THÈSE

PRÉSENTÉE
A L'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE
POUR OBTENIR LE GRADE DE DOCTEUR DE L'I.N.P.L.
PAR

Caroline JOSEPH

Sujet de la thèse :

APPLICATION DE L'ANALYSE DES MÉLANGES GAUSSIENS AU CALIBRAGE GÉOLOGIQUE DES DONNÉES SISMIQUES

Soutenue publiquement le 9 Septembre 1994 devant la Commission d'Examen :

MM. M. DEPAIX	*Président et rapporteur*
J.J. ROYER	*Directeur*
ER	*Rapporteur*
RNIER	*Examinateur*
LLET	*Examinateur*
A. HAAS	*Invité*

THÈSE

PRÉSENTÉE
A L'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE
POUR OBTENIR LE GRADE DE DOCTEUR DE L'I.N.P.L.
PAR

Caroline JOSEPH

Sujet de la thèse :

APPLICATION DE L'ANALYSE DES MÉLANGES GAUSSIENS AU CALIBRAGE GÉOLOGIQUE DES DONNÉES SISMIQUES

Soutenue publiquement le 9 Septembre 1994 devant la Commission d'Examen :

MM. M. DEPAIX	*Président et rapporteur*
J.J. ROYER	*Directeur*
R. BAYER	*Rapporteur*
Mme F. FOURNIER	*Examinateur*
MM. J.L. MALLET	*Examinateur*
A. HAAS	*Invité*

REMERCIEMENTS

Les travaux de ce mémoire se sont déroulés dans le service Géophysique de l'Institut Français du Pétrole, auquel j'adresse mes remerciements.

Plus particulièrement, je remercie Vincent Richard pour son accueil chaleureux au sein de sa petite "P.M.E. familiale". L'un des grands souvenirs de ma thèse restera le congrès de la SEG. Mais je n'oublierai pas non plus ces quelques moments où le "savoir (bien) vivre" prenait le pas sur le travail.

Toute mon amitié va à Frédérique Fournier sans qui cette thèse ne se serait certainement pas aussi bien déroulée. Toujours présente, même dans les moments difficiles, elle a su me laisser une grande liberté de travail, tout en m'apportant son concours et son dynamisme si nécessaire. Et malgré des conditions de travail parfois "étouffantes", j'ai énormément apprécié ces années de thèse, tant personnellement que professionnellement. Alors, Frédérique, merci pour tout, et surtout prends soin de ma jungle.

Je tiens de plus à remercier sincèrement Jean-Jacques Royer, qui a dirigé avec diligence ce travail depuis Nancy. Son suivi pendant cette thèse m'a beaucoup aidé, de même que ses nombreuses interventions pour aplanir des problèmes plus pratiques.

Je remercie également de façon collective tous les membres de la P.M.E. (dessinatrices, informaticiens, ingénieurs et secrétaires de choc), ceux qui sont encore là comme ceux qui ne le sont plus, et qui ont fait de ces années un grand éclat de rire.

Je souhaite exprimer ma reconnaissance à Monsieur Depaix, président de mon jury de thèse, et à Monsieur Bayer, qui ont tous deux accepté le rôle ingrat de rapporteur. Ce travail de thèse reliant les deux mondes de la Statistique et de la Géophysique, chacun d'eux a dû appréhender un nouveau domaine, ce qui, je l'espère, les aura intéressés.

Je remercie également Monsieur Mallet pour sa présence dans ce jury. Et j'adresse mes remerciements à André Haas pour les discussions que nous avons eu au sujet de ce travail (et pour celles à venir, je l'espère).

Je remercie enfin tous ceux qui m'ont soutenu, porté et supporté ces dernières années (famille, amis, chevaux). Leur calvaire prend fin...

Bonne chance à tous, et comme on dit là haut sur la montagne, arvi !

RÉSUMÉ

Dans le cadre de l'interprétation lithologique des données sismiques, des méthodes statistiques de calibrage permettent d'estimer des propriétés géologiques à partir d'attributs sismiques au niveau du réservoir. Le but de cette thèse est d'utiliser les mélanges gaussiens au sein de ces méthodes statistiques de calibrage.

Dans un premier temps, nous présentons la méthode de décomposition en classes gaussiennes que nous avons développée. Par la suite, nous décrivons les deux méthodes statistiques de calibrage que nous avons considérées (analyse canonique et régression non paramétrique), et présentons l'utilisation des mélanges gaussiens dans le cadre de ces méthodes.

Enfin, la méthodologie de calibrage fondée sur la régression non paramétrique a été appliquée aux données réelles d'un réservoir pétrolier. Ce réservoir est développé par une quarantaine de puits environ, et couvert par une sismique 2D à maillage dense. Des attributs sismiques, caractéristiques de la fenêtre temporelle correspondant au réservoir, ont été calibrés aux puits en terme d'épaisseurs cumulées de lithofaciès (grès et dolomies vacuolaires), permettant ainsi de prédire la distribution spatiale de ces lithofaciès entre les puits à partir de la sismique. Les résultats aux puits ont été comparés à ceux obtenus antérieurement sur ce champ par d'autres méthodes statistiques de calibrage, mettant ainsi en évidence l'intérêt de la méthodologie de calibrage par régression non paramétrique développée.

ABSTRACT

In the field of lithologic interpretation of seismic data, statistical calibration techniques allow to predict geological properties from seismic attributes at the reservoir level. The aim of this thesis is the application of gaussian mixture analysis to statistical calibration methods.

We first present the method of gaussian segmentation that we developed. Then, we describe two statistical calibration methods (canonical analysis and non parametric regression), and the application of gaussian mixture analysis to these methods.

Lastly, the methodology derived from non parametric regression is applied to an oil field. The reservoir under analysis is produced by over 40 wells, and covered by a dense 2D seismic survey. Seismic attributes which characterize the time window at the reservoir level are calibrated in terms of lithofacies cumulated thicknesses. It allows to predict the lithofacies spatial distribution between wells from the seismic data. The results at the wells are compared to those obtained by other statistical calibration methods : this enhances the interest of the non parametric regression approach.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

CHAPITRE 1

INTRODUCTION

I - CONTEXTE DES TRAVAUX

Les données sismiques constituent une source d'information très importante pour guider la reconnaissance géologique du sous-sol entre les puits. Ceci tient, d'une part, au coût limité d'acquisition et de traitement des données sismiques par rapport au coût des forages, et d'autre part à la bonne couverture spatiale du sous-sol que ce type de données assure (sismique 3D, en particulier).

Tout d'abord, les données sismiques sont utilisées dans le cadre de l'**interprétation structurale**, pour définir la géométrie des objets géologiques. On peut ainsi repérer les événements géologiques majeurs, tectoniques (faille, anticlinal...) ou sédimentaires (biseau stratigraphique...). L'**interprétation stratigraphique** des données sismiques consiste ensuite en une analyse qualitative des variations morphologiques du signal sismique entre deux horizons majeurs, afin d'en fournir une interprétation géologique. Ces deux types d'interprétation des données sismiques (interprétations structurale et stratigraphique) interviennent généralement aux stades de découverte ou d'appréciation d'un gisement.

Récemment, suite aux progrès effectués dans les domaines de l'acquisition et du traitement des données sismiques, un troisième type d'interprétation a été développé : l'**interprétation lithologique** des données sismiques. Cette interprétation, qui intervient généralement au stade de développement d'un gisement, est beaucoup plus fine que les interprétations structurale ou stratigraphique. Elle repose sur une analyse quantitative des amplitudes sismiques : on cherche à interpréter une petite portion de trace sismique, focalisée le plus souvent au niveau de l'intervalle réservoir, en terme d'information lithologique ou pétrophysique (porosité, contenu en fluides ...). Cette interprétation est souvent complexe, du fait des relations indirectes existant entre les propriétés géologiques d'un ensemble de couches constituant le réservoir et la réponse sismique de ce réservoir, ainsi que suite à la bande passante limitée des données sismiques. Mais compte tenu de la très bonne répartition spatiale des données sismiques, l'information obtenue est très intéressante, et peut notamment fournir des contraintes supplémentaires dans le cadre de simulations probabilistes de modèles de réservoir.

Les travaux présentés dans ce mémoire s'inscrivent dans le cadre de l'interprétation lithologique des données sismiques. Nous évoquerons donc les travaux antérieurs effectués dans ce domaine avant de discuter de l'objectif de nos travaux.

