Pertinence du découpage spatial produit par deux méthodes de classification (CHA et MIXMOD) ; application aux climats français

Daniel Joly, Florent Langrognet

Cybergeo : European Journal of Geography
Cartographie, Imagerie, SIG | 2016

Pertinence du découpage spatial produit par deux méthodes de classification (CHA et MIXMOD). Application aux climats français

Classifications and spatial segmentation relevance using two methods applied to the French climates

Daniel Joly et Florent Langrognet

Édition électronique
URL : http://journals.openedition.org/cybergeo/27414
DOI : 10.4000/cybergeo.27414
ISSN : 1278-3366

Éditeur
UMR 8504 Géographie-cités

Daniel Joly et Florent Langrognet

Pertinence du découpage spatial produit par deux méthodes de classification (CHA et MIXMOD). Application aux climats français

Devant la complexité du réel, la tentation est grande de le réduire aux tendances qui le structurent. Cette démarche modélisatrice permet d'extraire de l'information pertinente et de porter un jugement critique sur les partitions qui en découlent. La classification des climats, qui répond à cet objectif, est d'une grande utilité pour de nombreux domaines d'activité ou de recherche (tourisme, maîtrise de l'énergie, écologie…). De multiples solutions ont été proposées pour répondre à ces besoins. Par exemple, les agronomes, ayant besoin de mesurer l'importance des stress hydriques auxquels sont soumises les plantes, ont construit des modèles plus ou moins compliqués en ayant recours, entre autres, à la pluviométrie et à la température : il en découle plusieurs classifications des climats selon des indices, tels ceux de Gaussen, de De Martonne, de Köppen, etc. Ces classifications sont déductives en ce sens qu'elles obéissent à des règles de construction rigides et prédéfinies : le climat polaire de Köppen caractérise ainsi tous les lieux dont la température moyenne du mois le plus chaud est inférieure à 10 °C.

Un autre type de classification, commun en climatologie, repose sur le principe de l'induction, *« opération qui consiste à remonter, à partir de données singulières, à une information plus générale impliquant toutes les données initiales »* (Wikipedia). L'induction est probabiliste en ce sens que, à partir de faits observés, elle permet d'établir une loi générale à condition que d'autres faits ne viennent pas invalider le passage du particulier au général. Ainsi, la classification des climats français décrite dans l'article intitulé « Les types de climats en France, une construction spatiale » (Joly *et al.*, 2010) qui, dans la suite de cet article sera dénommé « Cybergéo1 », fournit une carte en tous points conforme, mais pas identique, à celle des climats de la France obtenue d'autres classifications effectuées sur le même espace (Arlery, 1979 ; Bessemoulin, 1987 ; Champeaux et Tamburini, 1994).

Toutefois, et même s'ils ne remettent pas en cause la qualité de la classification, il est apparu certains défauts qui méritaient d'être corrigés. Par exemple, la dispersion de pixels dans un ensemble appartenant majoritairement à un autre type peut poser problème. Lorsque cette imbrication concerne le passage d'une classe à l'autre, ce n'est pas gênant car ce processus se déroule à l'intérieur d'une petite frange de transition à l'intérieur de laquelle l'appartenance à l'une ou l'autre classe est hésitante. Cette particularité (discutée dans le paragraphe « *Incertitude des limites : apport de l'approche probabiliste* » de *Cybergéo1*) est normale car elle est due au fait que le paysage est composite et que les gradients de chacune des variables sur lesquelles porte la classification sont dissemblables (Durand-Dastès et Sanders, 1984). Ainsi, entre tel espace appartenant de manière homogène à une classe et tel autre appartenant, lui aussi de manière homogène, mais à une classe voisine, il existe une grande variété de lieux aux configurations climatiques voisines et, dans ces conditions, il faut peu pour que l'on bascule dans une classe plutôt que dans l'autre.

Cette incertitude pourrait être plus gênante lorsqu'un type est enclavé dans un autre. L'article *Cybergéo1* signale deux types (n° 7, le climat du bassin du Sud-Ouest, et le n° 6, le climat méditerranéen altéré) qui entraînent de telles apparentes contradictions. Le paragraphe « *La carte typologique* » explique ce phénomène par le fait que « *les espaces considérés, bien qu'éloignés l'un de l'autre, présentent les mêmes caractères climatiques. Si proximité spatiale est souvent synonyme de proximité statistique, deux points éloignés peuvent, parfois, présenter des ressemblances* ». Mais là encore, cette juxtaposition, dans le même espace du climat du bassin du Sud-Ouest et du méditerranéen altéré ne pose pas problème car ce sont en fait deux faciès assez voisins des climats du Midi. Dans une tonalité plus gênante, l'appartenance de certaines crêtes de l'Oisans au type 5 (climat océanique franc) est contestable d'un point de vue climatologique, même si cela s'explique par les raisons qui viennent d'être mentionnées, auxquelles il faut ajouter le fait que l'on se situe ici à l'articulation entre climat méditerranéen et climat de montagne. Aussi, les parties élevées de ce massif alpin ont en commun, avec la frange océanique, certains caractères climatiques tels que des températures moyennes et estivales fraîches, peu de jours chauds, des températures interannuelles stables, des précipitations interannuelles variables et des cumuls de précipitation assez élevés. Les valeurs de certaines autres variables sont différentes et introduisent des différences entre les deux espaces considérés, mais leur poids n'arrive pas à contrebalancer les ressemblances, de sorte que les pixels sont un peu plus proches du centre de gravité du type 5 que des types 1 (climats de montagne), 2 (climat des marges montagnardes), 4 (climat océanique altéré) ou 6 (méditerranéen altéré) qui, tous, sont également présents en différents endroits (versants, vallées) du Briançonnais et de l'Oisans, ce qui montre la complexité climatique de cet espace composite.

Une raison qui explique la juxtaposition, sur le même ensemble, de types aussi disparates provient sans doute aussi de la rigidité de l'algorithme de la classification hiérarchique ascendante (CHA) qui a été utilisé dans *Cybergéo1*. Cet algorithme, qui a recours à la distance euclidienne, présente l'inconvénient d'être beaucoup moins souple que ceux fondés sur les modèles de mélanges (Mclachlan, Peel, 2000). Par analogie, c'est comme si, pour calculer la distance entre deux villes, on se basait sur la distance « à vol d'oiseau » plutôt que de recourir à un algorithme permettant de sélectionner le plus court chemin en suivant la sinuosité des parcours. Aussi, l'intérêt de comparer les résultats fournis pas ces deux types de modèles tient-il dans la nature des individus à classer (pixels de la France continentale) structurés par de solides lois d'organisation spatiale.

Nous avons donc eu la curiosité de comparer les résultats obtenus par la CHA aux résultats apportés par l'outil logiciel MIXMOD (annexe 1) utilisant les modèles de mélange et proposant un large spectre de modèles et de critères (Biernacki, Celeux, Govaert, Langrognet, 2006 ; Langrognet, 2009). Le jeu de données à traiter, celui qui a été utilisé dans *Cybergéo1,* est identique dans les deux cas. Il s'agit de comparer les caractéristiques spatiales (extension, localisation) des classes fournies par la CHA et par MIXMOD, sachant que ces deux méthodes classent les pixels selon deux modèles distincts.

Données et méthodes

Les données

Du fait que les données à traiter par l'une et l'autre méthode de classification sont identiques à celles de *Cybergéo1*, seule une rapide présentation sera faite ici, le détail pourra être consulté dans l'article source. Les données sont enregistrées par Météo-France en 651 (température) et 2031

(pluviométrie) stations sur la normale 1971-2000. Elles sont constituées de deux paramètres (précipitations et températures), chacun décliné selon respectivement 6 et 8 variables.

Une première étape a consisté à interpoler les données stationnelles, variable par variable de manière à produire une information qui couvre tout l'espace. La méthode d'interpolation utilisée enchaîne deux phases d'analyse : d'abord des régressions à échelle locale suivies d'un krigeage[1] des résidus (Joly *et al.*, 2009 ; 2010, 2011, 2012). Les variables explicatives mobilisées pour mettre en œuvre les régressions sont le MNT de l'IGN à 250 m de résolution et la base européenne Corine Land Cover. De la première sont dérivées les six variables liées à la topographie (altitude, pente, orientation des versants, rugosité topographique, indice d'encaissement/surélévation topographique, rayonnement global théorique) et de la seconde trois variables liées à l'occupation du sol (indice de végétation, distance à la forêt et distance à l'océan ou à la mer). Au final, nous disposons des 14 images suivantes à la résolution de 250 m, chacune d'entre elles étant composée de 8 681 705 pixels considérés dans les analyses à venir comme individus statistiques à classer :

- 1 – Moyenne annuelle de température.
- 2 / 3 – Nombre de jours avec une température inférieure à -5°C / supérieure à 30°C.
- 4 – Amplitude thermique annuelle.
- 5 / 6 – Variabilité interannuelle de la température en janvier / en juillet.
- 7 – Cumuls annuels de précipitation.
- 8 / 9 – Écart des cumuls de janvier / de juillet par rapport à la moyenne annuelle des cumuls mensuels.
- 10/ 11 – Nombre de jours de précipitation en janvier / en juillet.
- 12 / 13 – Variabilité interannuelle des précipitations en janvier / en juillet.
- 14 – Rapport entre les abats d'automne (septembre + octobre) et ceux de juillet.

Méthode (Classification des données, métriques)

Le lecteur trouvera en annexe 2 une présentation plus complète des méthodes de statistiques utilisées dans cet article. Nous n'entrerons pas ici dans les détails mathématiques et donnerons une description des principales différences avec d'autres méthodes (dont la CHA). Ces méthodes, qui s'appuient sur les modèles de mélanges, sont disponibles dans MIXMOD, ensemble logiciel, diffusé librement (www.mixmod.org) depuis une quinzaine d4années.

Rappels sur la classification des données

La classification des données consiste à créer un regroupement en classes « homogènes » d'un ensemble d'individus. À partir de cette définition, se posent des questions importantes, notamment sur :

- le nombre de classes qui est souvent inconnu.
- le caractère homogène des classes : comment définit-on l'homogénéité ?
- la qualité d'une classification : sur quels critères peut-on affirmer qu'une classification est meilleure qu'une autre ?

