G. Charmet · F. Balfourier · P. Monestiez

Hierarchical clustering of perennial ryegrass populations with geographic contiguity constraint

Received: 10 April 1993 / Accepted: 28 June 1993

Abstract An algorithm of automatic classification is proposed and applied to a large collection of perennial ryegrass wild populations from France. This method is based on an ascendant hierarchical clustering using the Euclidian distance from the principal components extracted from the variance-covariance matrix between 28 agronomic traits. A contiguity constraint is imposed: only those pairs of populations which are defined as contiguous are grouped together into a cluster. The definition of contiguity is based on a geostatistical parameter: the range of the variogramme, i.e. the largest distance above which the variance between pairs of population no longer increases. This method yields clusters that are generally more compact than those obtained without constraint. In most cases the contours of these clusters fit well with known ecogeographic regions, namely, for macroclimatic homogeneous conditions. This suggests that selective factors exert a major influence in the genetic differentiation of ryegrass populations for quantitatively inherited adaptive traits. It is proposed that such a method could provide useful genetic and ecogeographic bases for sampling a core collection in widespread wild species such as forage grasses.

Key words Genetic resources · *Lolium perenne* L. · Ecotypic variation · Spatial variation · Core collection

Introduction

Plant species such as forage grasses, legumes or forest trees are present in nature as large numbers of wild populations. This is also seen in some wild relatives of domesticated crops like cereals and vegetables. These wild populations can represent useful genetic resources for the improvement of cultivated species and they have been intensively collected in last decades.

It is obviously unrealistic to collect and to conserve samples of every population. Furthermore, the size of a collection, which must be regularly multiplied in order to become freely available, is limited, in the outbred species, by the number of isolation facilities required. The concept of "core collection" proposed by Brown (1989) would be very useful for this reason. As forage plants are relatively undomesticated, many populations may have interesting characteristics. The aim of core sampling is to gather a representative sample of the ecotypic variation present in the whole collection.

Multivariate analyses and, in particular, clustering methods provide a scientific basis for sampling a core (Peeters and Martinelli 1989). In wild species, however, this variation is known to be influenced both by ecological and geographical factors, the former being a possible source of selection pressure while the latter may cause isolation-by-distance. Therefore, Peeters et al. (1990) have suggested the use of ecogeographical data in the exploitation of variation. These authors suggest that simple geographic data may offer an effective method for stratifying and sampling variation in germ plasm collections. In the same way, Weltzien (1989) found geographic grouping to be as effective as grouping based on similarity for morphological traits in distributing the variance for quantitative traits among and within groups of barley landrace populations from the Near East.

Several methods have been put forward to handle geographic data and quantitative traits together. Gabriel and Sokal (1969) proposed a method for partitioning an area into regions, with populations being homogeneous within each region but differing from one region to another. They defined the concepts of contiguity and connectedness. Other authors have proposed methods of hierarchical clustering with a geographic contiguity constraint (Lebart 1978; Monestiez 1978; Perruchet 1979). These methods lead to clusters more readable on maps than hierarchical clustering used alone.

Communicated by P. M. A. Tigerstedt

G. Charmet (✉) · F. Balfourier
INRA* Station d'Amélioration des Plantes, 63039 Clermont-Ferrand, France

P. Monestiez
INRA* Station de Biométrie, 84140 Montfavet, France

[*] Institut National de la Recherche Agrononique

They are useful in all situations where individuals are geographically defined and a contiguity can be defined, such as geology, economy or population genetics.

This paper describes the application of these methods to an extensive collection of perennial ryegrass populations from France.

Materials and Methods

Collection of populations and evaluation designs

The collection and evaluation designs have been fully described by Charmet et al. (1990). It could be useful, however, to recall the main steps. Eleven teams of private breeders or INRA workers collected seed samples throughout continental France in the summer of 1983 and 1984. Each year, each team methodically prospected one administrative region and sampled wild ryegrass as regularly as possible in order to obtain an extensive sample of the possible variation of the natural species. On each collecting site, seeds from at least 50 plants were taken from an ecologically homogeneous area of 100–1000 m^2. These conditions are supposed to yield a sample of seeds representative of the original panmictic population (Yonezawa 1985; Tyler 1987). Information about collection sites was taken at the time of collection: general habitat, grassland management, latitude, longitude and altitude. Seeds of all plants were bulked, without balancing for the contribution of each plant, then divided into 12 equal amounts for evaluation.

In total, 226 populations were collected in 1983 and evaluated from 1984 to 1986; 321 others were collected 1 year later and studied from 1985 to 1987. Wild ryegrass populations were evaluated in 12 different locations in France, either as turf plots or as spaced plant nurseries. This paper only deals with the spaced plant evaluation.

