STATISTIQUE ET ANALYSE DES DONNÉES

J. P. BRIANE
J. LACOURT
R. SALANON

Analyse des données volumineuses, application à la phytosociologie

Statistique et analyse des données, tome 2, n° 3 (1977), p. 15-23
<http://www.numdam.org/item?id=SAD_1977__2_3_15_0>

© Association pour la statistique et ses utilisations, 1977, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Statistique et Analyse des Données
3 - 1977

analyse des données volumineuses application à la phytosociologie

B R I A N E J.P.* , L A C O U R T J.*, S A L A N O N R.**

*Laboratoire de Taxonomie Végétale, Expérimentale et Numérique, Université de Paris Sud, F9I405 ORSAY

** Ecologie et Phytosociologie, UER Domaine Méditerranéen, Campus Universitaire Valrose, F 06034 NICE CEDEX

I. Méthodologie.

Introduction

Depuis longtemps les phytosociologues ont accumulé une grande masse d'information. Ces informations se présentent sous la forme de tableaux de releves de végetation. Un releve effectué sur le terrain consiste essentiellement en une liste d'espèces rencontrées, ainsi que d'autres renseignements de nature topographique édaphique . Le phytosociologue s'interresse surtout aux associations de plantes (indépendamment des relations au milieu ,au moins dans un premier stade). La donnee de base est un tableau de presence-absence constitue comme suit. Pour chaque relevé j et chaque espece i, k_{ij} = 1 ou 0 selon

que le relevé contient ou non l'espece. Il est desormais courant de soumettre ce type de tableau à l'analyse factorielle des correspondances. Cependant le besoin de synthese se fait sentir actuellement(au niveau de la systématique des associations végétales notamment). C'est ce besoin qui est à l'origine de la création de banques de donnees phyto-ecologiques. Aux problèmes propres à la centralisation des donnees s'ajoute celui du traitement de tableaux volumineux. L'objet de cette communication est de présenter une methode particulière, qui présente des avantages certains dans le domaine des synthèses phytosociologiques.

Analyse de données groupées.

La méthode mise au point comprend deux temps:

- D'abord une phase de réduction des données. A l'aide d'un programme de classification automatique on regroupe les relevés en un nombre fixé de classes.
- Dans une seconde phase on procède à l'analyse de données groupées (ADG par la suite).On fait l'analyse factorielle des correspondances (AFC) du tableau de contingence espècesxclasses de relevés(k_{iq} = nombre de relevés de la classe q contenant l'espèce i). Enfin on projette en éléments supplémentaires les relevés sur les axes du nuage des classes.

Nous avons utilisé deux programmes de classification automatique recherchant des partitions en un nombre fixé de classes. Un premier fondé sur l'algorithme des centres mobiles (cf. I), les écarts entre relevés et classes étant calculés avec la métrique du χ^2. Un second programme fondé sur l'algorithme d'échange (cf. I),avec comme critère optimisé le moment centré d'ordre deux de la partition (la métrique étant toujours celle du χ^2).

Avantages de la méthode.

Les résultats des programmes de classification (contenu en effectif de chaque classe , contribution de chaque espece à l'individualisation d'une classe (cf I)) sont deja pleins

d'enseignements. La plupart des regroupements précisant les unités "synsystématiques" ou écologiques. Cependant la partie qui nous interesse le plus et qui a apporté le plus au phytosociologue est l'ADG. Nous avons constaté un un certain nombre d'avantages.

