

REVUE DE STATISTIQUE APPLIQUÉE

MATHIEU VRAC

EDWIN DIDAY

ALAIN CHÉDIN

Décomposition de mélange de distributions et application à des données climatiques

Revue de statistique appliquée, tome 52, n° 1 (2004), p. 67-96

http://www.numdam.org/item?id=RSA_2004__52_1_67_0

© Société française de statistique, 2004, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) 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/>

DÉCOMPOSITION DE MÉLANGE DE DISTRIBUTIONS ET APPLICATION À DES DONNÉES CLIMATIQUES

Mathieu Vrac*, **, Edwin Diday*, Alain Chédin**

* Univ. Paris IX Dauphine, Place du Maréchal de Lattre-de-Tassigny, 75775 Paris
{vrac, diday}@pi.ceremade.dauphine.fr

** ARA/LMD, École Polytechnique, 91128 Palaiseau Cedex
{vrac, chedin}@lmd.polytechnique.fr

RÉSUMÉ

L'objectif de l'analyse des données symboliques est d'étudier des bases de données où les individus ont des variations internes. Cela permet de traiter des bases de taille considérable. Nous proposons dans cet article une méthode qui étend la décomposition de mélange de densités en permettant de classer des données de type distributions de probabilités (fonction de répartition), et en modélisant la loi de probabilité associée à ce type de données probabilistes. Cette méthode permet de tenir compte des dépendances qui existent entre les variables utilisées, ainsi que des dépendances à l'intérieur même d'une variable, entre différents points des fonctions de répartition des individus. Ces dépendances intra et inter variables sont prises en compte à l'aide de fonctions multidimensionnelles appelées copules (ou fonctions de dépendance). Les copules mettent en relation chacune des fonctions de répartition unidimensionnelles (dites « marginales ») avec la fonction de répartition multidimensionnelle (dite « jointe »). Nous présentons les notions utiles ainsi que la méthode que nous appliquons par la suite à une grande base de données climatiques.

Mots-clés : Décomposition de mélanges, Classification de distributions, Distributions de distributions, Copules

ABSTRACT

The goal of the analysis of symbolic data is the study of databases where the units have internal variations. It allows the treatment of databases with huge size. In this paper, we propose a method to classify distribution functions data and model the probability law associated to this kind of probabilistic data. This method extends densities mixture decompositions. It allows dependencies existing between the used variables and dependencies existing inside a variable (between different points of the cumulative distribution functions of the units) to be taken into account. These dependencies are modeled with multidimensional functions called copulas (or "dependencies functions"). Copulas link every unidimensional cumulative distribution (called "margin") to the multidimensional cumulative distribution. We introduce useful notions as well as the method and we apply it to a climatic database with huge size.

Keywords : Mixture model, Classification of distributions, Distributions of distributions, Copulas

1. Introduction

L'analyse des données symboliques (ADS, voir Bock, Diday, 2000, [2]), a pour but d'étendre l'analyse des données classiques à des données où les individus sont munis de variations internes. Ainsi en ADS, chaque variable peut prendre des valeurs multiples, intervalles, lois de probabilité, fonctions de répartition, munies parfois de règles et de taxonomies. Dans ce cadre général on cherche ici à étendre la décomposition de mélange de lois, où les variables sont à valeurs numériques, à des variables dont les valeurs sont des fonctions de répartition.

Le problème classique de décomposition de mélange consiste à estimer une densité de probabilité à partir d'un échantillon donné, en considérant que la densité cherchée est un mélange fini de K densités. Nous disposons donc d'un échantillon de N individus dans \mathfrak{R}^p dont la loi a pour densité

$$f(x_1, \dots, x_p) = \sum_{l=1}^K p_l f(x_1, \dots, x_p, \alpha_l) \quad (1)$$

avec

- $f(\cdot, \alpha)$ est une densité de probabilités de paramètre α appartenant à \mathfrak{R}^d (d est le nombre de coordonnées de α),
- $\forall l = 1, \dots, K, 0 < p_l < 1$ et $\sum_{l=1}^K p_l = 1$,
- et p_l est la probabilité qu'un point de l'échantillon suive la loi de densité $f(\cdot, \alpha_l)$.

Ce problème a été étudié par de nombreux auteurs sous deux approches différentes. L'approche la plus répandue consiste à traiter un problème d'estimation des paramètres (p_l, α_l) (« approche estimation ») (Dempster, Laird et Rubin, 1977, [7], Everitt et Hand, 1981, [10]). Les méthodes les plus classiques relèvent des techniques d'estimation par le maximum de vraisemblance. Les algorithmes d'optimisation de la vraisemblance sont, à des variantes près, de type EM (Estimation, Maximisation) (voir Dempster, Laird et Rubin, 1977, [7], Redner et Walker, 1984, [16], Shlezinger, 1968, [19]). La seconde approche (« approche classification ») considère la recherche d'une partition $P = (P_1, \dots, P_K)$ telle que chaque classe P_l soit assimilable à un sous-échantillon suivant la loi $f(\cdot, \alpha_l)$, le nombre K de composants du mélange est supposé connu (Diday, Ok et Schroeder, 1974, [8], Scott et Symons, 1971, [18], Symons, 1981, [22]). Dans ce cadre, les algorithmes utilisés sont de type nuées dynamiques. Dans cet article, nous nous intéressons à une généralisation de l'approche classification : la décomposition de mélange de lois lorsque les données se trouvent être des distributions de probabilités. Après une brève présentation de notions utiles telles que les distributions de distributions et les copules, nous explicitons la méthode proposée et l'appliquons à une grande base de données climatiques.

2. Décomposition de mélange appliquée aux distributions

2.1. Les entrées et les sorties

Soit $W = (w_1, \dots, w_N)$ un ensemble d'individus statistiques décrits par p variables. L'ensemble W est un échantillon de N individus d'une population totale inconnue Ω . Pour chaque variable, chacun des individus de Ω est décrit par une distribution de probabilité. Nous disposons donc de $\mathfrak{F} = (F_1, \dots, F_N)$ décrivant W , avec $F_i = (F_i^1, \dots, F_i^p)$. Chaque F_i^j correspond à la distribution de l'individu i pour la variable j . L'ensemble \mathfrak{F} (appelé «base de distributions») est un échantillon de N p -uplets de distributions provenant de $\Omega_F = \Omega_F^1 \times \dots \times \Omega_F^p$, avec Ω_F^j l'ensemble des distributions possibles décrivant les individus de Ω pour la variable j . Les données peuvent être résumées par un tableau où chaque case (intersection d'un individu i et d'une variable j) contient une distribution de probabilités (voir Figure 1).





	Variable 1	Variable p
Ind 1		
.....
Ind N		

FIGURE 1
Tableau de données distributions

Pour plus de clarté, nous supposons par la suite que nous ne disposons que d'une variable (le tableau ne contient qu'une colonne). Dans ce cas, F_i est la distribution de l'individu i pour cette variable.

Nous voulons décomposer la loi de probabilités des données (et obtenir ainsi une classification en K classes) et modéliser les dépendances qui existent dans une variable (*i.e.* une colonne) entre deux valeurs de distribution et entre les variables elles-mêmes. Nous obtenons donc :

- une partition $P = (P_1, \dots, P_K)$ de l'échantillon,
- les proportions $(p_k)_{k=1, \dots, K}$, les paramètres de copules et de FDD (voir ci-dessous) permettant de décrire les classes de la partition P.

2.2. Distributions de distributions

Pour travailler avec des données distributions de probabilités, nous introduisons la notion de «distributions de distributions» développée par E. Diday (2001, [9]) dans

le cas empirique, et nous lui donnons un contexte probabiliste plus général (voir Vrac, 2002, [24]).

Soit Ω une population d'individus w décrits par p variables continues dont le domaine est inclu dans \mathbb{R} . Soient V_j l'ensemble des valeurs possibles pour la variable j et $V = V_1 \times \dots \times V_p$. L'ensemble V_j étant un sous-ensemble de \mathbb{R} , nous notons ν_j la σ -algèbre des boréliens sur V_j et $\nu = \nu_1 \times \dots \times \nu_p$ la σ -algèbre produit sur V . Soit l'ensemble $\Omega_F = \Omega_F^1 \times \dots \times \Omega_F^p$ avec

$$\Omega_F^j = \{F : F \text{ est une fonction de répartition unidimensionnelle sur } (V_j, \nu_j)\}.$$

Nous définissons la σ -algèbre \mathcal{A}^j sur Ω_F^j ($j = 1, \dots, p$) par la σ -algèbre engendrée par les ensembles de la forme $A_T^x = \{F \in \Omega_F^j / F(T) \leq x\}$ pour tout $x \in [0, 1]$ et $T \in V_j$. Nous notons,

$$\mathcal{A}^j = \sigma^j(A_T^x).$$

Nous définissons \mathcal{A} la σ -algèbre de sous-ensembles de Ω_F par la σ -algèbre produit :

$$\mathcal{A} = \mathcal{A}^1 \times \dots \times \mathcal{A}^p.$$

Nous disposons alors de l'espace mesurable (Ω_F, \mathcal{A}) . Dans notre étude, nous disposons d'une variable aléatoire X , qui à tout w de Ω associe le p -uplet de fonctions de répartition $X(w) = F_w = (F_w^1, \dots, F_w^p) \in \Omega_F$:

$$\begin{aligned} X : (\Omega, \mathcal{M}, \mathbb{P}) &\longrightarrow (\Omega_F, \mathcal{A}) \\ w &\longmapsto F_w \in \Omega_F, \end{aligned}$$

avec \mathcal{M} une σ -algèbre de Ω et \mathbb{P} une mesure de probabilité sur (Ω, \mathcal{M}) . Soit $\mathfrak{F} = (F_1, \dots, F_N)$ un échantillon de N réalisations indépendantes et identiquement distribuées de la variable aléatoire à valeurs dans Ω_F . L'ensemble \mathfrak{F} décrit donc $W = (w_1, \dots, w_N)$, N individus de Ω avec $F_i = (F_i^1, \dots, F_i^p)$. Chaque F_i^j correspond à la fonction de répartition de l'individu i pour la variable j . L'ensemble \mathfrak{F} (appelé «base de distributions») est un échantillon de N p -uplets de fonctions de répartition provenant de Ω_F .

Pour plus de clarté, nous supposons que nous ne disposons que d'une variable. Dans ce cas, $F_w = F_w^1$ est la fonction de distribution de l'individu w pour cette variable, $V = V_1$ et $\Omega_F = \Omega_F^1$.