In each location, 30 plants of each origin were put into the field in the spring (April or May) in a randomized three-block design with six control cultivars (namely, cvs 'Mantilla', 'Sisu', 'Préférencé', 'Vigor', 'Manhatan', 'Idole') as elementary plots of 10 plants at 40-cm spacing. The nurseries were cut two or three times during the establishment year to allow diseases to develop, then frequently (every 3–4 weeks) in the second year to simulate a rotational grazing. Subsequently, all plants were cut the same day at each location whatever their developmental stage. As a result, most observations were made outside the sexual cycle, e.g. in early spring and later on during vegetative regrowth. Finally, the plants were allowed to flower in the last year to avoid seedling contamination.

Ten traits of agronomic interest were observed during the 3 years, most of them being scored visually on a 1–9 scale. These traits were described in the following ways:
During the sowing year

1) the ability to produce spikes in the year sown (1 = no spikes to 9 = many ears per plant)
2) growth habit in autumn: (1 = erect to 9 = prostrate)
3) crown rust susceptibility: (1 = resistant to 9 = heaviy damaged)

During the first year after sowing

4) frost susceptibility (1 = no change to 9 = dead)
5) early spring growth (1 = weak to 9 = vigorous)
6) aftermath heading (1 = none to 9 = many ears and few leaves in the regrowth)
7) summer aspect (1 = dried to 9 = green and vigorous)
8) autumn regrowth (1 = weak to 9 = vigorous)

In the second year after sowing

9) persistence: (1 = dead to 9 = fully living plant)
10) heading date: (date of ear emergence, fron January 1st).

Statistical development

An analysis of variance was performed using the model:

$$X = MU + YE + LO + YE.LO + YE.LO.BL + PO + LO.PO + EPS$$

where MU is the overall average, YE is the main year effect, LO is the main location effect, $YE.LO$ is the year-by-location interaction, $YE.LO.BL$is the block, hierachized within $YE.LO$, effect, PO is the main population effect, $LO.PO$ is the population × location interaction and EPS is the residual error.

This analysis allowed us to identify two kinds of agronomic traits. The first set of traits are those related to reproduction development: ability to flower in the year sown, growth habit, aftermath heading and date of ear emergence. These traits are quite stable from one location to another, and thus the overall population means $MU + PO$ (adjusted in this way for year effects) are meaningful descriptors. The other traits, i.e. vigour and susceptibility scores, show high levels of population × location interaction. A breakdown of these interactions using a multiplicative model (Mandel 1971) indicated that four locations out of nine summarize most of the population × location interactions.

Thus, a complete description of these sets of ryegrass populations requires the use of 28 initial variables, i.e. population means for four "stable" traits and 24 combinations of 4 locations × six traits for the interactive ones. Clearly this indicates that the use of multivariate methods such as principal component analysis can be useful in reducing the size of the data and therefore the complexity of their interpretation.

This has been done by Charmet et al. (1990), who carried out a hierarchical clustering method using the first six principal components in order to group the 547 wild populations into more homogeneous classes. However, most of the clusters obtained in this way group populations that are not located in a restricted geographic area. Such a "geographic" grouping can obviously be obtained by crossing the former hierarchical clustering with a regional partition.

Another method, called "hierarchical clustering with geographical contiguity constraint" has been developed by statisticians for use in geography, economy and geology (Monestiez 1978; Perruchet 1979). This problem of regional clustering has also been addressed by Fisher (1978), Lefkovitch (1980) and Felsenstein (1983). We used the computer package of Lebart (1978) in this study. The principles of this method can be summarized as follows. A contiguity matrix is built with every individual to classify by row (here the populations of ryegrass), and each row contains the list of individuals that are considered to be contiguous with the former population. This representation is equivalent to that of the theory of the graph, populations being thus the edges of a contiguity graph. A distance or similarity index is defined between every pair of populations. In our study the standard Euclidian distance was used. In the common hierarchical clustering the two closest populations are gathered into the first cluster. The contiguity constraint is quite simple: only those populations will be permitted to cluster that are declared to be contiguous in the matrix (i.e. along the graph). When two populations have been grouped into one cluster, both the distance matrix and the contiguity matrix are re-estimated. Here the incremental sum of squares is used as an aggregation strategy, and the new contiguity matrix simply consists of the fusion of the rows of the populations that have been grouped. In this way the contiguity is transmitted from step to step within the whole set of populations, which can thus be grouped in a unique cluster if no other constraint is imposed as, for instance, a maximum number of populations within each cluster.

