
8	c2	germ:7	germ:11
7	c1	germ:16	germ:12
6	XXX-1	germ:6	germ:9
5	germ:1	c3	germ:2
4	germ:4	c1	germ:10
3	germ:15	c2	germ:3
2	c3	germ:13	germ:8
1	XXX-1	germ:14	germ:5

Example of a regional farm design.

Describe the data

Once the data have been collected, a first step is to describe them with descriptive statistics and plots such as histograms and 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)



Improve the prediction of a target variable for selection

Analysis of agronomic traits

Quantitative or quantitative traits, multiple-trait approach

Number of plots per location: large

At least one environment (i.e. number location × nulber of year ≥ 1

Same entries in all locations; all entries are replicated at least twice in each location **D1 - fully-repicated**

M1 - Non parametric; multivariate regression; classification & regression trees; random forest

D1

M1

The problem is, given a set of p predictive attributes $X_1, X_2, \dots, X_p$, to estimate the value of a target attribute y. Denoting the estimative of y by $\hat{y}$, $\hat{y} = \hat{f}(X_1, X_2, \dots, X_p)$. An example would be to estimate the yield produced using the maize ear traits. The predictive attributes would be the maize ear traits. The $\hat{f}$ function can be obtained by any predictive algorithm.

We have worked only with algorithms that are able to predict quantitative target attributes. Moreover, we focused in interpretable algorithms, i.e., algorithms that can explain somehow how the value of $\hat{y}$ was predicted given the values of $X_1, X_2, \ldots, X_p$. Four different algorithms were used:

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

The four methods are described in the followinf sections.

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 define 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 - Multivariate Adaptive Regression Splines (Friedman 1991) was chosen because it has no assumption and has good interpretability (T., R., and Friedman 2001). MARS is quite similar to stepwise regression but the relations between each dependent variable and the independent one do not need to be linear, because each one of those relations is defined by a set of connected linear segments, instead of a single one. Like linear regression, MARS result is expressed as an equation typically a bit more complex than linear regression but also interpretable. MARS must be used as many times as the number of non-normal independent variables. At each time, only one variable is used. In all the experiments, the dependent variable is the selection cycle.

Random Forest

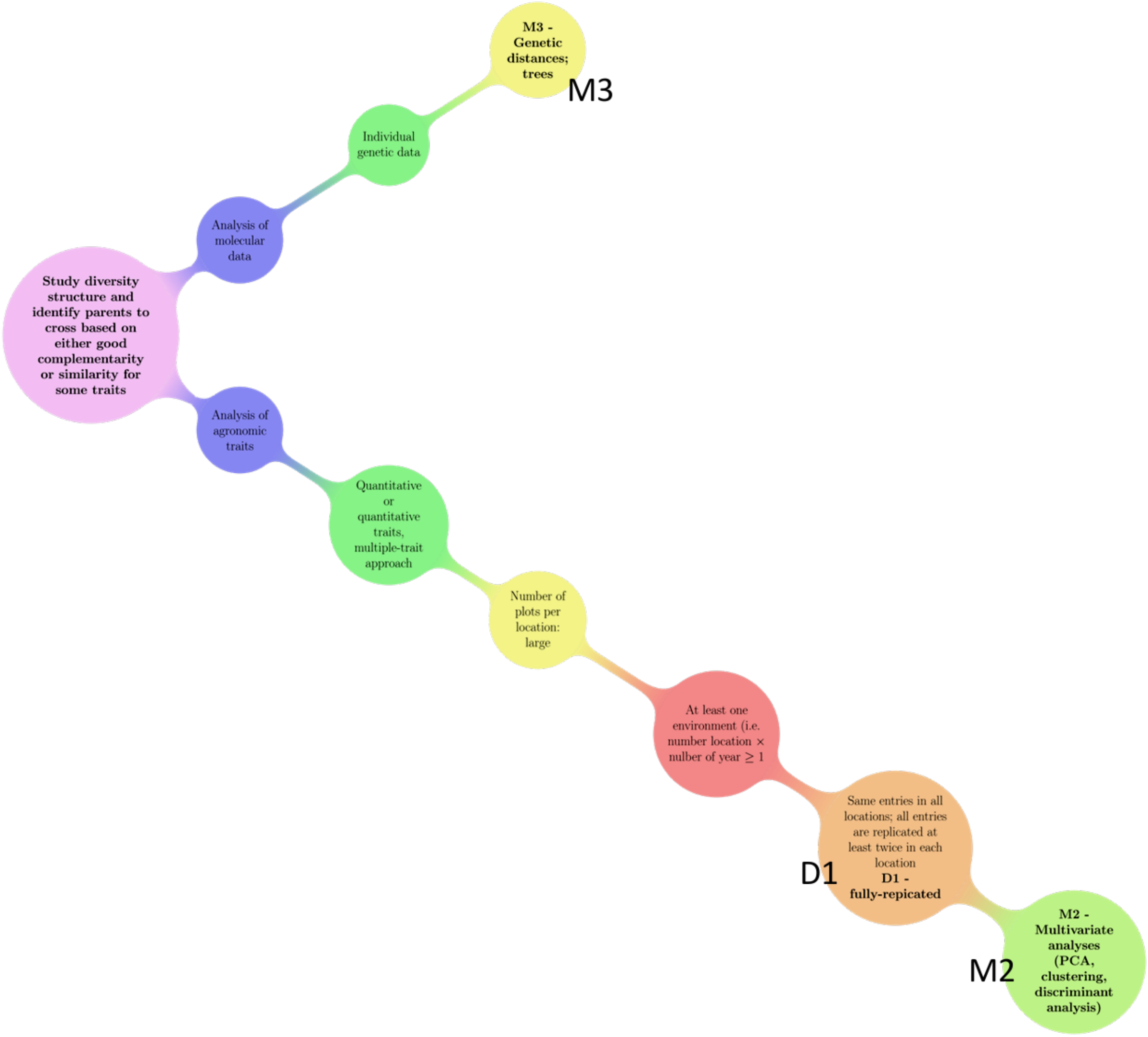
Random Forest (RF) (Breiman 2001) is a CART based approach, belonging to the family of ensemble methods, i.e., the use of a set of methods, instead of just one, in order to accomplish its task. RF generates several CART. Each generated CART is different because the tree is trained in a subset of the original set obtained using bagging (Breiman 1996) and using a random subset of the original set of features at each node. The interpretation of RF can be assessed using two different metrics (adapted for regression from (Kuhn J. 2008)):

-Mean Decrease Accuracy (% IncMSE): It is constructed by permuting the values of each variable of the test set (the test set is the out-of-bag subset that results from the bagging process), recording the prediction and comparing it with the unpermutated test set prediction of the variable (normalized by the standard error). It is the average increase in squared residuals of the test set when the variable is permuted. A higher % IncMSE value represents a higher variable importance.