Il existe de nombreuses méthodes de classification et l'objectif n'est pas ici de les lister toutes ni de présenter les avantages et inconvénients de chacune d'elles. Il s'agit plutôt de mettre à l'épreuve, sur le jeu de données climatiques disponibles, celle qui s'appuie sur des concepts probabilistes (les

[1] Le krigeage est une méthode d'estimation linéaire d'une variable généralisée (ici les résidus). Le krigeage est le plus souvent utilisé pour l'interpolation spatiale (pour plus de précisions, consulter Wikipedia : https://fr.wikipedia.org/wiki/Krigeage)

modèles de mélanges), et d'analyser les éventuelles plus-values qu'elle peut apporter par rapport à la CHA.

Tout d'abord, il convient de distinguer deux grandes catégories de méthodes de classification :

- Les méthodes qui ne cherchent qu'à extraire une classification à partir du jeu de données c'est-à-dire une affectation de chaque individu à une classe. La méthode CHA entre dans cette catégorie.
- Les méthodes qui cherchent à modéliser le phénomène qui a conduit à produire ces données. Il s'agit non seulement d'affecter les individus aux classes, mais également de caractériser ces classes. L'information produite est alors plus riche puisqu'on peut, par exemple, extraire un individu type par classe ou avoir une information sur la dispersion autour de cet individu type.

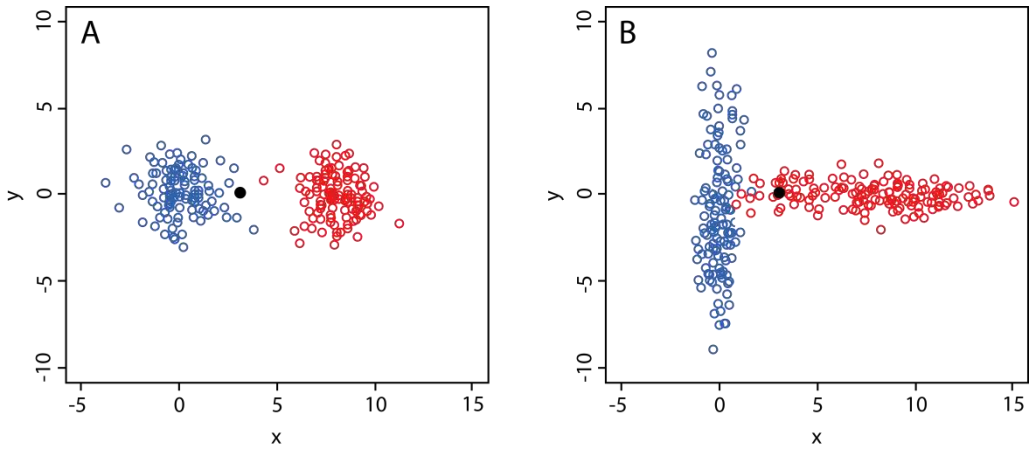
La distance euclidienne

La recherche d'une bonne partition passe par la création de classes homogènes, c'est-à-dire dont les individus sont proches. D'un point de vue mathématique, cela revient à minimiser l'inertie intra-classe : il apparaît clairement que la métrique utilisée joue un rôle majeur dans le calcul de l'inertie intra-classe, et ce faisant, dans la recherche de la meilleure classification. D'autre part, toute méthode de classification des données passe par une estimation de la distance entre deux individus. Cette distance est utilisée pour créer des classes homogènes dans lesquelles les individus sont « proches ». La définition de la distance entre deux individus joue alors un rôle crucial.

La plupart des méthodes de classification recourent à une distance euclidienne et donc isotrope. En d'autres termes, on suppose que la dispersion des individus à l'intérieur d'une classe est la même selon chaque variable (donc dans toutes les directions) et qu'elle est la même pour toutes les classes. Or, cette hypothèse, trop rigide et non vérifiée en pratique, peut conduire à des résultats erronés (la partition est non optimale). Pour s'en convaincre, il suffit de considérer d'une part un jeu de données avec deux classes dont la dispersion est effectivement isotrope (figure 1A) et un autre jeu de données avec deux classes de dispersion non isotrope (figure 1B). Ainsi, en utilisant une distance isotrope, le point X de coordonnées (3,0) sera classé dans le groupe bleu dans les deux situations (parce que plus proche du centre) alors que dans la deuxième il appartient, à l'évidence, au groupe rouge. Bien évidemment, il existe des techniques de prétraitement des données qui visent à éviter cette situation, mais elles ne sont pas toujours possibles ou efficaces, en particulier lorsque la non-isotropie est plus complexe (par exemple non parallèles aux axes).

Rappelons que c'est cette métrique qui a été utilisée dans la classification de *Cybergéo1* dont voici la démarche rapidement résumée. Les 14 variables sont discrétisées en cinq modalités puis soumises à une AFC qui, en générant des axes factoriels, donne lieu à un nouveau système de coordonnées dans lequel se positionnent individus et caractères. La matrice de distance ainsi construite sert de base à la CHA, méthode dont le principe est de regrouper itérativement les individus deux à deux selon un critère de distance euclidienne (parmi les quatre choix possibles, le critère de Ward a été retenu car il induit, à chaque étape de regroupement, une minimisation de la décroissance de la variance interclasse). Le barycentre des couples constitue, à chaque étape un nouvel individu qui est comparé à tous les autres selon la même métrique. Ainsi, les classes se constituent pas à pas, regroupant à chaque itération un nombre de plus en plus élevé d'individus. Le processus s'arrête lorsque tous les individus sont regroupés au sein d'une seule classe. L'arbre obtenu permet de choisir le nombre de classes.

Figure 1 : Données avec deux classes de dispersion isotrope (A) et non isotrope (B)



Les modèles de mélanges

Les modèles de mélanges sont des outils statistiques très utilisés et appréciés pour leur souplesse car ils permettent de modéliser un très large spectre de situations, y compris les plus complexes. En classification des données, ils autorisent le traitement des données multivariées, qualitatives, quantitatives (ou mixtes). Dans ce cadre, on considère que les individus d'une classe C_k sont des réalisations d'une loi de probabilité (de densité notée h_k), et que chaque classe a une probabilité p_k d'apparaître (il s'agit de la proportion de la classe C_k parmi toutes les classes). Dans ce contexte, la distance de Mahalanobis (Mahalanobis, 1936) est utilisée. En prenant en compte la variance et la corrélation des données, elle diffère de la distance euclidienne pour laquelle toutes les variables sont traitées indépendamment et de la même façon. On se dote ainsi d'un outil plus souple et plus précis pour mesurer la distance entre des individus.

La classification par modèles de mélanges de lois de probabilité vise à estimer le modèle sous-jacent, c'est-à-dire les paramètres des lois de probabilités et les proportions de chaque classe. On calcule ainsi la probabilité P_{ik} qu'un individu X_i appartienne à la classe C_k et on lui affecte la classe C_l pour laquelle P_{il} est maximale. Comme leur nom l'indique, les modèles de mélange consistent à recourir à plusieurs lois de probabilité pour modéliser la complexité. Parmi les lois de probabilité auxquelles il est possible de recourir pour classer des individus, les lois de distribution gaussienne (ou normales) sont les plus couramment utilisées pour traiter des données quantitatives. Pour le traitement de données qualitatives, on utilise habituellement des lois de distribution multinomiale.

Les modèles de mélanges gaussiens

Les lois de distribution gaussienne permettent de modéliser les données quantitatives y compris (et surtout) lorsque la distance entre les individus n'est pas isotrope. La classification des données dans ce cadre revient donc à modéliser :

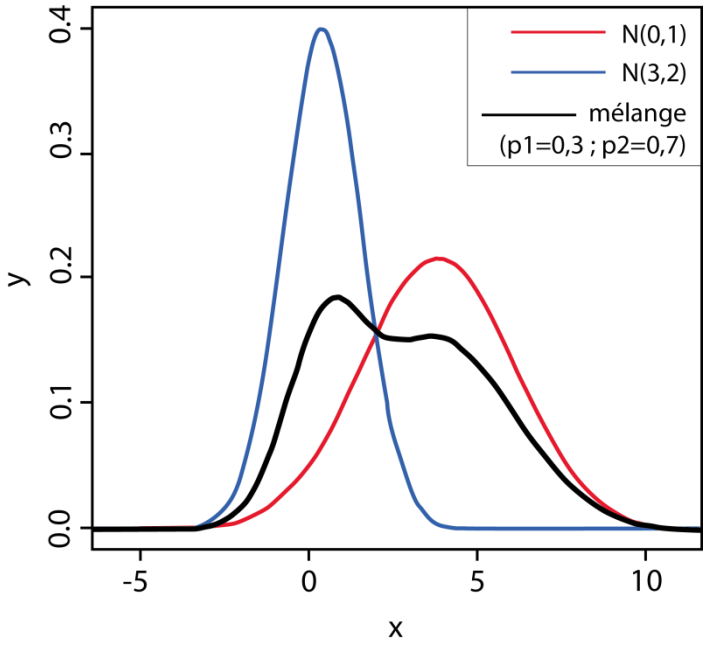
- les paramètres de chaque loi gaussienne (une par classe) : moyenne et variance.
- les proportions de chaque loi.

En dimension 1, pour deux classes, il s'agit donc d'estimer :

- la moyenne et la variance des deux classes (soit deux valeurs par classe).
- la proportion de chacune des classes (K-1 valeurs où K est le nombre de classes). La somme des proportions est égale à 1.

La figure 2 représente les fonctions de densité de deux lois gaussiennes (ayant pour moyenne respectivement 0 et 3 et pour variance 1 et 2) et du mélange de ces deux lois (de proportion respective 0,3 et 0,7).

Figure 2 : Fonctions de densité de deux lois gaussiennes (la 1ère de centre 0 et de variance 1, la 2e de centre 3 et de variance 2) et de leur mélange



Différents modèles gaussiens

Si les individus sont décrits par p variables, la moyenne de chaque classe est un vecteur de taille p et la variance une matrice de taille p x p. Le nombre de paramètres à estimer se décompose ainsi :

- proportions : K-1 valeurs (où K est le nombre de classes).
- moyenne de chaque classe : p valeurs, soit K_p valeurs.
- variance de chaque classe : p.(p+1)/2 valeurs (car la matrice est symétrique) soit K.p.(p+1)/2.