It should be noted that contiguity has not been defined here in the classical way used by geographers. In the case of entities such as districts, contiguity is easily defined when two districts have a common border. In the case of collection sites of natural populations such a definition cannot apply, since the area of each site is not well defined and the sites are often much smaller than districts. For this reason we have utilized another definition, where every population within a certain radius centred on a given population is declared to be contiguous to this population. The radius can either be chosen arbitrarily (for example to give quite a constant number of contiguous pairs at each row of the contiguity matrix) or be based on a geostatistical basis. Monestiez et al. (1994) established the variogrammes for some agronomic traits of this collection, i.e. the graphs of $V(X)$, the variance among all populations that are in a given class of distance x from each other, as a function of x. They found these variogrammes to have range of 120 km, that is the variance between populations reaches its maximum value for populations 120 km distant from each other.

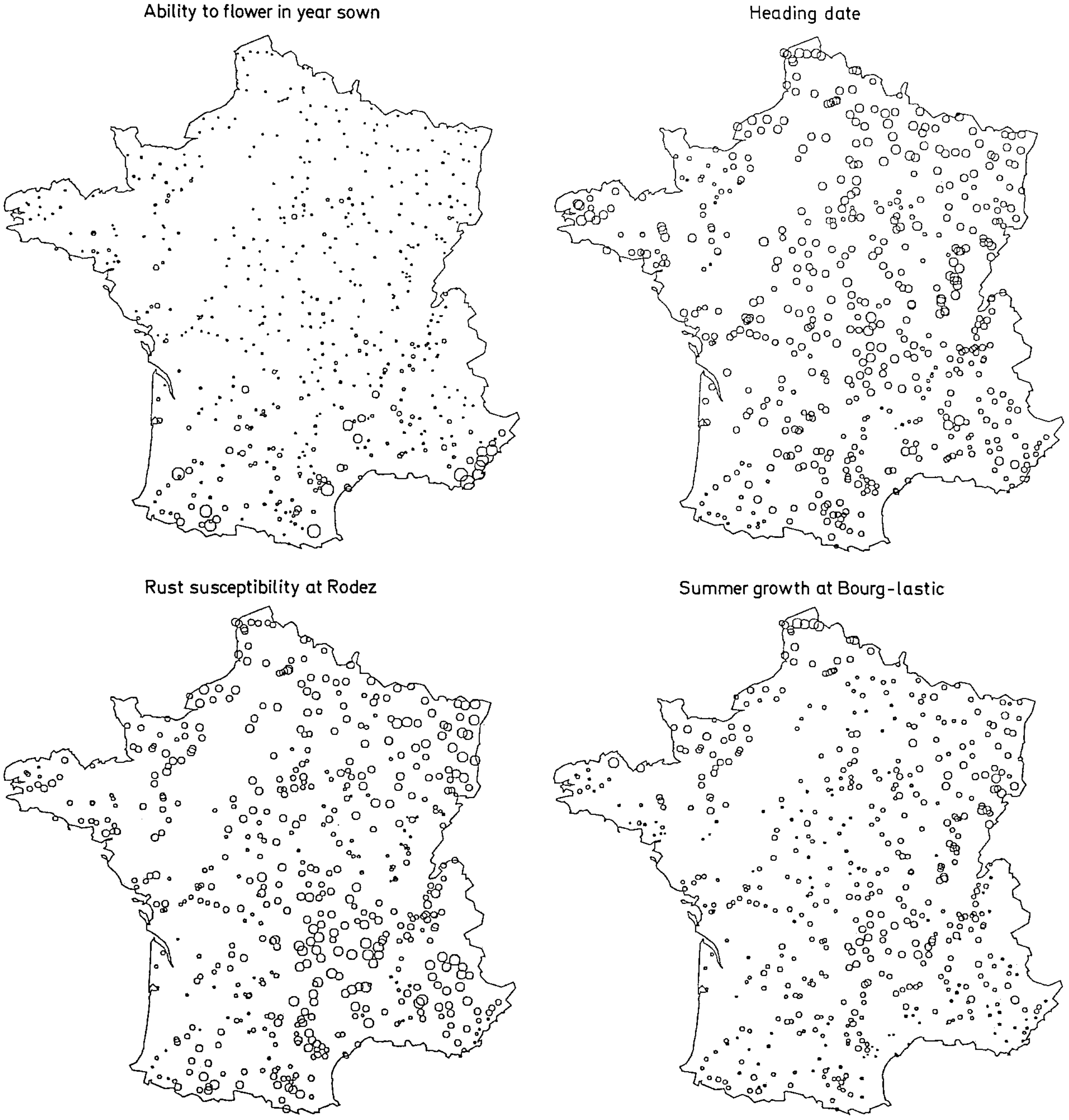
In the study discribed here, we have chosen a radius of 60 km, which is half the range of the variogramme. As a result, the variance between all pairs of contiguous populations is about half of the total variance between populations for most traits. In order to allow a comparison with the previous classification of this collection, the same ratio of between cluster variance to total variance (approximately 0.5) was used to determine the level of clustering. In fact, the number of clusters is greater than what was described in the previous study for the same variance ratio because of the contiguity constraint, which often leads to several clusters having similar agronomic characteristics but not being located contiguously on the map. This method resulted in 25 contiguous clusters, while the unconstrained hierarchical clustering method produced 11 clusters.