- Mean Decrease MSE (IncNodePurity): Measures the quality (NodePurity) of a split for every variable (node) 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 gives a fast variable importance that is often very consistent with the permutation importance measure. A higher IncNodePurity value represents a higher variable importance; i.e. nodes are much ‘purer’.

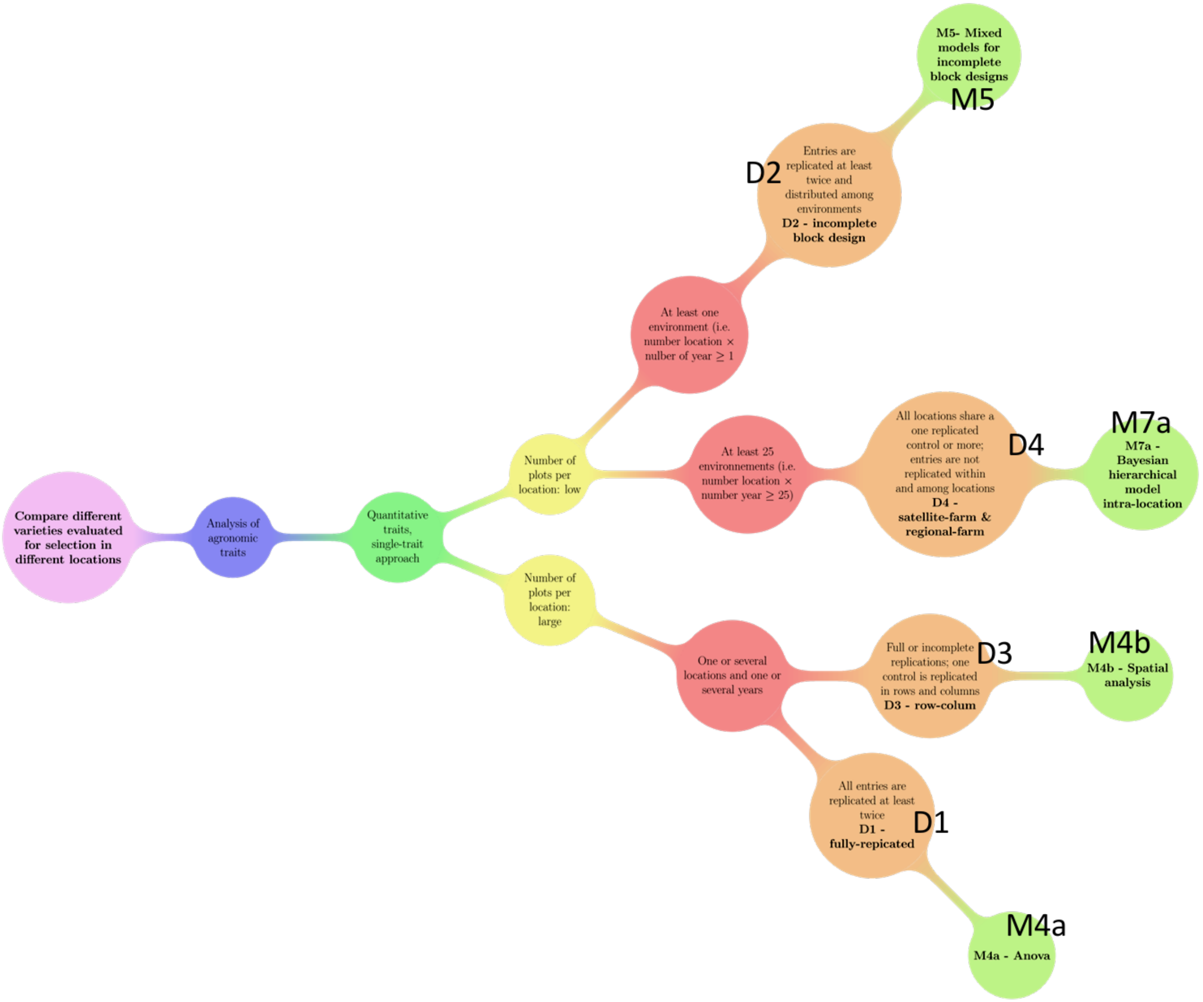
Multivariate analysis to study diversity structure and identify parents to cross based on either good complementarity or similarity for some traits (M2)

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.


M3 - Genetic distances; trees
M3
Individual genetic data
Analysis of molecular data
Study diversity structure and identify parents to cross based on either good complementarity or similarity for some traits
Analysis of agronomic traits
Quantitative or quantitative traits, multiple-trait approach
Number of plots per location: large
At least one environment (i.e. number location × nulber of year ≥ 1
Same entries in all locations; all entries are replicated at least twice in each location
D1 - fully-repicated
D1
M2 - Multivariate analyses (PCA, clustering, discriminant analysis)
M2

Analyses used to Ccompare 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 bayesian hierarchical model. Classic anova (**M4a**) is not explained here as it is a very classic analysis. Incomplete block design (**D2**) is presented but not the details of the mixed model (**M5**). Only spatial analysis (**M4b**) and the bayesian hierarchical model (**M7a**) are developed here.


Compare different varieties evaluated for selection in different locations
Analysis of agronomic traits
Quantitative traits, single-trait approach
Number of plots per location: low
At least one environment (i.e. number location × nulber of year ≥ 1
D2
Entries are replicated at least twice and distributed among environments
D2 - incomplete block design
M5
M5- Mixed models for incomplete block designs
At least 25 environnements (i.e. number location × number year ≥ 25)
D4
All locations share a one replicated control or more; entries are not replicated within and among locations
D4 - satellite-farm & regional-farm
M7a
M7a - Bayesian hierarchical model intra-location
Number of plots per location: large
One or several locations and one or several years
D3
Full or incomplete replications; one control is replicated in rows and columns
D3 - row-colum
M4b
M4b - Spatial analysis
D1
All entries are replicated at least twice
D1 - fully-repicated
M4a
M4a - Anova

Spatial analysis (M4b)

The experimental design used is the row-column design (**D3**). The following model is based on frequentist statistics. The model allows taking into account environmental variation within a block with few control 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):

$$Y_{ijk} = \alpha_i + r_j + c_k + f(u,v) + \varepsilon_{ijk} ; \varepsilon_{ijk} \sim N(0, \sigma^2)$$

With,

- Y_{ijk} the phenotypic value for germplasm i , row j and column k ,
- α_i the effect of germplasm i ,
- r_j the effect of row j ,

- c_k the effect of col k ,
- $f(u,v)$ the smooth bivariate function that simultaneously accounts for the spatial trend across both directions in the field (i.e. rows and columns),
- ε_{ijk} the residuals.