Définition 1 (Fonction de distributions de distributions). — Une «fonction de distributions de distributions» (FDD) au point T est la fonction définie par :

$$\begin{aligned} G_T : \overline{\mathbb{R}} &\longrightarrow [0, 1] \\ x &\longmapsto G_T(x) \end{aligned}$$

avec

$$G_T(x) = \mathbb{P}(\{F \in \Omega_F / F(T) \leq x\}) \forall x \in \overline{\mathbb{R}}.$$

Si cette fonction est modélisée de manière empirique à partir de \mathfrak{F} , la FDD est :

$$G_T^{emp}(x) = \mathbb{P}(\{F_i \in \mathfrak{F} / F_i(T) \leq x\}) \tag{2}$$

$$= \frac{\text{card}(\{F_i \in \mathfrak{F} / F_i(T) \leq x\})}{\text{card}(\mathfrak{F})}. \tag{3}$$

Il est évident que x n'a un sens que s'il appartient à $[0, 1]$. Pour $x \geq 1$, $G_T(x) = 1$ et pour $x \leq 0$, $G_T(x) = 0$. Par exemple, dans la Figure 2, dans le cas empirique, si $x = 0.4$, $G_{T_1}(x)$ est le pourcentage d'individus dont la distribution associée à la variable Y prend une valeur inférieure à 0.4 au point T_1 , soit $G_{T_1}(0.4) = 3/5$ (3 individus sur 5). L'équivalent en dimension n de cette définition est nécessaire.

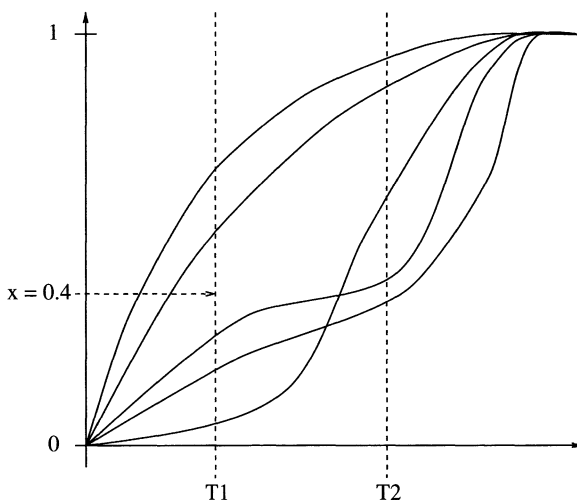


FIGURE 2
Points de distributions pour 2 valeurs de T

Définition 2 (Fonction de distributions jointes de n distributions). — Une «fonction de distributions jointes de n distributions» (FDJD) au point $T = (T_1, \dots, T_n)$ est la fonction définie par :

$$H_T : \begin{array}{ccc} \overline{\mathbb{R}}^n & \longrightarrow & [0, 1] \\ (x_1, \dots, x_n) & \longmapsto & H_T(x) \end{array}$$

avec

$$H_T(x_1, \dots, x_n) = \mathbb{P}(\{F \in \Omega_F / F(T_1) \leq x_1; \dots; F(T_n) \leq x_n\}).$$

On peut montrer que G_T est une distribution (voir Vrac, [24]) et que H_T est une fonction de répartition jointe de dimension n et de marginales G_{T_1}, \dots, G_{T_n} .

2.3. Modéliser la dépendance entre FDD à l'aide des copules

Une copule (Schweizer et Sklar, 1983, [17]) est une fonction mettant en relation la fonction de répartition multidimensionnelle avec chacune des fonctions de répartition marginales d'un n -uplet de variables aléatoires. Pour plus de clarté, nous rappelons la définition des fonctions copules (cf. [15]) ainsi que le théorème central de la théorie des copules, le théorème de Sklar.

Définition 3. — Une copule n -dimensionnelle (ou n -copule) C est une fonction de $[0, 1]^n$ dans $[0, 1]$ avec les propriétés :

1. Pour tout u de $[0, 1]^n$,

$$C(u) = 0 \text{ si au moins une coordonnée de } u \text{ est } 0, \quad (4)$$

$$\text{et si toutes les coordonnées de } u \text{ sont } 1 \text{ sauf } u_k \text{ alors } C(u) = u_k; \quad (5)$$

2. Pour tout a et b de $[0, 1]^n$ tels que $a \leq b$, $V_C([a, b]) \geq 0$, avec

$$V_C([a, b]) = \Delta_a^b C(t) = \Delta_{a_n}^{b_n} \Delta_{a_{n-1}}^{b_{n-1}} \dots \Delta_{a_1}^{b_1} C(t),$$

et

$$\Delta_{a_k}^{b_k} C(t) = C(t_1, \dots, t_{k-1}, b_k, t_{k+1}, \dots, t_n) - C(t_1, \dots, t_{k-1}, a_k, t_{k+1}, \dots, t_n)$$

est la différence d'ordre un de C pour la $k^{\text{ème}}$ composante.

Théorème 1 (Sklar, 1959, [21]). — Soit H une distribution n -dimensionnelle de marginales unidimensionnelles F_1, \dots, F_n . Alors il existe une copule C telle que pour tout (x_1, \dots, x_n) de \mathbb{R}^n :

$$H(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)). \quad (6)$$

Si les fonctions de distribution F_i sont continues, de domaines notés $\text{Dom}(F_i)$, C est unique ; sinon C est uniquement déterminé sur $\text{Dom}(F_1) \times \dots \times \text{Dom}(F_n)$. Inversement si C est une copule et F_1, \dots, F_n des distributions, la fonction H définie par (6) est une distribution jointe de marginales F_1, \dots, F_n .

La modélisation par copules des dépendances est basée sur la proposition suivante découlant du théorème de Sklar ([21] et [15]).

Proposition 1 ([9], [23]). — Soient H une FDJD et G_{T_1}, \dots, G_{T_n} , ses FDD marginales en T_1, \dots, T_n . Alors il existe une copule C telle que $\forall (x_1, \dots, x_n) \in \overline{\mathbb{R}}^n$,

$$H(x_1, \dots, x_n) = C(G_{T_1}(x_1), \dots, G_{T_n}(x_n)). \quad (7)$$

De plus, C est uniquement déterminé sur $\text{Dom}(G_{T_1}) \times \dots \times \text{Dom}(G_{T_n})$.

La décomposition de mélange peut s'appliquer du fait que H est une distribution jointe et nous avons

$$H(x_1, \dots, x_n) = \sum_{k=1}^K p_k H_k(x_1, \dots, x_n, \alpha_k) \quad (8)$$

avec $\forall k = 1, \dots, K$, $0 < p_k < 1$ et $\sum_{k=1}^K p_k = 1$, où $H_k(\dots, \alpha_k)$ est une fonction de répartition de paramètre α_k appartenant à \mathbb{R}^d (d est le nombre de coordonnées de α_k) et p_k est la probabilité qu'un point de l'échantillon suive la loi $H_k(\dots, \alpha_k)$.

À partir de la proposition précédente, la décomposition (8) devient :

$$H(x_1, \dots, x_n) = \sum_{k=1}^K p_k C_k(G_{T_1}^k(x_1, b_1^k), \dots, G_{T_n}^k(x_n, b_n^k), \beta_k), \quad (9)$$

avec

- β_k , le paramètre de la copule de la classe k ,
- $G_{T_i}^k(\cdot, b)$, la FDD au point T_i de la classe k et de paramètre b ,
- b_i^k , le paramètre de la FDD au point T_i pour la classe k .

Posons $h = \frac{\partial^n H}{\partial x_1 \dots \partial x_n}$ la densité correspondant à H et $h_k = \frac{\partial^n H_k}{\partial x_1 \dots \partial x_n}$ la densité correspondant à H_k . Avec un calcul non donné ici, nous pouvons montrer ([24]) que la densité h_k s'écrit :

$$h_k(x_1, \dots, x_n) = \left(\prod_{i=1}^n \frac{dG_{T_i}^k}{dx}(x_i) \right) \times \frac{\partial^n C_k}{\partial u_1 \dots \partial u_n}(G_{T_1}^k(x_1), \dots, G_{T_n}^k(x_n), \beta_k). \quad (10)$$

En travaillant sur les densités et non les distributions, à partir des équations (1) et (10), $h(x_1, \dots, x_n)$ s'écrit

$$\sum_{k=1}^K p_k \left(\prod_{i=1}^n \frac{dG_{T_i}^k}{dx}(x_i, b_i^k) \right) \times \frac{\partial^n C_k}{\partial u_1 \dots \partial u_n}(G_{T_1}^k(x_1, b_1^k), \dots, G_{T_n}^k(x_n, b_n^k), \beta_k).$$

Dans la suite, nous disposons de la base de distributions $\mathfrak{F} = \{F_1, \dots, F_N\}$ décrivant les individus $\{w_1, \dots, w_N\}$ pour une variable (*i.e.* F_i est la distribution de cette variable pour l'individu i). Nous calculons pour chaque individu i la valeur de sa distribution en T_1, \dots, T_n . Nous avons donc l'ensemble $\{(x_1^1, \dots, x_n^1), \dots, (x_1^N, \dots, x_n^N)\}$ avec $x_j^i = F_i(T_j)$ et $h_k(x_1^j, \dots, x_n^j)$ est notée $h_k(w_j)$.

3. L'algorithme de classification

Nous proposons ici une extension aux données distributions de la méthode de décomposition de mélange de lois par nuées dynamiques (Diday, Ok, Schroeder, 1974, [8]). À chaque étape, nous déterminons les paramètres de copules qui décrivent au mieux les classes de la partition courante, au sens d'un critère de qualité choisi. Pour cela, nous fixons un modèle de copules, soit non-paramétrique soit paramétrique. Il existe de nombreuses familles de copules paramétriques (appelées familles de copules Archimédienne). Nous n'utiliserons dans cette étude qu'une famille : la copule de Frank (définie dans la section 3.1.2 en dimension 2). Par ailleurs, nous avons besoins d'un critère d'adéquation entre une partition $(P_k)_{k=1,\dots,K}$ et un ensemble de copules $(C_{\beta_k})_{k=1,\dots,K}$. Le critère retenu est la log-vraisemblance classifiante

$$lvc(P, \beta) = \sum_{k=1}^K \sum_{i/w_i \in P_k} \log(h_k(x_1^i, x_2^i, \beta_k))$$

mais d'autres mesures d'adéquation peuvent donner de bons résultats. Après initialisation aléatoire d'une partition $P^0 = (P_1^0, \dots, P_K^0)$, l'algorithme est défini en deux étapes successives et itératives :

- Étape 1. Estimation des paramètres du mélange (mélange copules), qui maximisent le critère choisi (*i.e.* estimation de (p_1, \dots, p_K) , $(b_i^1, \dots, b_i^K)_{i=1,\dots,n}$ et $(\beta_1, \dots, \beta_K)$),
- Étape 2. Affectation des individus dans les nouvelles classes $(P_k)_{k=1,\dots,K}$:
 $P_k = \{\text{individus } w_i / p_k h_k(x_1^i, x_2^i, \beta_k) \geq p_m h_m(x_1^i, x_2^i, \beta_m) \forall m\}$,
avec $k < m$ en cas d'égalité.