Results

The spatial variation of some agronomic traits is illustrated in Fig. 1. While a trait such as "ability to flower in the year sown", which denotes an "annual habit", decreases from south to north along a cline, most of the other traits show a more "patchy" structure. In some instances it can be verified that the scale of this patchiness is related to the range of the variogrammes (Monestiez et al. 1994), which confirms the choice of the contiguity radius as an appropriate distance. It appears difficult, however, to establish a clear-cut geographic partition based on only one given trait. The multivariate approach seems therefore justified.

The results of the constrained hierarchical clustering method are illustrated in Fig. 2. While most of the clusters are located in quite compact areas on the map, some of them show a "chaining" effect that is due to the contiguity definition used.

Fig. 1 Spatial variation of some agronomic traits in perennial ryegrass populations. The size of the symbols is proportional to the population mean


Ability to flower in year sown
Heading date
Rust susceptibility at Rodez
Summer growth at Bourg-lastic

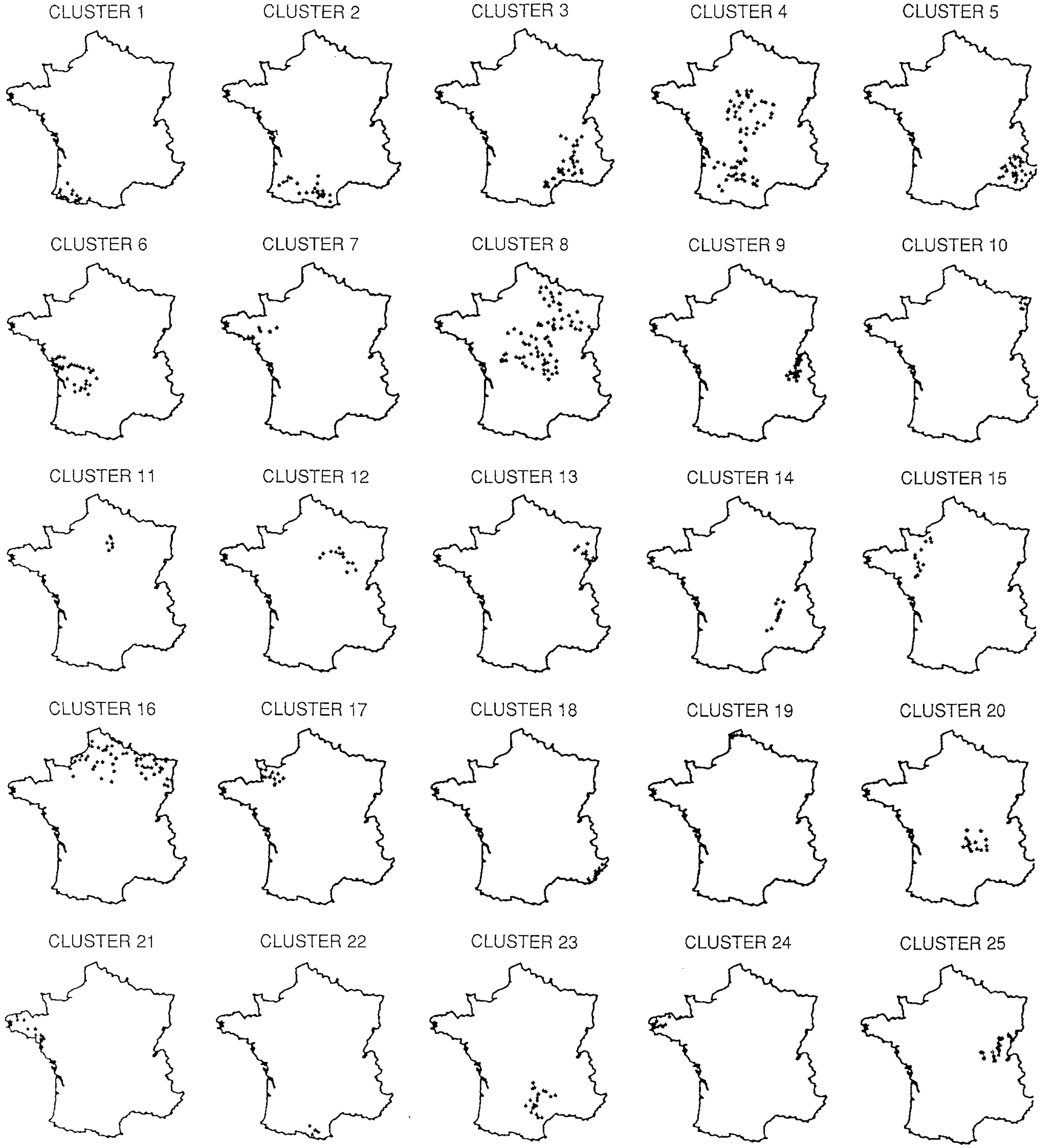


Fig. 2 Maps showing locations of 25 clusters obtained by hierarchical clustering with a contiguity constraint