Note that $f(u,v)$ is divided into 8 components excluding the intercept (Rodríguez-Álvarez et al. 2016):

- the linear effect of the rows (row),
- the linear effect of the columns (col),
- the linear interaction of rows and columns (row:col),
- the main row effect (f(row)),
- the main column effect (f(col)),
- the smooth varying coefficient term regarding rows (f(col):row),
- the smooth varying coefficient term regarding columns (row:f(col))),
- the smooth-by-smooth interaction component (f(col):f(row))

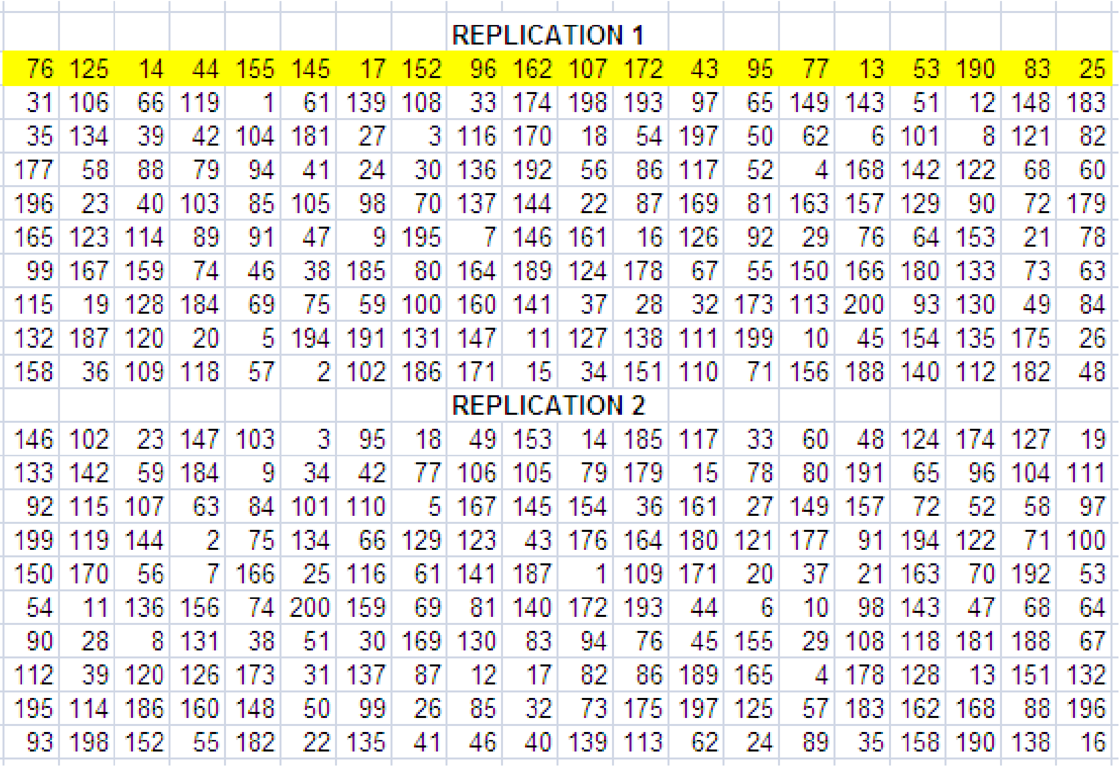
Much more information regarding the model as well as example of R package SpATS can be found in Rodríguez-Álvarez et al. (2016).

Incomplete block design and mixed model (M5)

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

Incomplete block designs have the objective of controlling the plot-to-plot variation and ideally they should allow the comparisons for all pairs of genotypes (Mead 1997); 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 √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 20 plots in the example shown in Fig. 26) is an independent unit and therefore can be allocated to a different field from each of the other incomplete blocks within the same location. An example of this is the incomplete block design with 20 bread wheat varieties in 3 replications (total 60 plots) and incomplete block size of 4 is shown below. The number of incomplete blocks which can be planted on each farm depends only on the farm size and therefore there can be farms with anywhere from 1 to 5 incomplete blocks. It is also possible that one full replication is planted by a large farmer and the 10 incomplete blocks of the other replication with 10 different farmers. The disadvantages of this layout are

1) 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 in which case the design is easier; however, there are certain number of entries for which *g* ≠ *sk* where the design is not easily available and 2) the loss of the row and column design which, as we will see later, 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).



								REPLICATION 1											
76	125	14	44	155	145	17	152	96	162	107	172	43	95	77	13	53	190	83	25
31	106	66	119	1	61	139	108	33	174	198	193	97	65	149	143	51	12	148	183
35	134	39	42	104	181	27	3	116	170	18	54	197	50	62	6	101	8	121	82
177	58	88	79	94	41	24	30	136	192	56	86	117	52	4	168	142	122	68	60
196	23	40	103	85	105	98	70	137	144	22	87	169	81	163	157	129	90	72	179
165	123	114	89	91	47	9	195	7	146	161	16	126	92	29	76	64	153	21	78
99	167	159	74	46	38	185	80	164	189	124	178	67	55	150	166	180	133	73	63
115	19	128	184	69	75	59	100	160	141	37	28	32	173	113	200	93	130	49	84
132	187	120	20	5	194	191	131	147	11	127	138	111	199	10	45	154	135	175	26
158	36	109	118	57	2	102	186	171	15	34	151	110	71	156	188	140	112	182	48
								REPLICATION 2											
146	102	23	147	103	3	95	18	49	153	14	185	117	33	60	48	124	174	127	19
133	142	59	184	9	34	42	77	106	105	79	179	15	78	80	191	65	96	104	111
92	115	107	63	84	101	110	5	167	145	154	36	161	27	149	157	72	52	58	97
199	119	144	2	75	134	66	129	123	43	176	164	180	121	177	91	194	122	71	100
150	170	56	7	166	25	116	61	141	187	1	109	171	20	37	21	163	70	192	53
54	11	136	156	74	200	159	69	81	140	172	193	44	6	10	98	143	47	68	64
90	28	8	131	38	51	30	169	130	83	94	76	45	155	29	108	118	181	188	67
112	39	120	126	173	31	137	87	12	17	82	86	189	165	4	178	128	13	151	132
195	114	186	160	148	50	99	26	85	32	73	175	197	125	57	183	162	168	88	196
93	198	152	55	182	22	135	41	46	40	139	113	62	24	89	35	158	190	138	16

An example of an incomplete block design randomized for 200 entries, 2 replications and incomplete blocks of size = 20 plots. One incomplete block is a sequence of 20 adjacent plots (like the sequence highlighted in yellow). The numbers indicate the entry number and their position the plot number: entry 76 is in plot 1 of rep 1, entry 125 is in plot 2 of rep 1, entry 25 is in plot 20 of rep 1, entry 31 is in plot 21 of rep 1, entry 146 is plot 1 of rep 2 or plot 201 if a unique plot number (from 1 to 400) is used.

ANOVAs of a hypothetical trial with 18 entries and various experimental designs, i.e. RCBD with 2 replications, alpha design (incomplete blocks) with 2 replications and with block size = 3, and with a row and column design (6 rows x 6 columns) and 2 replications.