Si nous choisissons une copule non-paramétrique, cet algorithme peut s'appliquer de la même manière (voir remarque, section 3.2).

3.1. Estimation des proportions du mélange et des paramètres de copules

3.1.1. Les proportions du mélange

Les proportions $(p_k)_{k=1,\dots,K}$ du mélange peuvent s'estimer très classiquement. Comme dans l'algorithme de base des nuées dynamiques, nous utilisons l'estimation $p_k = \frac{\text{card}(P_k)}{\text{card}(F)}$. Cependant, différentes variantes sont envisageables (Celeux, Govaert, 1993, [3]).

3.1.2 Les paramètres de copules

L'estimation des paramètres β de copules se fait par maximisation de $lvc(P, \beta)$, c'est-à-dire de

$$\sum_{\omega=(x_1,\dots,x_n) \in P_k} \log\left[\left(\prod_{i=1}^n \frac{dG_{T_i}^k}{dx}(x_i, b_i^k)\right) \times \frac{\partial^n C_k}{\partial u_1 \dots \partial u_n}(G_{T_1}^k(x_1, b_1^k), \dots, G_{T_n}(x_n, b_n^k), \beta_k)\right].$$

Même en dimension 2, ces équations sont assez complexes. Par exemple, la copule de Frank (cf. [15]) s'écrit :

$$C_{\beta}(u, v) = \frac{\log\left(1 + \frac{(\beta^u - 1)(\beta^v - 1)}{(\beta - 1)}\right)}{\log(\beta)},$$

avec β strictement positif et $\beta \neq 1$ et $\forall u, v \in [0, 1]$. Elle a les propriétés suivantes (cf. [11]) :

$$\lim_{\beta \rightarrow 0} C_{\beta}(u, v) = \min(u, v), \lim_{\beta \rightarrow 1} C_{\beta}(u, v) = uv \text{ (correspondant à l'indépendance)}$$

$$\text{et } \lim_{\beta \rightarrow \infty} C_{\beta}(u, v) = \max(u + v - 1, 0).$$

Cette copule Archimédienne ([15]) est engendrée par l'application

$$\phi_{\beta}(t) = -\log\left(\frac{1 - \beta^t}{1 - \beta}\right) \text{ et on a}$$

$$\frac{\partial^2 C_{\beta}}{\partial u \partial v}(u, v) = \frac{(\beta - 1) \log(\beta) \beta^{u+v}}{[(\beta - 1) + (\beta^u - 1)(\beta^v - 1)]^2}, 0 \leq u, v \leq 1.$$

Les paramètres β ne peuvent donc pas être déterminés par une formulation analytique. La résolution des équations de log-vraisemblance classifiante peut par exemple se faire par des techniques d'estimation numérique.

3.2. Modélisation et estimation des FDD

L'estimation des FDD se résume à l'estimation d'une fonction de distribution de probabilités. Les techniques de détermination sont donc très nombreuses. Dans cet article, nous nous sommes intéressés à 3 variantes.

La première variante est aussi la plus simple : l'estimation empirique.

$$G_T(x) = \frac{\text{card}(\{F_i \in \mathfrak{F} / F_i(T) \leq x\})}{\text{card}(\mathfrak{F})}.$$

Cette technique a le désavantage de ne donner que certaines valeurs pour $G_T(x)$.

La deuxième est une modélisation suivant la loi béta. Celle-ci correspond à la loi de Dirichlet en dimension 1. La fonction de densité est

$$f_{a,b}(x) = \frac{x^{a-1}(1-x)^{b-1}}{\int_0^1 y^{a-1}(1-y)^{b-1} dy},$$

avec $a > 0$ et $b > 0$ les deux paramètres de la loi béta. Sa particularité essentielle pour notre étude est d'aller de $[0,1]$ dans $[0,1]$. Dans l'expression $G_T(x)$, x doit

impérativement appartenir à $[0,1]$. La FDD est obtenue par intégration de la densité béta.

La dernière est la méthode de Parzen «tronquée». À partir d'un échantillon (X_1, \dots, X_N) de taille N , une estimation de la densité peut être

$$\hat{f}(x) = \frac{1}{c_N} \frac{1}{Nh} \sum_{i=1}^N Ke\left(\frac{x - X_i}{h}\right),$$

avec c_N tel que $\int \hat{f} = 1$, Ke est le noyau ou Kernel (généralement une fonction de densité de probabilités), h est le paramètre de forme (ou de lissage) également appelé «largeur de la fenêtre» (window width) par certains auteurs ([20]). Le noyau Ke choisi est la fonction de densité normale. Cette méthode nécessite tout de même l'estimation du paramètre h mais celle-ci est réalisée de manière automatique par la formule du h optimal au sens du MISE (Mean Integrated Square Error) $h = 1.06\sigma N^{-1/5}$ où σ est l'écart-type estimé de l'échantillon ([20]). Nous ne reviendrons pas ici sur le rôle de h . La FDD est obtenue par intégration de \hat{f} .

Remarque. — Une version totalement non-paramétrique de la méthode existe ([25], [24]). Elle utilise la notion de copules empiriques et de FDD empirique. Cette méthode peut donner des résultats satisfaisants lorsqu'on s'intéresse plus à la classification qu'à une description de la partition finale. Cependant elle a l'inconvénient de converger avec une lenteur rédhibitoire. Nous ne développerons pas davantage cette méthode dans cet article (cf. [24] et [25]).

4. Classification de distributions multidimensionnelles

Différentes écritures d'une copule en dimension $n \geq 3$ existent, pour une même copule multidimensionnelle. Elles ont toutes des avantages et des inconvénients mais le point commun est la complexité ([12]). Pour ne pas accentuer la difficulté des équations à résoudre, nous proposons deux méthodes pour traiter deux distributions différentes en même temps et/ou plus de deux T_i à la fois.

4.1. Couplage

La première est la technique par couplage. Nous disposons d'un échantillon de N individus $W = (w_1, \dots, w_N)$ décrit par deux variables distributions Y^1 et Y^2 . Déterminons deux décompositions de mélange de copules : l'une sur la variable Y^1 en K_1 classes (suivant 2 seuils de FDD : T_1^1 et T_2^1), l'autre sur Y^2 en K_2 classes (suivant 2 seuils de FDD : T_1^2 et T_2^2). Pour tout w de W et pour chaque variable $(Y^i)_{i=1,2}$, nous disposons donc de la valeur de la distribution jointe

$$H^{Y^1}(w) = \sum_{k=1}^{K_1} p_k^i C_{\beta_k^i} (G_{T_1^i}^k(\omega_1^i, b_{T_1^i}^k), G_{T_2^i}(\omega_2^i, b_{T_2^i}^k))$$

avec

- $(w_1^i, w_2^i) =$ valeurs de la distribution de la variable Y^i pour w respectivement en T_1^i et T_2^i (i.e. $w_j^i = \mathbb{P}(X_j^i \leq T_j^i)$, où X_j^i est la variable aléatoire représentant l'individu ω_j pour la variable Y^i),
- $\beta_k^i =$ paramètre de copule de la classe k pour la variable Y^i ,
- $b_{T_j^i}^k =$ paramètre associé à la FDD défini en T_j^i pour la classe k (dépend de la modélisation des FDD),
- $G_{T_j^i}^k(\omega_j^i, b_{T_j^i}^k) =$ valeur de la FDD associée à la variable Y^i au point T_j^i appliqué à la valeur ω_j^i .

Nous disposons maintenant d'un couple de valeurs de distributions (H^{Y^1}, H^{Y^2}) pour chacun des N individus de W . Posons que les valeurs $(H^{Y^1}(w_i))_{i=1, \dots, N}$ et $(H^{Y^2}(w_i))_{i=1, \dots, N}$ suivent respectivement les lois de distribution F_1 et F_2 que nous pouvons estimer. Supposons que le couple (H^{Y^1}, H^{Y^2}) soit de loi jointe H . Les conditions du théorème de Sklar sont vérifiées : Il existe une copule C telle que $\forall (x_1, x_2) \in [0, 1]^2, H(x_1, x_2) = C(F_1(x_1), F_2(x_2))$. Nous pouvons par conséquent appliquer à nouveau une décomposition de mélange de lois de lois à partir des N couples obtenus. Les paramètres de copules déterminés lors des decompositions de mélange de copules sur Y^1 et Y^2 fournissent des informations sur les dépendances présentes à l'intérieur des variables (entre T_1^1 et T_2^1 et entre T_1^2 et T_2^2). Les paramètres de copules obtenues au final permettront de caractériser les dépendances qui existent alors entre les distributions des deux variables Y^1 et Y^2 .

4.2. Arbre binaire

La seconde technique consiste en une méthode d'arbre binaire. À partir de la base de distributions complète (le haut de l'arbre), nous déterminons la meilleure partition en deux classes (noeuds-fils), puis la meilleure partition en deux classes des deux noeuds-fils et ainsi de suite (cf. Figure 3). À chaque étape, pour chaque noeud et pour chaque variable, nous déterminons les (T_1, T_2) optimaux. Le noeud N à couper est celui dont la variable de classification et les (T_1, T_2) associés maximisent le critère de qualité Q de la division. Nous pouvons utiliser $Q(N) = \sum_k \sum_{\omega \in P_k} \log h_{\beta_k}(\omega)$.

La méthode par arbre binaire présente l'intérêt de pouvoir travailler avec différentes variables et de pouvoir choisir les seuils (T_1, T_2) adaptés à chaque étape ainsi que les copules appropriées à chaque noeud, comme présenté en Figure 3. Elle a néanmoins l'inconvénient d'être assez longue. En effet, pour chaque nouveau (T_1, T_2) , la méthode doit calculer tous les $F(T_i(\omega))$ pour tous les individus ω de l'ensemble associé. De plus une partition en deux classes est calculée pour chaque noeud-fils. Par ailleurs, deux individus classés séparément à une étape ne pourront pas être classés ensemble par la suite.