II - INTERPRÉTATION LITHOLOGIQUE DES DONNÉES SISMIQUES

Deux types d'information peuvent être obtenus dans ce cadre :

- une information géologique qualitative (présence ou absence de corps sableux, passage d'un réservoir poreux à une formation compacte...),
- une information géologique quantitative (épaisseur cumulée de lithofaciès, porosité moyenne...).

On parlera respectivement d'une interprétation qualitative ou quantitative des données sismiques. En fait, cette interprétation implique presque toujours un calage de la réponse sismique sur l'information géologique reconnue aux puits. On parle alors de calibrage géologique.

1 Interprétation qualitative

L'interprétation qualitative (ou calibrage qualitatif) est associée à la notion de faciès sismique, formalisée pour la première fois par Mitchum *et al.* (1977). Un faciès est un ensemble de traces présentant des caractéristiques communes, du point de vue par exemple de l'énergie, de la fréquence apparente, de la continuité des réflexions... Dans beaucoup de cas, cette morphologie identique traduit un même environnement géologique de dépôt. Le calibrage qualitatif consiste donc à interpréter géologiquement (lorsque c'est possible) les faciès sismiques à l'aide des données géologiques disponibles aux puits. Ce genre d'étude a rapidement mis en évidence certains problèmes.

Tout d'abord, il est apparu nécessaire de mieux qualifier le caractère sismique en recherchant, en particulier, les **attributs sismiques** les plus discriminants par rapport au problème géologique posé.

D'autre part, il est important de pouvoir **automatiser la recherche des faciès sismiques**, du fait, entre autres, du grand volume de traces à analyser (sismique 3D, par exemple). Des techniques statistiques de reconnaissance des formes ont donc été utilisées. Les traces sismiques, analysées au niveau de l'objectif, sont représentées comme des points dans l'espace multivariable généré par les attributs sismiques. Des méthodes de classement, avec ou

sans apprentissage, sont ensuite appliquées dans cet espace pour regrouper les traces de morphologie voisine au sens des attributs considérés.

1.1 Recherche des attributs sismiques

Les attributs peuvent être calculés à partir de la représentation temporelle ou de la représentation fréquentielle des traces sismiques.

Il existe un certain nombre d'attributs classiques, comme les amplitudes à des temps donnés, l'énergie totale de la trace ou bien les quantiles de la distribution temporelle de l'énergie (Dumay et Fournier, 1988). D'autres attributs sismiques ont été développés, comme les composantes principales calculées sur un ensemble de paramètres sismiques (Hagen, 1982), les coefficients autorégressifs (Bois, 1980, 1981), les quantiles de la distribution des amplitudes du spectre de puissance (Sinvhal et Khattri, 1983, Sinvhal *et al.*, 1984)... On trouvera une synthèse assez complète des différents attributs sismiques utilisables dans l'article de Justice *et al.* (1985).

1.2 Recherche des faciès sismiques

Deux approches utilisant des méthodes statistiques de reconnaissance de formes ont été développées pour déterminer les faciès sismiques : la méthodologie sans apprentissage et la méthodologie avec apprentissage.

La **méthodologie sans apprentissage** consiste à appliquer des méthodes statistiques de classification automatique dans l'espace des attributs sismiques, afin de déterminer des classes de traces homogènes (Lendzionowski *et al.*, 1990). On utilise alors, *a posteriori*, l'information disponible aux puits pour interpréter géologiquement les faciès sismiques ainsi obtenus.

La **méthodologie avec apprentissage** consiste à appliquer des techniques d'analyse discriminante. On choisit tout d'abord, parmi toutes les traces sismiques, un ensemble de traces (dites traces d'apprentissage) typiques des diverses réalités géologiques connues sur le champ étudié. Ces traces peuvent être des traces sismiques réelles sélectionnées au voisinage de puits typiques (Dumay et Fournier, 1988). Mais on peut aussi prendre des traces sismiques synthétiques (Kubichek et Quincy, 1985).

Dans une première phase, on applique des techniques d'analyse discriminante sur l'ensemble des traces d'apprentissage, pour rechercher les attributs sismiques séparant le mieux les groupes (ou faciès) de traces typiques d'une même réalité géologique. Dans une seconde phase, les traces sismiques ne faisant pas partie de l'ensemble d'apprentissage sont affectées à ces différents faciès.

On obtient alors une carte de faciès sismiques directement interprétée géologiquement, l'information géologique ayant gouverné la formation des faciès.

2 Interprétation quantitative

Les donnés sismiques (amplitudes, impédances...) peuvent être directement intégrées dans des modèles de réservoir (Doyen *et al.*, 1988). Mais cette approche n'est applicable que dans le cas d'environnements géologiques simples. Le plus souvent, un **calibrage quantitatif** avec l'information aux puits permet l'extraction de paramètres réservoirs (tels que la porosité moyenne) à partir des traces sismiques. Remarquons au préalable qu'un calibrage qualitatif peut fournir de l'information quantitative : chaque faciès sismique peut être caractérisé par la porosité moyenne ou le pourcentage moyen de lithologies, calculés sur les puits correspondant à la réalité géologique décrite par ce faciès (Thadani *et al.*, 1987, Fournier, 1989).

Plusieurs méthodes peuvent être utilisées pour obtenir une telle information géologique quantitative. Une de ces méthodes consiste à appliquer des relations physiques existant entre les données sismiques et les variables géologiques pour prédire ces dernières. On peut ainsi utiliser la formule de Wyllie, qui fournit les porosités moyennes en fonction des temps de transit ou des impédances au niveau du réservoir (Angeleri et Carpi, 1982). Toutefois, cette formule n'est applicable que sous l'hypothèse que le contenu en fluide et la lithologie au niveau du réservoir sont connus entre les puits. Ceci n'est valide que dans le cas d'environnements géologiques très simples, ce qui limite l'intérêt de cette méthode.

Par ailleurs, des relations empiriques entre attributs sismiques et variables géologiques peuvent être établies par des techniques statistiques de calibrage. Ces techniques sont, en général, la régression linéaire (Stanulonis et Tran, 1992) ou l'analyse canonique (Fournier et Derain, 1992-a). Ces méthodes statistiques ne permettant que la recherche de relations linéaires entre les attributs sismiques et les variables géologiques, Fournier et Derain (1994) ont appliqué ces techniques sur chacun des faciès sismiques obtenus par calibrage qualitatif, à condition que le nombre de puits associés à ces faciès soit suffisant.

Enfin, des méthodes géostatistiques telles que le cokrigeage (Doyen, 1988) peuvent être appliquées, si le nombre de puits disponibles le permet.

III - OBJECTIFS

Les travaux de ce mémoire se placent dans le cadre de l'interprétation lithologique des données sismiques. Deux problèmes se posaient.

- Dans le cadre du calibrage qualitatif, peut-on quantifier les correspondances entre les faciès sismiques et les groupes de puits représentatifs d'une même réalité géologique ?
- Dans le cadre du calibrage quantitatif, les méthodes statistiques de calibrage comme la régression ou l'analyse canonique ne sont plus utilisables lorsque les relations entre attributs sismiques et variables géologiques sont non linéaires. On peut, dans certains cas, résoudre ce problème en effectuant au préalable un calibrage qualitatif, mais cela suppose un nombre de puits conséquent. Est-il possible de développer une méthodologie statistique d'interprétation lithologique des données sismiques pouvant s'affranchir de ces problèmes de linéarité ?

Pour résoudre ces problèmes, nous avons donc envisagé l'application d'une technique de décomposition d'une population en classes gaussiennes aux données géologiques et sismiques.

En ce qui concerne le premier problème, la technique de **décomposition en classes gaussiennes**, appliquée séparément sur les données géologiques et sismiques, permet de coder les données étudiées en fonction de l'appartenance aux classes obtenues. L'**analyse canonique sur données codées** devrait fournir une quantification des correspondances entre les classes géologiques et les classes sismiques.

Pour répondre au deuxième problème, la **décomposition en classes gaussiennes** de la population de calibrage générée par les données géologiques et les données sismiques permet d'obtenir une approximation de la fonction de densité empirique de cette population. L'application d'une **méthode de régression non paramétrique** permet alors de prédire les variables géologiques en fonction des attributs sismiques, sans hypothèse de linéarité.