- L'ADG a pour effet principal de "régulariser" les données. Pour des essais comparatifs entre ADG et AFC se reporter à une précédente publication(III).Pour expliquer ce point il nous faut signaler quelques particularités des données phytosociologiques. Tout d'abord le nombre des espèces est souvent du même ordre que celui des relevés, comme on pourra s'en rendre compte dans les applications. Beaucoup d'espèces ne sont présentes que dans un petit nombre de relevés , ces dernières contribuant à "brouiller" une analyse sans regroupement. On ne peut pas supprimer à priori ces espèces (comme cela s'est pratiqué)du tableau. Certaines espèces rares ont tout de même une signification importante. Il est fréquent dans une AFC qu'un premier axe soit déterminé par quelques relevés aberrants (appauvris ou au contraire riches en espèces rares). Ceci est d'autant plus fâcheux que les calculs sontlongs. Sur les diagrammes-espèces l'effet produit est semblable à celui du dédoublement des relevés (cf II). Les espèces importantes (de moyennes fréquences) ressortent à la périphérie du nuage, les autres rentrant dans le rang.

- Au rang des avantages de la méthode ,portons le faible côut marginal des analyses partielles. On a la possibilité de supprimer trés facilement un groupe de relevés globalement ou en partie. Ceci à deux niveaux , soit en recalculant les axes du sous-nuage, soit en projetant simplement ce dernier sur les axes du nuage global. L'ADG est essentiellement destinée à mettre en évidence le squelette général de données. Quand le nombre de relevés d'une analyse partielle descend à moins de 200 ,on revient à l'AFC sans regroupement (éventuellement avec dédoublement).

Conclusions.

Il reste à résoudre un certain nombre de problèmes. Tout d'abord au point de vue du temps de calcul , la phase de réduction des données est de loin la plus coûteuse. Avec les programmes que nous avons utilisés il faut de 5 à 10 minutes pour classifier environ 1000 relevés (IBM 370 et UNIVAC 1110). La capacité en nombre de relevés étant limitée à ce chiffre, compte tenu du grand nombre d'espèces. Maintenant que la méthode est rôdée c'est ce point que nous allons améliorer.

L'interprétation des cartes-espèces si elle est plus facile qu'en AFC, demeure délicate. Dans l'une des applications présentées(groupements des cultures),nous avons fait une classification sur les espèces à partir du tableau espèces × groupes de relevés. La partition en 10 classes obtenue, a beaucoup facilité l'exploitation des résultats de l'ADG.

II. Applications phytosociologiques

A. Etude synthétique de la végétation commensale des cultures d'Europe .

Introduction :

Pour réaliser cette synthèse, nous avons largement fait appel à des relevés publiés dans diverses revues européennes (l'origine des différents relevés de la bibliographie se trouve en annexe de notre thèse de spécialité- Lacourt 1977) ainsi qu'à 171 relevés personnels effectués principalement en région parisienne, dans le Berry et les Alpes-maritimes .

Nous avons ainsi étudié un ensemble de 2749 relevés comportant 694 espèces provenant de presque tous les pays d'Europe .

La première phase de la méthode employée consiste , nous l'avons vu précédemment, à réduire l'ensemble des données en

un petit nombre de groupes à l'aide d'un programme de classification automatique .Etant donné le volume des données, nous les avons au préalable réparties en 6 lots sur des bases essentiellement géographiques donc plus ou moins arbitraires. Chacun de ces 6 lots a été découpé en un nombre arbitraire de classes (5 en général) par classification automatique . Nous avons ainsi obtenu un tableau comportant 694 lignes (espèces) et 29 colonnes (groupes) qui a servi de base à l'ADG .

Principales conclusions pouvant être tirées de cette analyse

Nous n'étudierons ici que la carte des relevés en fonction des espèces suivant les axes (1) et (2), celle-ci étant reproduite très schématiquement sur les figures 1 et 2 . Mais il est bien évident que les principales conclusions résumées ci-dessous n'ont pu être révélées que par la comparaison des cartes de relevés suivant les axes (1) et (2) d'une part et (2) et (3) d'autre part ; cette dernière carte n'étant pas reproduite faute de place .