Sources of variation	RCBD	Sources of variation	Alpha	Sources of variation	Row and columns
Entries	17	Entries	17	Entries	17
Replications	1	Replications	1	Rows	5
		Blocks w replication	10	Columns	8
Error	17	Error	7	Error	8
Total	35	Total	35	Total	35

The different partitioning of the degrees of freedom in the ANOVA of and RCBD, an alpha design and a Row and Column design with 18 entries and 2 replications (Table 7), shows that in a RCBD all the variation which is not explained by entries and replications is contributing to the error. In an alpha design only the variation within incomplete blocks contributes to the error, but the price to pay is the reduction in the degrees of freedom of the error. In a row and column design there is a better control of the source of variation and a small gain in the degrees of freedom of the error.

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 unstable estimation of the residual variance and to a lack of power for comparing populations. **M7a** was implemented to improve the efficiency of mean comparisons. It is efficient with more than 20 environment (i.e. location x year) (Rivière et al. 2015). The model is based on bayesian statistics.

The model is described in Rivière et al. (2015). We restricted ourselves to analysing values at the plot level (the values may result from the average of individual plants measures). The phenotypic value Y_{ijk} for variable Y, germplasm i, environment j and block k is modelled as:

$$Y_{ijk} = \mu_{ij} + \beta_{jk} + \varepsilon_{ijk} ; \varepsilon_{ijk} \sim N\left(0, \sigma_j^2\right),$$

where

- μ_{ij} is the mean of germplasm i in environment j (note that this parameter, which corresponds to an entry, confounds the population effect and the population x environment effect);

- β_{jk} is the effect of block k in environment j satisfying the constraint[1] $\sum_{k=1}^{K} \beta_{jk} = 1$;
- ε_{ijk} is the residual error;
- $N\left(0,\sigma_j^2\right)$ denotes normal distribution centred on 0 with variance σ_j^2 , which is specific to environment j .

We take advantage of the similar structure of the trials in all environments of the network to assume that trial residual variances come from a common distribution:

$$\sigma_j^2 \sim \frac{1}{Gamma\left(\nu,\rho\right)} ,$$

where ν and ρ are unknown parameters. Because of the low number of residual degrees of freedom for each farm, we use a hierarchical approach in order to assess mean differences on farm. For that, we place vague prior distributions on the hyperparameters ν and ρ :

$$\nu \sim Uniform\left(\nu_{min},\nu_{max}\right);\rho \sim Gamma\left(10^{-6},10^{-6}\right) .$$

In other words, the residual variance of a trial in a given environment is estimated using all the informations available on the network rather than using the data from that particular trial only.

The parameters μ_{ij} and β_{jk} are assumed to follow vague prior distributions too:

$$\mu_{ij} \sim N\left(\mu_{.j},10^6\right);\beta_{jk} \sim N\left(0,10^6\right) .$$

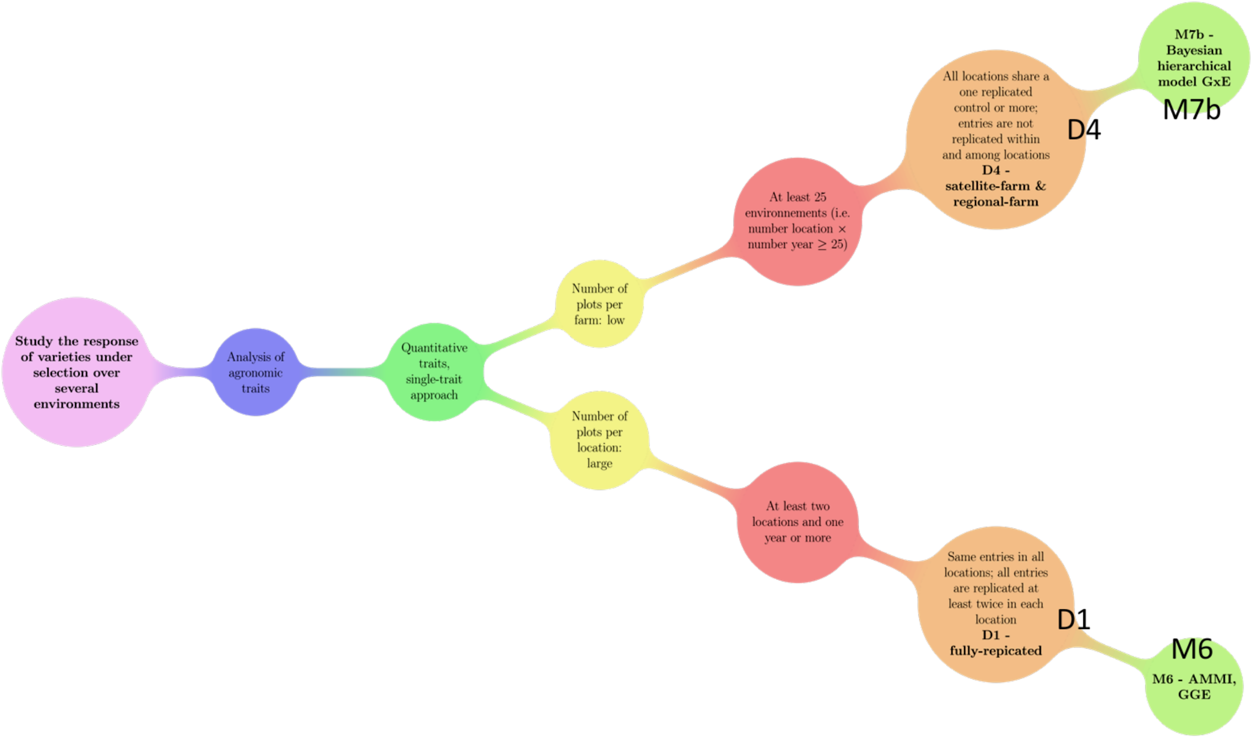
The inverse gamma distribution has a minimum value of 0 (consistent with the definition of a variance) and may have various shapes including asymmetric distributions. From an agronomical point of view, the assumption that trial variances are heterogeneous is consistent with organic farming: there are as many environments as farms and farmers leading to a high heterogeneity. Environment is here considered in a broad sense: practices (sowing date, sowing density, tilling, etc.), pedo climatic conditions, biotic and abiotic stress, ... (Desclaux et al. 2008). Moreover, the inverse gamma distribution has conjugate properties that facilitate MCMC convergence. This model is therefore a good choice based on both agronomic and statistical criteria.

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.

[1] Note that it is quite different from Rivière et al. (2015) where the model was written only for two blocks. Here there is no restriction on the number of blocks.

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 the bayesian hierarchical model (**M7b**).