Dans le chapitre 2 de ce mémoire, nous présentons donc la méthode de décomposition en classes gaussiennes retenue, les tests sur données synthétiques de cette méthode, puis la méthodologie de décomposition en classes gaussiennes que nous avons définie. Dans le chapitre 3, nous présentons la méthode d'analyse canonique, ainsi que son application sur des données codées. Dans le chapitre 4, nous développons la méthodologie de calibrage quantitatif non linéaire, basée sur une technique de régression non paramétrique. Cette méthodologie a ensuite été appliquée sur un champ pétrolier afin de prédire, au niveau du réservoir, des épaisseurs cumulées de lithofaciès à partir d'attributs sismiques ; nous discutons de cette application à des données

réelles dans le chapitre 5. Enfin, dans le chapitre 6, nous faisons une synthèse des travaux effectués, puis évoquons les différents travaux qui pourraient être envisagés dans le futur pour faire suite à ceux-ci.

CHAPITRE 2

DÉCOMPOSITION D'UNE POPULATION QUELCONQUE EN CLASSES GAUSSIENNES

I - TRAVAUX ANTÉRIEURS

Le principe de décomposition d'une population quelconque en classes gaussiennes a fait l'objet de nombreux travaux. Jusqu'à présent, plusieurs voies ont été explorées.

La première méthode utilisée fut la **méthode des moments** (Pearson, 1894), consistant à calculer les moments de la fonction de densité observée, et à les égaler aux moments de la fonction de densité théorique du mélange de lois normales. Dans le cadre de deux sous-populations normales monovariables, la fonction de densité théorique du mélange de lois s'écrit :

$$f(x;m_1,m_2,\sigma_1,\sigma_2) = p \cdot f_1(x;m_1,\sigma_1) + (1-p) \cdot f_2(x;m_2,\sigma_2)$$

où :

- p est le poids de la sous-population 1 qui suit une loi normale $N(m_1,\sigma_1)$,
- $1-p$ est le poids de la sous-population 2 qui suit une loi normale $N(m_2,\sigma_2)$.

Ce problème nécessite donc d'identifier cinq inconnues, qui sont les paramètres p, m_1, m_2, σ_1, σ_2.

Si on note V_r le moment observé centré d'ordre r et W_r le moment théorique centré d'ordre r, la méthode des moments consiste à résoudre le système :

$$(1) \quad \begin{cases} V_r = W_r \quad , \quad \forall\, r = 1 \ldots 5 \\ V_r = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^r \\ W_r = \displaystyle\int (x - E(x))^r \cdot f(x;m_1,m_2,\sigma_1,\sigma_2)\, dx \end{cases}$$

où :

- $x_1 \ldots x_n$ sont les valeurs de x observées sur les n individus,
- $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est la moyenne empirique,
- $E(x) = p.m_1 + (1-p).m_2$ est l'espérance mathématique de x.

Il est aussi possible de considérer les moments théoriques non centrés d'ordre r dans la méthode des moments.

Dans le cas de moments centrés, le système (1) de 5 équations à 5 inconnues ($p, m_1, m_2, \sigma_1, \sigma_2$) se développe sous la forme :

$$(2) \quad \begin{cases} 0 = p\delta_1 + (1-p)\delta_2 \\ V_2 = p\left(\sigma_1^2 + \delta_1^2\right) + (1-p)\left(\sigma_2^2 + \delta_2^2\right) \\ V_3 = p\left(3\delta_1\sigma_1^2 + \delta_1^3\right) + (1-p)\left(3\delta_2\sigma_2^2 + \delta_2^3\right) \\ V_4 = p\left(3\sigma_1^4 + 6\sigma_1^2\delta_1^2 + \delta_1^4\right) + (1-p)\left(3\sigma_2^4 + 6\sigma_2^2\delta_2^2 + \delta_2^4\right) \\ V_5 = p\left(15\sigma_1^4\delta_1 + 10\sigma_1^2\delta_1^3 + \delta_1^5\right) + (1-p)\left(15\sigma_2^4\delta_2 + 10\sigma_2^2\delta_2^3 + \delta_2^5\right) \end{cases}$$

où :

- $\delta_1 = m_1 - E(x)$,
- $\delta_2 = m_2 - E(x)$.

En fait, Pearson a montré que résoudre le système (2) revient à résoudre l'équation suivante faisant intervenir un polynôme d'ordre 9 :

$$(3) \quad \sum_{j=0}^{9} b_j u^j = 0$$

les coefficients b_j étant des fonctions de V_1, V_2, V_3, V_4 et V_5. Alors, si le système (2) admet une solution, elle peut être calculée à partir de la racine réelle négative de l'équation (3) ci-dessus. A partir de cette racine $\hat{u}$, on peut en déduire les valeurs de δ_1 et δ_2, puis celles des paramètres p, m_1, m_2, σ_1^2, σ_2^2, car leur expression en fonction de $\hat{u}$, V_1, V_2, V_3, V_4 et V_5 est connue. En pratique, on obtient parfois des valeurs négatives de σ_1^2 ou σ_2^2 : il faut alors ajouter des contraintes sur ces variances dans le système (1).

Finalement, on constate que même dans le cas très simple d'une décomposition monovariable en deux sous-populations gaussiennes, les problèmes calculatoires engendrés sont importants. Doetsch (1928) a généralisé cette méthode à la décomposition d'une population monovariable en plus de deux sous-populations gaussiennes, mais cela se révèle difficilement utilisable en pratique. Enfin, Charlier et Wicksell (1924) et Day (1969) ont utilisé cette méthode pour la décomposition d'une population 2D en deux sous-populations gaussiennes, mais en faisant des hypothèses sur les matrices de variance-covariance des classes gaussiennes. En pratique, les applications de la méthode des moments sont donc très restreintes.

Dans un second temps, la **méthode du maximum de vraisemblance** a été utilisée pour la décomposition d'une population. Cette méthode, présentée pour la première fois par Rao en 1948 (cas d'une population monovariable à décomposer en deux classes gaussiennes), a été par la

suite appliquée au cas le plus général : la décomposition d'une population multivariable en plusieurs classes gaussiennes de matrices de variance-covariance différentes.

Dans le cas **monovariable**, si on note $x_1 \ldots x_n$ la population étudiée, le problème de décomposition en K sous-populations gaussiennes (de poids p_k, de moyenne m_k et d'écart-type σ_k, pour $k = 1 \ldots K$) consiste à approximer au mieux la densité observée de chaque individu par sa densité théorique $f(x_i)$ issue du mélange de sous-populations. Comme cette fonction de densité théorique dépend de p_k, m_k et σ_k (pour $k = 1 \ldots K$), le problème consiste à estimer ces paramètres avec les contraintes suivantes :

$$\begin{cases} f(x_i) = \sum_{k=1}^{K} p_k f_k(x_i; m_k, \sigma_k) \quad , \quad \forall\, i = 1 \ldots n \\ f_k(x_i; m_k, \sigma_k) = \dfrac{1}{\sqrt{2\pi}\,\sigma_k} \cdot \exp\left[-\dfrac{(x_i - m_k)^2}{2\sigma_k^2} \right] \\ \sum_{k=1}^{K} p_k = 1 \end{cases}$$

Par la méthode du maximum de vraisemblance, on cherche donc à résoudre le problème suivant :

$$\begin{cases} \text{Max} \prod_{i=1}^{n} f(x_i) \\ \sum_{k=1}^{K} p_k = 1 \end{cases} \quad \Leftrightarrow \quad \begin{cases} \text{Max} \quad \sum_{i=1}^{n} \log(f(x_i)) \\ \sum_{k=1}^{K} p_k = 1 \end{cases}$$

En utilisant la méthode du Lagrangien, ce problème revient à maximiser L avec :

$$L = \sum_{i=1}^{n} \log(f(x_i)) - \lambda \left(\sum_{k=1}^{K} p_k - 1 \right)$$