Ainsi le nombre de paramètres à estimer est Kp + K.p.(p+1)/2 + K-1. Sur les données traitées dans cet article (14 variables) et pour 8 classes, 959 paramètres sont alors à estimer. Cette situation correspond au cadre le plus général, permettant de modéliser tout type de distance entre les individus (en particulier les situations non isotropes).

Afin de modéliser plus efficacement des situations moins contraintes, 14 modèles gaussiens différents sont disponibles dans MIXMOD. Les dessins de la figure 3 permettent de visualiser très schématiquement en deux dimensions comment les modèles gaussiens permettent de modéliser la dispersion des individus pour chaque classe en imposant (ou non) des contraintes sur leur dispersion, orientation et forme. Il s'agit bien entendu d'une simplification théorique qui vise à montrer le résultat des options disponibles dans MIXMOD portant sur le volume, l'orientation ou la forme des classes. On pourrait par exemple savoir, concrètement, le résultat d'une classification qui imposerait aux classes d'avoir la même forme et la même orientation tout en laissant libre leur volume et leur proportion. Pour cela, il faudrait appliquer l'une des 14 solutions proposées par la méthode à un jeu de données composé de deux variables seulement et observer le résultat ; puis reproduire la même démarche aux 13 autres modèles. Ce serait sans doute intéressant, mais cela nous éloignerait beaucoup trop de l'objectif de cette contribution.

Pour bien comprendre cette figure 3 et l'intérêt de ces modèles, il faut revenir à la décomposition de la matrice de variance Σ_k en valeurs singulières proposées par Celeux et Govaert (Celeux, Govaert, 1995 ; Banfiel, Raftery, 1993) :

$\Sigma_k = \lambda_k . D_k . A_k . D'_k$ avec :

$$\lambda_k = |\Sigma_k|^{(1/d)}$$

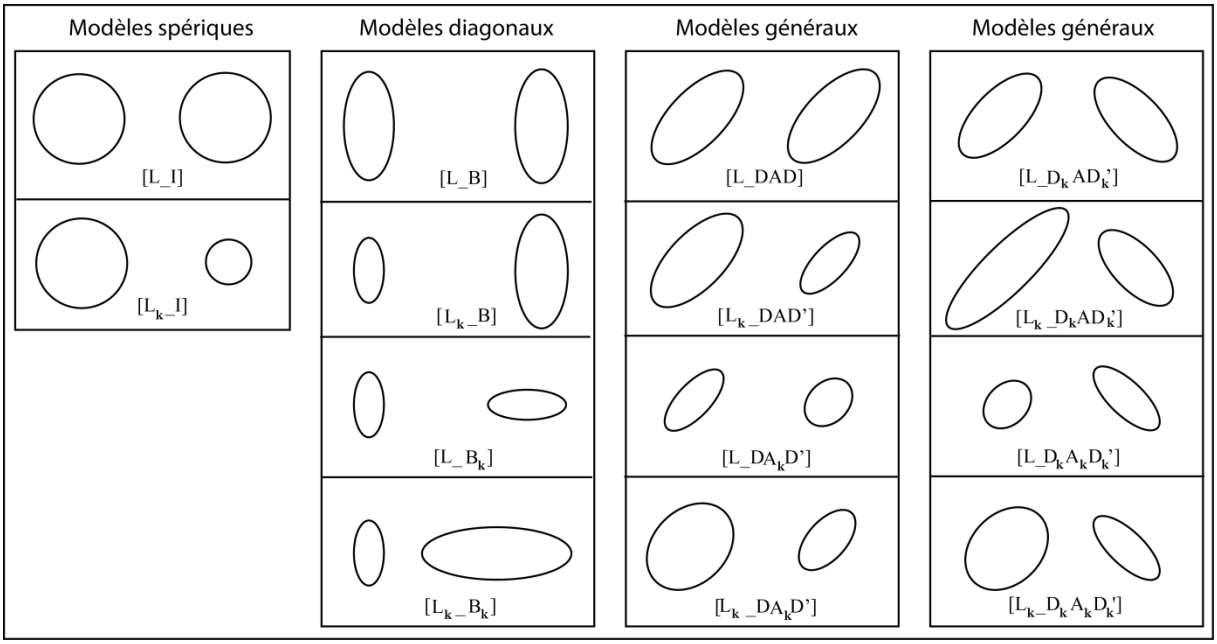
D_k, la matrice orthogonale des vecteurs propres de $\Sigma_k$et D'_k sa transposée

A_k: $diag(a_{(1k)}, \ldots, a_{(dk)})$, la matrice des valeurs propres normalisées de $\Sigma_k$avec les $a_{(ik)}$ classés par ordre croissant sur la diagonale et $|A_k| = 1$

Cette décomposition permet de donner une interprétation géométrique aux 3 valeurs :

- $\lambda_k$représente le **volume** occupé par la classe G_k
  Note : le volume représente la 'place' qu'occupent les individus d'une classe et non sa proportion (le pourcentage d'individus)
- $D_k$représente l'**orientation** de la classe G_k
- A_k représente la **forme** de la classe G_k

Figure 3 : Les 14 modèles gaussiens disponibles dans Mixmod
La forme, le volume et l'orientation des classes sont représentés schématiquement en 2 dimensions (exemple d'une classification portant sur 2 variables). En réalité, ces contraintes sont appliquées aux 14 variables de la matrice des données climatiques



En imposant ou non des contraintes aux paramètres λ_k,D_k et A_k, on obtient des modèles différents, destinés à traiter des situations dans lesquelles les volumes, les orientations, les formes des classes peuvent être égaux ou libres (différents). Ce faisant, Celeux et Govaert proposent 14 modèles (intégrés dans Mixmod), du plus parcimonieux au moins parcimonieux.

Les modèles utilisant un indice *k* pour l'un de ces trois paramètres laissent libre ce paramètre et l'absence d'indice *k* impose une contrainte d'égalité sur le paramètre concerné.

Exemple : dans le modèle $[\lambda_k D_k A D'_k]$ les volumes et les orientations sont libres (elles peuvent être différentes selon les classes) et les formes sont identiques quelle que soit la classe.

Les 14 modèles peuvent être regroupés en trois grandes familles :

- **famille sphérique** : la matrice forme est l'identité ($A = I$) et donc la matrice orientation ne joue aucun rôle. En dimension 2, l'ellipse de dispersion est un cercle. Ici, la contrainte ne peut porter que sur le volume (soit deux modèles) :

- Modèle L_I : les matrices de variances sont toutes identiques et sphériques.
- Modèle L_k_I : les matrices de variances sont sphériques mais les volumes des classes sont libres.

- **famille diagonale** : $B = D.A.D'$est diagonale (et donc Σ aussi). En dimension 2, l'ellipse de dispersion est orientée suivant les axes. Ici, la contrainte peut porter sur le volume et l'orientation (soit 2*2=4 modèles) :
  - Modèle L_B : les matrices de variance ont la même orientation et le même volume (mais elles ne sont pas sphériques).
  - Modèle L_k_B : les matrices de variance ont la même orientation mais leur volume est libre.
  - Modèle L_B_k : les matrices de variance n'ont pas la même orientation mais le même volume.
  - Modèle L_k_B_k : les matrices de variance n'ont ni la même orientation, ni le même volume.

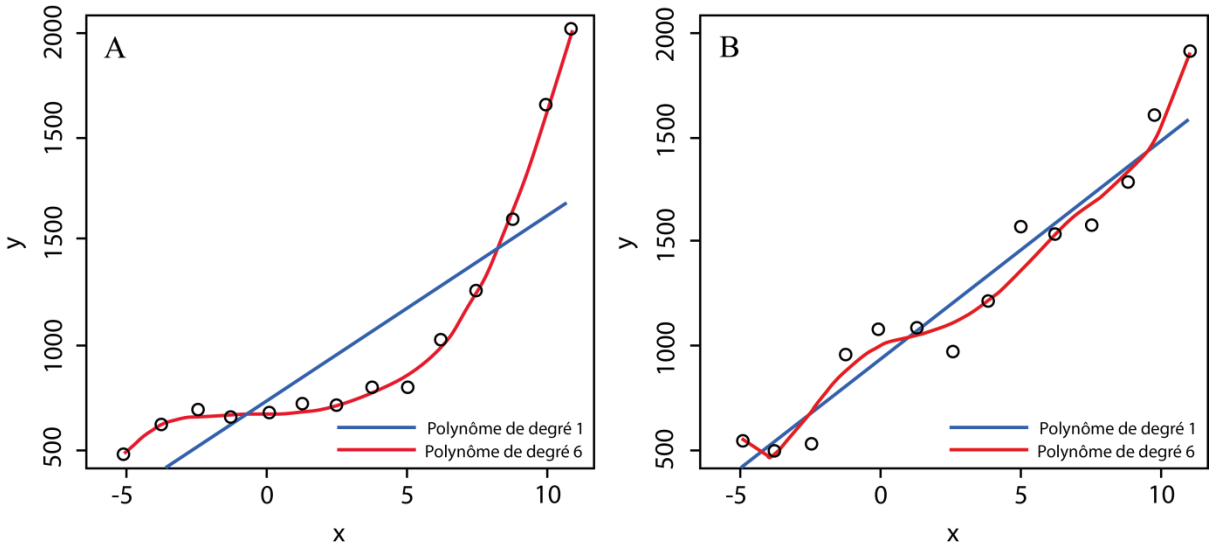
- **famille générale** : elle regroupe tous les autres modèles. Ici, la contrainte peut porter sur le volume, l'orientation et la forme (soit 2*2*2=8 modèles). Ils permettent de modéliser des situations de moins en moins contraintes ; jusqu'au modèle [L_k_$D_kA_kD'_k$] (aucune contrainte).

De plus, les proportions des classes peuvent être libres (dans c_e c_as, l_e nom des modèles est préfixé par « p_k ») ou égales (préfixés par «p »). Ainsi, on dispose de 28 modèles différents dans MIXMOD.

En intégrant dans les calculs de distance entre deux individus les matrices de variances, les modèles de mélanges gaussiens permettent de modéliser toutes les situations allant de la distance isotrope à la distance la plus générale, la moins contrainte.

Quel modèle choisir ?