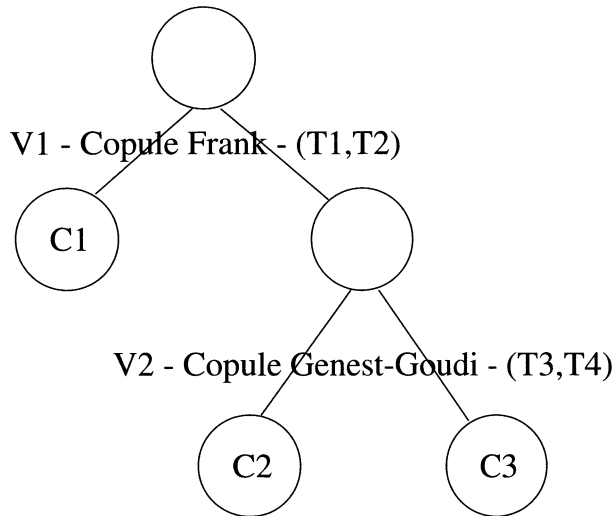


FIGURE 3
Exemple de la méthode par arbre binaire

5. Application à une base de données climatiques

L'étude du climat est un axe important de la recherche mondiale. Que ce soit du point de vue de la prévision du temps à court terme ou de celui de l'évolution dynamique du climat à longue échéance, les statistiques et l'analyse de données jouent un rôle essentiel (Davis and Walker, 1992, [6], Kalstein and al., 1993, [13]). Par exemple, dans l'algorithme d'inversion de l'équation de transfert radiatif, qui permet entre autre d'interpréter des observations satellitaires en terme de variables thermodynamiques atmosphériques (Chédin *et al*, 1985, [4], Achard, 1991, [1], Chevallier, 1998, [5]), une partition de différents types de profils atmosphériques est utilisée dans le but de déterminer une solution initiale proche de la solution vers laquelle l'algorithme va converger. La partition utilisée doit donc regrouper les profils ayant des propriétés physiques proches à l'intérieur d'une classe et bien distinctes entre les classes. La méthode présentée dans cet article permet en plus de connaître les lois de probabilités des variables et donc les probabilités d'occurrence des profils atmosphériques. Nous disposons de données atmosphériques provenant du centre Européen ECMWF (European Center for Medium range Weather Forecasting) de Reading (U.K.). Un maillage du globe terrestre est réalisé, chaque maille correspondant à un degré de latitude et un degré de longitude. Ce maillage est étendu en altitude sur 50 niveaux appelés «coordonnées sigma», relatives à la pression au sol. Pour une longitude et une latitude données, l'ensemble des valeurs en altitude est appelé «profil» vertical. Pour chaque point du maillage de l'atmosphère (une longitude, une latitude, une altitude), nous disposons des valeurs de la pression, de la température, de l'humidité, du vent,... Ces valeurs sont celles des prévisions («forecast») à six heures d'intervalle et ceci 4 fois par jour (0h, 6h, 12h, 18h,

temps universel) sur une période allant de décembre 1998 à décembre 1999. Nous possédons par conséquent un « quadrillage » tridimensionnel de l'atmosphère de la terre, représentant un an de son état thermodynamique complet 4 fois par jour. Nous souhaitons appliquer la décomposition de mélange de copules sur les données de température et/ou d'humidité des profils atmosphériques du 15 décembre 1998 à 0h. À partir de ces données, nous obtenons alors une classification des profils, ainsi que l'estimation de la loi de probabilité (marginales et jointes) des variables utilisées. Les données étant fournies sous la forme de valeurs numériques, une estimation des fonctions de distribution de probabilité de la température et de l'humidité est tout d'abord réalisée pour chacun des profils. L'estimation est faite en supposant que les données numériques d'un profil atmosphérique sont d'une même loi de probabilité. Ces données sont « mises à plat » sur la droite réelle et la distribution de probabilité est estimée en appliquant la méthode des noyaux de Parzen. Chaque individu « profil atmosphérique » (situé à chaque intersection longitude × latitude) est caractérisé par deux distributions, l'une de température, l'autre d'humidité. Deux valeurs de température T_1 et T_2 et deux valeurs d'humidité H_1 et H_2 ont été fixées par une connaissance *a priori* pour l'estimation des FDD $(G_{T_i}(x))_{i=1,2}$ et $(G_{H_i}(x))_{i=1,2}$. Les valeurs choisies sont $T_1 = 225K$, $T_2 = 265K$ ($K = \text{degrés Kelvin}$) et $H_1 = 0.00003Kg/Kg$, $H_2 = 0.006Kg/Kg$ ($Kg/Kg = Kg \text{ de vapeur d'eau par } Kg \text{ d'air}$). Le nombre de profils étant élevé (360 longitudes × 180 latitudes = 64800 profils), nous ne prenons qu'une longitude sur deux et une latitude sur deux, soient (360/2) longitudes × (180/2) latitudes = 16200 profils.

5.1. Résultats

Nous classons tout d'abord les profils de température en 7 classes à partir de la copule de Frank et avec une FDD modélisée suivant une loi bêta de paramètres ν_1 et ν_2 . Nous disposons d'un tableau d'une colonne et 16200 lignes. À partir des 16200 distributions, l'algorithme converge en 2 itérations vers la classification représentée en Figure 4 où chaque carré en pointillés représente une étendue de 30° de longitude et 30° de latitude. Chaque pixel est grossi artificiellement de manière à obtenir un effet visuel continu. Les paramètres de copules et de FDD sont donnés dans le Tableau 1.

TABLEAU 1
Paramètres de la classification en 7 classes en température

Classes	β	ν_1 en T_1	ν_2 en T_1	ν_1 en T_2	ν_2 en T_2
1	0.000001	6.836969	14.342546	12.208704	2.217218
2	0.300001	11.408380	69.945442	21.956680	14.064272
3	0.004093	12.180901	70.0	61.601810	70.0
4	0.000001	12.651747	70.0	56.703354	70.0
5	0.000001	13.335871	70.0	11.891472	11.261731
6	0.030567	6.040135	25.066311	8.938780	3.687328
7	0.007445	8.839353	22.021719	19.165813	2.168266

7 classes en température par DMC, 15/12/98 0H

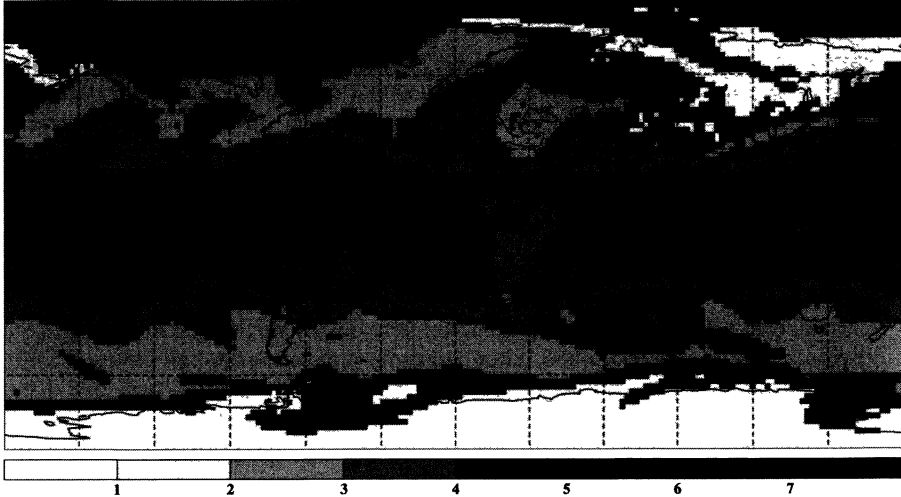


FIGURE 4

*Classification en 7 classes par DMC sur la température
($T_1 = 225\text{ K}$, $T_2 = 265\text{ K}$) pour le 15 décembre 1998 à 0H*

Les classes de la partition semblent cohérentes et posséder des propriétés climatiques distinctes et réalistes. On peut retrouver par exemple une grande classe dite «tropicale» (classe 4), deux classes «polaires» correspondant à l'été dans l'hémisphère sud et l'hiver dans l'hémisphère nord (classes 1 et 7), deux classes «tempérées» (classes 2 et 5). La classe 3 fait le lien entre les zones tempérées et les zones tropicales tandis que la classe 6 fait le lien entre les zones tempérées et les zones polaires. De plus, les classes 1, 4 et 5 ont des paramètres de copules identiques (0.000001) ce qui signifie que leur copule est proche de la copule *min*. Autrement dit, les distributions à l'intérieur de chacune des classes ont tendance à évoluer parallèlement sans se couper.

Par ailleurs, une comparaison intéressante peut être faite avec le tracé de la température moyenne entre 500 et 700 hPa ($1\text{hPa} = 1\text{millibar}$), présentée en Figure 5.

Nous voyons que les classes de transitions entre les classes tempérées et tropicales de la Figure 4 correspondent bien à des zones de transition de la température 500-700 hPa de la Figure 5. Les «lagues» (incursions d'air chaud dans des masses d'air plus froid ou inversement) sont bien identifiées. Le disque situé à 60 degrés N \times 60 degrés E (classe 7) sur la Figure 4, s'explique parfaitement par la carte de la Figure 5 et correspond à une dépression. L'accord avec l'analyse synoptique de la situation est très bon (formes des incursions, position des dépressions creuses, etc.).

Un autre manière de décrire les classes est de regarder la fonction de densité de probabilité pour la variable température à différents niveaux de pression de l'atmosphère. Ces densités, tracées par la méthode des noyaux de Parzen, donnent des

15/12 0H Temperature moyenne (700-500 hPa)

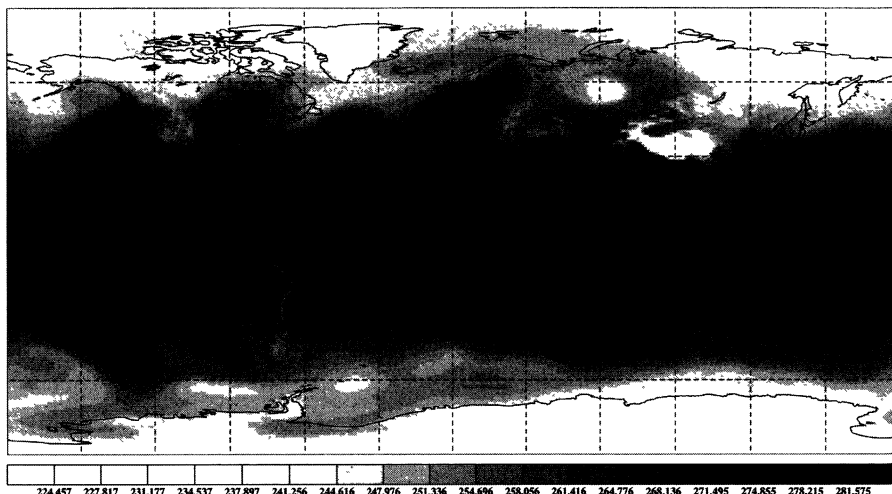


FIGURE 5

Température moyenne entre 500 et 700 hPa pour le 15 décembre 1998 à 0H

informations sur la répartition des températures dans chaque classe et donc sur leur discrimination. Nous voyons par exemple sur les Figures 7 et 8 (correspondant aux niveaux 900 hPa et 500 hPa respectivement) que les classes sont bien discriminées (l'association classes-densités est donnée en Figure 6).

