Comparaison de différentes stratégies spatiales d'échantillonnage d'une core collection parmi des populations naturelles d'espèces fourragères

François Balfourier, Gilles Charmet, Jean-Marie Prosperi, Goulard Michel, Pascal P. Monestiez

Genet. Sel. Evol. 30 (Suppl. 1) (1998) S215–S235

Original article

Comparison of different spatial strategies for sampling a core collection of natural populations of fodder crops

François Balfourier[a*], Gilles Charmet[a], Jean-Marie Prosperi[b], Michel Goulard[c], Pascal Monestiez[c]

[a] Institut national de la recherche agronomique, Station d'amélioration des plantes, 234, avenue du Brézet, 63039 Clermont-Ferrand cedex 2, France

[b] Station de génétique et amélioration de plantes, Institut national de la recherche agronomique, Domaine de Melgueil 34130 Mauguio, France

[c] Station de biométrie, Institut national de la recherche agronomique, BP 91, 84143 Montfavet cedex, France

Abstract – Seven different strategies for sampling a core collection are compared in large collections of natural populations of two fodder crop species (*Lolium perenne* and *Medicago truncatula*) previously evaluated for agronomic traits. The first five strategies consist of random or stratified sampling methods with one level of classification of the population (classification based on agronomic traits with or without a geographic contiguity constraint, based on the country or region of origin, or on spatial structures of populations). The last one of these strategies is based on geostatistical analysis, which allows studies of the spatial pattern of genetic diversity. The two remaining strategies are based on the principal component score strategy (PCSS), which consists of maximising a distance criterion between populations, this criterion being applied to agronomic traits of the populations or to their spatial components after geostatistical analysis. The different strategies are compared using two criteria: i) their capacity to capture the phenotypic diversity of the whole collection, as measured by the Shannon index; ii) their ability to restore the spatial structuration of the initial collection, measured by distance between maps for agronomic traits. The comparisons for the two species show that strategies that take into account spatial structuration of diversity give the best results for both criteria. © Inra/Elsevier, Paris

core collection / geostatistics / *Lolium perenne* / *Medicago truncatula* / spatial structuration

* Correspondence and reprints

Résumé – Comparaison de différentes stratégies spatiales d'échantillonnage d'une core collection parmi des populations naturelles d'espèces fourragères. Sept stratégies d'échantillonnage d'une core collection sont comparées sur deux collections de populations naturelles d'espèces fourragères (*Lolium perenne* et *Medicago truncatula*) préalablement évaluées pour des caractéristiques agronomiques. Les cinq premières stratégies consistent en des tirages aléatoires simples ou stratifiés selon un seul niveau de classification des populations: (i) aléatoire simple (stratégie Random); (ii) stratifié selon une classification basée sur les caractéristiques agronomiques (stratégie Stratagro); (iii) stratifié selon la région ou le pays d'origine (stratégie Stratreg); (iv) stratifié selon les caractéristiques agronomiques avec contrainte de contiguïté géographique (stratégie Straconti); (v) stratifié selon les structures spatiales des populations (stratégie Stratspa), déterminées par les méthodes de la géostatistique. Il s'agit de méthodes statistiques adaptées à l'analyse de données spatiales et à la mise en évidence de structure ou distributions spatiales non aléatoires. Enfin, deux autres stratégies déterministes sont testées; elles sont basées sur un critère de maximisation de distance entre populations, ce critère étant appliqué aux données agronomiques de base des populations (stratégie PCSS) et à leurs composantes spatiales (stratégie SPPCSS). Les différentes stratégies sont comparées selon deux critères: leur capacité à capturer la diversité phénotypique contenue dans la collection de départ, mesurée par un indice de Shannon, et leur aptitude à restituer la structuration spatiale de la collection globale, mesurée par une distance entre cartes établies pour des caractères agronomiques. Les comparaisons montrent que, chez les deux espèces, les stratégies tenant compte des structures spatiales (Stratspa et SPPCSS) donnent les meilleurs résultats, aussi bien pour l'indice de diversité globale que pour la représentativité des cartes.

core collection / géostatistique / *Lolium perenne* / *Medicago truncatula* / structuration spatiale

1. INTRODUCTION

Genetic resources from natural populations are valuable materials for plant breeders, especially in fodder crops for which adaptation to environmental and farming conditions are fundamental factors of diversity. However, the increasing size of genetic resources collections often limits their distribution, optimal utilisation and sometimes conservation. In addition to practical difficulties that may be encountered by curators in the multiplication and regeneration of accessions, users wish to obtain information about the genetic material available. For this, evaluation of the genetic resources for agronomic traits is required, but is almost impossible in very large collections. For this reason, several authors (Frankel, 1984; Brown, 1989b) proposed the concept of 'core collections', which consists of identifying a subsample of manageable size from a large collection, as representative of the total genetic diversity as possible. In this case, evaluation and pre-breeding work may be managed with the core collection, while the rest of the introductions may be maintained, at a lower cost, in a reserve collection. Reviews on the topic of core collections have been published by Hamon et al. (1994) and Brown (1995).

Recent studies in France on fodder species of *Lolium* (Balfourier and Charmet, 1994; Monestiez et al., 1997) and *Medicago* genera (Chaulet and Prosperi, 1994; Bonnin et al., 1996a, b) showed that, whatever the species, traits of agronomic interest submitted to natural selection, as well as some presumably neutral markers, often exhibit strong geographical (i.e. spatial) structuration.

The aim of the present study was to demonstrate the value, for spontaneous species such as fodder crops, of considering the spatial pattern of genetic diversity of collected populations as a sampling criterion for a core collection, using geostatistical techniques. This involves a series of statistical methods applied to spatial data analysis and used to display spatial structures and nonrandom distributions.

2. MATERIALS AND METHODS

2.1. Plant material

The study was carried out on two model species of fodder crops: an outbreeding grass, *Lolium perenne* L., perennial ryegrass, and an inbreeding legume, *Medicago truncatula* Gaertn, which is a Mediterranean annual *Medicago*. These two species were chosen because it is well known that reproductive behaviour has considerable effects on gene flow, and hence on spatial structuration of diversity. Perennial ryegrass is widespread in Europe, while annual *Medicago* or medic is a species mainly localised in warm Mediterranean areas or at low altitudes. These species, which exhibit a large diversity in their natural populations, are considered as models for population genetics studies on fodder crops. Furthermore, medic is also one of the model species used for studies of plant–microorganisms relationships (Barker et al., 1990).

A collection of 550 natural populations, previously collected from all over France (21 administrative regions), was used for perennial ryegrass (Charmet and Balfourier, 1991), while, for medic, a sample of 110 natural populations from six countries or regions over the Mediterranean area (Algeria: 33, Spain: 38, Portugal: 5, France: 19, Greece: 11 and Crete: 4) was used. Because of the varying abundance of this species from one country to another, the distribution of populations may not have been homogeneous for the whole area analysed.

2.2. Agronomic traits

Agronomic evaluation was carried out on the 550 ryegrass populations, using spaced plant nurseries in a multilocal evaluation network. Several traits were observed and scored according to a visual 1 to 9 scale: seasonal vigour traits (spring, summer, autumn), frost and rust susceptibilities, persistency, alternativity (i.e. ability to produce spikes in sowing year), heading date, aftermath heading and growth habit. To describe diversity for these traits, taking into account genotype by environment interactions, 28 agronomic variables (Charmet et al., 1990) were considered (*table I*).