Figure 4 : Ajustement de deux séries par des polynômes d'ordre 1 et 6



Choisir un modèle trop contraint (euclidien) ne permettra pas d'obtenir les bons paramètres des classes. En effet, si l'on considère – à tort – que la distance est isotrope, les affectations des individus aux classes seront erronées (figure 1). Choisir le modèle le plus général peut être tentant, mais il faut garder à l'esprit le nombre très élevé de paramètres qui sont alors à estimer. Il convient parfois de préférer un modèle avec quelques contraintes réalistes permettant de diminuer de manière significative le nombre de paramètres à estimer et ainsi obtenir une meilleure modélisation.

Une analogie avec des polynômes d'interpolation peut aider à comprendre cette alternative. Avoir recours à un modèle complexe peut être indispensable pour des situations complexes. C'est le cas pour interpoler par un polynôme les points représentés dans la figure 4A. Une interpolation de degré 1 ne peut donner de bons résultats et il convient d'augmenter le degré du polynôme pour obtenir des résultats satisfaisants (degré 6). En revanche, dans des situations plus simples (figure 4B), la qualité d'une interpolation linéaire est tout à fait comparable (voire meilleure car plus simple) qu'avec un polynôme de degré 6. Estimer deux paramètres (interpolation de degré 1) peut suffire pour modéliser correctement des situations relativement simples. Augmenter la complexité du modèle n'apporte pas de plus-value et augmente le nombre de calculs.

Des critères pour choisir le bon modèle, le bon nombre de classes

Comparer la qualité de deux classifications est un problème très compliqué surtout lorsque le nombre de classes et le nombre de variables sont grands. Aussi un autre atout des modèles de mélange réside dans leur capacité à proposer des critères objectifs et quantifiés pour comparer les résultats de deux classifications. En effet, les algorithmes utilisés dans MIXMOD entrent dans la famille des algorithmes de type EM (espérance-maximisation, en anglais *Expectation-Maximisation*[2]) qui permettent de maximiser une quantité statistique : la vraisemblance. Ces algorithmes itératifs convergent vers le maximum de vraisemblance. Ainsi comparer deux résultats de classification revient, au moins dans un premier temps, à comparer les valeurs de vraisemblance obtenues.

Lorsque les connaissances a priori sur les données ne permettent pas de choisir un seul modèle idéal pour les représenter, il est commun de recourir à des méthodes qui permettent de sélectionner les modèles les plus performants (Lebarbier et Mary-Huard, 2004) ou en privilégiant les modèles en phase avec l'objectif de l'utilisateur. Ainsi le critère « Bayesian Information Criterion » (BIC) (Schwarz, 1978) prend en compte la complexité du modèle en ajoutant à la vraisemblance un terme pénalisant les modèles les plus complexes de sorte que les modèles plus simples seront privilégiés (à qualité de résultats équivalents). Le critère « Integrated Classification Likelihood » (ICL) (Biernacki et Govaert, 1999) permet de privilégier les situations dans lesquelles les classes seront bien séparées en ajoutant à la vraisemblance un terme d'entropie intra-classe.

Résultats

À l'issue de la CHA, nous avons obtenu une typologie où les huit climats suivants ont été identifiés et cartographiés (figure 5A) sur le territoire métropolitain :

- type 1 : les climats de montagne,
- type 2 : le climat semi-continental et le climat des marges montagnardes,
- type 3 : Le climat océanique dégradé des plaines du Centre et du Nord,
- type 4 : Le climat océanique altéré,
- type 5 : Le climat océanique franc,
- type 6 : Le climat méditerranéen altéré,
- type 7 : Le climat du bassin du Sud-Ouest,
- type 8 : Le climat méditerranéen franc.

[2] Précisons que la traduction littérale de *« Expectation-Maximisation »* n'est pas « Espérance-Maximisation ». C'est cependant ce qui est couramment utilisé car l'étape E consiste à calculer l'espérance mathématique (de la vraisemblance statistique). C'est pour cela que l'équivalence anglais-français du terme « EM » n'est pas une simple traduction.

Le paramétrage de MIXMOD a été effectué avec l'objectif de fournir une partition spatiale de l'espace français qui puisse être comparée à celle issue de la CHA. La recherche du nombre K de classes devait d'abord être sinon égale, du moins voisine de 8. Ensuite, un appariement des classes issues de la CHA devait pouvoir être fait avec les classes issues de MIXMOD. Compte tenu des multiples possibilités de classification proposées par MIXMOD, un grand nombre de résultats a été produit. La pertinence de chacun d'eux a été évaluée par les statistiques associées : valeur de la vraisemblance et du critère BIC.

Les résultats présentés ci-dessous sont extraits d'un ensemble beaucoup plus complet de simulations réalisées avec MIXMOD. Avec 28 modèles différents, et pour un nombre de classes variant de 4 à 12, il eût fallu tester 252 configurations différentes (28*9), pour un temps de calcul allant de 30 min à 3 h selon la configuration. Une stratégie alternative a été adoptée : estimer la qualité des résultats pour un nombre restreint de configurations décrivant un large spectre, sélectionner les meilleures (selon les valeurs de la vraisemblance et du critère BIC) et enfin, affiner la recherche autour de ces configurations. Parmi ce spectre, seuls les modèles n'imposant pas de contraintes sur les proportions (en « p_k ») ont été utilisés (les zones climatiques n'ont à l'évidence pas le même nombre d'individus.)

Dans un premier temps, les configurations suivantes ont été testées :

- un nombre de classes égal à 4, 6, 8, 10 et 12.
- les modèles (1) $p_k\_L_k\_I$, $p_k\_L_k\_B_k$, (2) $p_k_L_k$_DAD' et (3) $p_k\_L_k\_D_kA_kD'_k$ (la signification des modèles est donnée dans le paragraphe « différents modèles gaussiens ».

Les valeurs de vraisemblance obtenues varient de $-7{,}7.10^8$ (pour la moins bonne) à $-5{,}12.10^8$ (pour la meilleure) et les valeurs du critère BIC (à minimiser) de $1{,}02.10^9$ (pour la meilleure) à $1{,}57.10^9$ (pour la moins bonne).

Tableau 1 : Valeurs de la vraisemblance et de BIC pour les deux meilleures configurations

Modèle – nombre de classes	**Vraisemblance**	**BIC**
$p_k_L_k$_DAD' – 10 classes	$-5{,}12.10^8$	$1{,}02.10^9$
$p_k\_L_k\_D_kA_kD'_k$ – 10 classes	$-5{,}15.10^8$	$1{,}03.10^9$

Deux configurations se distinguent clairement des autres tant pour la vraisemblance que pour le critère BIC : il s'agit des modèles $p_k_L_k$_DAD' à 10 classes et $p_k\_L_k\_D_kA_kD'_k$ à 10 classes (voir les valeurs dans le tableau 1.)

Il est important de noter que les meilleures classifications selon MIXMOD (selon le maximum de vraisemblance et le critère BIC) se sont révélées être également, les plus pertinentes au niveau de l'interprétation climatique, ce qui accrédite l'hypothèse que la statistique a bien analysé le phénomène. Pour montrer l'effet, sur la typologie et l'organisation spatiale des classes, qu'apportent les contraintes de forme et de volume lors de la constitution de ces dernières, les résultats de deux modèles seront présentés. Il s'agit du modèle le plus général et d'un autre modèle, voisin du précédent quoique moins général. Ces deux configurations ne pouvant être facilement départagées au regard des valeurs très proches de la vraisemblance et de BIC, il nous a semblé pertinent de présenter leurs résultats respectifs et de les confronter aux résultats de la CHA.

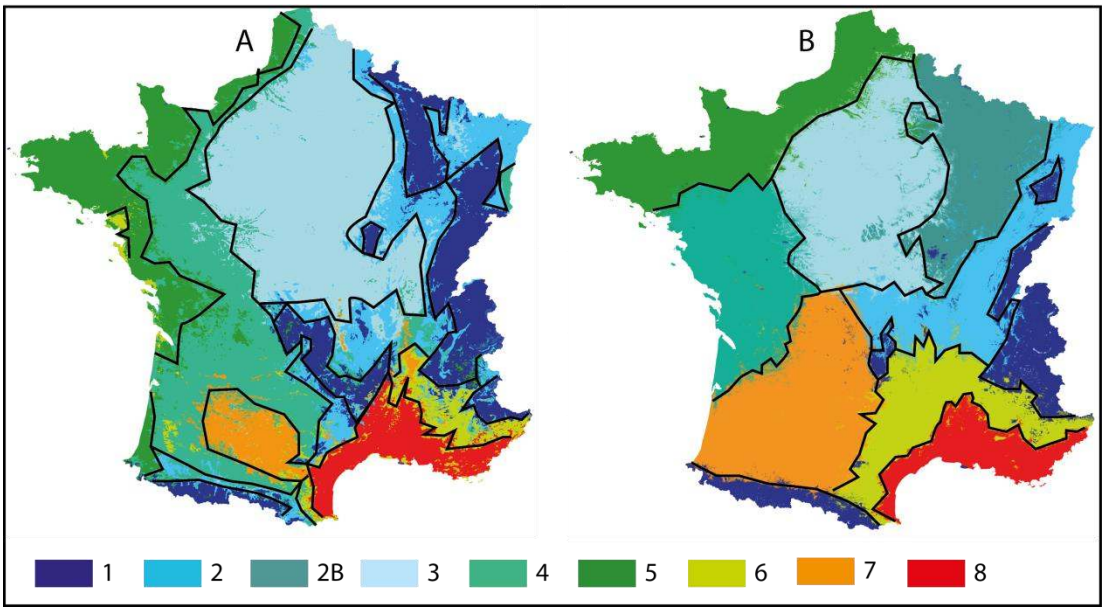
L'organisation spatiale donnée par la CHA et le modèle général de MIXMOD

Modèle « $p_k\_L_k\_D_kA_kD'_k$»

Comme nous l'avons vu plus haut, ce modèle correspond à la situation pour laquelle on n'impose aucune contrainte sur la forme, la dispersion, le volume de chaque classe et qui peut ainsi modéliser le plus librement possible la réalité. Les proportions des classes sont également non contraintes (ce sont les indices « k » qui le spécifient). Une représentation de ce modèle en dimension 2 et pour deux classes est disponible sur la figure 3 (modèle aussi appelé $L_k\_D_kA_kD_{k'}$, représenté dans la dernière case située en bas et à droite de la figure).