classe 1	2230 ind	— —
classe 2	2655 ind	---×---
classe 3	1411 ind	---*---
classe 4	4426 ind□.....
classe 5	1866 ind	---■---
classe 6	1584 ind	---○---
classe 7	2208 ind	---●---

FIGURE 6

Association 7 classes - densités température

Densités de la var $0.5(T46+T47)$ - 900 hPa, 7 classes en Temp par copules

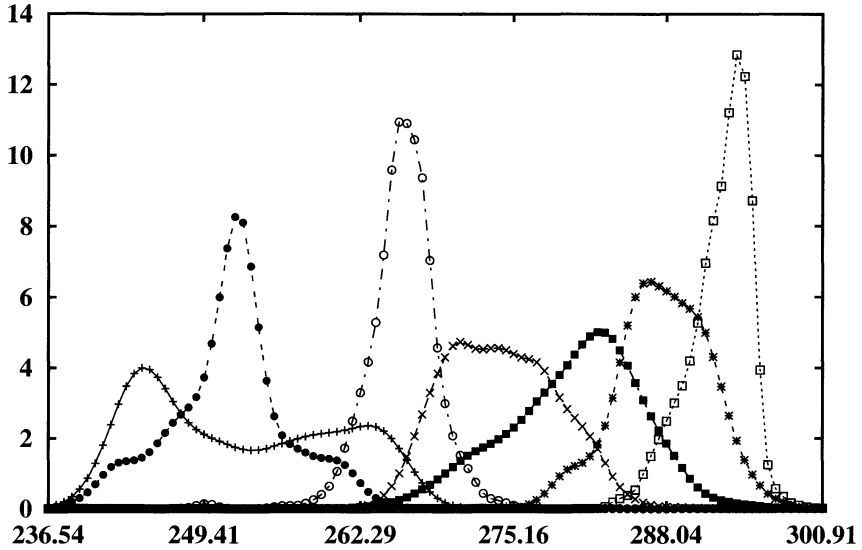


FIGURE 7

Densités de la température (Kelvin) par classe à 900 hPa

Densités de la var T37 - 500 hPa, 7 classes en Temp par copules

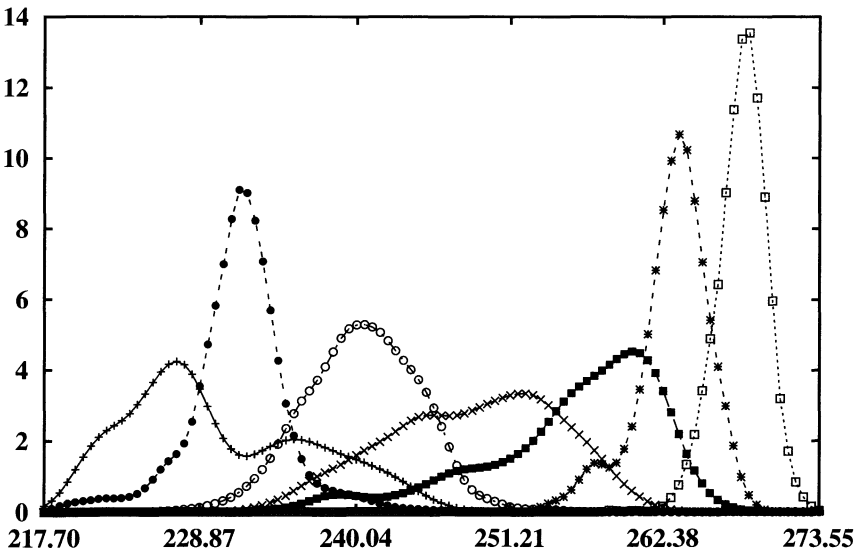


FIGURE 8

Densités de la température (Kelvin) par classe à 500 hPa

Nous retrouvons également que les classes 1 et 7 sont les deux classes froides de la classification et que la classe 4 est la plus chaude de toutes. Nous voyons parfaitement les différences entre les classes dites de transition (ou classes plus tempérées). D'après la classification de la Figure 4, nous pouvons trier ces classes par température moyenne croissante, de la plus proche des classes polaires (1 et 7) à la plus proche de la classe tropicale (classe 4) : classe 6, classe 2, classe 5, classe 3. Cet ordre est totalement vérifié par le calcul des densités.

Cependant, plus nous montons en altitude (*i.e.* plus la pression diminue), moins les classes se discriminent entre-elles : les densités empiètent davantage les unes sur les autres. C'est en particulier le cas au-dessus de la tropopause (niveau de recroissance de la température), par exemple à 70 hPa (voir Figure 9). La classe 4, tropicale, reste très distincte, et, cette fois, devient la plus froide, comme attendu.

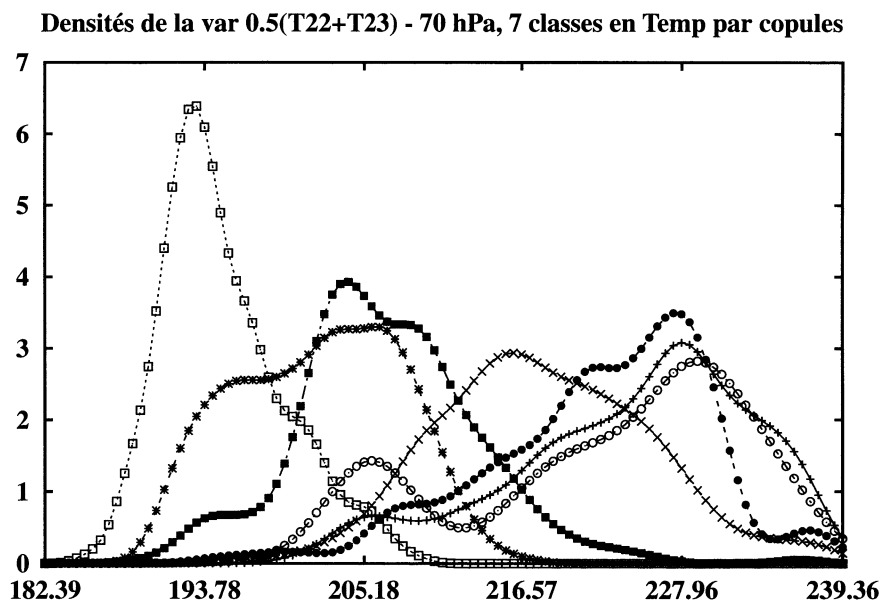


FIGURE 9
Densités de la température (Kelvin) par classe à 70 hPa

Une classification similaire en sept classes a été lancée sur les 16200 distributions de la variable humidité. Les paramètres obtenus se trouvent dans le tableau 2 et la classification représentée sur la Figure 10.

Le résultat visuel est moins organisé que pour la température. Cet effet était prévisible du fait de la plus grande variabilité de l'humidité.

Cette partition peut être décrite par référence à la quantité totale de vapeur d'eau intégrée verticalement par profil et donné par la Figure 11.

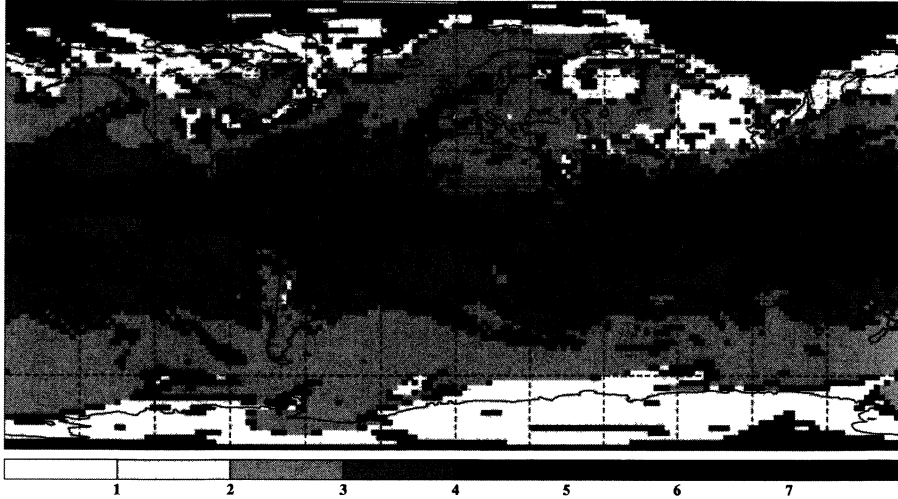
7 classes en humidite par DMC, 15/12/98 0H

FIGURE 10

*Classification en 7 classes par DMC sur l'humidité
($H_1 = 0.00003 \text{ kg/kg}$, $H_2 = 0.006 \text{ kg/kg}$) pour le 15 décembre 1998 à 0H*

TABLEAU 2

Paramètres de la classification en 7 classes en humidité

Classes	β	ν_1 en T_1	ν_2 en T_1	ν_1 en T_2	ν_2 en T_2
1	0.200001	1.772749	36.192062	70.0	24.484861
2	0.016939	6.292977	742.659424	30.004604	13.469539
3	0.619099	0.000001	12.617126	16.619267	20.368376
4	0.445017	0.000001	12.617126	29.424589	48.242210
5	0.100001	0.000001	12.617126	6.890862	5.887025
6	0.020641	0.000001	12.617126	38.931557	14.840351
7	0.017804	2.215375	23.266142	70.0	18.847424

Les points communs entre la carte de la Figure 10 et celle de la Figure 11 sont multiples. Les bras d'incursions d'air humide sont précisément définis et la plupart des formes sont retrouvées avec une précision étonnante. On peut remarquer que la classe tropicale précédente est scindée en deux classes (3 et 4). La classe 4 correspond aux zones les plus humides de la figure 11. De plus, toute la frontière entre la classe 3 et une masse d'air moins humide (classe 2) est bordée par la classe 5 qui correspond à

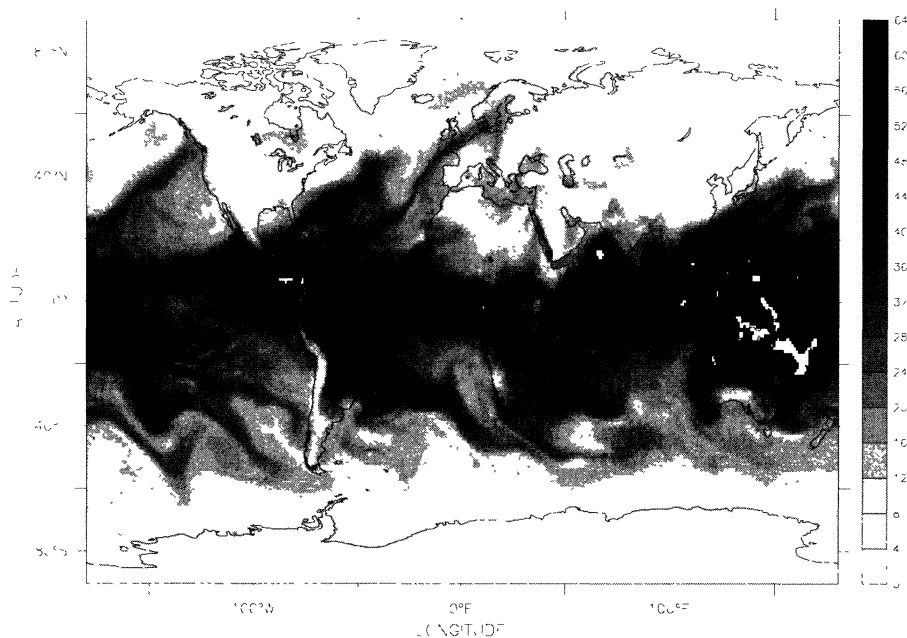


FIGURE 11

Total Column Water Vapor (Quantité totale de vapeur d'eau intégrée, Kg/m^2) pour le 15 décembre 1998 à 0 H

des incursions d'air humide dans un milieu plus sec. Par ailleurs, la méthode semble avoir identifié deux classes différentes de très faible humidité (classes 1 et 7).