Study the response of varieties under selection over several environments
Analysis of agronomic traits
Quantitative traits, single-trait approach
Number of plots per farm: low
At least 25 environnements (i.e. number location × number year ≥ 25)
All locations share a one replicated control or more; entries are not replicated within and among locations
D4 - satellite-farm & regional-farm
D4
M7b - Bayesian hierarchical model GxE
M7b
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-repicated
D1
M6
M6 - AMMI, GGE

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 (Gauch 2006):

1. An **ANOVA** with the following model:

$$Y_{ijk} = \mu + \alpha_i + \theta_j + rep_k(\theta_j) + (\alpha\theta)_{ij} + \varepsilon_{ijk} ; \varepsilon_{ijk} \sim N(0, \sigma^2)$$

With,

- Y_{ijk} the phenotypic value for replication *k*, germplasm *i* and location *j*,
- μ the general mean,
- α_i the effect of germplasm *i*,
- θ_j the effect of location *j*,

- $rep_k(\theta_j)$ the effect of the replication *k* nested in location,
- $(\alpha\theta)_{ij}$ the interaction effect of germplasm × location,
- ε_{ijk} the residuals.

Or, if there are several years in the data set:

$$Y_{ijkl} = \mu + \alpha_i + \theta_j + \beta_l + (\alpha\theta)_{ij} + (\alpha\beta)_{il} + (\theta\beta)_{jl} + rep_k(\theta\beta_{jl}) + \varepsilon_{ijk}; \varepsilon_{ijk} \sim N(0, \sigma^2)$$

With,

- Y_{ijkl} the phenotypic value for replication *k,* germplasm *i*, location *j* and year *l*,
- β_l the year l effect,
- $(\alpha\beta)_{il}$ the germplasm × year interaction effect,
- $(\theta\beta)_{jl}$ the location × year interaction effect,
- ε_{ijk} the residuals,
- and all other effects are the same as in the previous model.

2. A **PCA** that focuses on the germplasm × location interaction term:

$$(\alpha\theta)_{ij} = \sum_{n}^{N} \lambda_n \gamma_{in} \omega_{jn}$$ [2]

which can also be written:

$$(\alpha\theta)_{ij} = \sum_{n}^{N} \left(\sqrt{\lambda_n}\gamma\right)\left(\sqrt{\lambda_n}\omega_{jn}\right)$$

With,

- $(\alpha\theta)_{ij}$ the interaction between germplasm *i* and location *j*,
- N the number of dimensions (PCA components) which has as maximum value the number of location,
- λ_n the eigen value for component *n*,

[2] Note than in PPBstats, the PCA is done on $\eta_i \nu_j + \varepsilon_{ijk}$. The residuals is in the last dimension of the PCA

- γ_{in} the eigen vector for germplasm *i* and component *n*,
- ω_{jn} the eigen vector for location *j* and component *n*.

The data are double centered on location and germplasm. The PCA studies the structure of the interaction matrix. The locations are the variables and the germplasms are the individuals.

This PCA allows to detect :

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

GGE (M6)

The experimental design used is the fully replicated design (**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.

The GGE model can be written as followed:

$$Y_{ijk}=\mu+\theta_j+rep_k(\theta_j)+\sum_n^N \lambda_n \gamma_{in}\omega_{jn}+\varepsilon_{ijk};\varepsilon_{ijk}\sim N(0,\sigma^2)$$

with,

- Y_{ijk} the phenotypic value for replication *k,* germplasm *i* and location *j*,
- μ the general mean,
- θ_j the effect of location *j*,
- $rep_k(\theta_j)$ the effect of replication *k* nested in location,
- N the number of dimension (PCA components) which has as maximum value the number of location,
- λ_n the eigen value for component *n*,
- γ_{in} the eigen vector for germplasm *i* for component *n*,
- ω_{jn} the eigen vector for location *j* for component *n*.
- ε_{ijk} the residuals.

The location effect (*θ*) can be replaced by a more general environment effect that would include both location and year effects.

In case the location effect is the focus, the year effect can also be taken into account as follows:

$$Y_{ijkl} = \mu + \theta_j + \beta_l + (\theta\beta)_{jl} + rep_k(\theta\beta_{jl}) + \sum_n^N \lambda_n \gamma_{in} \omega_{jn} + \varepsilon_{ijk} ; \varepsilon_{ijk} \sim N(0, \sigma^2)$$

With,

- Y_{ijkl} the phenotypic value for replication *k,* germplasm *i*, location *j* and year *l*,
- β_l the effect of year l ,
- $(\theta\beta)_{jl}$ the location × year interaction effect,
- ε_{ijk} the residuals,
- and all other effects are the same as in the previous model.

Bayesian hierarchical model (M7b)

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

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. In these conditions, it is recommended to implement the **M7b** method.

Implementing the **M7b** method requires to base on at least around 75 environments and 120 germplasms present in at least two environments (95% of missing $G \times E$ combinations). It is based on bayesian statistics.

Theory of the model

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

The phenotypic value Y_{ij} for a given variable Y , germplasm *i* and environment *j*, is modeled as:

$$Y_{ij} = \alpha_i + \theta_j + \eta_i \theta_j + \varepsilon_{ij} ; \varepsilon_{ij} \sim N(0, \sigma_e^2)$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where

- I is the number of germplasms,
- J is the number of environments,
- α_i is the main effect of germplasm *i*,
- θ_j is the main effect of environment *j*,

- ε_{ij} is the residual and
- $N\left(0,\sigma_e^2\right)$ is the normal distribution with mean 0 and variance σ_e^2. The interaction between germplasm *i* and environment *j* is modelled as a multiplicative term $\eta_i\theta_j$, where η is a regression coefficient that depends on the germplasm *i*, and a remaining term that contributes to the residual ε_{ij}.

This model is further written as:

$$Y_{ij}=\alpha_i+\beta_i\theta_j+\varepsilon_{ij};\varepsilon_{ij}\sim N\left(0,\sigma_\varepsilon\right),$$

Where $\beta_i=\left(1+\eta_i\right)$ is the sensitivity of germplasm *i* to environments. This model is known as the Finlay Wilkinson model or as 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 $\beta_i>1$ ($\beta_i<1$) is more (less) sensitive to environments than a germplasm with the average sensitivity (Nabugoomu, Kempton, and Talbot 1999).

Given the string disequilibrium of the dataset and the large amount of data, this model is implemented with a hierarchical Bayesian approach.

We use hierarchical priors for α_i, β_i and θ_j and a vague prior for σ_ε.

$$\alpha_i\sim N\left(\mu,\sigma_\alpha^2\right),\beta_i\sim N\left(1,\sigma_\beta^2\right),\theta_j\sim N\left(0,\sigma_\theta^2\right),\sigma_\varepsilon^{-2}\sim Gamma\left(10^{-6},10^{-6}\right),$$

where μ, σ_α^2, σ_β^2 and σ_θ^2 are unknown parameters. The mean of β_i is set to 1 (Nabugoomu, Kempton, and Talbot 1999).