A table showing the cluster means for 28 agronomic traits compared with overall means and the means of control cultivars is available on request as its size is too large to be published. This information can either be presented graphically as in Fig. 3 for some traits or summarized by multivariate techniques such as principal component analysis. In Fig. 4, the barycentres of the 25 clusters are plotted on the Cartesian graph of the first two principal components, which explain 40% of the total variance. The first component is positively correlated with heading date and negatively correlated with non-vernalized flowering ability, aftermath heading and frost susceptibility at the mountain evaluation site. The variance for component two can be explained by frost susceptibility at different sites (negative correlation), spring growth at Rodez and Bourg Lastic and most of the summer and autumn growth and persistence scores. This component could possibly be used as an alternative selection index (Godshalk and Timothy 1988) as it enables the rapid identification of the most promis-

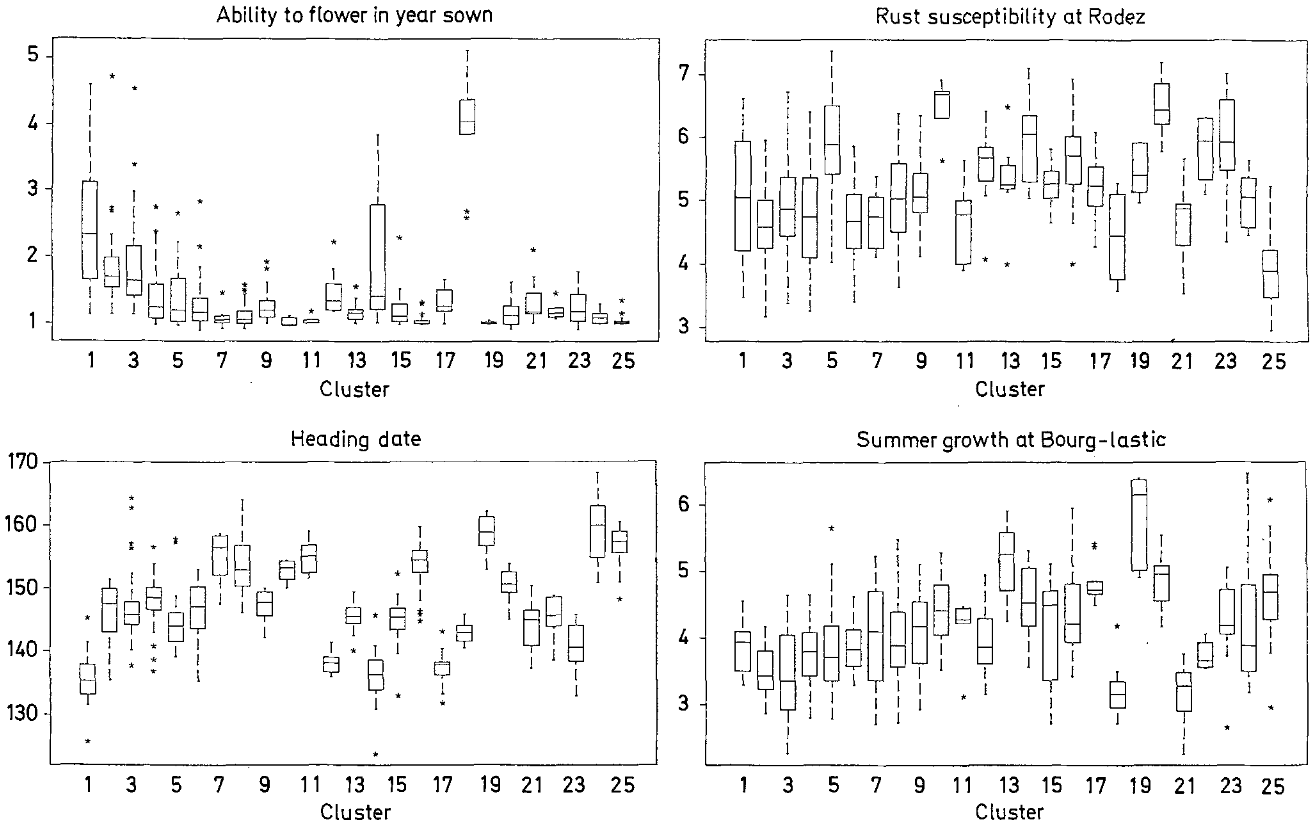


Fig. 3 Within and between cluster variation of some agronomic traits

ing clusters for use in breeding of new varieties. When only considering the late-flowering material (corresponding to cluster 10 in our previous study (Charmet et al. 1990) no fewer than 5 clusters have valuable populations: cluster 11, 16, 19, 24 and 25).