Quatre remarques et conclusions principales peuvent être retenues :

- a) Signification des axes factoriels :

Si dans un premier temps l'on repère à la fois l'origine géographique et le type de culture pour chacun des 2749 relevés on constate deux faits très significatifs :

- L'axe (1) représente indiscutablement un axe de température . En effet, en abcisses négatives se trouvent des relevés provenant des pays du Nord et du Nord-Ouest de l'Europe exclusivement . Par contre, plus l'abcisse augmente et plus les relevés correspondants proviennent soit de pays méditerranéens, soit de pays continentaux .(cf. fig.1). Il y a donc une opposition très nette entre pays à étés ou fin de printemps frais en abcisses négatives et pays à étés très chauds en abcisses positives .
- L'axe (2) représente un axe de type decultures. En effet, en ordonnées positives, c'est-à-dire au-dessus de l'axe (1), se retrouvent tous les relevés de moissons, tandis qu'en

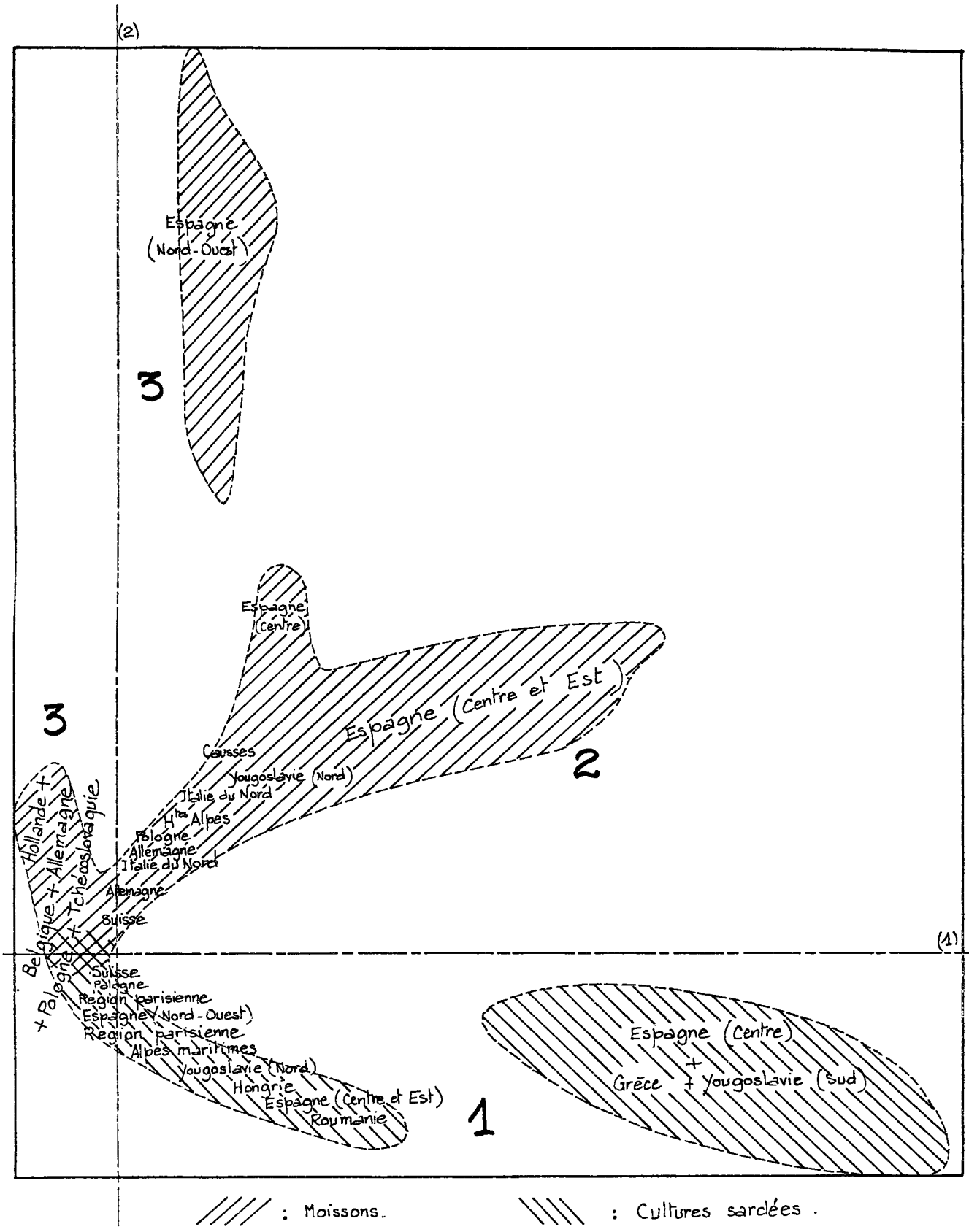

Fig. 1 - Représentation schématique de la carte des relevés suivant les axes (1) et (2), où sont figurés l'origine géographique et le type de culture

ordonnées négatives (au-dessous de l'axe (1)) ce ne sont que des relevés de cultures sarclées ou de vignes (fig.1) .

- b) L'étude de la forme générale du nuage de points (de relevés) permet de montrer qu'à partir d'un noyau central, correspondant à un mélange de relevés de moissons et de cultures sarclées du nord et du nord-ouest de l'Europe, s'individualise 3 groupements suivant 3 directions .

- La direction 1 correspond aux groupements de mauvaises herbes des cultures sarclées méditerranéennes et continentales mais aussi, à sa base, des cultures sarclées les plus thermophiles du nord de l'Europe sur sol sableux .

- La direction 2 correspond aux moissons des régions méditerranéennes et steppiques mais comprend aussi les moissons médio-européennes sur sol calcaire .

- La direction 3 enfin,représente l'ensemble des associations messicoles, calcifuges et acidiphiles principalement du domaine atlantique .