For the medic, a trial plot of two 1-m rows per population, in triplicate, was planted in 1995 at Montpellier, according to a randomised block design. Observations were made of plant vigour, illustrated by plot density and size, phenological stages, pathogenic susceptibility, pod and seed production and natural regeneration scored 3 weeks after the first autumn rainfalls. Thus, a total of 29 agronomic traits was observed (*table II*).

Table I. List of 28 agronomic variables analysed in 550 ryegrass populations.

Abbreviation	Variables observed	Scale
FRO1 to FRO4	Frost susceptibility, measured in four locations (1 = Rodez, 2 = Mont de Marsan, 3 = Bourg-Lastic, 4 = Le-Pin-au-Haras)	1 (low) to 9 (high)
SPV1 to SPV4	Spring vigour growth, measured in four locations	1 (weak) to 9 (very vigorous)
SUM1 to SUM4	Summer regrowth, measured in four locations	1 (weak) to 9 (very vigorous)
AUT1 to AUT4	Autumn regrowth, measured in four locations	1 (weak) to 9 (very vigorous)
PER1 to PER4	Persistency, measured in four locations	1 (low) to 9 (high)
RUS1 to RUS4	Rust susceptibility, measured in four locations	1 (low) to 9 (high)
ALT	Mean alternativity at the four locations	1 (low) to 9 (high)
GH	Mean growth habit at the four locations	1 (erect) to 9 (prostrate)
AH	Mean aftermath heading at the four locations	1 (low) to 9 (high)
HD	Mean heading date at the four locations	Number of days after 1st January

Table II. List of 29 agronomic variables analysed in 110 medic populations.

Abbreviation	Variables observed	Scale
SD	Seedling density 1 month after sowing	% of ground coverage
H1 to H7	Height of the plot*	cm
W1 to W7	Width of the plot*	cm
DFL	Days to first flowering**	Number of days from
DMP	Days to first mature pod**	sowing
ROO	Susceptibility to root-rot (association of several pathogens)	1 (susceptible) to 9 (tolerant)
SPP	Susceptibility to *Phoma medicagenis* and *Pseudopeziza medicagenis*	1 (susceptible) to 9 (tolerant)
HEA	General healthy state of plot	1(bad) to 9 (good)
WP	Weight of harvested pods on 1.5 m^2	g
WS	Weight of viable seeds on 1.5 m^2	g
W100p	Weight of 100 pods	g
W1000s	Weight of 1 000 seeds	g
Ws-Wp	Seed/pod yield ratio	%
Ns-Np	Average number of seeds per pod	–
Np-m^2	Number of pods per m^2	–
Ns-m^2	Number of viable seeds per m^2	–
Rege	Regeneration on 24 September 1995	% of ground coverage

* Date of measures: 4 December 1995 and one measure every 15 days from 15 February to 15 May; ** number of days from sowing date to first open corolla (or first mature pod separating plant) on three separate plants.

2.3. Statistical development

2.3.1. Principal component analysis (PCA) of agronomic data and clustering of populations

A basic PCA was carried out on the two sets of centred and scaled agronomic data, i.e. 28 variables for 550 ryegrass populations, and 29 variables for 110 medic populations. Following this, a multivariate hierarchical clustering analysis was carried out on the first components of each PCA whose eigenvalue was greater than 1 (i.e. nine axes for ryegrass, noted *comp1r* to *comp9r*, and five axes for medics, *comp1m* to *comp5m*). After cutting trees, this makes it possible to define agronomic clusters among populations for the two species. Using the same ratio of between-cluster variance over total variance = 0.5, 11 agronomic clusters were identified for ryegrass (Charmet et al., 1990) and four clusters for medic populations.

Finally, a multivariate hierarchical clustering analysis, using a geographical contiguity constraint (Charmet et al., 1994), was applied to the ryegrass data set only, which led us to consider 25 'contiguous' clusters for the same ratio of 0.5. This method permits grouping of populations according to both their similarity for agronomic traits and to their geographical proximity.

2.3.2. Geostatistical analyses

Spatial structuration of diversity for agronomic traits was analysed using geostatistical methods (Cressie, 1991; Wackernagel, 1995). Application to ryegrass populations has already been reported (Monestiez et al., 1994). Most computations, figures and maps were made using the S-Plus programming package (Becker et al., 1988).

2.3.2.1. Univariate analysis

In geostatistics, stochastic modelling is used to describe the spatial data. Suppose that the studied variable is a realisation of a random field and that the observed data are a spatial sample of some sites of this realisation. Let $Z(x)$ denote the random field at location x and $z(x)$ denote the observed data at the same location x, which is a realisation of $Z(x)$. From different observations $(z(x_1), \ldots, z(x_n))$, we can define for each couple (i, j), the dissimilarity between two points x_i and x_j by:

$$g^*(x_i, y_j) = 1/2[z(x_i) - z(x_j)]^2 \quad (1)$$

Assuming stationarity of order two (mean and variance of the variable are supposed to be invariant by translation), we define the variogram:

$$g(h) = 1/2E[(Z(x+h) - Z(x))^2] \quad (2)$$

where $x + h$ is the translation of x, and E denotes the expectation. For the isotropic case, the variogram $g(h)$ can be estimated, using previous dissimilarities, by:

$$g^*(h) = 1/2n_c\Sigma_{i,j}[z(x_j) - z(x_i)]^2 \quad (3)$$

where n_c is the number of couples, and the distance $d(x_j, x_i) = h$.

The analysis of spatial structuration of a variable will be made with its variogram which measures the semivariance of the difference between points separated by a given distance h. Structural analysis consists of describing and modelling the estimated variogram $g^*(h)$. Classically, three different items are described by such functions:

- First, the discontinuity at the origin. The variogram equals zero at the origin by definition. For a distance close to zero, a significant value usually called the 'nugget effect' is often observed. This initial variance can be seen as measurement error or microscale variation.
- Second, the variogram usually increases with increasing distances. The shorter the distances between the sites, the more dependent the observations.
- Third, the 'sill' is the steady-state value reached by the function. When the distance between sites is larger than a given value, the variogram function becomes constant and the sites should be considered to be independent. This distance value, called the 'range', is an essential parameter in spatial description. The value of the sill is close to that of the global population variance.

For a two-dimensional space, spatial structuration of a variable can be visualised on maps. To obtain smoothed maps, we need to predict values at unsampled locations. In geostatistics, the prediction method is called 'kriging'. Kriging uses the observed values of the variable in conjunction with the information provided by the variogram (Wackernagel, 1995).

In the univariate case, four variables, considered as realisations of random variables, were independently analysed in the present study: the first two components of PCA of ryegrass data (*comp1r, comp2r*) and of medics (*comp1m, comp2m*).

2.3.2.2. Multivariate spatial method

Multivariate spatial analysis is a generalisation, in the multivariate case, of the study of variograms and respective cross-variograms for each variable. The cross-variogram between two variables Z_1 and Z_2 is defined as:

$$g_{12}(h) = 1/2E[(Z_1(x+h) - Z_1(x))(Z_2(x+h) - Z_2(x))] \qquad (4)$$

These cross-variograms provide information about the spatial relations of dependence between different variables. A linear co-regionalisation model is used as a tool for simultaneously modelling variograms and cross-variograms. The basic physical idea is that the variables under study are generated by different processes acting additively with various specific spatial structures. Finally, in 'kriging analysis', these different structures are themselves decomposed by PCA into spatial components, which are all mutually independent. As in PCA, the relationships between original variables and spatial components can be described and spatial components can be mapped. A more detailed description of the model and its fittings is given by Goulard and Voltz (1992), while Monestiez et al. (1994) provide a useful description of kriging analysis.