Discussion and Conclusion

The compactness of the clusters

The method presented here clearly succeeded in obtaining more geographically "compact" clusters than did the usual non-constrained method. It is important to note that this compactness cannot be obtained with the non-constrained method simply by requiring a higher number of classes.

In most cases (22 clusters out of 25), the size of each cluster is of the same magnitude as the scale of patchy variation of individual variates evidenced by MONESTIEZ et al. (1994). However there are two remaining weaknesses with the proposed classification. As contiguity is transmitted from place of place in a "stepping-stone" model, widespread clusters are obtained in some instances (cluster 4, 8 and to a lesser extent, cluster 16). Clearly, clusters 4 and 8 could have been divided if more clusters had been required, or if an upper limit on the size of a cluster had been imposed. We note that these "large clusters" group the less differentiated populations as shown by the central location of their barycentres in Fig. 4. It could, therefore, be somewhat artificial to obtain only compact clusters when a background of few differentiated (i.e. from the overall mean) widespread populations reflects the actual situation. Our only guideline for choosing the clustering level was the ratio between cluster variance and total variance, which was about 0.5.

Interpretation of contiguous clusters

The findings of this study are less easy to interpret than when clusters obtained "freely" are closely associated with a specific geographic factor, as in the case of barley landrace populations from the Near East (Weltzien 1989) or, to a lesser extent, the case of perennial ryegrass populations from Italy (Falcinelli et al. 1988). Another approach would be the use of multivariate techniques such as canonical analysis to confirm the usefulness of an a priori geographical grouping, for example by country of origin (Erskine et al. 1989). None of these approaches has proven to be valuable in this collection of perennial ryegrass populations from France. Keeping in mind that the formation of the clusters from contiguous populations is based on their similarity for agronomic traits, it is however remarkable that many cluster limits fit some macroclimatic borders extremely well: e.g. wet-mild southwest (cluster 1), Mediterranean and Rhone valley (cluster 3), French Riviera (cluster 18). Thus, a strong effect of natural selection caused by the macro-environment on a regional scale, especially by climatic factors, may be postulated to explain most of the 25 contiguous clusters. This is more evident for some clusters than others, for example, clusters 20 and 22, in which the sinuous border contours actually relate to a specific altitude.

The average size of cluster area may be related to the maximum distance of gene flow that leads to the homogenization of the genetic background of the populations (Sokal and Oden 1978).

The probable influence of natural selection on the agronomic traits of this ryegrass collection has been reported by Balfourier and Charmet (1991). Why this influence did not appear at the multivariate level on the location of non-