On peut voir que malgré des valeurs d'humidité du même ordre, ces deux classes ont des paramètres de copules différents (0.2 et 0.018) traduisant un comportement différent de leurs distributions. On voit également une « spirale » située à (60° N, 60° E) qui correspond parfaitement à la dépression centrée sur cette zone (cette particularité se retrouve sur la classification en température sous la forme d'un disque rouge : classe 7).

Le tracé des densités de chaque classe à différents niveaux de pression (altitude) nous donne également des informations sur la manière dont les classes se discriminent. Nous voyons dans la Figure 13, correspondant aux densités à 900 hPa, que la classe 1, avec son « aile » de densité, plus humide correspond à des situations plus fréquentes que la classe 7, très sèche et associée à des situations de type « hiver polaire ». De plus, cette figure montre que la discrimination est très nette pour les niveaux bas de l'atmosphère.

Pour les niveaux plus élevés, l'humidité diminue extrêmement rapidement, conduisant les densité à se superposer progressivement, diminuant ainsi leur pouvoir discriminant. Cependant, nous retrouvons parfaitement ce que la comparaison avec la quantité totale de vapeur d'eau nous laissait penser :

- les classes 1 et 7 ont les densités dont le mode statistique est le plus faible (*i.e.* ce sont les classes les plus sèches),
- la classe 4 est la classe la plus humide avec une densité ayant un pic vers 0.015 kg/kg,
- les classes tempérées et de transition peuvent s'ordonner de la plus sèche à la plus humide : classe 3, classe 5, classe 2, classe 6.

classe 1 2614 ind	—+—
classe 2 4397 ind	- - - x - - -
classe 3 2006 ind	- - - * - - -
classe 4 3507 ind	- ···· □ ····
classe 5 653 ind	- - - ■ - - -
classe 6 968 ind	- ···· ○ ····
classe 7 2235 ind	- ···· ● ····

FIGURE 12
Association classes – densités humidité

Densités de la var 0.5(H46+H47) - 900 hPa, 7 classes en Hum par copules

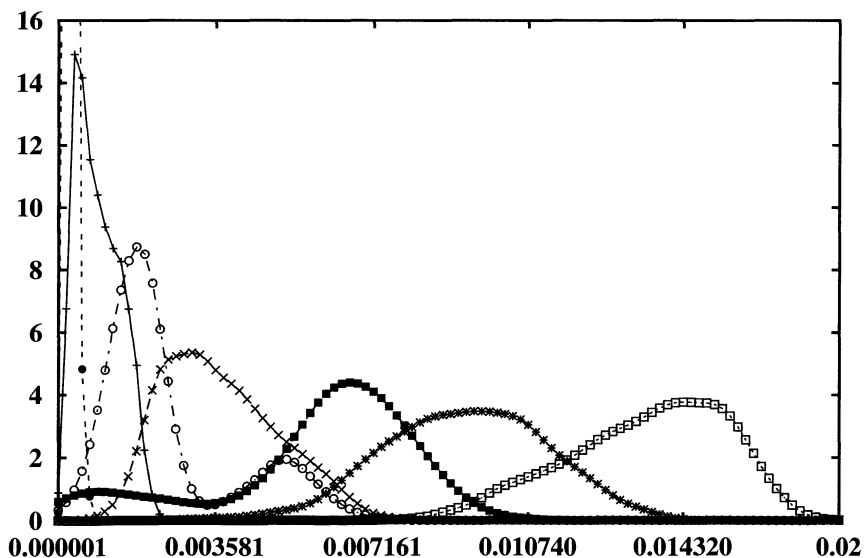


FIGURE 13
Densités de l'humidité spécifique (kg/kg) par classe à 900 hPa

Les classifications obtenues par décomposition de mélange de copules (DMC) sur la variable température, puis sur la variable humidité, sont très bonnes au regard de la connaissance *a priori* dont nous disposons et de la cohérence physique des classes. Qu'en est-il d'une classification couplant les deux variables ?

À partir des deux classifications précédentes en température et en humidité, nous appliquons la méthode de couplage afin d'obtenir une partition qui tienne compte des deux variables physiques. La modélisation est faite par copules de Frank et par lois béta. La classification est projetée Figure 14 et ses paramètres sont donnés dans le Tableau 3.

7 classes en température et humidité par couplage DMC, 15/12/98 0H

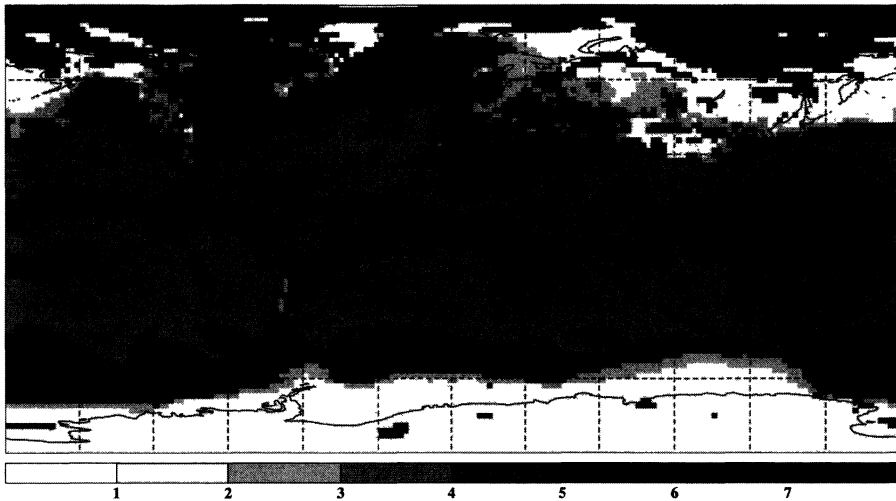


FIGURE 14
 Classification en 7 classes par couplage de DMC sur la température
 et l'humidité pour le 15 décembre 1998 à 0H

TABLEAU 3
 Paramètres de la classification en 7 classes en température et humidité

Classes	β	ν_1 en T_1	ν_2 en T_1	ν_1 en T_2	ν_2 en T_2
1	0.000001	6.712667	2.140064	5.703492	5.222391
2	0.100001	70.0	70.0	10.42458	14.541202
3	0.200001	18.965822	88.125916	8.056098	145.218979
4	0.050867	19.533854	112.066284	6.489847	357.520905
5	0.362295	12.315609	31.493969	5.033236	18.545059
6	0.126157	0.86489	7.177879	3.316219	7.178005
7	0.003896	23.222773	4.773149	13.366582	3.108013

La méthode de couplage semble donner de bons résultats : mélange cohérent des deux précédentes classifications. Nous retrouvons une différence entre l'été de l'hémisphère sud et l'hiver de l'hémisphère nord (classes 1 et 7) induite par la variable température avec des variations dues à la variable humidité. De plus, deux classes tropicales sont identifiées (classe 4 très humide et classe 3 un peu moins). La classe 4 correspond beaucoup mieux que la classe 4 de la figure 10, aux zones les plus humides de l'analyse. L'accord avec le contenu intégré en vapeur d'eau est quasiment parfait. Les autres classes décrivent la transition des classes tropicales (chaudes et humides) aux classes polaires (froides et sèches). La « spirale » (60° N, 60° E) est toujours présente et est significative de la dépression centrée sur cette zone.

De manière générale, les détails, induits par la température ou par l'humidité, restent d'une grande précision : nous pouvons voir, par exemple, une zone plus froide et plus sèche que son voisinage (classes 5 et 6), dans le sud Australien, parfaitement visible sur la figure 11.

Si on compare les densités de probabilité de la variable température à 900 hPa de la classification en température (Figures 4 et 7) avec celle des classes obtenues par couplage (Figures 14 et 16), ces dernières ont un pouvoir discriminant un peu inférieur.

classe 1 3229 ind	—+—
classe 2 937 ind	---×---
classe 3 4146 ind	---*---
classe 4 2402 ind□.....
classe 5 2263 ind	---■---
classe 6 1453 ind	---○---
classe 7 1950 ind	---●---

FIGURE 15

Association classes – densités couplage (température, humidité)

Cet effet était prévisible du fait que la partition par couplage tient compte des deux variables (température et humidité). Les classes restent cependant relativement bien distinctes en température et ceci jusqu'à la tropopause ; les densités à 900 hPa (Figure 16) et à 300 hPa (Figure 17) empiètent davantage les unes sur les autres que sur la Figure 7 mais chaque classe ressort clairement.

Au-dessus de la tropopause (*i.e.* dans la stratosphère), les densités sont un peu moins distinctes les unes des autres. La Figure 18 des densités à 70 hPa illustre cet effet.