To summarise the spatial information given by numerous original agronomic variables, kriging analysis was not carried out on the original variables, but rather on the first nine components of PCA for ryegrass (*comp1r* to *comp9r*), and on the first five components of PCA for medics (*comp1m* to *comp5m*).

2.3.3. Sampling strategies

Seven core collection sampling strategies were tested, using various proportions, p, of the whole collection (p = 5, 10, 20 and 40 %), for the two species. For the first five strategies, 200 random samples were generated by computer for each of the four proportions. These strategies consisted of simple or stratified random sampling:

- random sampling (RANDOM): simple random sampling, without replacement, of a proportion p in the whole collection of 550 ryegrass populations and 110 medic populations;
- random sampling stratified by agronomic clustering (STRATAGRO): random sampling of a proportion p in each of the q agronomic clusters, previously defined after PCA of the original agronomic variables (q = 11 for ryegrass and = 4 for medics);
- random sampling stratified by regions of origin (STRATREG): random sampling of a proportion p in each French administrative region for ryegrass (21), or each country of origin for medics (6);
- random sampling stratified by contiguous and agronomic clustering (STRATCONTI): random sampling of a proportion p in each of the 25 contiguous clusters, as previously defined for ryegrass only (Charmet et al., 1994);
- random sampling stratified by spatial structures (STRATSPA): random sampling of a proportion p in each of the ten clusters derived from clustering on the first component of PCA of spatial structures, as previously obtained by kriging analysis.

Thus, all these stratified samples were generated according to the p strategy of Brown (1989b), i.e. were proportional to cluster size.

The last two strategies are based on the principal component score strategy (PCSS) (Hamon et al., 1995; Noirot et al., 1996). PCSS consists of three steps: the application of PCA to the quantitative data of the collection; the computerisation of the weighted Euclidean distance between individuals, using new coordinates on PCA axes; and the selection of the accession that has the highest relative contribution (RC) to cloud inertia. Then, step by step, we add to the core the accession that maximises the sum of RC within the core. Finally, all the accessions of the collection are classified according to a maximisation distance criterion, then in selecting the first p%. It is therefore a deterministic method. We applied this method first using components of PCA of original agronomic variables, and second using components of PCA of spatial structures:

- PCSS: selection, in the whole collection, of a sample of p accessions, according to their RC score after PCA of original agronomic variables;
- spatial principal component score strategy (SPPCSS): selection, in the whole collection, of a sample of p accessions according to their RC score after PCA of spatial structures.

2.4. Evaluation criterion and comparison of the different strategies

At first, the seven sampling strategies were compared for their capacity to capture the initial diversity of original collection, which was quantified by a Shannon diversity index computed for all agronomic traits (Shannon and

Weaver, 1963). For each undeterministic method, 200 samples were computed to calculate Shannon indices. Then, to make comparisons, only the lower 95 % confidence limit of the index distribution was taken into account. Considering the importance of spatial structuration of diversity for agronomic traits, the different strategies were then tested for their ability to produce subsamples of populations, which make it possible to estimate a kriged map which is the closest to the original map obtained for the whole collection.

The reference maps used here were those of the first two components of PCA of original agronomic variables for ryegrass (*comp1r*, *comp2r*) and medics (*comp1m*, *comp2m*). The ryegrass reference maps were obtained, from the 550 observations of the whole collection, by kriging on a regular grid of 1 264 equally-spaced points covering France; for medics, the kriging method was performed from the 110 observed values on a grid of 2 479 points covering the six Mediterranean countries.

These reference maps were then compared with maps obtained from subsamples of $p = 5$, 10, 20 and 40 % of the whole collection, and resulting from the different strategies. Comparisons were made through a map distance d simply defined by $d = 1/N\ \Sigma_i |Zrefe_i - Ztest_i|$, where N is the number of regular grid points, $Zrefe_i$ the reference map value at location i, calculated with the whole set of accessions, and $Ztest_i$ the value of the tested map, at the same location i, calculated with only the core accessions.

The first five strategies being undeterministic, we calculated a distance map for each of the 200 samples of each strategy and each proportion p. We then kept the *dsup* distance, which was the upper 95 % confidence limit of the distance distribution.

In the same way, to calculate a confidence interval on the reference map, 200 bootstrap resamplings were made among the 550 ryegrass and 110 medic accessions, and used, as previously, to calculate the *dsup* distance for each reference map.

3. RESULTS

3.1. PCA of original agronomic variables

The first nine components of PCA of 28 agronomic variables of ryegrass summarise 65 % of the total variance. However, from *table III*, giving correlations between initial variables and the first two components of PCA (*comp1r*, *comp2r*), we observe that the first axis is positively correlated with alternativity (ALT), aftermath heading (AH) and frost susceptibility in several locations (FRO1, FRO3), and negatively correlated with persistency and vigour traits. This axis can be globally interpreted as representative of vigour. The second axis is rather negatively correlated with heading date (HD), and positively with rust susceptibility (RUS1, RUS4); it is thus an axis linked with earliness and plant physiology.

As for medics, the first five components of PCA of the original data set account for 78 % of the total variance. *Table IV* gives correlation coefficients between the 29 variables studied and the first two components, *comp1m* and *comp2m*. The first axis (42 %) can be easily interpreted as a vigour axis since all

Table III. Correlation coefficients between the 28 initial ryegrass variables and the first two components (*comp1r*, *comp2r*) of PCA.

Variable	*comp1r*	*comp2r*	Variable	*comp1r*	*comp2r*
FRO1	0.613	−0.485	AUT1	−0.142	−0.068
FRO2	0.341	−0.495	AUT2	−0.345	−0.650
FRO3	0.540	−0.278	AUT3	−0.491	−0.359
FRO4	0.332	−0.384	AUT4	−0.565	−0.214
SPV1	−0.535	0.480	PER1	−0.300	0.326
SPV2	−0.176	0.093	PER2	−0.441	−0.033
SPV3	−0.572	0.244	PER3	−0.373	−0.320
SPV4	−0.496	−0.042	PER4	−0.458	−0.321
SUM1	−0.398	0.022	RUS1	−0.165	0.541
SUM2	−0.259	−0.659	RUS2	−0.101	0.241
SUM3	−0.626	−0.057	RUS3	−0.178	0.237
SUM4	−0.320	−0.089	RUS4	0.310	0.406
ALT	0.573	0.120	GH	0.027	0.321
AH	0.475	0.143	HD	−0.241	−0.441

* Date of measures: 4 December 1995 and one measure every 15 days from 15 February to 15 May; ** number of days from sowing date to first open corolla (or first mature pod separating plant) on three separate plants. See *table I* for abreviations.

height plots (H1 to H6) and width plots (W1 to W7) are positively correlated with it. The second axis (15 %) is linked with reproductive traits, such as pod and seed weights (Ws, Wp), or pod and seed number produced per square meter (Np-m^2, Ns-m^2). Flowering date (DFL) correlation coefficients with the two axes are quite similar, in contrast to the seed weight/pod weight ratio (Ws-Wp), hence approximately seed quality.

3.2. Geostatistical analysis

Although the univariate analysis was carried out on the first two components of each original PCA, we present here only results concerning the first component for each species.

Figure 1 shows variograms of these components, respectively for ryegrass (*comp1r*) and medics (*comp1m*). The *comp1r* variogram has been fitted with a combination of two spherical models, corresponding to two different spatial structures with ranges of 120 and 300 km, respectively. In addition, since we observe a discontinuity at the origin (nugget effect), a third pepitic structure was added to the model. For medics (*comp1m*), two models were also used: a spherical model, with a range of 600 km, and a linear model, in addition to a pepitic structure, leading also to three specific spatial structures.