- c) Les discontinuités observées dans le nuage de points, en particulier dans la direction 3, correspondent à des manques de données .Ainsi,pour la végétation messicole acidiphile du domaine atlantique, nous avons un manque de données très important qui concerne l'ouest et le sud-ouest de la France et qui se traduit sur le diagramme par une discontinuité importante .Nous pouvons donc prévoir théoriquement la composition floristique des associations végétales des régions non inventoriées .Ceci est bien entendu lié à la chorologie (répartition géographique) des divers groupements végétaux et à celle des espèces les constituant .

- d) Enfin, d'un point de vue purement phytosociologique (c'est-à-dire exclusivement systématique) cette analyse synthétique a permis de remettre en question toutes les classifications existantes et d'établir une liste de synonymes souvent très importante (qui ne figure que très partiellement sur la fig.2) pour chacun des 4 syntaxons (groupements végétaux) mis en évidence:

- "noyau central" du nuage = Polygono-Chenopodion polyspermi .
- direction 1 = Diplotaxion .
- direction 2 = Secalion .
- direction 3 = Scleranthion annuae .

B. Etude synthétique des "pineraies".

Le compte-rendu de cette application, seule présentée au colloque, n'a pu parvenir dans les délais trés courts qui nous étaient impartis. Nous renvoyons le lecteur interressé par le sujet, à une publication antérieure détaillée (cf.III).

Bibliographie.

I. BENZECRI J.P. 1973 : L'analyse des données.
2 volumes, 1236 p., Dunod, Paris.

II. BRIANE J.P., LAZARE J.J., ROUX G., SASTRE C. 1974 :
L'analyse factorielle des correspondances et l'arbre de longueur minimum; exemples d'application.
Adansonia, Sér.2, 14 (1). pp. 111-137.

III. BRIANE J.P., LAZARE J.J., SALANON R. 1977 :
Le traitement des tres grands ensembles de données en analyse factorielle des correspondances- Proposition d'une méthodologie appliquée à la phytosociologie.
Département de Mathématiques, Université de Nice. 38p.

IV. LACOURT J. 1977 : Essai de synthese sur les syntaxons commensaux des cultures d'Europe.
Thèse de 3° cycle, Université Paris-Sud, Orsay. 249 p.