Nous pouvons voir sur cette figure que la classe 4, qui est la plus chaude entre 900 et 300 hPa, est la plus froide à 70 hPa. Ce phénomène est également présent pour la classe 3 : classe plus chaude que la plupart des classes dans les parties basses

Densités de la var 0.5(T46+T47) - 900 hPa, 7 classes par couplage temp et hum

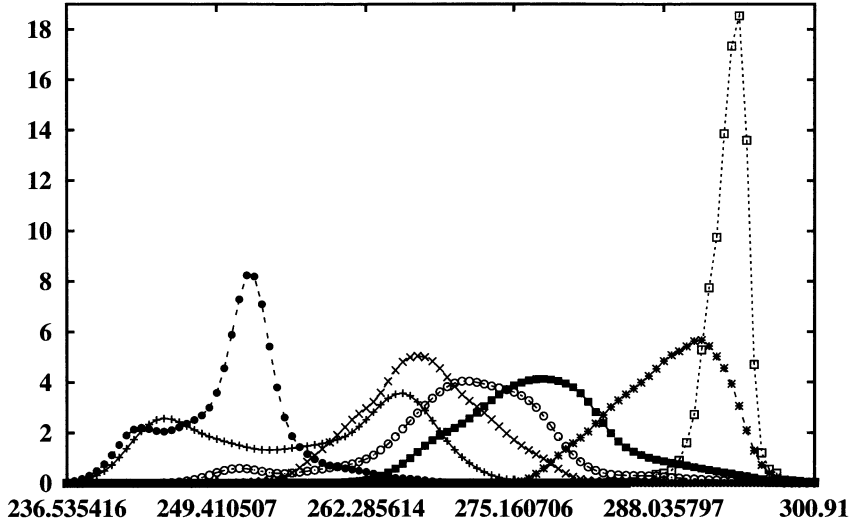


FIGURE 16
Densités de la température (Kelvin) par classe à 900 hPa

Densités de la var T32 - 300 hPa, 7 classes par couplage temp et hum

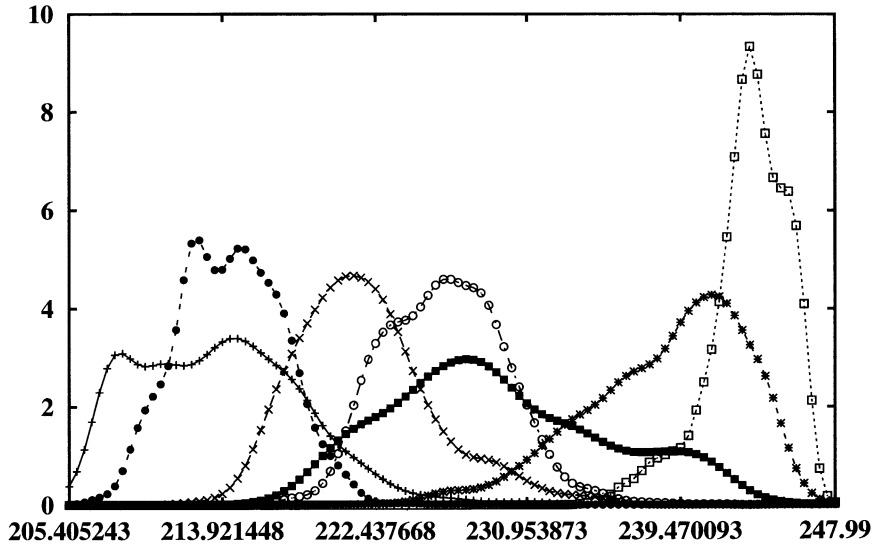


FIGURE 17
Densités de la température (Kelvin) par classe à 300 hPa

Densités de la var $0.5(T22+T23)$ - 70 hPa, 7 classes par couplage temp et hum

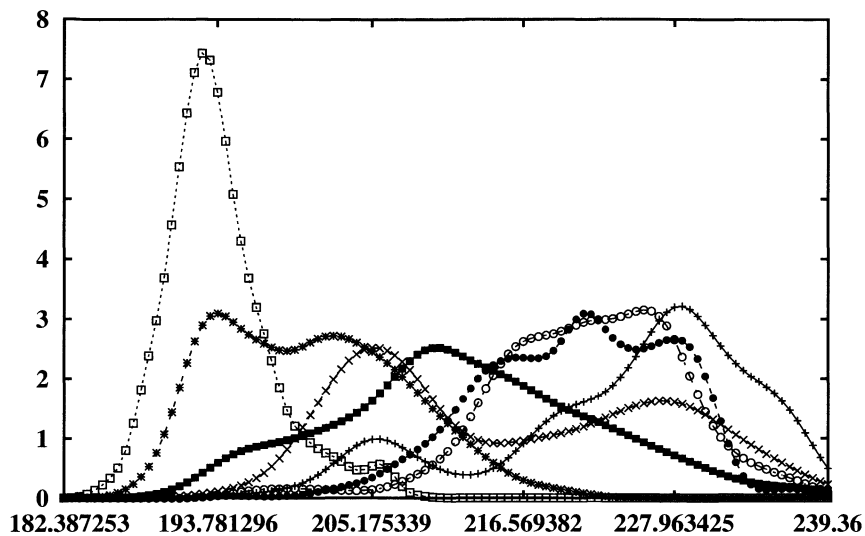


FIGURE 18

Densités de la température (Kelvin) par classe à 70 hPa

Densités de la var $0.5(H46+H47)$ - 900 hPa, 7 classes par couplage temp et hum

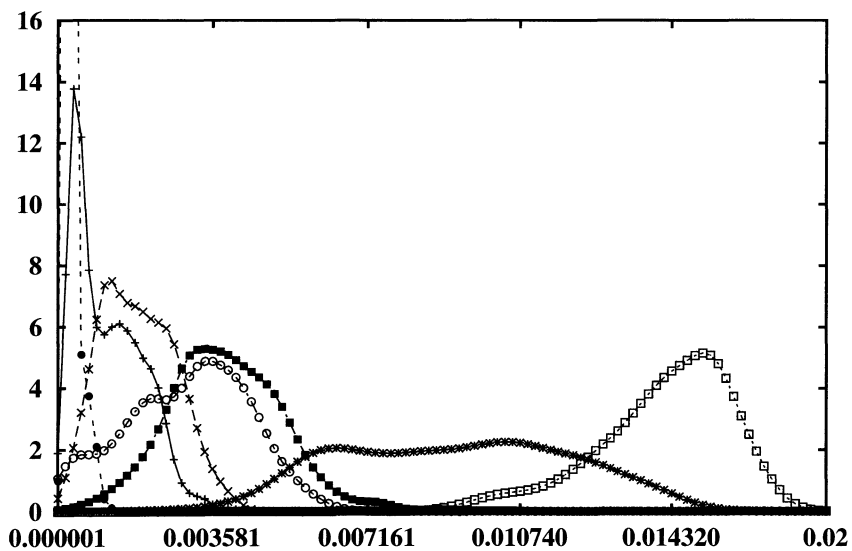


FIGURE 19

Densités de l'humidité spécifique (kg/kg) par classe à 900 hPa

de l'atmosphère, plus froide que la plupart lorsque la pression diminue. L'inversion, attendue pour ces classes tropicales, est parfaitement mise en évidence.

Par ailleurs, de même que pour la température, les densités de probabilité de l'humidité à 900 hPa (Figure 19) des classes par couplage ont une discrimination inférieure à celles des classes en humidité seule (Figure 13).

Elles restent cependant, pour la plupart, relativement distinctes, et décrivent bien les caractéristiques d'humidité des classes (classe 4 humide, classes 1 et 7 sèches, etc.). Ces classes, issues de la méthode par couplage, sont aussi beaucoup plus riches puisqu'elles décrivent des comportements liant la température et la vapeur d'eau, les deux variables essentielles pour la description d'une situation synoptique. La classification ainsi obtenue est en effet en très bon accord avec celle du 15 décembre 1998.

5.2. Comparaisons

Afin de démontrer les performances de la méthode par décomposition de mélange de copules et la cohérence de ses résultats, une autre méthode plus «classique» de classification est appliquée sur les données climatiques : la méthode dite de «classification mixte». Cette stratégie de classification, développée par Mollière (1985, [14]) et décrite par la figure 20, est traditionnellement bien adaptée à la classification de très grands ensembles de données en groupes homogènes. Elle consiste en la combinaison d'une Analyse en Composantes Principales (ACP), de la méthode des nuées dynamiques avec auto-validation, suivie de la méthode de classification ascendante hiérarchique des voisins réciproques dont les résultats sont eux-mêmes stabilisés par nuées dynamiques (pour davantage de détails, voir Vrac, 2002, [24]). Cette procédure, assez complexe, est considérée comme très efficace.

La classification mixte est donc lancée sur toutes les données (1 degré par 1 degré) du 15 décembre 1998 à 0H. Les individus sont les profils atmosphériques décrits par 62 valeurs numériques : les 37 premières coordonnées sigma de la variable de température (jusqu'à $P = 10$ hPa si $P_{sol} = 1013$ hPa) et les 24 premières coordonnées sigma de la variable d'humidité spécifique (jusqu'à $P = 155$ hPa si $P_{sol} = 1013$ hPa). Nous ne prenons que ces variables car les données de base sont des prévisions, et l'exactitude des prévisions, passée une certaine altitude, est contestable. Les valeurs sont centrées et normées par variable. Les données sont pondérées pour donner un poids égal à 1 à l'ensemble des données température et un poids de 1 à l'ensemble des données humidité. L'algorithme tient donc autant compte de la température que de l'humidité. La partition en 7 classes ainsi obtenue se trouve Figure 21. Cette classification apparaît immédiatement beaucoup plus «rustique» que celle de la Figure 14. Les classes de type tropical semblent être retrouvées (classes 5, 6 et 7). Cependant, nous pouvons voir sur cette classification un bras d'air chaud et humide partant du sud de la Floride vers le nord-est. Un coup d'oeil à la carte du contenu total intégré en vapeur d'eau du 15/12/98 à 0H (Figure 11) nous permet de voir que ce bras existe, mais qu'il en existe également un second parallèle plus à l'est qui est totalement absent. Ces deux incursions sont, par contre, bien présentes dans la classification couplée en 7 classes par la méthode par copules (Figure 14).