Ceci dit, estimer correctement les paramètres de ce type de modèle n'est possible que si l'on dispose d'une grande quantité d'information, comme c'est le cas ici (plus de huit millions d'individus et 14 variables). En effet, le nombre de paramètres (représentant la moyenne, la matrice variance, et la proportion de chaque classe) à estimer pour ce modèle non contraint, pour 10 classes et en dimension 14 est de 1199 (pour plus d'information, se reporter à la documentation statistique de MIXMOD disponible sur www.mixmod.org).

Figure 5 : Typologie climatique du territoire français issue de la CHA (A) et de MIXMOD (B) : modèle pk_LK_DkAkD'k



La structure spatiale des climats vus par la CHA et MIXMOD

Globalement, MIXMOD (figure 5B) reproduit la structure spatiale des climats de la CHA. Les climats océaniques de la façade ouest s'opposent au méditerranéen et au climat de montagne avec les dégradés qui ont été décrits dans *Cybergéo1*. Il est important de souligner que la classification MIXMOD est moins bruitée que celle résultant de la CHA avec des classes plus compactes, plus uniformes et moins enchevêtrées les unes dans les autres contrairement aux classes 1 et 2 de la CHA. Il apparaît donc clairement que ce problème de l'imbrication problématique de certains types au sein d'ensembles climatiquement différents lié à la CHA est solutionné par MIXMOD.

MIXMOD ajoute deux classes. La première (classe 2B de la figure 5) correspond au triangle renversé dont le sommet est, au sud, la vallée de la Saône et la base, au nord, la Lorraine et les Ardennes. Ce type s'intercale entre le climat océanique dégradé des plaines du Centre et du Nord (type 3) et le

climat semi-continental (type 2). La seconde classe ajoutée, qui regroupe des pixels tous situés dans certaines vallées des Alpes centrales (Oisans, Champsaur, Maurienne), n'est pas pratiquement pas visible sur la figure 4 car elle concerne un nombre infime de pixels. Ce point sera examiné plus en détail ultérieurement avec le paragraphe consacré au « cas des Alpes centrales ».

D'autres différences apparaissent concernant cette fois-ci l'étendue et l'emplacement des classes communes aux deux classifications :

- L'océanique franc (type 5) de la CHA concerne la Vendée, la Bretagne et la Normandie avec des extensions vers la Flandre maritime et le Pays Basque en un fin liseré côtier. Avec MIXMOD, il concerne une frange assez large le long de la Manche et du Pas-de-Calais, entre Bretagne et frontière belge.
- L'océanique altéré (type 4) ne se présente plus sous la forme d'un long ruban parallèle au précédent (figure 4A), mais sous la forme d'un bloc compact et homogène calé entre le Sud Bretagne et le bassin d'Arcachon à l'Ouest et le Maine et l'ouest du Massif central à l'Est.
- Le climat du bassin du Sud-Ouest (type 7) couvre, avec MIXMOD, quatre à cinq fois plus de surface, occupant tout l'espace situé entre le sud-ouest du Massif central et le pied des Pyrénées, jusqu'au littoral.
- Le climat océanique dégradé des plaines du Centre et du Nord (type 3) concerne des espaces analogues, quoique un peu plus réduits avec MIXMOD (la limite est de ce type reste en deçà du Morvan).
- Le climat semi-continental et le climat des marges montagnardes (type 2), présent presque partout en France sous la forme de petites enclaves (figure 4A) est, avec MIXMOD (figure 4B), très concentré à l'est, montagne exclue, avec une petite extension vers le Centre, au nord de l'Auvergne et du Forez.
- Le climat de montagne (type 1), très extensif avec la CHA, est beaucoup plus réduit avec MIXMOD. Il n'apparaît plus que sur dans les Alpes et les Pyrénées au-dessus de 900 m, dans le Morvan et le Massif central au-dessus de 800 m et dans les Vosges et le Jura au-dessus de 700 m.
- Le climat méditerranéen altéré (type 6) ne se limite plus aux Alpes et Préalpes du sud mais se poursuit à l'Ouest en basse vallée du Rhône et au sud du Massif central (Cévennes, Minervois, Corbières).
- Enfin, l'extension du climat méditerranéen franc (type 8) fournie par les deux classifications est pratiquement identique, s'étendant le long du littoral éponyme en une bande mince vers les frontières italienne et espagnole, avec un renflement aux abords de la vallée du Rhône et de la Provence.

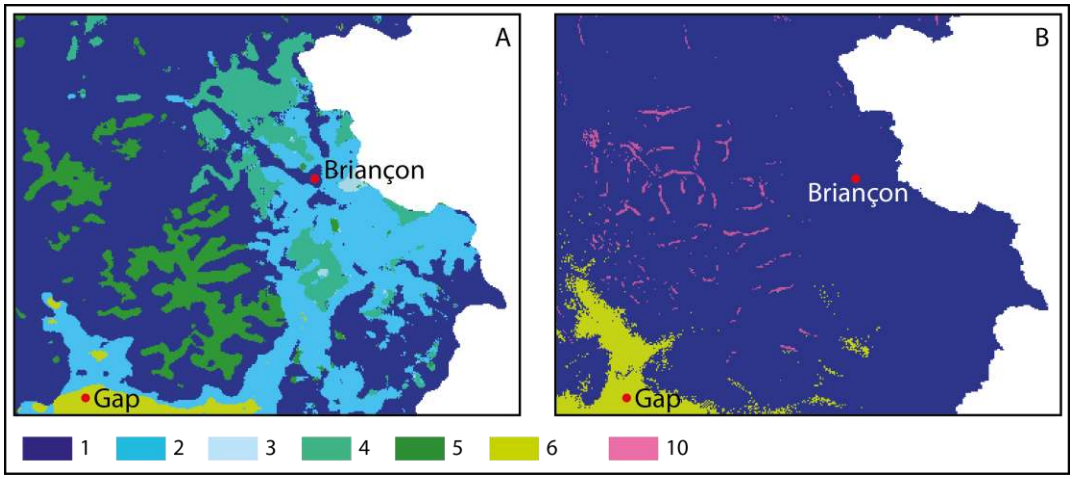
Dans l'une et l'autre carte, les classes s'organisent selon des dispositions qui tiennent tout à la fois aux composantes zonales et méridiennes. Cela est bien mis en évidence avec la classification MIXMOD qui juxtapose des types climatiques compacts selon une espèce de carroyage disposé du nord au sud et d'ouest en est.

Le cas des Alpes centrales

Le cas particulier évoqué dans *Cybergéo1* au sujet de la multiplicité des classes présentes dans le Briançonnais et l'Oisans est présenté dans la figure 6A qui montre que cet espace réduit à la topographie complexe et tourmentée est composé de six des huit types de la CHA. Seuls manquent

le méditerranéen franc et le climat du bassin du Sud-Ouest. Avec MIXMOD (figure 6B), la situation est beaucoup plus simple puisque seuls deux types majeurs apparaissent. Le méditerranéen dégradé est présent dans les vallées proches de Gap tandis que le type 1, le climat de montagne, concerne l'essentiel de l'espace. Le type 10, qui apparaît uniquement dans cette petite région alpine, est confiné au fond des vallées de l'Oisans et du Champsaur où, semble-t-il, le climat est très particulier : pas assez froid pour être classé dans le type 1, mais pas assez chaud et sec durant l'été pour rejoindre le méditerranéen de transition. Dans la mesure où la proportion n'est pas contrainte dans la constitution des classes, ce type original émerge en tant qu'une dixième classe très peu fournie sans que, sur le plan climatique, il soit loisible d'en dire beaucoup plus.

Figure 6 : Typologie climatique du Briançonnais et de l'Oisans issue de la CHA et de MIXMOD



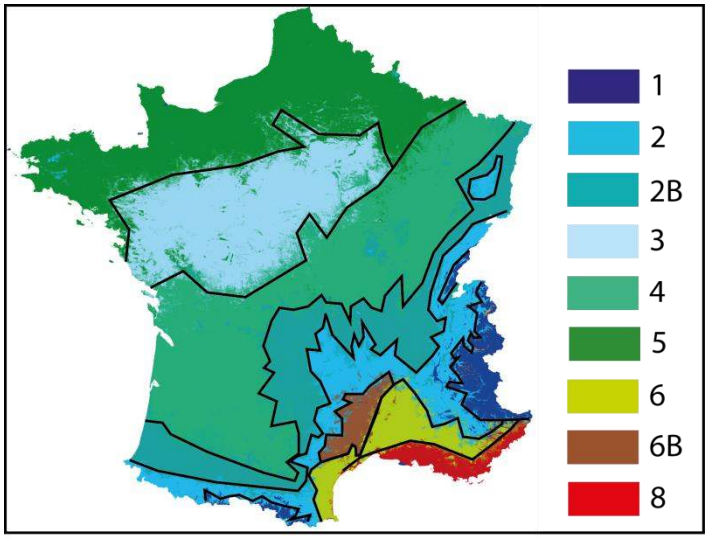
L'organisation spatiale donnée par une autre classification de MIXMOD

À titre indicatif, nous avons reproduit une autre représentation des climats français obtenue de MIXMOD pour montrer qu'en matière de classification, rien n'est figé et que les modèles choisis ont une énorme importance sur les résultats. Il s'agit du modèle « $p_k\_L_K\_DAD'$ » à 10 classes qui se différencie du précédent en imposant aux classes d'avoir la même forme et la même orientation tout en laissant libre leur volume et leur proportion. Une représentation de ce modèle en dimension 2 et pour deux classes est disponible sur la figure 3 (modèle aussi appelé L_k_DAD', deuxième modèle général représenté dans la case située à la 2e ligne et la 3e colonne de la figure). Le nombre de paramètres à estimer pour ce modèle (pour la moyenne, la matrice variance, et la proportion de chaque classe) tombe à 263 (à comparer à 1199 pour le modèle $p_k\_L_k\_DAD'$).