On peut ensuite calculer les équations normales. En constatant que :

$$\sum_{k=1}^{K} \left(p_k \cdot \frac{\partial L}{\partial p_k} \right) = 0 = n - \lambda$$

on obtient les équations suivantes pour $k = 1 \ldots K$:

$$\frac{\partial L}{\partial p_k} = 0 = \sum_{i=1}^{n} \frac{f_k(x_i; m_k, \sigma_k)}{f(x_i)} - n$$

$$\frac{\partial L}{\partial m_k} = 0 = \sum_{i=1}^{n} \left[\frac{p_k f_k(x_i; m_k, \sigma_k)}{f(x_i)} \cdot \frac{(x_i - m_k)}{\sigma_k^2} \right]$$

$$\frac{\partial L}{\partial \sigma_k} = 0 = \sum_{i=1}^{n} \left[\frac{p_k f_k(x_i; m_k, \sigma_k)}{f(x_i)} \cdot \frac{\left((x_i - m_k)^2 - \sigma_k^2 \right)}{\sigma_k^3} \right]$$

On peut alors déduire des équations normales les quatre équations suivantes :

$$(4) \quad p_k = \frac{1}{n}\sum_{i=1}^{n} p(k/x_i)$$

$$(5) \quad m_k = \frac{1}{n\,p_k}\sum_{i=1}^{n} p(k/x_i)\cdot x_i$$

$$(6) \quad \sigma_k^2 = \frac{1}{n\,p_k}\sum_{i=1}^{n} p(k/x_i)\cdot (x_i - m_k)^2$$

$$(7) \quad p(k/x_i) = \frac{p_k \cdot f_k(x_i; m_k, \sigma_k)}{\sum_{j=1}^{K} p_j \cdot f_j(x_i; m_j, \sigma_j)}$$

où $p(k/x_i)$ s'interprète comme la probabilité *a posteriori* que x_i appartienne à la classe k (cf. FIG. 1).

La fonction du maximum de vraisemblance n'étant pas bornée, les équations normales n'ont pas de solution analytique. Il faut donc résoudre ce problème de façon itérative. Ainsi, l'utilisation de la méthode du maximum de vraisemblance nécessite le choix d'une solution initiale $p^0(k/x_i)$, pour $k = 1 \ldots K$, à laquelle on applique de façon itérative jusqu'à convergence les équations (4) à (7) ci-dessus. A l'itération (s) on aura pour toute classe k :

$p^{(s-1)}(k/x_i)$ connu

$$p_k^{(s)} = \frac{1}{n}\sum_{i=1}^{n} p^{(s-1)}(k/x_i)$$

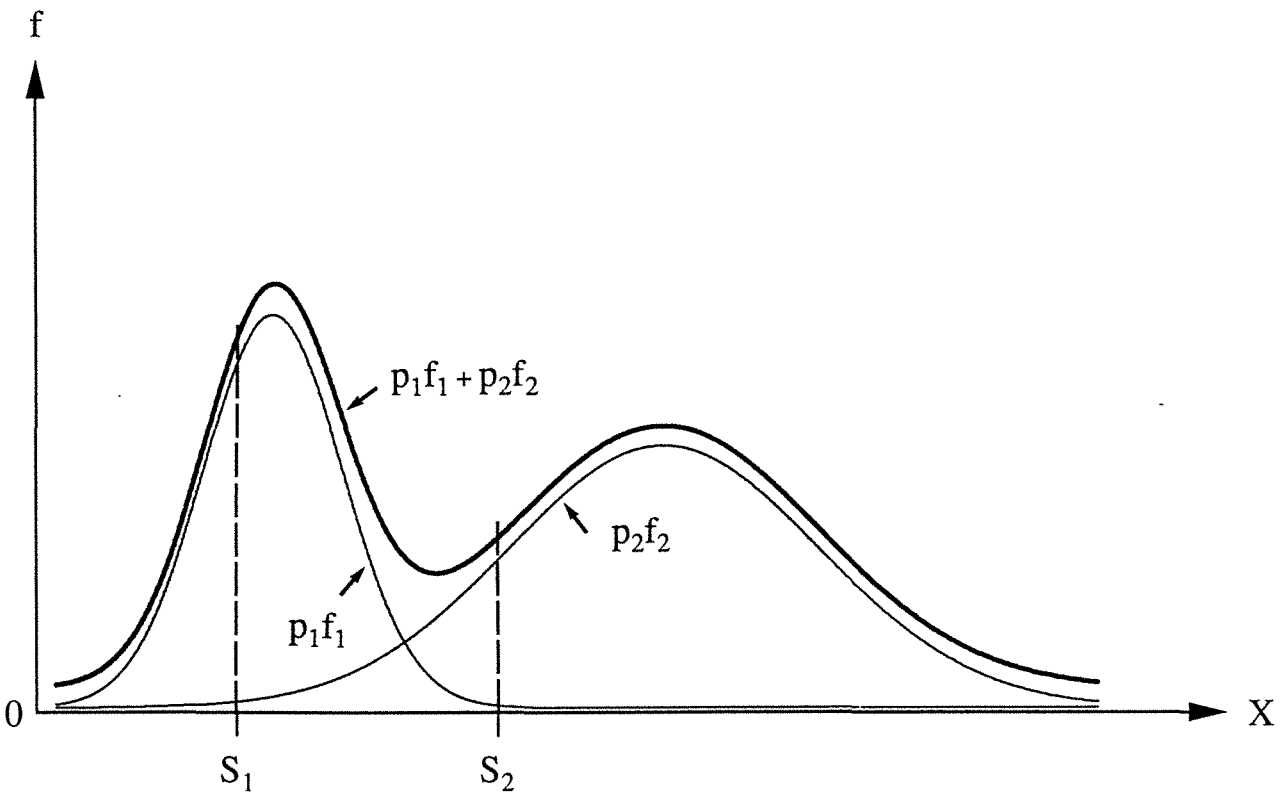
$$m_k^{(s)} = \frac{1}{n\,p_k^{(s)}}\sum_{i=1}^{n} p^{(s-1)}(k/x_i)\cdot x_i$$

$$\sigma_k^{2(s)} = \frac{1}{n\,p_k^{(s)}}\sum_{i=1}^{n} p^{(s-1)}(k/x_i)\cdot \left(x_i - m_k^{(s)}\right)^2$$

$$p^{(s)}(k/x_i) = \frac{p_k^{(s)} \cdot f_k\left(x_i; m_k^{(s)}, \sigma_k^{(s)}\right)}{\sum_{j=1}^{K} p_j^{(s)} \cdot f_j\left(x_i; m_j^{(s)}, \sigma_j^{(s)}\right)}$$

Nous pouvons faire deux remarques concernant la méthode du maximum de vraisemblance :

- elle nécessite de connaître *a priori* le nombre K de gaussiennes sous-jacentes à la population, ce qui est difficile à déterminer ;
- il y a autant de solutions potentielles que de maxima locaux au problème du maximum de vraisemblance ; la solution obtenue dépend en fait de la solution initiale retenue.



$K = 2$

$$p(k/x) = \frac{p_k f_k(x)}{p_1 f_1(x) + p_2 f_2(x)} \qquad \forall\ k = 1,2$$

- si $x \leq S_1$, $p(1/x) = 1$, $p(2/x) = 0$
- si $x \geq S_2$, $p(1/x) = 0$, $p(2/x) = 1$
- sinon, $p(1/x) \in [0,1]$, $p(2/x) \in [0,1]$

Fig. 1 Probabilité *a posteriori* p(k/x)

Le choix de cette solution initiale $p^0(k/x_i)$ est donc très important. En pratique, elle est souvent obtenue par des méthodes de classification du type "nuées dynamiques". On pourra se reporter aux articles de Hasselblad en 1966 (cas d'une population monovariable) ou de Wolfe en 1970 (cas général), qui traitent tous deux ce problème.

Dans la littérature, il est souvent fait mention de la **méthode EM** ou Estimation-Maximisation (Dempster *et al.*, 1977). Cette méthode est utilisée pour des populations comportant des données manquantes. Dans le cadre de la décomposition d'une population en classes gaussiennes, elle est en fait identique à la méthode du maximum de vraisemblance : l'étape d'estimation correspond au calcul des probabilités *a posteriori* et l'étape de maximisation à celui des paramètres de chaque classe (poids, moyennes et matrices de variance-covariance).

Celeux et Diébolt (1986) ont proposé un algorithme directement adapté de la méthode EM : l'**algorithme SEM** (Stochastique-Estimation-Maximisation). Cet algorithme devait résoudre les deux problèmes liés à la méthode EM (ou à la méthode du maximum de vraisemblance), à savoir :

- la connaissance *a priori* du nombre de classes gaussiennes sous-jacentes à la population,
- l'influence de la solution initiale sur la solution finale.