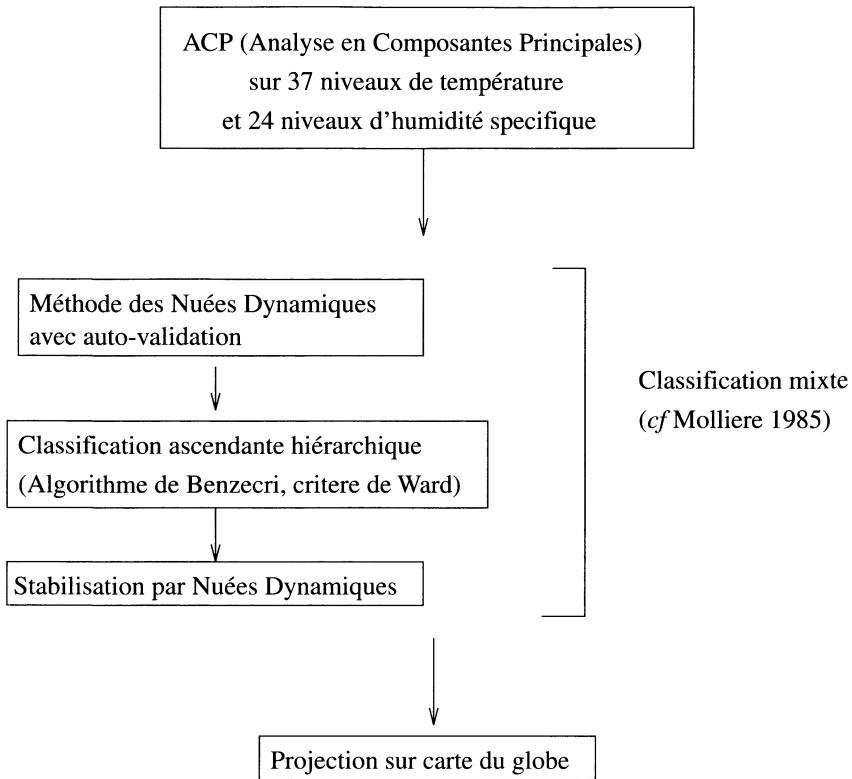


FIGURE 20
Stratégie de classification mixte

De plus, les classes autres que tropicales sont assez grossières et proches d'un comportement zonal. Par exemple, la différence entre l'été de l'hémisphère sud et l'hiver de l'hémisphère nord est nettement moins marquée. Aucun détail précis ne semble présent au-dessus de 30° nord et au-dessous de 30° sud. Nous notons également, avec les tracés des densités de probabilité des variables température et humidité par classe (non présentés dans cet article), que le pouvoir discriminant de ces classes est inférieur à celui des classes de la partition par couplage de copules.

Une conclusion claire est que la partition obtenue par classification mixte sur les données numériques est inférieure à celle obtenue par DMC avec les données fonctions de répartition, au sens de la cohérence physique des classes.

Une autre comparaison a été effectuée avec le résultat de la méthode EM appliquée aux données numériques de température et d'humidité. Ce résultat (non présenté ici) est assez similaire à celui de la classification mixte et possède des défauts communs. La partition obtenue est de qualité inférieure à celle obtenue par couplage avec DMC, les incursions d'air sont très mal définies et ont totalement disparu dans l'hémisphère sud. De plus, la classe tropicale est grossière et le comportement zonal

7classes 15/12 0H (37T(1), 24H(1))

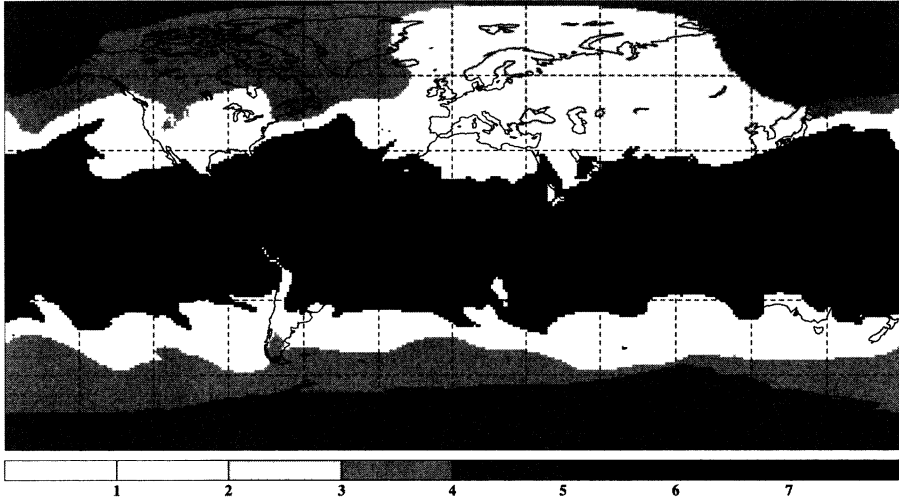


FIGURE 21

*Classification en 7 classes sur 37 températures et 24 humidités
pour le 15 décembre 1998 à 0H*

de la classification mixte sur les données numériques brutes est également présent (essentiellement dans l'hémisphère sud).

Certains points sont tout de même cohérents : la différence entre l'été et l'hiver des deux hémisphères et les incursions d'air de l'hémisphère nord, bordant la frontière entre la zone tropicale et la zone tempérée, sont réalistes.

Cependant, cette partition semble avoir perdu son aspect dynamique. Par exemple, la dynamique de la spirale est avalée par une classe énorme (classe 6) qui se retrouve sur la classification de la méthode mixte.

6. Conclusion

Cette approche originale de la décomposition de mélange généralise la décomposition classique de mélange en utilisant un niveau d'abstraction supplémentaire (voir [9]) et permet de travailler sur des données de taille considérable. Les premiers résultats obtenus sur un grand ensemble complexe de données sur l'atmosphère terrestre conduisent à une classification réaliste au plan climatologique et semblent très prometteurs des lors que d'autres variables seront considérées. Différents algorithmes sont déjà généralisés de la même façon : l'extension théorique des algorithmes EM, SEM, SAEM, MCEM et CEM est faite (*cf.* [24]).

Par ailleurs, l'estimation des paramètres de copules est ici réalisée par la technique du maximum de vraisemblance mais d'autres techniques d'estimations sont envisageables telles que le τ de Kendall ou le ρ de Spearman ([15]).

Les FDD peuvent aussi être déterminées différemment. La loi bêta, les calculs empiriques ou non-paramétriques peuvent être remplacés par une décomposition classique de mélange de lois gaussiennes par exemple.

Une technique de choix automatique des T_i optimaux est en étude et divers approches sont déjà proposées dans [24]. Les résultats peuvent en effet être sensiblement différents lors d'une légère modification des T_i .

De plus, le cas multidimensionnel doit être un axe de recherche fort. La méthode proposée dans ce papier s'étend à la dimension p (voir Vrac, 2002, [24]), cependant, la complexité des formules au-delà de la dimension deux fait que le recours à des astuces de simplification des calculs est indispensable.

Les perspectives d'études et d'applications de la décomposition de mélange de copules s'inscrivent dans un cadre de recherche pouvant intéresser différentes disciplines désireuses d'utiliser l'analyse et la modélisation cohérente des données probabilistes. Elles sont donc vastes et beaucoup reste encore à faire.

Remerciements

Les auteurs aimeraient remercier tout le groupe ARA du LMD et plus particulièrement R. Armante, P. Naveau, N. A. Scott et S. Serran pour leur aide, leur soutien et leur amitié.

Références

- [1] V. ACHARD. *Trois problèmes clés de l'analyse 3D de la structure thermodynamique de l'atmosphère par satellite : mesure du contenu en ozone; classification des masses d'air; modélisation hyper rapide du transfert radiatif*. Thèse de doctorat, Université Paris VII, 1991.
- [2] H.H. BOCK and E. DIDAY. *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*. Springer-Verlag, Heidelberg, 2000.
- [3] G. CELEUX and G. GOVAERT. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of statist. computer*, 47, 127-146, 1993.
- [4] A. CHÉDIN, N. SCOTT, C. WAHICHE, P. MOULINIER. The improved initialization inversion method : a high resolution physical method for temperature retrievals from satellites of the TIROS-N series, *J. Clim. Appl. Meteor.*, 24, 128-143, 1985.
- [5] Frédéric CHEVALLIER. *La modélisation du transfert radiatif à des fins climatiques : une nouvelle approche fondée sur les réseaux de neurones artificiels*. Thèse de doctorat, Université de Paris 7, 1998.

- [6] R.E. DAVIS and D.R. WALKER. An upper-air Synoptic Climatology of the Western United States, *American Meteorological Society*, 5, 1449-1467, 1992.
- [7] A.P. DEMPSTER, N.M. LAIRD and D.B. RUBIN. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society*, 39, 1-38, 1977.
- [8] E. DIDAY, Y. OK and A. SCHROEDER. The dynamic clusters method in pattern recognition. In *Proceeding of IFIP congress*, Stockholm, 1974.
- [9] Edwin DIDAY. A generalisation of the mixture decomposition problem in the symbolic data analysis framework, *Cahiers du CEREMADE*, (0112), 2001.
- [10] B. EVERITT and D. HAND. *Finite mixture distributions*, Chapman and Hall, London, 1981.
- [11] C. GENEST. Frank's family of bivariate distributions. *Biometrika*, 74, 549-555, 1987.
- [12] Younès HILLALI. *Analyse et modélisation des données probabilistes : capacités et lois multidimensionnelles*. Thèse de doctorat, Université de Paris IX Dauphine, 1998.
- [13] Laurence S. KALSTEIN, J. SCOTT GREENE, Michael C. NICHOLS, C. DAVID BARTHEL. A new spatial synoptic climatological procedure. In *AMS Eight Conference on Applied Climatology*, pages 169-173, Anaheim, California, 17-22 January 1993.
- [14] J.L. MOLLIERE. What is the real number of clusters? In *9th meeting of the German Classification Society*, 1985.
- [15] Roger B. NELSEN. *An introduction to Copulas*. Springer Verlag, Lectures Notes in Statistics, 1998.
- [16] R.A. REDNER and H. WALKER. Mixture densities, maximum likelihood and the EM algorithm. *SIAM*, 26, 195-239, 1984.
- [17] B. SCHWEIZER and A. SKLAR. *Probabilistic Metric Spaces*. Elsevier North-Holland, New-York, 1983.
- [18] A. SCOTT and M. SYMONS. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27, 1971.
- [19] I. SHLEZINGER. An algorithm for solving the self organization problem. *Cybernetics*, 2, 1968.
- [20] B.W. SILVERMAN. *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
- [21] A. SKLAR. Fonction de répartition à n dimensions et leurs marges. *Inst. Statist. Univ. Paris Pub.*, 8 :229-231, 1959.
- [22] M. SYMONS. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37, 1981.
- [23] M. VRAC, E. DIDAY and A. CHÉDIN. Mixture decomposition of copulas and application to climatology. In *IPMU 2002*, Université de Savoie, Annecy, 2002.

- [24] Mathieu VRAC. *Analyse et modélisation de données probabilistes par Décomposition de Mélange de Copules et Application à une base de données climatologiques*. Thèse de doctorat, Université Paris IX Dauphine, décembre 2002.
- [25] Mathieu VRAC, Edwin DIDAY, Alain CHÉDIN and Philippe NAVEAU. *Mélange de distributions de distributions*. In *SFC'2001 8^{èmes} Rencontres de la Société Francophone de Classification*, Université des Antilles et de Guyane, Guadeloupe, 17-21 décembre 2001.