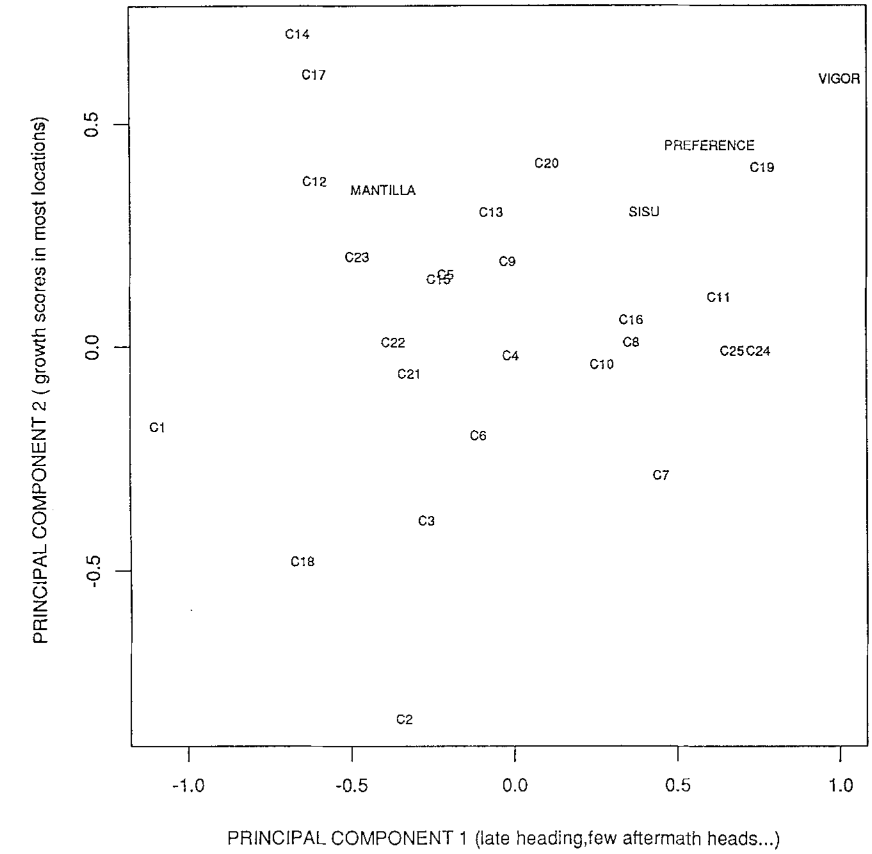


Fig. 4 Barycentres of the 25 contiguous clusters plotted on the graph with principal component 1 as the X axis and principal component 2 as the Y axis

constrained clusters may be explained by evolutive convergence leading to similar phenotypes amongst genetically differentiated populations from separate regions.

The influence of non-macroclimatic selective factors, such as the type of habitat, soil or grassland managent, cannot be excluded. It would be very interesting, for example, if such a factor discriminates between clusters 4 and 8 in central France. The chief influence of a regional macroclimatic factor would explain why these "micro-environment" factors show little significant relationship with agronomic traits when test-ed on the whole collection. Unfortunately we found few relationships between clusters and microenvironmental factors, with the notable exception of clusters 12 and 17, which group populations from grassland only, while populations from the other clusters come either from grassland or from unmanaged areas such as roadsides.

In conclusion, our method leads to clusters of populations with similar agronomic characteristics, which in most cases present a quite compact grouping and fit to known macroclimatic entities. This regional grouping was not obtained arbitrarily, but is based on the Euclidian distance from agronomic traits. Since the distance is computed from principal components that are extracted from a variance-covariance matrix, it provides an estimate of genetic divergence of populations for quantitatively inherited traits (Humphreys 1991). Moreover, the choice of the agronomic traits takes into account the genotype × environment interaction.

The geographic contiguity constraint helps to find areas inhabited by a unique cluster of populations, thus clearing the map of small-scale variation that can be considered to be background noise. Indeed, a larger number of classes has to be considered than when non-constrained clustering is used. This constraint may appear to be somewhat artificial. However, it is based on a geostatistical basis, the maximum distance for linking populations by the contiguity relationships being half of the range of the variogrammes for agronomics traits. This choice is consistent with the choice of between cluster variance over total variance of 0.5 for the clustering level. Indeed, this definition of contiguity is both wide (in comparison with the usual definition) and rigid, having a yes/no basis. Recently, Oliver and Webster (1989) proposed another more sophisticated method of spatially constrained classification where contiguity is replaced by a spatial weighting using the parameters of the variogrammes.

The use of this classification

The geographically constrained clustering succeeded in organizing a large, extensive collection of a widespread spontaneous forage species, perennial ryegrass, into a set number of clusters that are clearly identified by both their agronomic behaviour and their geographic location. As stated in the introduction, such a classification provides a useful basis for setting up a core collection. Genebank curators require that a

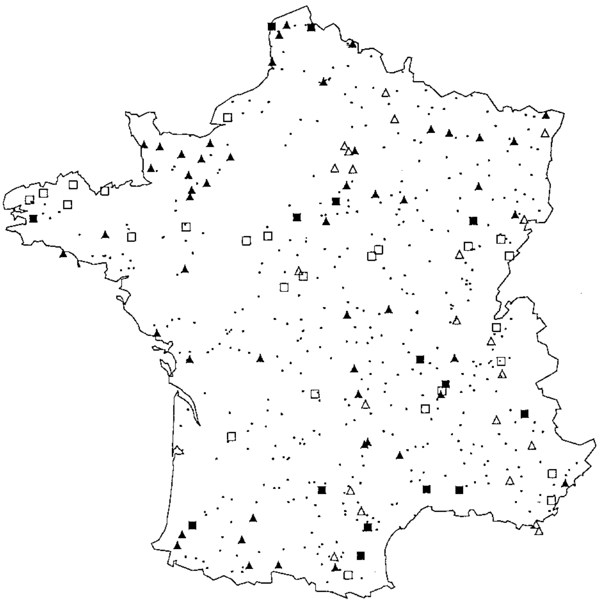


Fig. 5 Map of sites and habitats of the populations from the core collection. Habitat: □ roadside, △ path, ■ fallow, ▲ meadow

core collection is as representative as possible with regard to agronomic traits (conventional clustering) and in the case of wild species, to geographic origin. The proposed method makes core sampling easier as an adequate representation of every cluster will ensure that both of the above requirements are met.