Pour cela, Celeux et Diébolt font intervenir une étape d'apprentissage probabiliste (ou étape stochastique) avant les deux étapes de maximisation et d'estimation de la méthode EM.

L'algorithme comporte tout d'abord une phase d'initialisation avec le choix de la solution initiale $p^0(k/x_i)$, pour $k = 1...K$. Une valeur doit aussi être affectée aux deux paramètres suivants :

- K, un majorant supposé du nombre de classes gaussiennes sous-jacentes à la population,
- C dans $]0,1[$, le poids minimal de chaque classe gaussienne.

Cette phase d'initialisation se poursuit par l'application itérativement et jusqu'à convergence des trois étapes mentionnées précédemment. A l'itération (s), on effectue donc :

- l'étape **stochastique**, avec tirage pour tout x_i d'une loi multinomiale de paramètres 1 et $\left(p^{(s-1)}(k/x_i), \forall\, k = 1...K\right)$, ce qui fournit la réalisation $\left(N_1^{(s)}(x_i), ..., N_K^{(s)}(x_i)\right)$ à valeurs dans $\{0,1\}$; s'il existe k tel que :

$$\frac{1}{n}\sum_{i=1}^{n} N_k^{(s)}(x_i) < C$$

on supprime la classe k, et on réitère l'étape stochastique sur les (K-1) classes restantes ;

- l'étape de **maximisation** de la méthode EM, en remplaçant dans les équations les probabilités *a posteriori* $p^{(s-1)}(k/x_i)$ par les réalisations $N_k^{(s)}(x_i)$;
- l'étape d'**estimation** de la méthode EM.

Celeux et Diébolt ont constaté que cet algorithme fournit en général de meilleurs résultats que la méthode EM, sauf dans le cas de populations de petite taille (n<100) pour lesquelles la méthode EM est plus efficace.

Notons que la méthode du maximum de vraisemblance (ou la méthode EM) et la méthode SEM, que nous avons présentées dans le cas monovariable, ont été aussi développées dans le cas d'une population multivariable.

Parallèlement, pour décomposer une population monovariable en classes gaussiennes, de nombreuses **méthodes graphiques** s'appuyant sur les propriétés de la loi normale ont été développées. Entre autres, Harding (1949) a utilisé le fait que, représentée sur un papier à échelle gaussienne, la fonction de répartition d'une loi normale est une droite. Ses travaux ont été repris pour la première fois dans le domaine minier par Sinclair (1976). La méthode consiste à définir visuellement les points d'inflexion de la fonction de répartition, ces points d'inflexion définissant les bornes des classes gaussiennes sous-jacentes à la population. La fonction de répartition de chaque classe est ensuite représentée graphiquement sur le papier à échelle gaussienne, ce qui permet d'en calculer les caractéristiques (moyenne et écart-type). Bridges et McCammon (1980) ont développé le programme DISCRIM à partir de cette méthode.

Ultérieurement, Royer et Mezghache (Mezghache, 1989) ont automatisé la méthode **graphique** monovariable de Harding, et ont utilisé la solution fournie par cette méthode comme solution initiale de la méthode du **maximum de vraisemblance**. Par la suite, ils ont adapté cette initialisation au cas d'une population multivariable (Mezghache, 1989). L'avantage de cette initialisation par rapport à celles généralement utilisées pour la méthode du maximum de vraisemblance est qu'elle ne nécessite pas de connaître *a priori* le nombre de classes gaussiennes sous-jacentes à la population.

Pour avoir un point de vue plus général sur les différentes méthodes de décomposition d'une population en classes gaussiennes, on pourra se reporter à l'ouvrage de Everitt et Hand (1981) ou à l'article de Holgersson et Jorner (1978). Dans le cadre de ce travail de recherche, la méthode développée s'apparente à la méthode du maximum de vraisemblance, avec différentes initialisations dont celle de Royer et Mezghache.

II - ÉTUDE DE LA MÉTHODE DE DÉCOMPOSITION D'UNE POPULATION EN CLASSES GAUSSIENNES RETENUE

Dans cette partie, nous présentons la méthode de décomposition que nous avons retenue : il s'agit de la méthode du maximum de vraisemblance étendue au cas d'une population multivariable. Puis nous développons deux des initialisations envisagées pour la méthode du maximum de vraisemblance. L'une est l'initialisation proposée par Royer et Mezghache (ou méthode de Harding automatisée). L'autre consiste à appliquer une méthode de classification multivariable sur la population.

1 Méthode du maximum de vraisemblance dans le cadre multivariable

Soit une population multivariable de n individus dans un espace de dimension p. Notons :

- x_i^j la valeur observée de l'individu i pour la variable j, pour tout i de 1 à n et j de 1 à p,
- $\left(x^j, \forall\, j = 1...p\right)$ les vecteurs des variables (vecteurs de taille nx1),
- $\left(x_i, \forall\, i = 1...n\right)$ les vecteurs des individus (vecteurs de taille 1xp).

Nous cherchons à décomposer la population étudiée en K classes gaussiennes multivariables, la classe k ayant pour moyenne $M_k = \left[M_k(1),...,M_k(p)\right]$ (vecteur de taille 1xp) et pour matrice de variance-covariance Σ_k (matrice pxp). Le nombre de classes K est supposé connu.

Le problème de décomposition peut se poser sous la forme d'un problème du maximum de vraisemblance, comme dans le cadre monovariable. En utilisant la méthode du Lagrangien, on aboutit aux quatre équations suivantes :

$$p_k = \frac{1}{n}\sum_{i=1}^{n} p(k/x_i)$$

$$M_k(j) = \frac{1}{n\,p_k}\sum_{i=1}^{n} p(k/x_i)\cdot x_i^j \quad ,\forall\, j = 1\ldots p$$

$$\Sigma_k(j_1,j_2) = \frac{1}{n\,p_k}\sum_{i=1}^{n} p(k/x_i)\cdot\left[x_i^{j_1} - M_k(j_1)\right]\cdot\left[x_i^{j_2} - M_k(j_2)\right] \quad ,\forall\, j_1 = 1\ldots p, \forall\, j_2 = 1\ldots p$$

$$p(k/x_i) = \frac{p_k\, f_k(x_i)}{\sum_{j=1}^{K} p_j\, f_j(x_i)}$$

où :

$$f_k(x_i) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma_k)}}\cdot\exp\left[-\frac{1}{2}(x_i - M_k)\cdot\Sigma_k^{-1}\cdot{}^t(x_i - M_k)\right]$$

Ces quatre équations sont les équations de la méthode du maximum de vraisemblance généralisée au cas multivariable. Elles n'ont pas de solution analytique. La démarche utilisée est donc identique à celle du cas monovariable. Supposons connue une solution initiale $p^0(k/x_i)$, pour $k = 1\ldots K$. En lui appliquant de façon itérative et jusqu'à convergence les quatre équations ci-dessus, nous pouvons calculer le poids p_k de la classe k, sa moyenne M_k et sa matrice de variance-covariance Σ_k, puis les probabilités *a posteriori* $p(k/x_i)$ pour tout individu x_i.

La méthode du maximum de vraisemblance développée dans le cadre multivariable est celle que nous avons retenue pour la décomposition d'une population en classes gaussiennes. Nous allons maintenant présenter deux initialisations possibles, permettant d'obtenir une solution initiale $p^0(k/x_i)$, pour $k = 1\ldots K$.

2 Initialisation proposée par Royer et Mezghache

L'initialisation proposée par Royer et Mezghache (Mezghache, 1989) pour la méthode du maximum de vraisemblance est issue d'une méthode de décomposition graphique monovariable : la méthode de Harding. Nous allons donc la présenter, avant de parler des travaux de Royer et Mezghache.

2.1 Principe de la méthode de Harding

La méthode de Harding (1949) est une méthode **graphique monovariable** pour la décomposition d'une population quelconque en un mélange de populations gaussiennes. Elle utilise le résultat suivant ; sur le graphe de Henry (papier à échelle gaussienne), la fonction de répartition d'une loi normale de moyenne m et d'écart-type σ est une droite d'équation :

$$Y = \frac{X - m}{\sigma}$$

(cf. FIG. 2, cas 1 et 2).