La disposition des climats fournie par le modèle $p_k\_L_K\_DAD'$ (figure 7) semble, de prime abord, très différente de celle qui viennent d'être présentée. Ce nouveau modèle classe en effet les climats selon une logique où s'impose une opposition entre deux pôles extrêmes : les espaces littoraux du nord et du nord-ouest représentés par le type 5 (océanique frais) d'une part et le méditerranéen du sud-est (type 8) d'autre part. Entre ces deux pôles, les types s'allongent selon une orientation générale dirigée du sud-ouest vers le nord-est et se succèdent les uns aux autres du nord au sud : le type 3 succède vers le sud au type 5 qui vient d'être évoqué, arrivent ensuite les types 4, 2B, 2 puis 6 et enfin 8. Cette disposition est localement gauchie par le relief. La montagne se superpose sur ce dispositif de sorte qu'il est possible de rapprocher les Vosges du type 4, le Jura et les Alpes du Nord du type 2B, les Alpes centrales et les Pyrénées du type 2 et les Alpes du Sud (Mercantour) du type 6. Le type 6B, qui correspond au Vivarais, aux Cévennes et au Minervois, est la nuance montagnarde du

méditerranéen altéré (type 6). On retrouve également ici la classe 10 correspondant à certaines vallées des Alpes centrales.

Figure 7 : Typologie climatique du territoire français issue du modèle p_k_L_K_DAD' en 10 classes de MIXMOD



Il peut paraître surprenant que l'Alsace ou la vallée du Doubs appartiennent au même type climatique (2B) que le Pays Basque ou la Chalosse. Il en va de même pour le plateau Lorrain et les Landes (classe 4). En fait, si ces régions éloignées appartiennent au même type, c'est parce qu'elles se situent à la même distance le long du gradient nord-ouest/sud-est : les premières un peu plus près du sud-est, les secondes un peu plus près du nord-ouest.

Conclusion

L'opération qui consiste à classer un jeu de données est délicate car elle est assujettie à la méthode utilisée pour y parvenir. Pour nous en convaincre, nous avons recouru à deux méthodes distinctes, l'une fondée sur un calcul de distance euclidienne (CHA), l'autre reposant sur des calculs de distance non linéaire et des modèles de mélange (MIXMOD). Les résultats obtenus sont intéressants car ils segmentent l'espace français selon la même logique nord-sud/est-ouest, ce qui est cohérent dans la mesure où les individus à classer sont structurés par les mêmes lois d'organisation spatiale. Pour autant, les cartes obtenues présentent des types climatiques dont l'étendue et la disposition spatiales sont loin d'être identiques. Les deux modèles de MIXMOD qui ont été présentés apportent des résultats différents dans la disposition des classes les unes par rapport aux autres : classes compactes dans le premier cas, classes allongées dans le second. Mais tous deux reflètent aussi les mêmes contraintes d'organisation spatiale. Précisons que l'interprétation d'un résultat statistique multivarié tel que peut l'être une classification portant sur 14 variables est souvent difficile. Parmi les nombreux modèles proposés par MIXMOD, seuls deux ont été retenus, les autres, notamment les modèles très contraints, ne l'ont pas été car ils fournissent des résultats difficiles à interpréter.

La pertinence d'une classification est difficile à établir et rend délicat l'évaluation de la qualité de cette dernière : y a-t-il un algorithme meilleur qu'un autre ? Vaste question qui n'a pas de réponse absolue tant elle dépend de la nature des données, de l'objectif assigné à la classification et de bien d'autres contraintes. Il est toutefois communément admis qu'il n'y a pas d'algorithme supérieur à un

autre concernant les classifications fondées sur le critère de distance euclidienne en l'absence d'information a priori sur le problème à traiter. Le choix de la CHA est donc tout aussi pertinent qu'un autre. Pour autant, contrairement à la plupart des populations statistiques dont les individus ne sont pas localisés, celle qui a été traitée ici a permis, grand avantage, de juger visuellement de la cohérence spatiale des classes obtenues. Ainsi, malgré une bonne répartition dans l'espace français des types climatiques, les résultats fournis par la CHA semblent moins pertinents que ceux de MIXMOD compte tenu de l'imbrication de classes hétérogènes au sein d'un même type, qui contribue à bruiter la qualité des cartes. Les classes de MIXMOD sont plus homogènes, ce phénomène d'interpénétration de classes disparates au sein d'une même région étant supprimé. Seule subsiste la présence, au sein d'un type donné, de pixels appartenant à un type statistiquement proche, phénomène qui, on l'a vu, est normal.

Pour toutes ces raisons, et si nous devions désigner la meilleure classification parmi les trois qui ont été présentées, c'est sans nul doute le modèle $p_k\_L_K\_D_KA_KD'_k$ à 10 classes de MIXMOD qui l'emporterait; avec, en seconde position la CHA. Mais nous sommes conscients que cet avis est empreint d'une certaine subjectivité. D'autres choix sont possibles car les trois méthodes retenues conduisent à des classifications satisfaisantes et il n'y a pas de règle qui permette de trancher de manière rigide pour l'une ou pour l'autre. Notre choix reflète un certain a priori sur la disposition des climats dans l'espace français. Le deuxième modèle de MIXMOD qui a été présenté (figure 7) est séduisant en ce sens qu'il ouvre sur une classification moins classique.

Bibliographie

Arlery R., 1979, Le Climat de la France, Ministère des Transports, Secrétariat Général de l'Aviation Civile et Direction de la Météorologie Nationale (Eds.), Paris, 131 p

Banfield J.D., Raftery A.E., 1993, "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics*, Vol. 49, No.3, 803-821.

Bessemoulin J., 1987, Atlas Climatique de la France Edition Réduite, Ministère des Transports, Secrétariat Général de l'Aviation Civile et Direction de la Météorologie Nationale (Eds.), Paris, 29 planches

Biernacki C., Celeux G., Govaert G., Langrognet F., 2006, "Model-Based Cluster and Discriminant Analysis with the MIXMOD Software", *Computational Statistics and Data Analysis*, Vol. 51, 587-600.

Biernacki C., Govaert G.,1999, "Choosing models in model-based clustering and discriminant analysis", *Journal of Statistical Computation and Simulation*, Vol. 64, No.1, 49-71.

Celeux G., Govaert G.,1992, "A Classification EM Algorithm for Clustering and Two Stochastic Versions", *Computational Statistics and Data Analysis*, Vol.14, No.3, 315-332.

Celeux G., Diebolt J., 1985, "The SEM Algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem", *Computational Statistics Quarterly*, Vol. 2, 73-82

Celeux G., Govaert G.,1995, "Gaussian parsimonious clustering models.", *Pattern Recognition*, Vol. 28, No.5, 781-793.

Champeaux J.-L., Tamburini A., 1994, “Zonage climatique de la France à partir des séries de précipitations quotidiennes du réseau climatologique d’état”, *Publications de l’Association Internationale de Climatologi*e, Vol. 7, 92-98.

Dempster A., Laird N., Rubin D., 1977, “Maximum Likelihood from Incomplete Data with the EM Algorithm (with discussion)”, *Journal of the Royal Statistical Society*, Series B, Vol. 39, No.1, 1-38.

Durand-Dastes F., Sanders L., 1984, “Les contrastes climatiques : noyaux et zones de transition”, in Theo Quant, *Géoscopie de la France*, *Paradigme*, Paris, 91-101.

Joly D., Brossard T., Cardot H., Cavailhès J., Hilal M., Wavresky P., 2009, “Interpolation par régressions locales : application aux précipitations en France”, *L’Espace géographique*, Vol. 38, No.2, 157-170.

Joly D., Brossard T., Cardot H., Cavailhès J., Hilal M., Wavresky P., 2010, “ Les types de climats en France, une construction spatiale”, *Cybergeo : Revue européenne de géographie,* document 501, http://cybergeo.revues.org/index23155.html

Joly D., Brossard T., Cardot H., Cavailhès J., Hilal M., Wavresky P., 2011, “Temperature Interpolation by local information; the example of France”, *International Journal of Climatology*, Vol. 31, No.14, 2141-2153.

Joly D., Bois B., Zaksek K., 2012, “Rank-ordering of topographic variables correlated with temperature”, *Atmospheric and Climate Science*, Vol. 2, No.2, 139-147. http://www.scirp.org/journal/PaperInformation.aspx?paperID=18815

Langrognet F., 2009, “MIXMOD : un logiciel de classification supervisée et non supervisée pour données quantitatives et qualitatives“, *La Revue de Modulad*, No.40, 23-40.

Lebarbier E., Mary-Huard T, “Le critère BIC : fondements théoriques et interprétation. [Research Report] RR-5315, INRIA. 2004, pp.17. <inria-00070685>

Mclachlan G, Peel D., 2000, *Finite Mixture Models, Wiley Series in Probability and Statistics*, Wiley-Interscience.Mahalanobis P. C., “On the generalised distance in statistics”, *Proceedings of the National Institute of Sciences of India*, Vol. 2, No.1, 1936, 49-55Schwarz G., 1978, “Estimating the Dimension of a Model,” *The Annals of Statistics*, Vol. 6, 461-464.

Wikipedia. https://fr.wikipedia.org/wiki/Induction_(logique)

Annexes

Annexe 1 : MIXMOD : un ensemble logiciel de classification des données par modèles de mélanges

Diffusion

Le site web (www.mixmod.org) du projet Mixmod contient un ensemble de rubriques (documentations, articles, captures d'écrans…) et, bien évidemment, met à disposition les composants logiciels Mixmod. Au nombre de quatre, ils sont tous diffusés sous licence libre (GNU GPL – voir https://www.gnu.org/copyleft/gpl.html) ce qui favorise leur diffusion, les échanges avec la communauté scientifique, la possibilité de les adapter à des besoins spécifiques. De plus, ce type de licence (avec le code source) favorise les retours et augmente la fiabilité des composants logiciels.

Initialement destiné à la communauté scientifique, Mixmod a rapidement fait l'objet d'intérêt d'utilisateurs d'horizons variés et hétérogènes. Pour répondre à ces attentes différents composants logiciels ont été développés et diffusés. Ces outils intègrent de nombreuses fonctionnalités de classification des données par modèles de mélanges dont :

- La classification supervisée et non supervisée,
- le traitement de données qualitatives ou qualitatives, ou mixtes (qualitatives et quantitatives),
- les modèles spécifiques pour les données de grande dimension (décrites par plusieurs centaines voire milliers de caractéristiques),
- les algorithmes EM, CEM, SEM,
- plusieurs méthodes d'initialisation.