Figures 2 and *3* show interpolated maps of *comp1r* and *comp1m*, i.e. maps resulting from kriging of the first component of each PCA on regular grid points. Thus, *figure 2*, for ryegrass, displays the strong spatial structuration of component *comp1r*: in brown colour, we find populations, rather localised in the south of France, with high scores of alternativity, aftermath heading, frost susceptibility and low vigour scores. In contrast, northern populations

Table IV. Correlation coefficients between the 29 initial medic variables and the first two components (*comp1m*, *comp2m*) of PCA.

Variable	*comp1m*	*comp2m*
SD	0.526	−0.197
H1	0.794	0.028
W1	0.787	−0.067
H2	0.796	0.030
W2	0.906	−0.113
H3	0.704	0.100
W3	0.911	−0.087
H4	0.824	0.064
W4	0.940	−0.017
H5	0.867	0.124
W5	0.947	0.026
H6	0.716	−0.400
W6	0.912	−0.100
H7	0.476	−0.560
W7	0.846	−0.208
DFL	−0.578	−0.525
DMP	−0.076	−0.266
ROO	0.095	0.109
SPP	0.494	0.299
HEA	−0.475	−0.212
WP	0.254	0.785
WS	0.305	0.856
W100p	0.545	−0.138
W1000s	0.622	−0.084
Ws-Wp	0.353	0.673
Ns-Np	0.547	0.152
Np-m^2	−0.371	0.726
Ns-m^2	−0.077	0.941
Rege	−0.081	0.240

See *table II* for abreviations.

appear more vigorous (green colour). As a general rule, in addition to a clear north–south cline, some patches of homogeneous values and various radius may be noted, explained by the addition of two specific processes of spatial structuration: one with a range of 120 km, the other with a 300 km range.

Figure 3 displays values of *comp1m* for medics. The regionalisation of this new variable is clearly perceptible on the map. French and south Algerian populations exhibiting low growth scores (brown colour) contrast with Spanish and Cretan populations, and to a lesser extent with those from the coastal area of Algeria, which express a stronger growth (green colour). It seems that populations originating from borders of the distribution area, whether in the north (France: areas with high rainfall, but cold climate) or in the south (dry Algerian area, with very low rainfall), clearly contrast with the populations from more favourable areas, with moderate temperatures and sufficient rainfall.

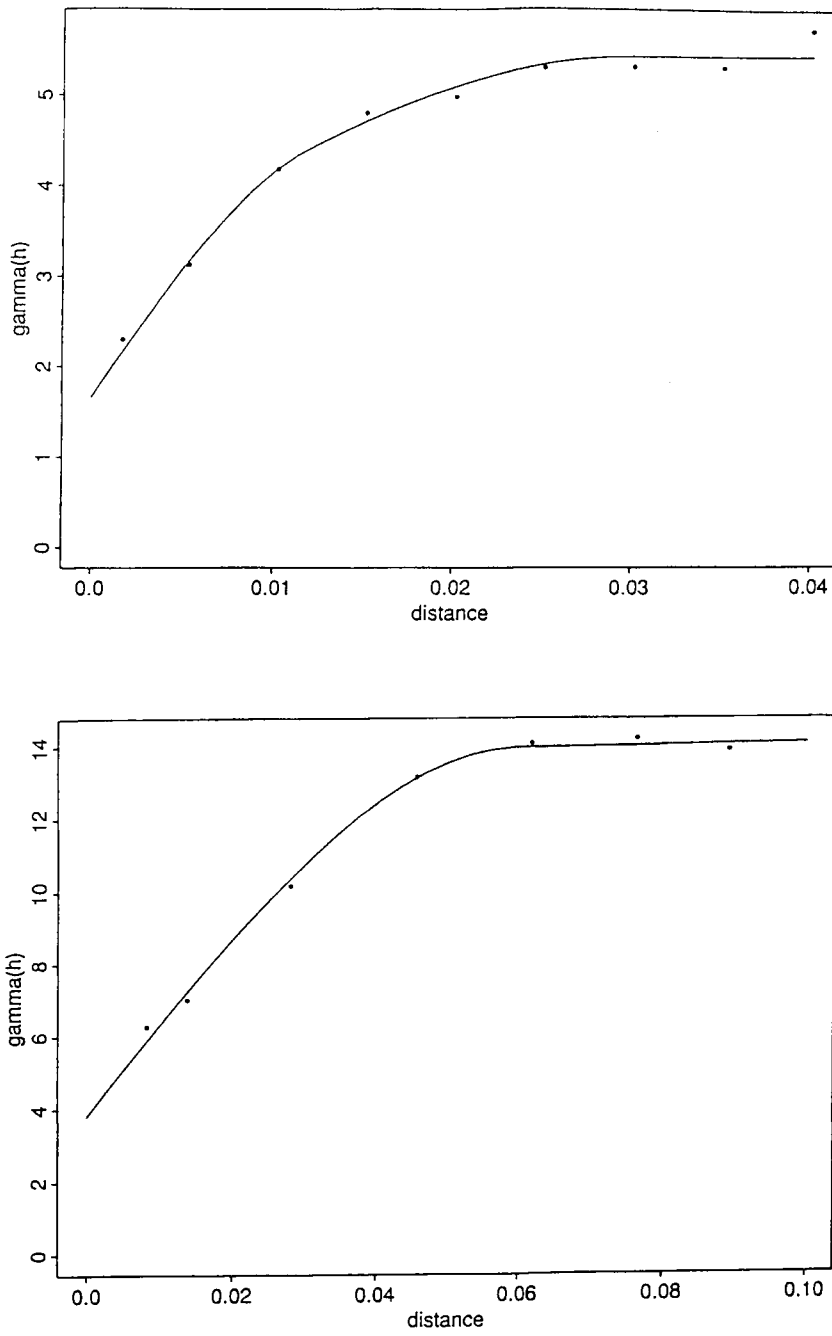

Figure 1. Variograms of first principal component for ryegrass (top part) and medics (bottom part). (x-axis = distance between pairs of populations, in 10^{-4} km; y-axis = semivariance between population in each class of distance).