De même, sur ce graphe, la fonction de répartition d'un mélange de K populations gaussiennes est une courbe qui peut être approximée par K segments de droites séparés par des points d'inflexion (cf. FIG. 2, cas 3 et 4, pour un mélange de deux populations gaussiennes).

Décrivons maintenant la méthode développée par Harding.
Soit une variable aléatoire X prenant les valeurs $x_1 \ldots x_n$ telles que $x_1 \leq x_2 \leq \ldots \leq x_n$, et soit F sa fonction de répartition.
Le problème de la décomposition de la population $(x_1 \ldots x_n)$ en sous-populations gaussiennes consiste à décomposer F sous la forme :

$$F(x_i) = p_1 F_1(x_i) + \ldots + p_K F_K(x_i), \quad \forall\, i = 1 \ldots n$$

où :

- K est le nombre de sous-populations, inconnu *a priori*,
- F_k est la fonction de répartition gaussienne de la kième sous-population,
- p_k est le poids de la kième sous-population.

Pour résoudre ce problème, on calcule la fonction de répartition empirique $\hat{F}$ de la variable X. On recherche visuellement les points d'inflexion $s_1 \ldots s_{K-1}$ de $\hat{F}$. Ces (K-1) points d'inflexion, ou seuils, définissent K classes (ou sous-populations) supposées gaussiennes.

On calcule ensuite le poids et la fonction de répartition de chaque classe k de la façon suivante :

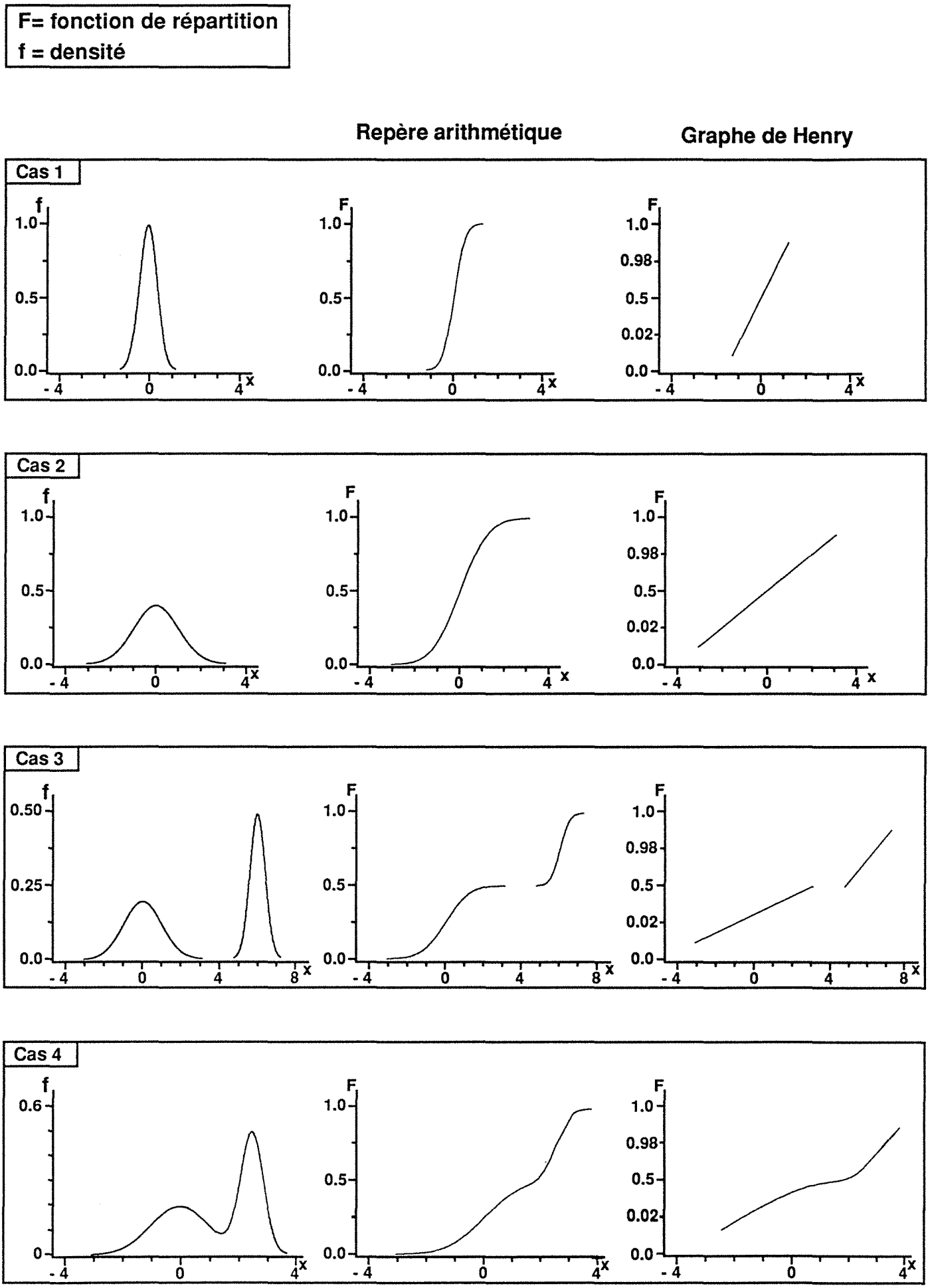


Fig. 2 Loi normale et graphe de Henry

pour $k = 1$

$$p_1 = \hat{F}(s_1)$$

$$F_1(x_i) = \begin{cases} \dfrac{\hat{F}(x_i)}{p_1} & \text{si } x_i \leq s_1 \\ 1 & \text{sinon} \end{cases}$$

pour $k = 2 \ldots (K-1)$

$$p_k = \hat{F}(s_k) - \hat{F}(s_{k-1})$$

$$F_k(x_i) = \begin{cases} 0 & \text{si } x_i \leq s_{k-1} \\ \dfrac{\hat{F}(x_i) - \hat{F}(s_{k-1})}{p_k} & \text{si } s_{k-1} < x_i \leq s_k \\ 1 & \text{sinon} \end{cases}$$

pour $k = K$

$$p_K = \hat{F}(x_n) - \hat{F}(s_{K-1})$$

$$F_K(x_i) = \begin{cases} 0 & \text{si } x_i \leq s_{K-1} \\ \dfrac{\hat{F}(x_i) - \hat{F}(s_{K-1})}{p_K} & \text{sinon} \end{cases}$$

On reporte alors la fonction de répartition de chaque classe k sur le graphe de Henry. Sous l'hypothèse que la classe k est gaussienne, on sait que sa fonction de répartition F_k sur ce graphe est une droite d'équation :

$$Y = A_k X + B_k$$

où

$$\begin{cases} A_k = \dfrac{1}{\sigma_k} \\ B_k = -\dfrac{m_k}{\sigma_k} \end{cases}$$

En faisant un ajustement linéaire sur le graphe de Henry, on peut donc en déduire la moyenne m_k et l'écart-type σ_k de la classe k, pour tout k de 1 à K.

Comme on peut le constater, la méthode de Harding est très simple à utiliser en pratique, mais elle présente toutefois les inconvénients suivants.

- La recherche du nombre et de la position des points d'inflexion sur la fonction de répartition empirique $\hat{F}$ de la variable considérée se fait de façon visuelle, ce qui est peu précis.
- Si le nombre de points d'inflexion et la position de ces points sont mal estimés, l'approximation obtenue peut être de mauvaise qualité ; pour l'améliorer, il faudra alors reprendre la méthode depuis le début.
- L'estimation des paramètres de chaque sous-population (moyenne et écart-type) peut manquer de précision car on les obtient par un ajustement graphique.

2.2 Automatisation de la méthode de Harding par Royer et Mezghache

2.2.1 Cas monovariable

Royer et Mezghache (Mezghache, 1989) ont travaillé sur la méthode de Harding afin de minimiser les inconvénients mentionnés ci-dessus.