Then, we place weakly-informative priors on the hyperparmeters μ, σ_α^2, σ_β^2 and σ_θ^2:

$$\mu\sim N\left(v,v^2\right),\sigma_\alpha\sim Uniform\left(0,v\right),\sigma_\beta\sim Uniform\left(0,1\right),\sigma_\theta\sim Uniform\left(0,v\right),$$

where v is the arithmetic mean of the data: $v=\sum_{ij}Y_{ij}/n$ with *n* the number of observations. Uniform priors are used for σ_α^2, σ_β^2 and σ_θ^2 to reduce the influence of these priors on posterior results (**???**). The support of these priors take into account the prior knowledge that σ_α^2, σ_β^2 and σ_θ^2 are expected to be respectively smaller than v, 1 and v.

Initial values for each chain are taken randomly except for μ, σ_α and σ_θ whose initial values are equal to their posterior median from additive model (i.e. model with $\forall i,\beta_i=1$).

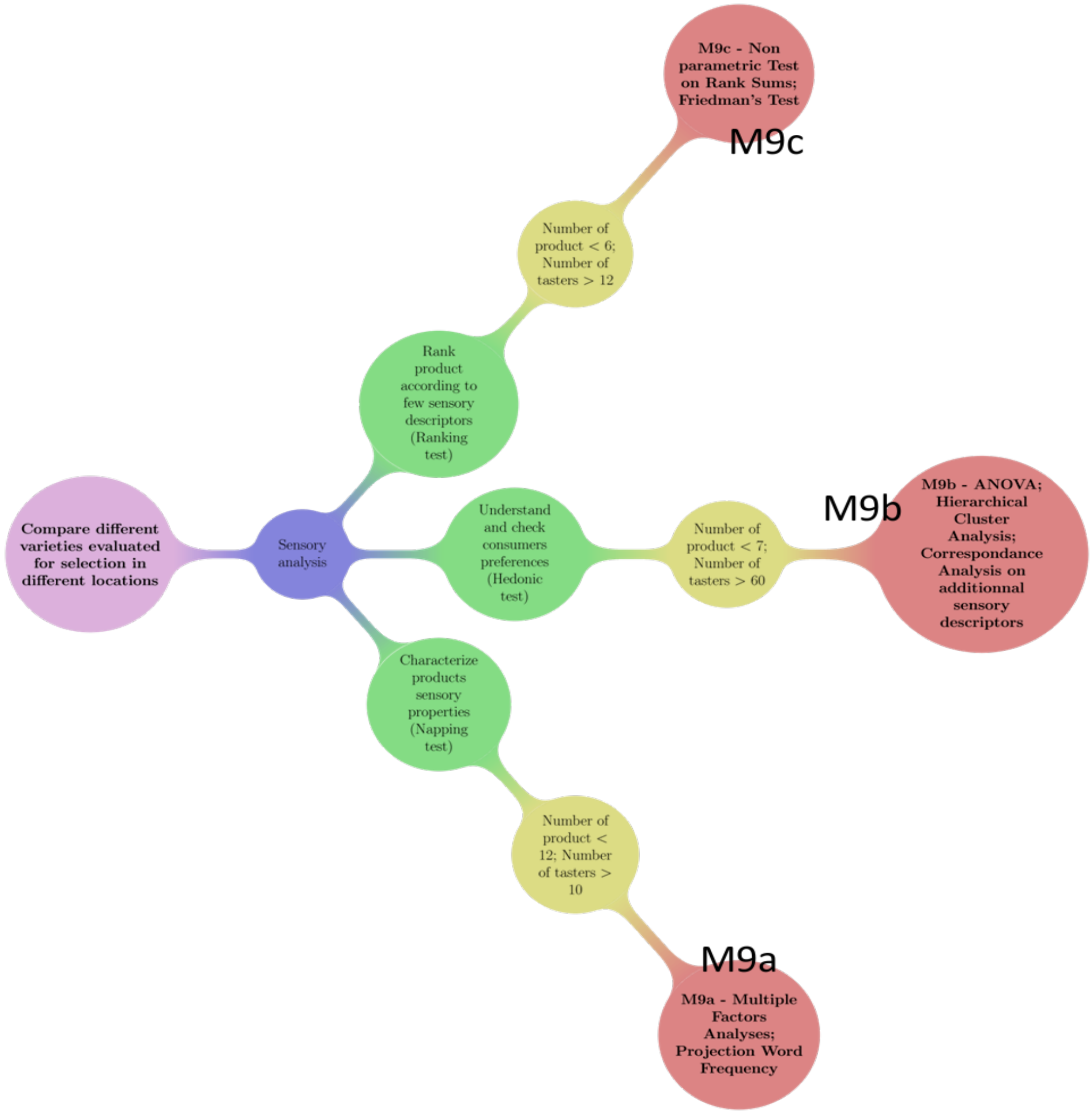
The main parameter of interest are the germplasm main effects ($\alpha_i,i=1,\ldots,I$), the environment main effects ($\theta_j,j=1,\ldots,J$) and the germplasm sensitivities ($\beta_i,i=1,\ldots,I$). For

α_i , the average posterior response of each germplasm over the environments of the network is calculated as: $\gamma_i = \alpha_i + \beta_i \bar{\theta}$,

where $\bar{\theta} = \frac{1}{J}\sum_{j=1}^{J} \theta_j$.

To simplify, the α_i notation is kept instead of γ_i (i.e. $\alpha_i = \gamma_i$). But keep in mind it has been corrected.

Organoleptic analysis


Compare different varieties evaluated for selection in different locations
Sensory analysis
Rank product according to few sensory descriptors (Ranking test)
Number of product < 6; Number of tasters > 12
M9c - Non parametric Test on Rank Sums; Friedman's Test
M9c
Understand and check consumers preferences (Hedonic test)
Number of product < 7; Number of tasters > 60
M9b
M9b - ANOVA; Hierarchical Cluster Analysis; Correspondance Analysis on additionnal sensory descriptors
Characterize products sensory properties (Napping test)
Number of product < 12; Number of tasters > 10
M9a
M9a - Multiple Factors Analyses; Projection Word Frequency

Napping Test (M9a)

The Napping allows looking 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 allows a 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 panellists 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 from 12 to 25 tasters according to the judge's experience with the product and to the objective of the experiment. For example ten farmers-bakers 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. The table presents the number of panellists (*i*) sets (one set for each panellist) of two columns corresponding to the horizontal and vertical coordinates (*X*,*Y*). The two columns are completed for each subject (i.e. person that taste) j : the X-coordinate (X_j) and the Y-coordinate (Y_j) for each product.

Sensory descriptors are coded through a "products × 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" is a qualitative variables with the number of words cited as modalities.

To analyse this kind of data, a Multiple Factor Analysis (MFA) should be performed. Each subject constitute a group of two un-standardised variables. The MFA leads to a synthesis of the panellist's tablecloth. Two products are close if all judges consider them close on the napping. The more the two first components of MFA explain the original variability, the more the judges are in agreement.

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 correlation with the factors of MFA are calculated and represented as in usual PCA.

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 characterise 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 does not 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.