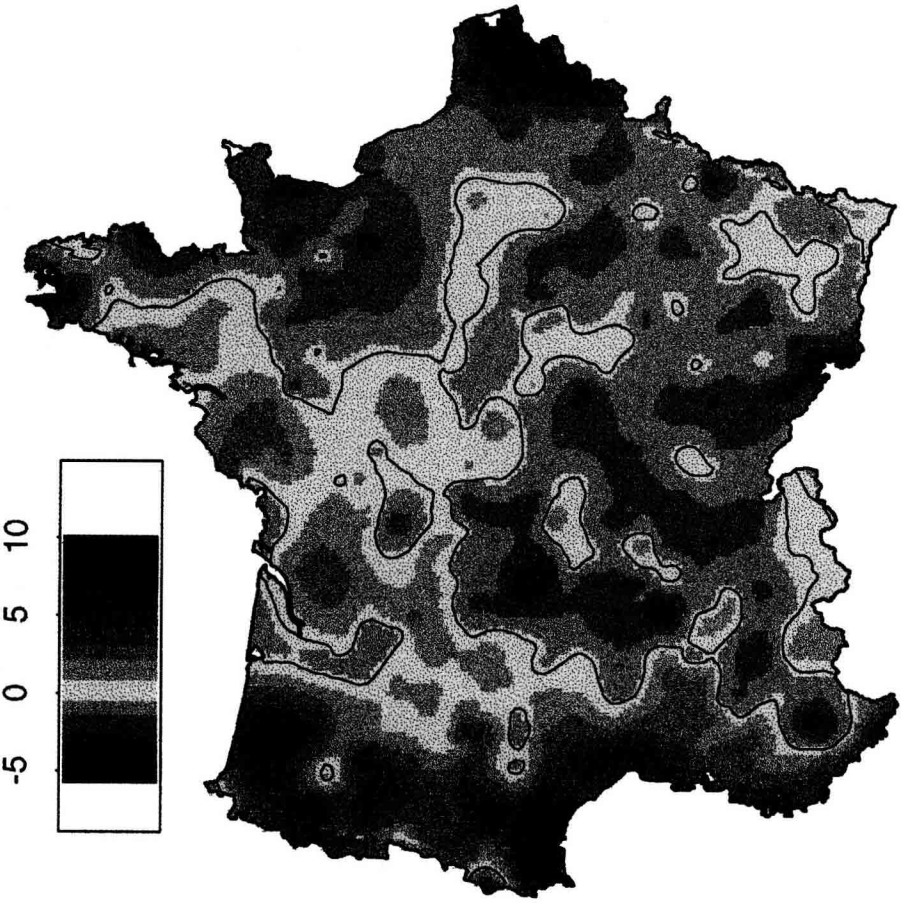

Figure 2. Interpolated map resulting from the kriging of first principal component (*comp1r*) of ryegrass populations (green = strong vigour, yellow = medium, brown = weak).

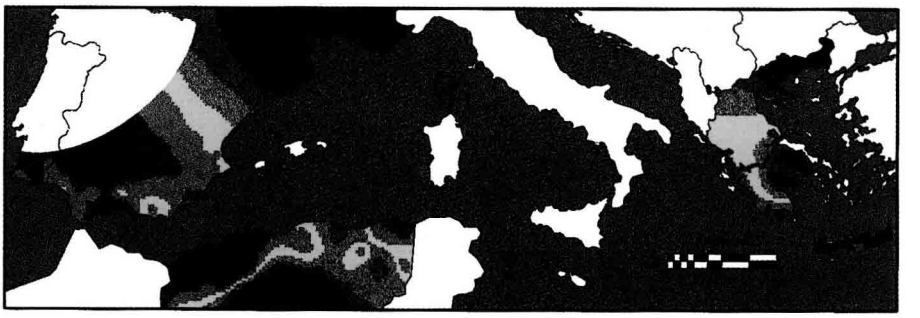

Figure 3. Interpolated map resulting from the kriging of first principal component (*comp1m*) of medic populations (green = strong vigour, yellow = medium, brown = weak).

Medicago truncatula is generally encountered at a high frequency in the latter regions (Prosperi et al., 1991, 1996). These observations are all the more interesting because, although trial plots were established in only one location at Montpellier, populations originating from this area did not show any favourable interaction with the experimental location.

These two maps were then used as reference maps for *comp1r* and *comp1m* variables, in the evaluation and the comparison of the different sampling strategies.

A multivariate analysis, or kriging analysis, permitted the decomposition of the first nine components of initial ryegrass PCA of the specific spatial structures previously described. This analysis indicated a total inertia contribution of 25 % for the first spherical structure (120 km range), 36 % for the second spherical structure (300 km range) and 38 % for the pepitic structure. These three structures were themselves decomposed by PCA into mutually independent components. Finally, the first four components of the two spherical structures, weighted by the respective relative weight of each two structures (25 and 36 %), i.e. eight 'spatial components', were used as input variables in STRATSPA and SPPCSS strategies.

In the same way, the decomposition of the first five components of initial medic PCA of their specific spatial structures, then the decomposition of these structures by PCA, led us to retain the first two components by structures, i.e. four weighted spatial components which were finally used in STRATSPA and SPPCSS strategies applied to medics.

3.3. Sampling strategy efficiencies

Results concerning relative Shannon indices of diversity are presented in *figure 4* for ryegrass and medics. The '100' value corresponds to the index of the whole collection. For the two species, the efficiency of the different sampling strategies decreases with the collection size.

For ryegrass, the Shannon index still remains above 90 %. All the stratified sampling methods give very similar results, scarcely superior to simple random sampling, except for $p = 5$ %. The two deterministic methods produce samples with relative indices of diversity higher than those of other methods. In particular, the SPPCSS strategy is clearly superior for $p = 20$ and 10 %.

In contrast, for medics, the relative Shannon index decreases more drastically with sample size, especially for samples below 20 %, with a loss of initial diversity that may exceed 30 %. In contrast to ryegrass, this decrease is faster with the PCSS strategy, which is unexpected. The SPPCSS method still appears the most efficient for sample sizes of 20 and 10 %, the values for 5 % (i.e. six populations) being probably without significance.

Figure 5 displays map distances with respect to the reference map, with the different sampling strategies and for principal components *comp1r* and *comp1m*, for ryegrass and medic, respectively.

For ryegrass, RANDOM, STRATAGRO, STRATREG and STRATCONTI methods give very similar results for $p = 40$, 20 and 10 %, with a slight superiority of STRATCONTI for a sample size of 5 %. The two strategies using information on spatial structuration, SPPCSS and STRATSPA, appear more efficient overall than the previous ones, but with quite erratic behaviour of

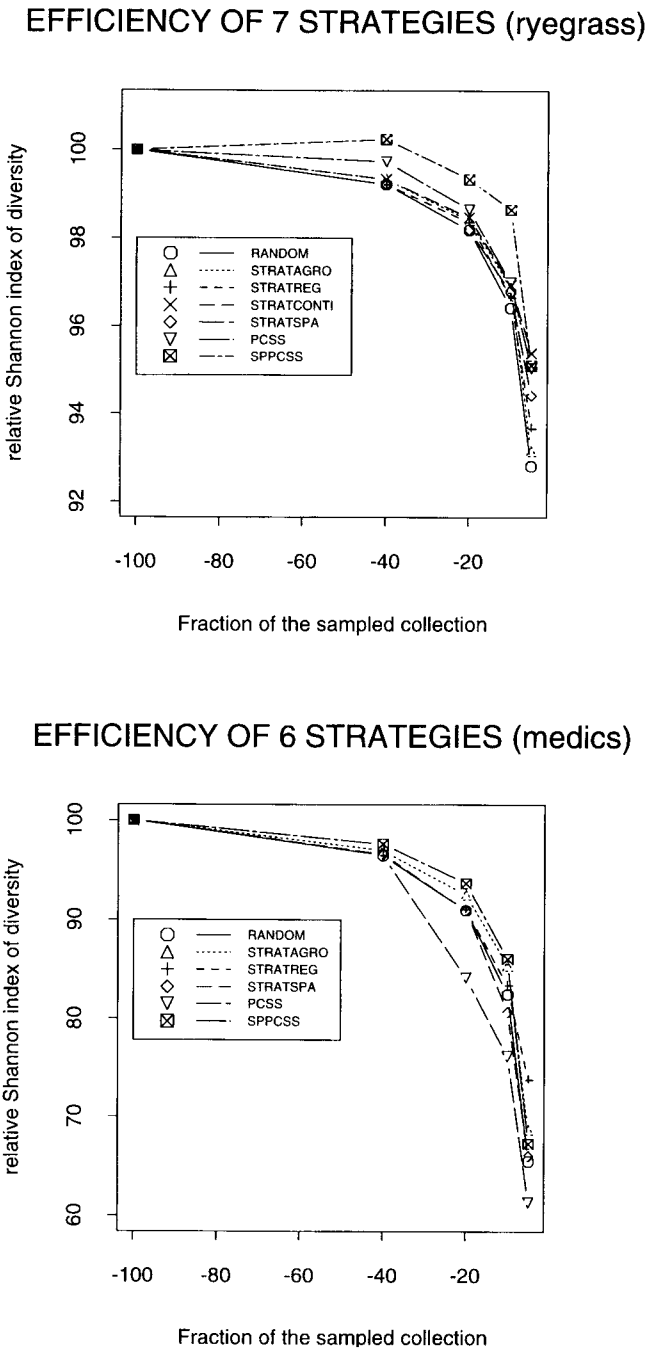

Figure 4. Lower 95 % confidence limit of the proportion of the total diversity represented in samples of various size, for the different sampling strategies, in ryegrass (top part) and medic collections (bottom part).