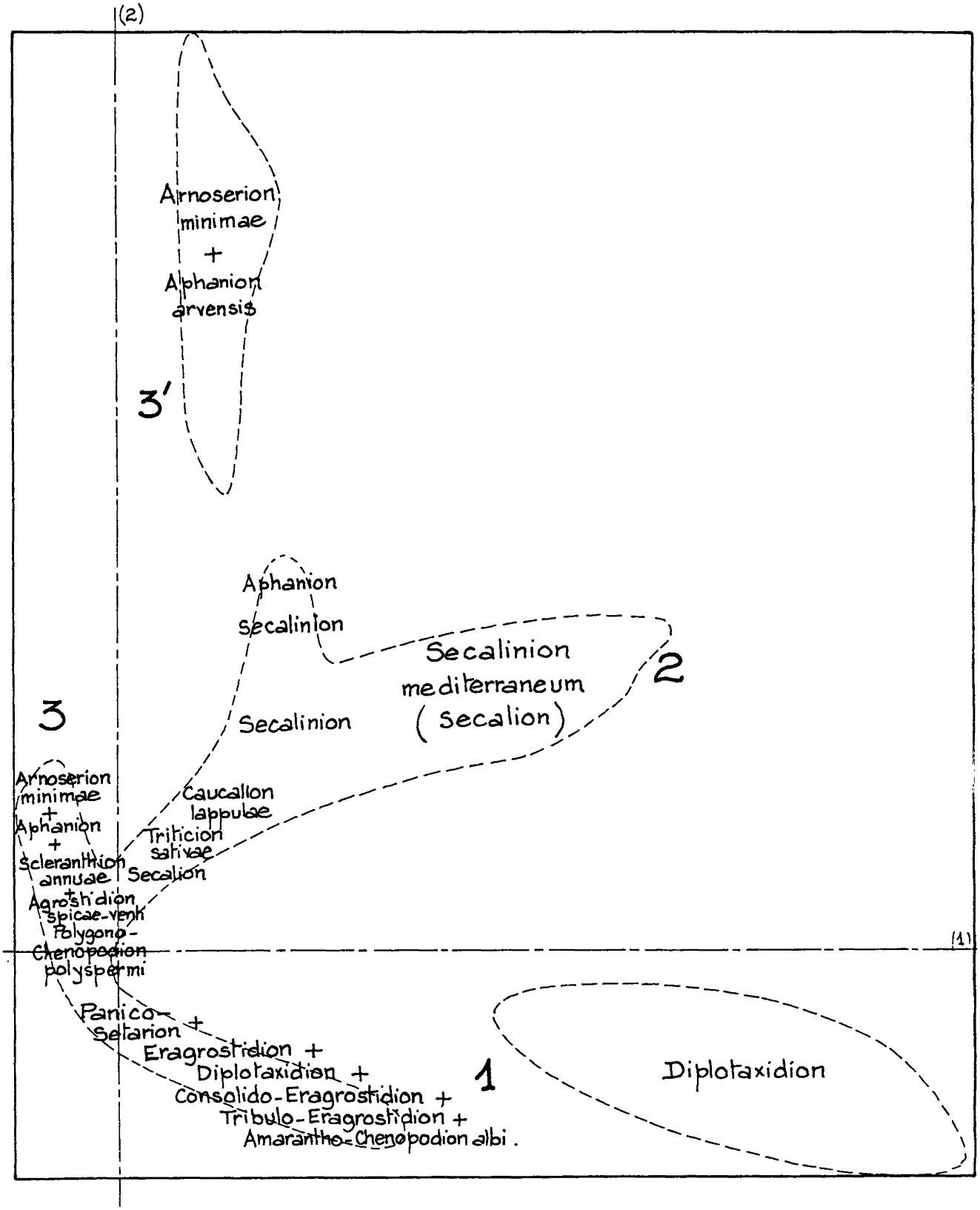

Fig. 2 - Représentation schématique de la carte des relevés (fig. 5) où sont figurées les différentes alliances données par les auteurs des relevés.