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 comprises 12 semi-naive assessors (consumers initiated to sensory analyses) according to the ISO 8587 standard[3], 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. The null hypothesis (H0) is that all varieties have exactly the same responses (rank means are equal) and a Friedman's test (non parametric

[3] 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.

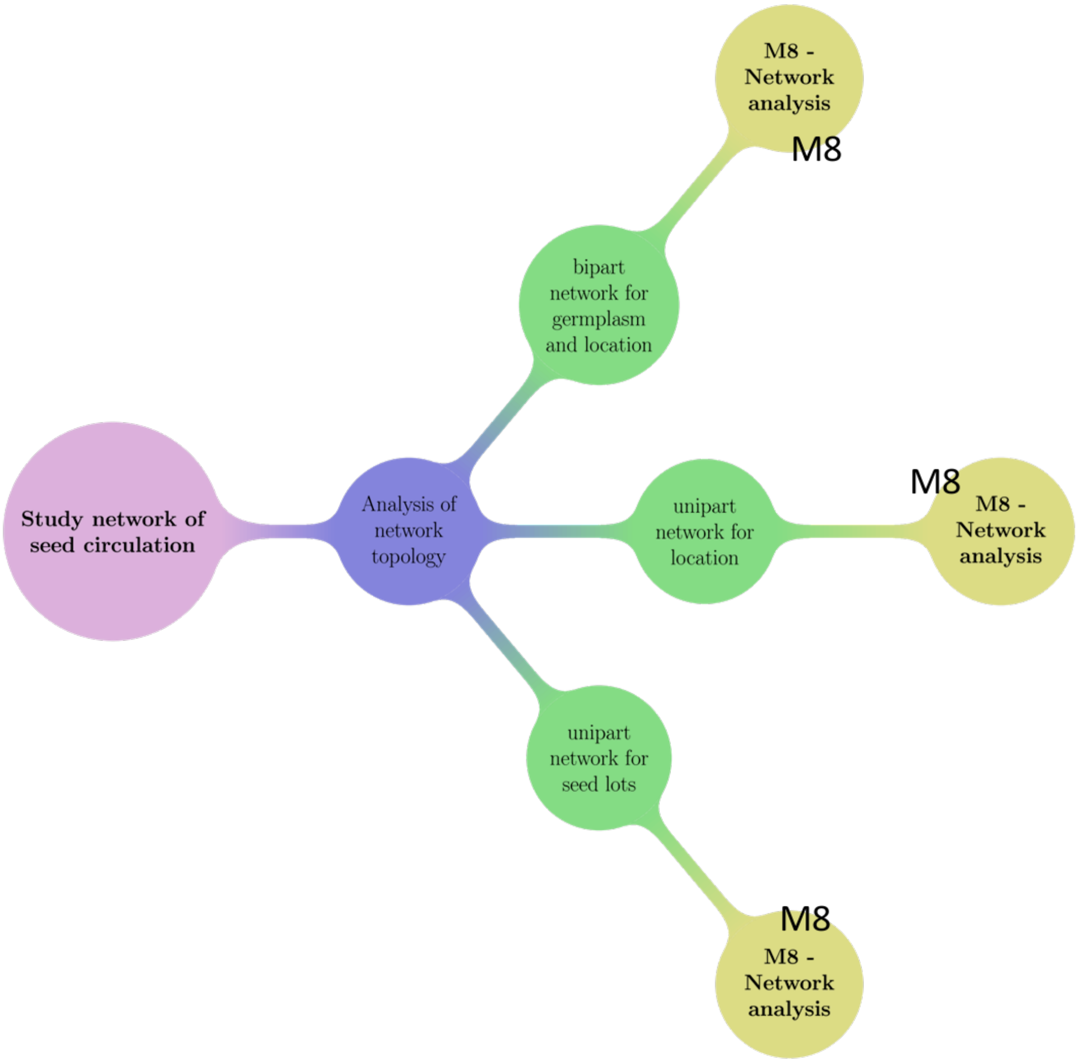
test on k independent samples) leads to the rejection or acceptance of this hypothesis, based on α value (<0.05).

Molecular analysis (M3)

Molecular analyses 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.

See Figure in Section for method **M2**.

Network analysis (M8)


Study network of seed circulation
Analysis of network topology
bipart network for germplasm and location
unipart network for location
unipart network for seed lots
M8 - Network analysis
M8
M8
M8 - Network analysis
M8
M8 - Network analysis

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, (a) where a node can be a seed lot (i.e. a combination of a germplasm in a given location a given year) and edges are relationships such as diffusion, mixture, reproduction, crosses or selection for example ; or (b) 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.

Data format

Three formats are possible:

1. unipart networks that represent the relationships between seed lots: a data frame with the following compulsory columns:

   - "seed_lot_parent": name of the parent seed lot in the relationship,
   - "seed_lot_child" ; name of the child seed lots in the relationship,
   - "relation_type": the type of relationship between the seed lots,
   - "relation_year_start": the year when the relationship starts,
   - "relation_year_end": the year when the relationship stops,
   - "germplasm_parent": the germplasm associated to the parent seed lot,
   - "location_parent": the location associated to the parent seed lot,
   - "year_parent": the year of the last relationship of the parent seed lot,
   - "germplasm_child": the germplasm associated to the child seed lot,
   - "location_child": the location associated to the child seed lot,
   - "year_child": represents the year of the last relation event of the child seed lot.

The possible options are: "alt_parent", "long_parent", "lat_parent", "alt_child", "long_child", "lat_child" to get map representation, supplementary variables with tags: "_parent", "_child" or "_relation".

2. unipart networks that represent relationship of germplasm circulation between locations: a data frame with the following compulsory columns (same as above): "location_parent", "location_child", "relation_year_start", "relation_year_end". Possible options are: "germplasm_parent", "year_parent", "germplasm_child", "year_child". Other possibles option are: "alt_parent", "long_parent", "lat_parent", "alt_child", "long_child", "lat_child" to get a map representation.

3. bipart networks represent "which location has which germplasm which year": a data frame with the followin compulsory columns: "germplasm", "location", "year". Possible options are: "alt", "long", "lat" to get a map representation.

Descriptive analysis

Based on these types of dataset, descriptive analyses can be carried out to better understand how exchanges are organized within a CSB or a breeding programme. Unipart networks of seed-lots can be displayed in the chronological order. Barplots can be used to show the distribution of germplasm 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).

References

Bernardo, R. 2002. *Breeding for quantitative traits in plants*. Stemma Press, Woodbury, Minnesota.

Breiman, L. 1996. "Bagging Predictors." *Machine Learning* 26: 123–40.

———. 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.

Finlay, K., and G. Wilkinson, 1963 The analysis of adaptation in a plant breeding programme. Crop Pasture Sci. 14: 742–754.

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).

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.