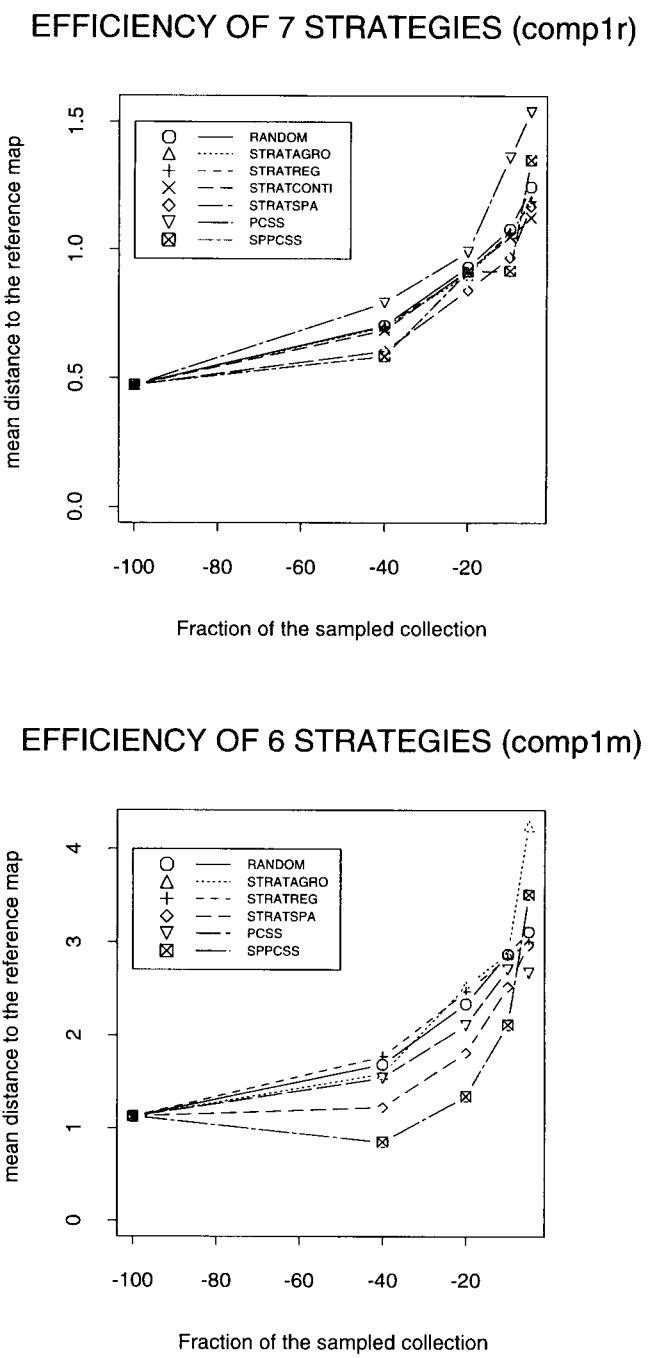

Figure 5. Upper 95 % confidence limit of map distance compared with reference map distance, for the ryegrass variable *comp1r* (top part), and the medic variable *comp1m* (bottom part) and the different sampling strategies.

the SPPCSS method for $p = 20$ % and $p = 5$ %. The PCSS strategy shows the poorest results, which demonstrates that the maximisation of diversity, without taking into account its spatial structuration, does not allow the optimum representation of these structures. It therefore seems that the STRATSPA method is the most recommendable. However, for all the methods, the quality of map reconstruction, which measures efficiency to represent spatialised variability, decreases quite rapidly for $p < 40$ %, with an upper limit of confidence interval roughly higher than the bootstrapping value of the whole collection. In contrast with the conclusions concerning the Shannon index, this criterion would lead us to recommend a sample size of 220 populations for a core collection.

In the case of medics, conclusions are very similar. There are very small differences between all classical stratified sampling strategies and the PCSS strategy, whereas spatialised methods are more efficient. This time, it is the SPPCSS strategy which appears the best. For this species, the distance map criterion leads us to suggest a core collection size of at least 20 % of the whole collection.

4. DISCUSSION AND CONCLUSION

From computations based on the theory of neutral genes, Brown (1989a) showed that about 70 % of the allelic richness should be included in a core subset of 5 to 10 % of the whole collection. He also suggested that every subspecies or geographic origin should be represented in the core, which should then be developed through stratified sampling methods. Subsequently, Brown (1989b) compared three sampling strategies, where the number of accessions retained in each cluster was constant (C), proportional (P) to the cluster size or to its logarithm (L). He concluded that the latter method was the most efficient. A theoretical explanation could be drawn from the expectation of the number of alleles that can be maintained in a sample of size N according to the neutral theory:

$$E(k) = \theta * \log(N/\theta) + 0.6$$

where $\theta = 4\ Ne\ \mu$, Ne being the effective size and μ the neutral mutation rate. If θ is considered to be constant, $E(k)$ is proportional to $\log(N)$. Alternatively, Schoen and Brown (1995) proposed that each cluster should be represented by a number of accessions proportional to $\theta = 4\ Ne\ \mu$. However, as these parameters are not easy to estimate accurately, Schoen and Brown suggested using a rough estimate of θ by $h/(1-h)$, where h is the Nei (1977) diversity index, and named this 'H strategy'. This method appears to be close to the 'G strategy' proposed by Yonezawa et al. (1995), which differs only by the diversity measure.

Several methods have been used to devise core collections in perennial relatives of soybean (Brown et al., 1987), groundnut (Holbrook et al., 1993), tulip (Loos and van Duin, 1989), annual (Diwan et al., 1994, 1995) or perennial *Medicago* (Basigalup et al., 1995) and cassava (Cordeiro et al, 1995). Most of these studies used hierarchical clustering based on either passport or evaluation data (Peeters and Martinelli, 1989; Crossa et al., 1995). Some studies also used agroecological data on the site of origin (Tohme et al., 1995).

Methodological studies comparing different sampling strategies have also been reported. The efficiency criteria used diversity indices based either on

neutral markers or on quantitative traits such as descriptors or agronomic traits. Erksine and Muehlbauer (1991) compared three sampling methods, namely random (R) or stratified (C and P), but they did not find any significant differences for the number of isozyme alleles captured. Spagnoletti-Zeuli and Qualset (1993) compared five methods for the phenotypic variance of durum wheat. They reported that the most efficient strategy is double clustering, based on both the geographic origin of lines and canonical variables from evaluation data, as in the STRATAGRO method reported here. This result seems to be quite general, as similar conclusions have been reported for ryegrass populations (Casler, 1995), barley landraces (van Hintum et al., 1995) and for the United States Department of Agriculture core collections of annual (Diwan et al., 1994, 1995) and perennial (Basigalup et al., 1995) *Medicago*.