L'algorithme EM (en anglais Expectation-maximisation), proposé par Dempster *et al.* (1977), permet de trouver le maximum de vraisemblance des paramètres de modèles probabilistes. EM se compose de deux étapes :

- une étape d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées
- une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

On utilise ensuite les paramètres trouvés à l'étape M comme point de départ d'une nouvelle phase d'évaluation de l'espérance (étape E), et l'on itère ainsi. Cet algorithme converge vers un maximum (local) de vraisemblance.

Les algorithmes CEM et SEM sont des variantes de EM en ajoutant entre l'étape E et l'étape M, une troisième étape (dite 'classifiante' pour CEM et 'stochastique' pour SEM)

Les composants logiciels

mixmodLib : la bibliothèque de calcul. Cette bibliothèque est la pierre angulaire du projet, les autres composants s'appuyant dessus. Elle est constituée d'une centaine de classes C++ et fait l'objet d'une attention particulière tant sur le plan de l'optimisation des performances que de la robustesse. Des classes de « haut niveau » permettent d'interfacer cette bibliothèque à d'autres composants logiciels.

mixmodForMatlab. Ce composant logiciel a été développé et diffusé dès 2002 pour permettre d'utiliser les fonctionnalités de Mixmod dans l'environnement scientifique Matlab, assez couramment utilisé dans le milieu de la recherche. Il est constitué d'un ensemble de fonctions Matlab servant d'interface pour mixmodLib et pour visualiser les résultats. Il est important de préciser que les calculs sont effectués dans mixmodLib.

MixmodGUI (Graphical User Interface). A la demande d'utilisateurs peu familiers avec Matlab et aussi pour ne pas être obligé de se doter de cet outil pour bénéficier d'une interface graphique pour Mixmod, mixmodGUI a été développé à partir de 2010. Cette interface graphique, disponible pour Linux et Windows est l'outil le plus simple à prendre en main parmi les composants logiciels Mixmod mais, il faut le préciser, n'intègre pas toutes les fonctionnalités de mixmodLib.

RMixmod. Enfin, pour répondre aux demandes de la communauté de statisticiens habitué à l'environnement scientifique R, le package Rmixmod a été développé à partir de 2012. Comme pour mixmodForMatlab, il s'agit d'un ensemble de fonctions R permettant de profiter de la bibliothèque mixmodLib dans R.

Annexe 2 : précisions statistiques

Classification des données, métriques

Rappels sur la classification

Il convient tout d'abord de faire quelques rappels sur les problématiques de classification des données (vocabulaire, définitions, objectifs).

- Un jeu de données est composé de *n* individus décrits par *d* caractéristiques (les variables). D'un point de vue mathématique, ces individus (ou observations) sont notés x_i avec $x_i \in \mathbb{R}^d$(i=1...n).
- Une partition d'un ensemble *X* est un ensemble *P* de sous-ensembles non vides de *X* deux à deux disjoints et qui forment un recouvrement de *X*.
- La classification des données consiste à trouver une partition du jeu de données x_i.
- Les sous-ensembles seront appelés des classes que l'on notera G_k

Notons que toute partition est une classification et que la principale difficulté est de trouver la partition la plus pertinente (en fonction de l'objectif). Comparer plusieurs classifications (/partitions) n'a de sens qu'à travers un critère objectif (classes bien séparées, limites nettes entre les classes, …).

Chercher une partition revient donc à :

- chercher le nombre *K* de classes (/de sous-ensembles),
- affecter à chaque individu une classe (/un sous-ensemble).

On notera Z la matrice partition de *X*,

avec, pour chaque individu x_i, :

$z_{ik} = 1$ pour *k* tel que $x_i \in G_k$ et 0 sinon.

Exemple avec n=4, d=3, K=2 (4 individus décrits par 3 variables et pour 2 classes) :

$$X = \begin{pmatrix} 0.2 & 0.5 & 0.8 \\ 0.3 & 0.1 & 0.2 \\ 0.9 & 2.0 & -1.2 \\ 0.4 & 1.2 & -0.2 \end{pmatrix} \text{ et } Z = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \text{ donc } \begin{matrix} x_1 \in G_2 \\ x_2 \in G_1 \\ x_3 \in G_2 \\ x_4 \in G_2 \end{matrix}$$

Quelle métrique pour classer ?

La recherche d'une bonne partition passe par la création de classes homogènes c'est-à-dire dont les individus sont proches. Nous reviendrons sur cette notion très importante de distance (et donc de proximité). D'un point de vue mathématique, cela revient à minimiser l'inertie intra-classe définie par :

$$W_M(z) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \|x_i - \overline{x_k}\|_M$$

avec :

$\|.\|_M$ la distance euclidienne avec M comme métrique dans $\mathbb{R}^d$

$\overline{x_k}$ la moyenne de classe G_k avec

$$\overline{x_k} = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik}\, x_i$$

$$n_k = \sum_{i=1}^{n} z_{ik}\, x_i$$

Il apparaît clairement que la métrique *M* joue un rôle majeur dans le calcul de l'inertie intra-classe, et ce faisant, dans la recherche de la meilleure classification.

La métrique identité

Figure 8 : 1er jeu de données (A) ; classification du 1er jeu de donnée avec une métrique identité (B)

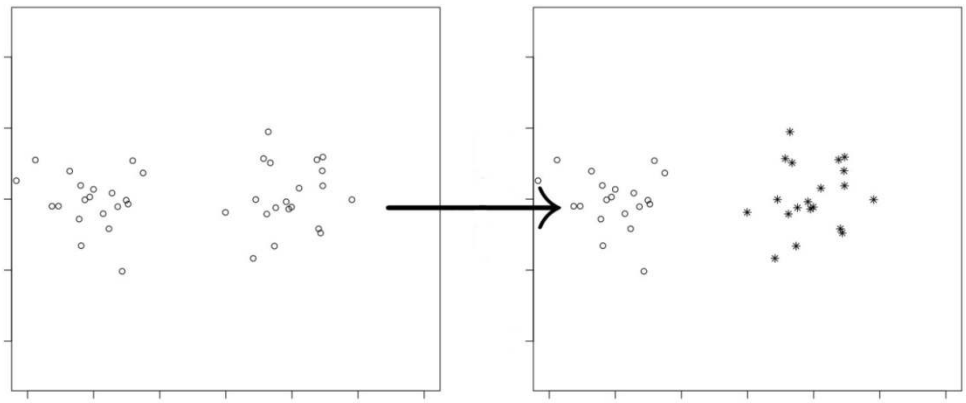
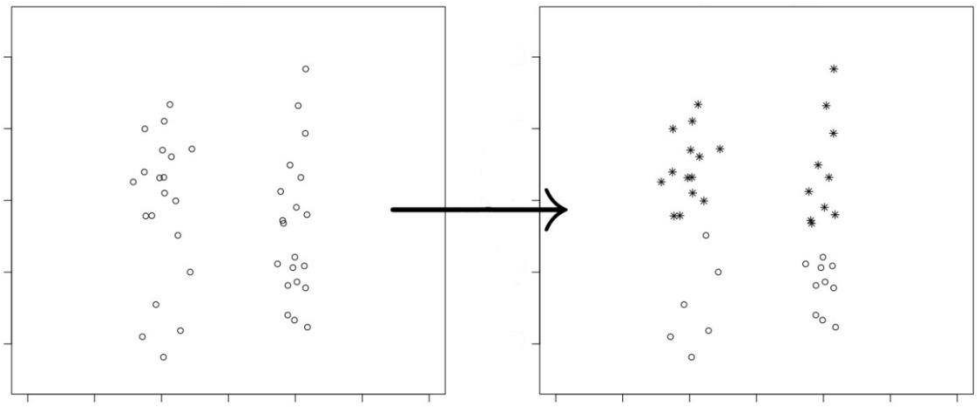


Figure 9 : 2e jeu de données (A) ; B : classification du 2e jeu de donnée selon une métrique identité (B)



La plupart des méthodes de classification de données repose sur l'utilisation de la métrique identité. En d'autres termes, on suppose que la dispersion des individus à l'intérieur d'une classe est la même selon chaque variable (donc dans toutes les directions). Cette hypothèse, à l'évidence non justifiée en général, peut conduire à des résultats incorrects (la partition est non optimale). À titre d'illustration, prenons les deux jeux de données suivants (figures 8A et 9A). Si l'utilisation de la métrique identité mène à une bonne classification dans le 1er cas (classification fig. 8B), elle conduit à un mauvais choix dans le second : alors que le bon sens conduit à proposer une classe à gauche et une classe à droite, la métrique identité ne peut que proposer une classe en haut et une classe en bas (figure 9B) pour des raisons de proximité liée au choix de la métrique identité.

C'est cette métrique qui a été utilisée dans la classification de *Cybergéo1* dont voici la démarche rapidement résumée. Les 14 variables sont discrétisées en 5 modalités puis soumises à une AFC qui, en générant des axes factoriels, donne lieu à un nouveau système de coordonnées dans lequel se positionnent individus et caractères. La matrice de distance ainsi construite sert de base à la CHA, méthode dont le principe est de regrouper itérativement les individus deux à deux selon un critère de distance euclidienne (parmi les quatre choix possibles, le critère de Ward a été retenu car il induit, à chaque étape de regroupement, une minimisation de la décroissance de la variance interclasse). Le barycentre des couples constitue, à chaque étape un nouvel individu qui est comparé à tous les autres selon la même métrique. Ainsi, les classes se constituent pas à pas, regroupant à chaque itération un nombre de plus en plus élevé d'individus. Le processus s'arrête lorsque tous les individus sont regroupés au sein d'une seule classe. L'arbre obtenu permet de choisir le nombre de classes.

Les modèles de mélanges

Les modèles de mélanges sont des outils probabilistes couramment utilisés et appréciés pour leur adaptabilité à modéliser un grand nombre de situations. Comme nous allons le voir, ils peuvent être utilisés pour traiter des problèmes de classification des données (Mclachlan, Peel, 2000). On se dote ainsi d'un cadre rigoureux pour répondre à un certain nombre de questions (choix du nombre de classe, métrique…) plutôt que de faire des hypothèses non réalistes et/ou d'utiliser des heuristiques discutables. Pour ce faire, on considère les individus d'une classe comme un échantillon d'une loi de probabilité. Chaque classe sera alors modélisée par une loi de probabilité, loi gaussienne pour des données quantitatives (comme c'est le cas ici).