Ils ont tout d'abord mis en place deux procédures permettant de résoudre numériquement la méthode de Harding. La **première procédure** concerne la **recherche des points d'inflexion**. Elle a été développée à partir d'une approximation de la dérivée seconde de $\hat{F}$:

$$\hat{F}''(x_i) \approx \frac{\hat{F}(x_i + h) + \hat{F}(x_i - h) - 2\hat{F}(x_i)}{h^2}$$

Alors, s est point d'inflexion de $\hat{F}$ si :

$$\begin{cases} \hat{F}''(s) = 0 \\ \hat{F}''(s - \varepsilon) < 0 \\ \hat{F}''(s + \varepsilon) > 0 \\ \varepsilon > 0 \end{cases}$$

Le nombre et la position des points d'inflexion sont donc déterminés de façon automatique par cette procédure.

D'autre part, ils ont mis au point une **seconde procédure** portant sur l'**ajustement linéaire** sur le graphe de Henry, basée sur l'idée suivante : il est équivalent de représenter la fonction de répartition d'une variable sur le graphe de Henry (repère à échelle gaussienne) ou de représenter l'inverse de Gauss de cette fonction de répartition dans un repère arithmétique. Ainsi, si une variable suit une loi normale, l'inverse de Gauss de sa fonction de répartition est aussi une droite dans un repère arithmétique.

Dans le cadre de la méthode de Harding, pour chaque sous-population k, Royer et Mezghache travaillent donc sur l'inverse de Gauss FI_k de la fonction de répartition F_k, et non sur F_k elle-même. Ils utilisent pour cela une approximation rationnelle de l'inverse de Gauss, fournie par Abramowitz et Stegun (1972) :

$$\begin{cases} FI_k(x_i) = z - \dfrac{a_0 + a_1 z + a_2 z^2}{b_0 + b_1 z + b_2 z^2 + b_3 z^3} + \varepsilon \\ z = \sqrt{\ln\left(\dfrac{1}{F_k(x_i)^2}\right)} \end{cases}$$

avec :

- $|\varepsilon| \leq 4.5 \times 10^{-4}$
- $a_0 = 2.515517$
- $a_1 = 0.802853$
- $a_2 = 0.010328$
- $b_0 = 1.$
- $b_1 = 1.432788$
- $b_2 = 0.189269$
- $b_3 = 0.001308$

Ils peuvent ensuite appliquer de façon automatique la régression linéaire sur FI_k pour chaque valeur de k afin de calculer A_k et B_k, et ils en déduisent les valeurs de m_k et σ_k. Ils calculent alors la densité de tout individu x_i pour chaque classe, soit :

$$f_k(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma_k} \cdot \exp\left[-\frac{(x_i - m_k)^2}{2\sigma_k^2}\right]$$

puis sa probabilité *a posteriori* d'appartenir à la classe k :

$$p(k/x_i) = \frac{p_k f_k(x_i)}{\sum_{j=1}^{K} p_j f_j(x_i)}$$

Cette probabilité *a posteriori* peut nous permettre d'attribuer tout individu x_i à la classe j qui la maximise :

$$p(j/x_i) = \max\{p(k/x_i), \forall\, k = 1...K\} \quad \Rightarrow \quad x_i \text{ appartient à la classe j}$$

L'ajout de ces deux procédures permet donc d'automatiser totalement la méthode de Harding. Toutefois, cette méthode ne peut fournir qu'une unique décomposition, les seuils des classes ne pouvant être modifiés. De plus, elle n'est applicable que dans le cadre monovariable, ce qui est très limitatif dans la pratique. Nous allons donc présenter au paragraphe suivant l'extension de cette méthode au cas multivariable, proposée par Royer et Mezghache.

2.2.2 Extension au cas multivariable

Soit un tableau de p variables et n individus, $(x^j, \forall\, j = 1...p)$ les vecteurs des variables et $(x_i, \forall\, i = 1...n)$ les vecteurs des individus.

La méthode proposée dans le cas multivariable comporte deux étapes (Mezghache, 1989). La **première étape** reprend la méthode de Harding automatisée monovariable présentée ci-dessus. On applique donc cette méthode à chaque variable x^j, pour $j = 1...p$. On obtient $K(j)$ seuils pour la variable x^j, notés $S^j(k)$, pour $k = 1...K(j)$.

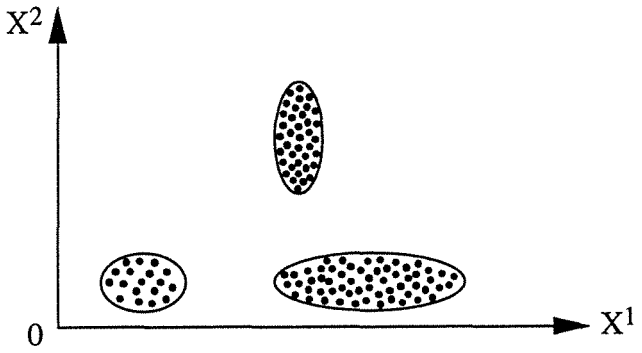
Dans une **seconde étape**, on fait le produit cartésien des classes monovariables obtenues, pour se replacer dans le cadre multivariable. On obtient ainsi K classes multivariables, et on peut calculer les probabilités *a posteriori* $p(k/x_i)$ qu'un individu appartienne à la classe k, pour tout k de 1 à K.

2.3 Utilisation de la méthode de Harding automatisée comme initialisation à la méthode du maximum de vraisemblance

La méthode de Harding automatisée présente un avantage par rapport à de nombreuses techniques de décomposition : il n'est pas nécessaire de connaître *a priori* le nombre de classes sous-jacentes à la population. De plus, elle a pu être étendue au cas multivariable. Royer et Mezghache l'ont donc utilisée comme initialisation à la méthode du maximum de vraisemblance. On pourra se reporter à l'exemple suivant pour avoir une idée de la succession des étapes.

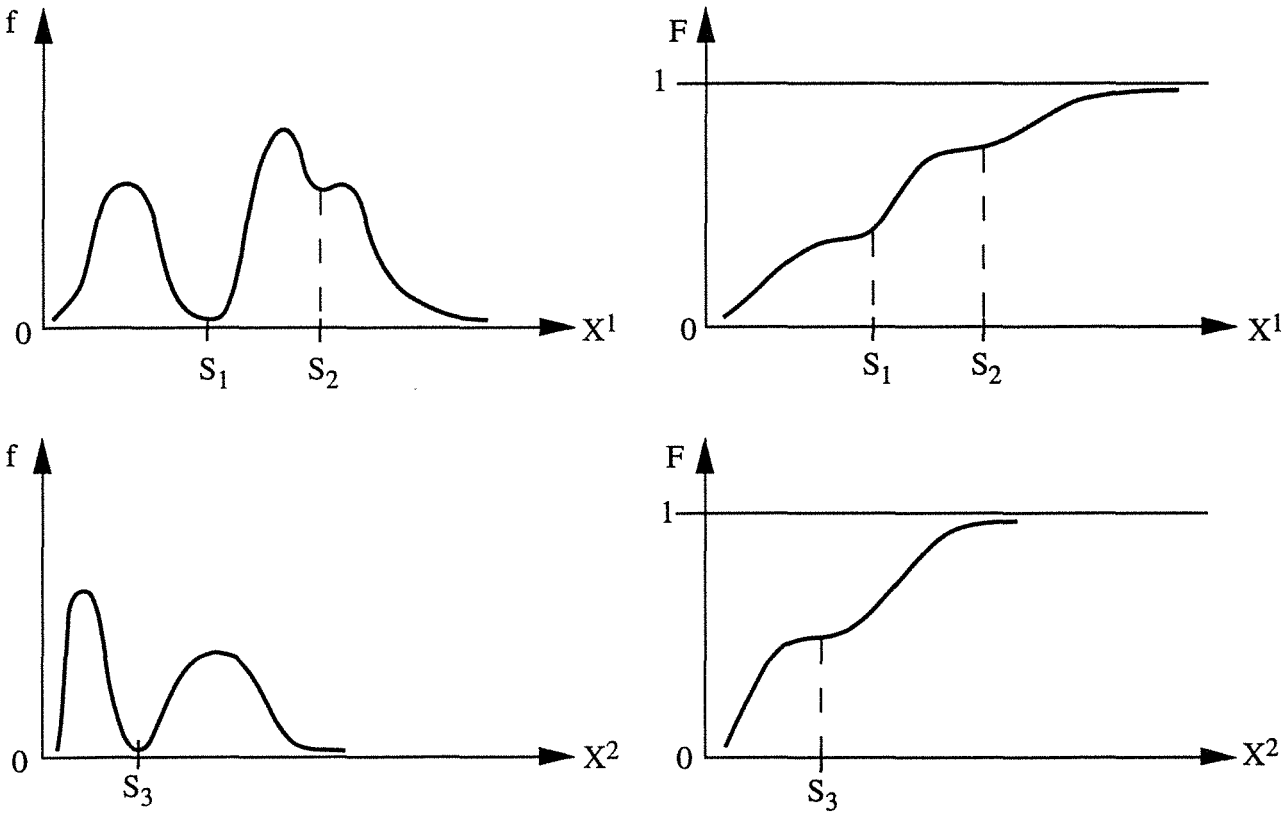
EXEMPLE

Soit deux variables X^1 et X^2, et soit une population dont la représentation graphique dans l'espace généré par ces variables est la suivante :