The geographic origin of the material, either primary origin for natural populations, or secondary adaptation as in the case of landraces, appears to be an efficient criterion for organising a collection before sampling a core. However, the usual clustering by country or administrative region, although easily available, has few theoretical bases, as natural selection pressures rarely stop at administrative boundaries, except perhaps as a result of some agricultural practices. Therefore, the use of methods which explicitly take into account the geographic distance between sites may be more justified. In a previous study, we suggested the use of hierarchical clustering with a geographic contiguity constraint, based on the range of the variogram (Charmet et al., 1994), and showed that this method was the most efficient in representing the phenotypic diversity of the whole collection of ryegrass populations, as estimated by a Shannon index (Charmet and Balfourier, 1995).

The criteria used to assess the degree of diversity in the core are most often based on phenotypic values (mean, range, variance, Shannon index, etc.; see Galwey, 1995), and less frequently on the genotypic value of accessions – except in one study on maize, which used combining abilities (Radovic and Jelovac, 1994). It should be kept in mind that clustering phenotypes may be misleading, particularly when phenotypes are evaluated in a limited range of environments. Such misclassifications may cause loss of true genetic variability. Alternative procedures such as the use of genetic markers have been proposed (van Hintum, 1994; Schoen and Brown, 1995, 'M strategy') or the use of coefficients of parentage (van Hintum and Haalman, 1994). However, the implementation of these methods requires either the knowledge of pedigrees or the availability of biochemical or molecular analyses of the whole collection. Clearly these conditions are often unrealistic.

The methods proposed here are based on the following assumption: the spatial structures identified by geostatistic analyses are the combined result of selective pressure by the environment (sensu lato) and isolation-by-distance (Monestiez et al., 1994). Consequently, populations that are located close to each other are more likely to share a common genetic background, and this was already taken into account by clustering with a contiguity constraint. Reciprocally, populations that are located far from each other, and which nevertheless are phenotypically similar, are likely to possess different genetic combinations, selected independently under similar pressures, leading to similar adaptations. Thus, the maintenance of optimum genetic diversity implies that every spatial cluster be sampled, and not only every phenotypic cluster. This

is usually done empirically when gene-bank managers cross, for example, a phenotypic clustering with the country of origin. The proposed method of sampling within every spatial structure appears to be more accurate. The consequence is that the sample obtained is that which best represents the map constructed from the whole set of data. Therefore, we proposed that the criterion to be minimised is the distance between the two maps.

The SPPCSS deterministic strategy may be considered, a priori, as the most suitable for our purposes. This method yielded the best results in the two species studied, for the map distance criterion as well as for the global diversity index. The decline of the Shannon diversity index as the sample size decreases is faster in medics than in perennial ryegrass. This is probably related to the reproductive biology of medics (selfing), which leads to a higher rate of between-population differentiation. However, the STRATSPA strategy is not significantly less efficient than SPPCSS for the 40 % cores. In perennial ryegrass, the two methods give very similar results for 40 %, the behaviour of SPPCSS being somewhat erratic for smaller core sizes. It does not appear to show dramatic differences between the two model plants. For both species and for the map distance criterion, the optimal sampling fraction seems to be 20 to 40 % of the whole collection, whatever its total size (110 for medics, 550 for ryegrass). Thus, the 10 % threshold, established by Brown (1989a) for neutral genes, appears to be too small for agronomic traits. Similar conclusions have already been reported from experiments on medics (Diwan et al., 1995), and through simulations (Yonezawa et al., 1995).

The main advantage of stratifying on spatial components is that it makes it possible to combine the clustering with other passport data such as microenvironment factors, or complementary data available on any subset of populations. In particular, isozyme frequencies at eight to ten loci are available for all the *Medicago* and for 60 ryegrass populations. These data will be used to refine the core collections based on spatial strategies. In self-pollinated species, population sizes are expected to be more variable than in cross-pollinated species, so an 'H strategy' may be used to sample within each cluster obtained by STRATSPA. Similarly, in outbreeding species such as ryegrass, a more precise study of gene flow through Wright *Fst* analysis may allow a weighting of within-cluster sampling according to the average degree of between-population differentiation in each region.

The core collections, which are being devised, will be multiplied with the help of private breeding companies, in the framework of the national charter on conservation of fodder crop genetic resources. They will also be included in part in European core collections which are being defined by ECPGR, which is currently supporting a pilot programme on perennial ryegrass, with the aim to extend it to other forage species (Gass et al., 1995).

REFERENCES

Balfourier F., Charmet G., Geographic patterns of isozyme variation in Mediterranean populations of perennial ryegrass, Heredity 72 (1994) 55–63.

Barker D.G., Bianchi S., Blondon F., Dattée Y., Duc G., Flament P., Gallusci P., Génier P., Guy P., Muel X., Tourneur J., Dénarié J., Huguet T., *Medicago truncatula*,

a model plant for studying the molecular genetics of the *Rhizobium*–legume symbiosis, Plant Mol. Biol. Rep. 8 (1990) 40–49.

Basigalup D.H., Barnes D.K., Stucker R.E., Development of a core collection for perennial *Medicago* plant introductions, Crop Sci. 35 (1995) 1163–1168.

Becker R.A., Chambers J.M., Wilks A.R., The New S language. A Programming Environment for Data Analysis and Graphics, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA, 1988.

Bonnin I., Huguet T., Gherardi M., Prosperi J.M., Olivieri I., High level of polymorphism and spatial structure in a selfing plant species, *Medicago truncatula* (Leguminosae), shown using RAPD markers, Am. J. Bot. 83 (1996a) 843–855.

Bonnin I., Prosperi J.M., Olivieri I., Genetic markers and quantitative genetic variation in *Medicago truncatula* (Leguminosae): a comparative analysis of population structure, Genetics 143 (1996b) 1795–1805.

Brown A.H.D., The case for core collection, in: Brown A.H.D., Frankel O.H., Marshall D.R., Williams J.T. (Eds.), The Use of Plant Genetic Resources, Cambridge University Press, Cambridge, 1989a, pp. 136–156.

Brown A.H.D., Core collection: a practical approach to genetic resources management, Genome 31 (1989b) 818–824.

Brown A.H.D., The core collection at the crossroads, in: Hodgkin T., Brown A.H.D., van Hintum T.J.L., Morales E.A.V. (Eds.), Core Collection of Plant Genetic Resources, IPGRI-Wiley & Sons, Chichester, UK, 1995, pp. 77–92.

Brown A.H.D., Grace J.P., Speer S.S., Designation of a "core" collection of perennial *Glycine*, Soybean Genetics Newsletter 14 (1987) 59–70.

Casler M.D., Patterns of variation in a collection of perennial ryegrass accessions, Crop Sci. 35 (1995) 1169–1177.

Charmet G., Balfourier F., Collecte et évaluation de populations naturelles de ray grass anglais en France, C.R. Acad. Agric. Fr. 77 (1991) 53–64.

Charmet G., Balfourier F., The use of geostatistics for sampling a core collection of perennial ryegrass populations, Genet. Res. Crop Evol. 42 (1995) 303–309.

Charmet G., Balfourier F., Bion A., Agronomic evaluation of a collection of French perennial ryegrass populations: multivariate analysis using genotype x environment interactions, Agronomie 10 (1990) 807–823.

Charmet G., Balfourier F., Monestiez P., Hierarchical clustering of perennial ryegrass populations with geographic contiguity constraint, Theor. Appl. Genet. 88 (1994) 42–48.