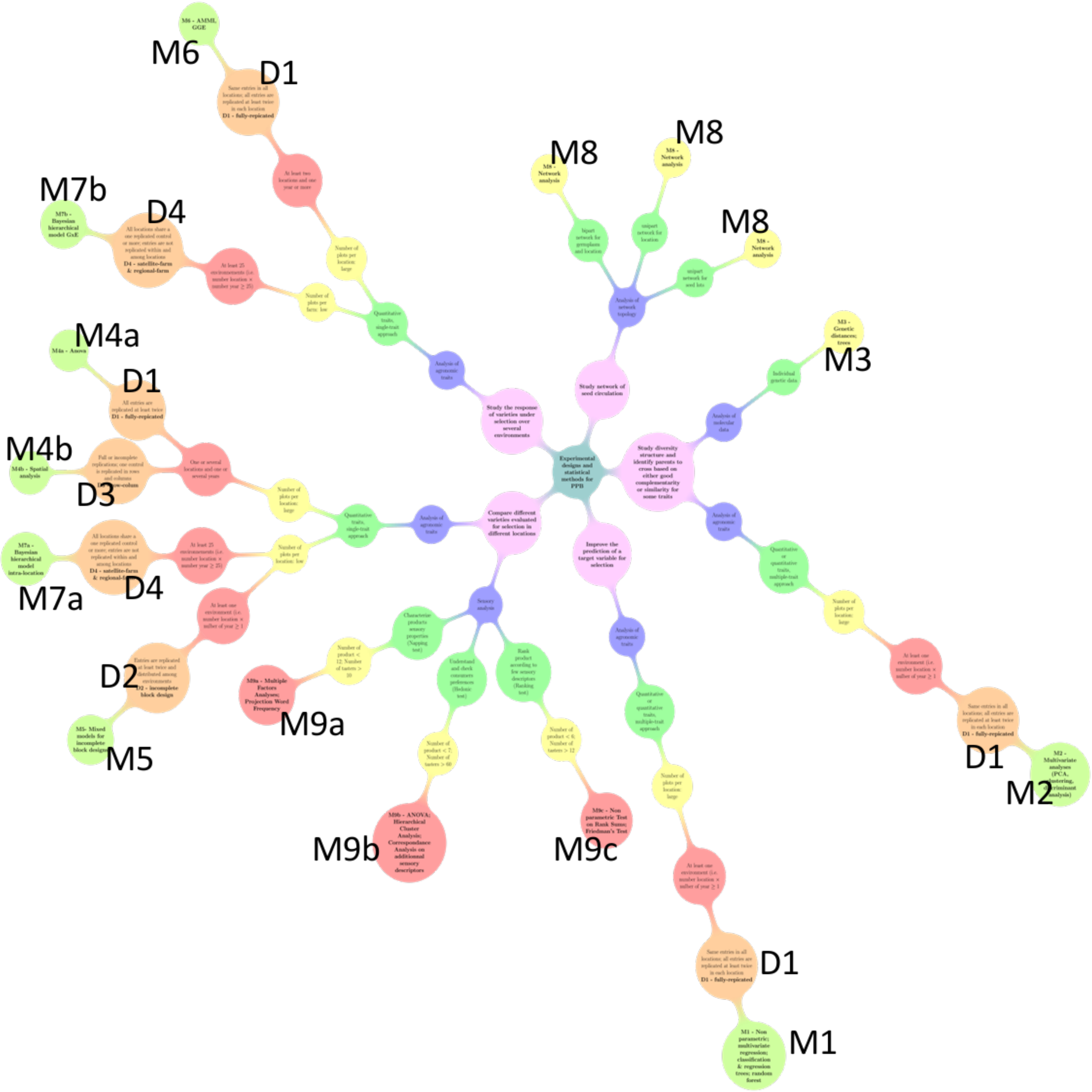
Singh, M., and K. El-Shama’a. 2015. *Experimental Designs for Precision in Phenotyping*.

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.

Appendix


M6 - AMMI, GGE
M6
D1
Same entries in all locations; all entries are replicated at least twice in each location
D1 - fully-repicated
At least two locations and one year or more
Number of plots per location: large
M7b
M7b - Bayesian hierarchical model GxE
D4
All locations share a one replicated control or more; entries are not replicated within and among locations
D4 - satellite-farm & regional-farm
At least 25 environnements (i.e. number location × number year ≥ 25)
Number of plots per farm: low
Quantitative traits, single-trait approach
Analysis of agronomic traits
M8
M8 - Network analysis
M8
M8 - Network analysis
M8
M8 - Network analysis
bipart network for germplasm and location
unipart network for location
unipart network for seed lots
Analysis of network topology
Study network of seed circulation
M3 - Genetic distances; trees
M3
Individual genetic data
Analysis of molecular data
M4a
M4a - Anova
D1
All entries are replicated at least twice
D1 - fully-repicated
Study the response of varieties under selection over several environments
Experimental designs and statistical methods for PPB
Study diversity structure and identify parents to cross based on either good complementarity or similarity for some traits
M4b
M4b - Spatial analysis
Full or incomplete replications; one control is replicated in rows and columns
D3 - row-colum
D3
One or several locations and one or several years
Number of plots per location: large
Quantitative traits, single-trait approach
Analysis of agronomic traits
Compare different varieties evaluated for selection in different locations
Improve the prediction of a target variable for selection
Analysis of agronomic traits
M7a - Bayesian hierarchical model intra-location
All locations share a one replicated control or more; entries are not replicated within and among locations
D4 - satellite-farm & regional-farm
M7a
D4
At least 25 environnements (i.e. number location × number year ≥ 25)
Number of plots per location: low
Quantitative or quantitative traits, multiple-trait approach
Sensory analysis
Characterize products sensory properties (Napping test)
Understand and check consumers preferences (Hedonic test)
Rank product according to few sensory descriptors (Ranking test)
At least one environment (i.e. number location × nulber of year ≥ 1)
Number of plots per location: large
D2
Entries are replicated at least twice and distributed among environments
D2 - incomplete block design
M9a - Multiple Factors Analyses; Projection Word Frequency
Number of product < 12; Number of tasters > 10
M9a
M5- Mixed models for incomplete block design
M5
Number of product < 7; Number of tasters > 60
Number of product < 6; Number of tasters > 12
Analysis of agronomic traits
Quantitative or quantitative traits, multiple-trait approach
Number of plots per location: large
At least one environment (i.e. number location × nulber of year ≥ 1)
Same entries in all locations; all entries are replicated at least twice in each location
D1 - fully-repicated
D1
M2 - Multivariate analyses (PCA, clustering, discriminant analysis)
M2
M9b
M9b - ANOVA; Hierarchical Cluster Analysis; Correspondance Analysis on additionnal sensory descriptors
M9c - Non parametric Test on Rank Sums; Friedman's Test
M9c
At least one environment (i.e. number location × nulber of year ≥ 1)
D1
Same entries in all locations; all entries are replicated at least twice in each location
D1 - fully-repicated
M1 - Non parametric; multivariate regression; classification & regression trees; random forest
M1