Si $x_i \in G_k$, alors $x_i$est une réalisation de la loi gaussienne décrivant $G_k$dont la densité est :

$h_k(x) = (2\pi)^{-d/2}.|\Sigma_k|^{-1/2}.exp\left(\frac{1}{2}(x_i-\mu_k)'\Sigma_k^{-1}(x_i-\mu_k)\right)$ où :

$\mu_k$est la moyenne de G_k

$\Sigma_k$est la matrice variance de G_k

De plus, si l'on définit $p_k$comme la probabilité qu'un individu soit dans la classe G_k, alors la densité de la loi probabilité du mélange est :

$$f(x) = \sum_{k=1}^{K} p_k \,.\, h_k(x)$$

En utilisant les modèles de mélanges, on cherchera à estimer les meilleurs paramètres du modèle sous-jacent :

μ_k, moyenne de la classe G_k
Σ_k, matrice variance de G_k
p_k, proportion de la classe G_k

Pour ce faire, comme on le verra plus bas, on utilisera des outils mathématiques éprouvés qui permettent de mettre un sens à une relation d'ordre entre des modèles, et ce faisant, expliquer pourquoi on peut considérer un modèle meilleur qu'un autre.

Des métriques différentes via des matrices de variances différentes pour modéliser des situations différentes

La métrique naturelle dans ce cas est la distance de Mahalanobis :

$D(x, \mu_k) = \sqrt{\left((x-\mu_k)'.\Sigma_k^{-1}.(x-\mu_k)\right)}$ qui décrit la distance entre x et μ_k(le centre de la classe k) compte tenu de la dispersion Σ_k de la classe k.

Ainsi, non seulement on ne fait pas l'hypothèse de métrique identité mais en plus, on se dote d'un spectre large de métriques obtenues avec des matrices de variances différentes. On utilisera la décomposition de la matrice Σ_k en valeurs singulières proposées par Celeux et Govaert (Celeux, Govaert, 1995 ; Banfiel, Raftery, 1993) :

$\Sigma_k = \lambda_k.D_k.A_k.D'_k$ avec :

$\lambda_k = |\Sigma_k|^{(1/d)}$
D_k, la matrice orthogonale des vecteurs propres de $\Sigma_k$et D'_k sa transposée
A_k: $diag\left(a_{(1k)},\ldots,a_{(dk)}\right)$, la matrice des valeurs propres normalisées de $\Sigma_k$avec les $a_{(ik)}$ classés par ordre croissant sur la diagonale et $|A_k| = 1$

Cette décomposition permet de donner une interprétation géométrique aux trois valeurs :

- λ_k représente le volume occupé par la classe G_k
- D_k représente l'orientation de la classe G_k
- A_k représente la forme de la classe G_k

En imposant ou non des contraintes aux paramètres λ_k,D_k et A_k, on obtient des modèles différents, destinés à traiter des situations dans lesquelles les volumes, les orientations, les formes des classes peuvent être égaux ou libres (différents). Ce faisant, Celeux et Govaert proposent 14 modèles (intégrés dans Mixmod), du plus parcimonieux au moins parcimonieux. Les modèles utilisant un indice *k* pour l'un de ces trois paramètres laissent libre ce paramètre et l'absence d'indice *k* impose une contrainte d'égalité sur le paramètre concerné.

Exemple : dans le modèle $[\lambda_k D_k A D'_k]$ les volumes et les orientations sont libres (elles peuvent être différentes selon les classes) et les formes sont identiques quelle que soit la classe.

Les 14 modèles peuvent être regroupés en trois grandes familles :

- **famille sphérique** : la matrice forme est l'identité ($A = I$) et donc la matrice orientation ne joue aucun rôle. En dimension 2, l'ellipse de dispersion est un cercle
- **famille diagonale** : $B = D.A.D'$est diagonale (et donc Σ aussi). En dimension 2, l'ellipse de dispersion est orientée suivant les axes.
- **famille générale** : elle regroupe tous les autres modèles. On notera $C = D.A.D'$

A titre d'illustration, la figure 3 représente ces 14 modèles dans le cas de deux classes en dimension 2. Enfin, en imposant ou non l'égalité des paramètres p_k(proportion de la classe k), on obtient 28 modèles gaussiens différents, capables de modéliser un spectre très large de situations.

Recours à des algorithmes éprouvés

Disposant d'un tel cadre probabiliste, il convient alors de chercher le paramètre
$\Theta = (p_{1,}\dots, p_k, \mu_{1,}\dots, \mu_k, \Sigma_{1,}\dots, \Sigma_k)$

maximisant la fonction de vraisemblance :

$$L(\Theta, x_{1,}\dots, x_n) = \sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} p_k.h_k(x_i)\right)$$

L'algorithme EM (Dempster, Laird, and Rubin, 1997) (mais aussi ses variantes CEM (Celeux et Govaert 1992) et SEM (Celeux et Diebolt 1985)) constitue alors un outil de choix pour la recherche du maximum de vraisemblance. Cet algorithme itératif fait croître la vraisemblance à chaque étape et converge vers un maximum. Il présente cependant l'inconvénient de ne pas forcément converger vers le maximum global (la fonction de vraisemblance peut naturellement avoir plusieurs maxima locaux). Pour éviter cet écueil, un certain nombre d'outils sont proposés dans Mixmod dont :

- la possibilité d'enchaîner plusieurs algorithmes (par exemple commencer par SEM qui génère une chaîne de Markov autour du maximum et terminer par EM pour converger vers ce maximum) ;
- un large choix de stratégies d'initialisation pour démarrer l'algorithme prenant en compte d'éventuelles informations complémentaires

Recours à des critères pour sélectionner la meilleure classification

Atteindre le maximum de vraisemblance ne suffit pas toujours pour choisir le meilleur modèle, et ce pour deux raisons :

Les modèles les plus généraux seront privilégiés au détriment des modèles les plus contraints (qui sont aussi les plus robustes). C'est pour cela que le critère BIC (*Bayesian Information Criterion*) (Schwarz, 1978) a été introduit. Il permet de pénaliser la fonction de vraisemblance par un terme qui prend en compte la complexité du modèle :

$$BIC_{(m,K)} = -2L_{(m,k)} + \nu_{(m,K)}.ln(n)$$

où :

$BIC_{(m,K)}$ représente la valeur du critère BIC pour le modèle m et pour K classes
$L_{(m,k)}$ représente la log-vraisemblance

$\nu_{(m,K)}$ représente le nombre de paramètres libres du modèle m

Mais ceci n'est pas toujours suffisant pour trouver « la bonne » classification si l'on souhaite intégrer l'objectif de l'utilisateur (par exemple, privilégier des solutions avec des classes bien séparées). L'idée est de pénaliser encore la fonction de vraisemblance par un terme qui représente l'objectif de l'utilisateur. On peut ici citer le critère ICL (*Integrated Completed Likelihood*) (Biernacki, Govaert, 1999), qui privilégiera les classifications présentant des classes bien séparées défini par :

$$ICL_{(m,K)} = BIC_{(m,k)} - 2\sum_{i=1}^{n}\sum_{k=1}^{K} z_{(ik)}.ln(t_{(ik)})$$

où :

$t_{(ik)}$ représente la probabilité que $x_i$appartienne à G_k

$z_{(ik)}$ vaut 1 pour la valeur de k maximisant $t_{(ik)}$, 0 sinon

---

Pour citer cet article

Référence électronique

Daniel Joly et Florent Langrognet, « Pertinence du découpage spatial produit par deux méthodes de classification (CHA et MIXMOD). Application aux climats français », *Cybergeo : European Journal of Geography* [En ligne], Cartographie, Imagerie, SIG, document 761, mis en ligne le 08 janvier 2016, consulté le 11 janvier 2016. URL : http://cybergeo.revues.org/27414 ; DOI : 10.4000/cybergeo.27414

À propos des auteurs

Daniel Joly
Directeur de recherche au CNRS
Laboratoire ThéMA, UMR 6049 CNRS et Université Bourgogne Franche-Comté
daniel.joly@univ-fcomte.fr

Florent Langrognet
Ingénieur de recherche au CNRS
Laboratoire de Mathématiques de Besançon, UMR 6623 CNRS et Université Bourgogne Franche-Comté
florent.langrognet@univ-fcomte.fr

Droits d'auteur

Résumés

L'espace climatique de la France continentale a d'abord été segmenté grâce à la classification hiérarchique ascendante (CHA), méthode fondée sur un calcul de distance linéaire. Ensuite, nous avons eu recours à MIXMOD, ensemble logiciel qui propose une méthode de classification basée sur une approche probabiliste dont les principes méthodologiques sont détaillés. Les résultats issus de deux modèles de MIXMOD sont comparés à celui produit par la CHA. L'espace français est segmenté de manière analogue par le premier modèle de MIXMOD et par la CHA dans un

dispositif général où les climats de l'Ouest s'opposent à ceux du Sud, les montagnes ressortant de manière autonome. L'apport de MIXMOD apparaît dans son aptitude à segmenter l'espace de manière plus homogène que ne le fait la CHA. Le second modèle de MIXMOD présente une composition spatiale des climats originale, opposant l'océanique frais le long de la Mer du Nord au méditerranéen chaud confiné aux abords de la Côte d'Azur ; entre eux, les climats intermédiaires s'allongent de la côte atlantique aux frontières de l'Est.

Classifications and spatial segmentation relevance using two methods applied to the French climates

The climates of continental France are classified by using two models. The first one (ascending hierarchical classification: AHC) is based on a linear distance calculation while the second (MIXMOD), is nonlinear. The principles of the AHC are briefly presented, but those on which MIXMOD is based are more detailed. The two classifications from MIXMOD are compared to the one from AHC. The French space is segmented in a similar manner by the first model of MIXMOD and AHC in a general framework where the climates of the West opposed to the South, the mountains emerging independently. The contribution of MIXMOD appears in its ability to segment the space more evenly than does the CHA. The second model of MIXMOD presents an original structure of climates, opposing the climate of the region along the North Sea to the Mediterranean one confined near the Côte d'Azur; between them, intermediate climates extend from the Atlantic coast to the eastern borders.

Entrées d'index

Mots-clés : classification, climat, France
Keywords : classification, climate, France