Chaulet E., Prosperi J.M., Genetic diversity of *Medicago truncatula* Gaertn collected in Algeria, in: Proceedings of the 7th Genetic Resources section meeting of Eucarpia, March 1994, Inra, Clermont-Ferrand, France, 1994, pp. 255–257.

Cordeiro C.M.T., Morales E.A.V., Ferreira P., Rocha D.M.S., Costa I.R.S., Valois A.C.C., Silva S., Towards a Brazilian core collection of cassava, in: Hodgkin T., Brown A.H.D., van Hintum T.J.L., Morales E.A.V. (Eds.), Core Collection of Plant Genetic Resources, IPGRI-Wiley & Sons, Chichester, UK, 1995, pp. 155–168.

Cressie N., Statistics for Spatial Data, Wiley, New York, 1991.

Crossa J., De Lacy I.H., Taba S., The use of multivariate methods in developing a core collection, in: Hodgkin T., Brown A.H.D., van Hintum T.J.L., Morales E.A.V. (Eds.), Core Collection of Plant Genetic Resources, IPGRI-Wiley & Sons, Chichester, UK, 1995, pp. 77–92.

Diwan N., Bauchan G.R., McIntosh M.S., A core collection for the United States annual *Medicago* germplasm collection, Crop Sci. 34 (1994) 279–285.

Diwan N., McIntosh M.S., Bauchan G.R., Methods of developing a core collection of annual *Medicago* species, Theor. Appl. Genet. 90 (1995) 755–761.

Erksine W., Muehlbauer F.J., Allozyme and morphological variability, outcrossing rate and core collection formation in lentil germplasm, Theor. Appl. Genet. 83 (1991) 119–125.

Frankel O.H., Genetic perspectives of germplasm conservation, in: Arber W., Llimensee K., Peacock W.J.P. (Eds.), Genetic Manipulation: Impact on Man and Society, Starlinger Cambridge University Press, Cambridge, 1984.

Galwey N.W., Verifying and validating the representativeness of a core collection, in: Hodgkin T., Brown A.H.D., van Hintum T.J.L., Morales E.A.V. (Eds.), Core Collection of Plant Genetic Resources, IPGRI-Wiley & Sons, Chichester, UK, 1995, pp. 187–198.

Gass T., Sackville-Hamilton R., Kolshus K., Frison E., Report of a working group on forages. Fifth Meeting of ECP/GR working group, Hissar, Bulgaria, 31 March–2 April, IPGRI, Rome, Italy, 1995.

Goulard M., Voltz M., Linear coregionalization model: tools for estimation and choice of cross-variogram matrix, Math. Geol. 24 (3) (1992) 269–286.

Hamon S., Hodgkin T., Dussert S., Anthony F., Noirot M., Core collection: theoretical and applied aspects, in: Proceedings of the 7th Genetic Resources section meeting of Eucarpia, March 1994, Inra, Clermont-Ferrand, France, 1994, pp. 73–82.

Hamon S., Noirot M., Antony F., Developing a coffee core collection using the principal components score strategy with quantitative data, in: Hodgkin T., Brown A.H.D., van Hintum T.J.L., Morales E.A.V. (Eds.), Core Collection of Plant Genetic Resources, IPGRI-Wiley & Sons, Chichester, UK, 1995, pp. 117–126.

Hintum T.L.J. van, Comparison of marker systems and construction of a core collection in a pedigree of European barley, Theor. Appl. Genet. 89 (1994) 991–997.

Hintum T.L.J. van, Haalman D., Pedigree analysis for composing a core collection of modern cultivars, with examples from barley (*Hordeum vulgare* s. lat.), Theor. Appl. Genet. 88 (1994) 70–74.

Hintum T.L.J. van, Bothmer R. von, Visser D.L., Sampling strategies for composing a core collection of cultivated barley (*Hordeum vulgare* S. lat.) collected in China, Hereditas 122 (1995) 7–17.

Holbrook C.C., Anderson W.F., Pittman R.N., Selection of a core collection from the U.S. germplasm collection of peanut, Crop Sci. 33 (1993) 859–861.

Loos B.P., Duin P.J.W. van, Establishing a core collection representing genetic variation in tulip, Plant Genetic Resources Newsletter 78–79 (1989) 11–12.

Monestiez P., Goulard M., Charmet G., Geostatistics for spatial genetic structure: study of wild populations of perennial ryegrass, Theor. Appl. Genet. 88 (1994) 33–41.

Monestiez P., Goulard M., Charmet G., Balfourier F., Analysing spatial genetic structures by multivariate geostatistics: study of wild populations of perennial ryegrass, in: Baasi E.Y., Schofield N. (Eds.), Geostatistics Wollongong, vol. 2, Kluwer Academic Publishers, Dordrecht, 1997, pp. 1197–1208.

Nei M., F-statistics and analysis of gene diversity in subdivided populations, Ann. Hum. Genet. 41 (1977) 225–233.

Noirot M., Hamon S., Anthony F., The principal component scoring: a new method of constituting a core collection using quantitative data, Genet. Res. Crop Evol. 43 (1996) 1–6.

Peeters J.P., Martinelli J.A., Hierarchical cluster analysis as a tool to manage variation in germplasm collection, Theor. Appl. Genet. 78 (1989) 42–48.

Prosperi J.M., Boumard P., Angevain M., Mansat P., Répartition et adaptation écotypique de *Medicago* spp. annuelles en Méditerranée occidentale, in: Proceedings of IV International Congress of Rangelands, Montpellier, France, avril 1991, Cirad Montpellier, 1991, pp. 413–416.

Prosperi J.M., Angevain M., Bonnin I., Chaulet E., Genier G., Jenczewski E., Olivieri I., Ronfort J., Genetic diversity, preservation and use of genetic resources of Mediterranean legumes: Alfalfa and Medics, Cah. Options Méditerranéennes 18 (1996) 71–89.

Radovic G., Jelovac D., The possible approach in maize 'core collection' development, in: Proceedings of the 7th Genetic Resources section meeting of Eucarpia, March 1994, Inra, Clermont-Ferrand, France, 1994, pp. 109–115.

Shannon C.E., Weaver W., The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL, 1963.

Schoen D.J., Brown A.H.D., Maximising genetic diversity in core collections of wild relatives of crop species, in: Hodgkin T., Brown A.H.D., van Hintum T.J.L., Morales E.A.V. (Eds.), Core Collection of Plant Genetic Resources, IPGRI-Wiley & Sons, Chichester, UK, 1995, pp. 55–77.

Spagnoletti-Zeuli P.L., Qualset C.O., Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat, Theor. Appl. Genet. 87 (1993) 295–304.

Tohme J., Jones P., Beebe S., Iwanaga M., The combined use of agroecological and characterization data to establish the CIAT *Phaseolus vulgaris* core collection, in: Hodgkin T., Brown A.H.D., van Hintum T.J.L., Morales E.A.V. (Eds.), Core Collection of Plant Genetic Resources, IPGRI-Wiley & Sons, Chichester, UK, 1995, pp. 95–107.

Wackernagel H., Multivariate Geostatistics, Springer-Verlag, Berlin, 1995.

Yonezawa K., Nomura T., Morishima H., Sampling strategies for use in stratified germplasm collections, in: Hodgkin T., Brown A.H.D., van Hintum T.J.L., Morales E.A.V. (Eds.), Core Collection of Plant Genetic Resources, IPGRI-Wiley & Sons, Chichester, UK, 1995, pp. 35–53.