In order to obtain a core of 20% from the whole collection (111 out of 547 populations), from 2 to 12 populations have been sampled from each cluster (depending upon its size) and multiplied in field isolation between 1988 and 1993 (Fig. 5). The list and passport data of populations of this core are available on request, and seeds will be freely distributed at the end of the regeneration scheme. Maps or graphic representations similar to Fig. 4 for all agronomic traits can also be obtained. An in-depth evaluation of this core collection, including yield and adaptation trials at several locations will be carried out, hopefully in the form of a European co-operative programme.

An electrophoretic survey of isozyme loci is being generated from the core populations (Charmet et al. 1993). The purpose of this survey is to demonstrate the structure of within and between population differentiation for neutral genes, and it would be particularly useful in elucidating the genetic basis of conti- guous clusters by computing Wright's statistics (Wright 1978).

Other uses of the core collection include the formation of bulk populations either for conservation purposes (Guy et al. 1990) or for genetic improvement by recurrent selection. Further results will involve both isozyme markers and quantitative genetics parameters.

Acknowledgements This work was supported by a grant no. 1344A from the French Ministry of Scientific Research and Technology.

References

Balfourier F, Charmet G, (1991) Relationships between agronomic characters and ecogeographical factors in a collection of french perennial ryegrass populations. Agronomie 11:645–657

Brown AHD (1989) The case for core collection. In: Brown AHD, Frankel OH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, UK, pp 136–156

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

Charmet G, Balfourier F, Ravel C (1993) Isozyme polymorphism and goegraphic differentiation in a collection of French perennial ryegrass populations. Genet Resources Crop Evol (in press)

Erskine W, Adham Y, Holly L (1989) Geographic distribution of variation in quantitative traits in a world lentil collection. Euphytica 43:97–103

Falcinelli M, Veronesi F, Lorenzetti S (1988) Evaluation of an Italian germplasm collection of *Lolium perenne* L. through a multivariate approach. In: Poisson C (ed) Natural variation and breeding for adaptation. Proc EUCARPIA Fodder Crops Sect Meet. Lusignan, France, pp 23–35

Felsenstein J (1983) Numerical taxonomy. Springer, Berlin Heidelberg New York Tokyo

Fisher MM (1978) Zur Lösung funktionaler regionaltaxonomischer Probleme auf der Basis von Interaktionmatrizen: ein neuer graphentheoretischer Ansatz. Karlsruher Manuskr Math Theor Wirtsch. Sozialgeogr no. 25

Gabriel KR, Sokal RR (1969) A new statistical approach to geographic variation analysis. Syst Zool 18:259–270

Goledshalk EB, Timothy DH (1988) Factor and principal component analyses as alternatives to index selection. Theor Appl. Genet 76:352–360

Guy P, Ghesquiere M, Charmet G, Prosperi JH (1990) Pooling accessions. Advantages and disadvantages. In: Report working group on forages. IBPGR, Rome, pp 35–49

Humphreys MO (1991) A genetic approach to the multivariate differentiation of perennial ryegrass (*Lolium perenne* L.) populations. Heredity 66:437–443

Lebart L (1978) Programme d'agrégation avec contraintes (C.A.H. contiguité). Cah Anal Données 3:275–287

Lefkovitch LP (1980) Conditional clustering. Biometrics 36:43–58

Mandel J (1971) A new analysis of variance model for non-additive data. Technometrics 13:1–18

Monestiez P (1978) Présentation de deux méthodes utilisant la notion de contiguité pour l'analyse des données géographiques. PhD thesis, University of Pierre et Marie Curie, Paris

Monestiez P, Goulard M, Charmet G (1994) Geostatistics for spatial genetic structures: study of wild populations of perennial ryegrass. Theor Appl Genet (in press)

Oliver MA, Webster R (1989) A geostatistical basis for spatial weighting in multivariate classification. Math Geol 21:19–35

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

Peeters JR, Wilkes HG, Galwey NW (1990) The use of ecogeographical data in the exploitation of variation from gene banks. Theor Appl Genet 80:110–112

Perruchet C (1979) Classification sous contrainte de contiguité continue: application aux sciences de la terre. Thèse 3ème cycle, University of Pierre et Marie Curie, Paris

Sokal PR, Oden NL (1978) Spatial autocorrelaation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. Biol J Linn Soc 10:229–249

Tyler BF (1987) Collection, characterization and utilization of genetic resources of temperate grass and clover. IBPGR Training Courses, Lecture Ser1, p 66. IBPGR, Rome, Italy

Weltzien E (1989) Differentiation among barley landrace populations from the Near East. Euphytica 43:29–39

Wright S (1978) Evolution and the genetics of populations, vol 4: variability within and among natural populations. University of Chicago Press, Chicago, USA

Yonezawa K (1985) A definition of the optimal allocation of effort in conservation of plant genetic resources with application to sample size determination for field collection. Euphytica 34:345–354