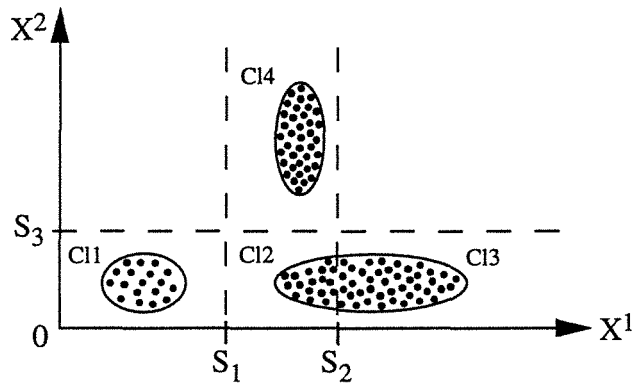


Les étapes 1 et 2 correspondent à l'initialisation par la méthode de Harding automatisée. L'étape 3 correspond à la méthode du maximum de vraisemblance.

Étape 1 : recherche des seuils sur la fonction de répartition de chaque variable, d'où les seuils S_1 et S_2 pour la variable X^1, et le seuil S_3 pour la variable X^2 ;



Étape 2 : passage au cadre multivariable, d'où quatre classes Cl1, Cl2, Cl3 et Cl4 ;



Étape 3 : agrégation des classes Cl2 et Cl3 par le processus itératif de la méthode du maximum de vraisemblance.

3 Initialisation par classification multivariable

Dans le cas de la décomposition d'une population multivariable, l'initialisation proposée par Royer et Mezghache repose sur l'étude de la fonction de répartition de chacune des p variables. Or les lois marginales d'une population ne correspondent pas forcément à la loi de la population dans son ensemble : les seuils obtenus par initialisation monovariable ne sont peut-être pas optimaux dans l'espace multivariable (cf. FIG. 3).

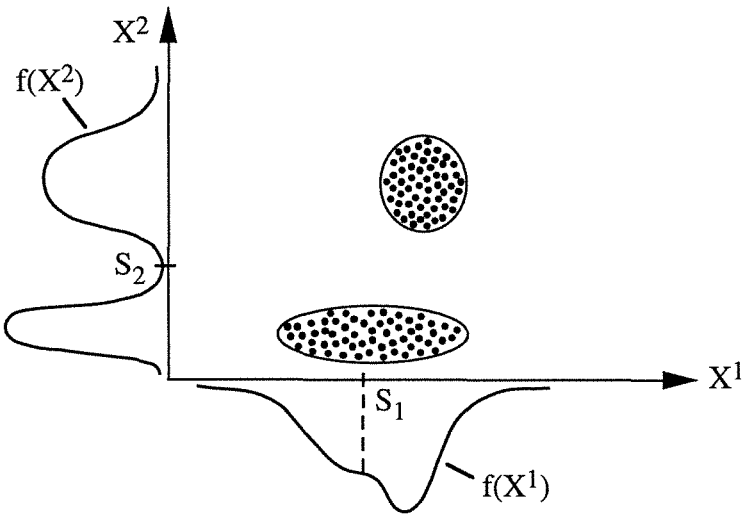
Nous avons donc décidé d'obtenir une solution initiale à la méthode du maximum de vraisemblance par classification multivariable de la population (avec une méthode de partitionnement de type "nuées dynamiques").

Plusieurs méthodes de partitionnement peuvent être utilisées, notamment la méthode des centres mobiles développée par Forgy (1965) ou celle des nuées dynamiques présentée par Diday (1979). Nous avons finalement retenu une méthode issue de la méthode "k-means" de MCQueen (1967) et implémentée dans le logiciel statistique SAS sous le nom de "FASTCLUS".

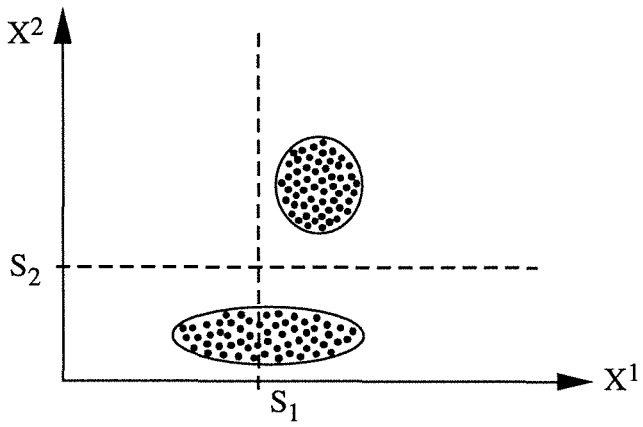
Supposons qu'on recherche une partition en K classes. La méthode utilisée comprend les trois étapes suivantes.

1- On choisit K individus dans la population multivariable étudiée. Ces K individus sont assimilés aux centres de gravité des K classes recherchées.

Application de l'initialisation monovariable à X^1 et X^2 :



Les seuils obtenus par initialisation monovariable sont S_1 et S_2 :



Le seuil optimal dans R^2 est S_2 :

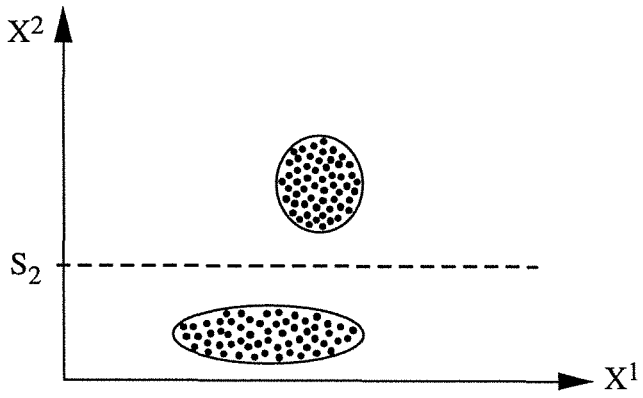


Fig. 3 Exemple de population pour laquelle les seuils obtenus par initialisation monovariable ne sont pas optimaux

2- On affecte tous les individus au centre de gravité le plus proche (avec la distance euclidienne), sachant que les centres de gravité sont recalculés après l'attribution de chaque individu.

3- On calcule les centres de gravité des K classes ainsi formées. Puis on réitère les étapes 2 et 3 jusqu'à convergence.

En final, cette méthode de partitionnement nous fournit K classes. On peut alors calculer les probabilités *a posteriori,* que nous utiliserons comme solution initiale de la méthode du maximum de vraisemblance.

III - IMPLÉMENTATION DE LA MÉTHODE DE DÉCOMPOSITION

1 Introduction

Nous avons retenu comme méthode de décomposition en classes gaussiennes la méthode du maximum de vraisemblance. Concernant la phase d'initialisation, notre choix a été :

- d'implémenter la méthode de Harding automatisée, en lui ajoutant différentes options permettant de travailler soit sur la fonction de répartition empirique, soit sur une estimation lissée de la fonction de répartition ;
- d'implémenter deux autres initialisations monovariables ;
- d'utiliser l'initialisation multivariable obtenue par des méthodes statistiques de classification.

Cela augmente le nombre de solutions initiales potentielles, sachant que de la solution initiale dépend la décomposition finale.

Par ailleurs, nous avons ajouté une contrainte portant sur le poids des classes, afin de supprimer les classes de trop faible poids. Cette contrainte joue à la fois pendant l'initialisation et pendant la partie itérative.

Enfin, le fait d'utiliser différentes initialisations risque de fournir plusieurs décompositions en classes gaussiennes. Nous avons donc implémenté des critères permettant de quantifier la qualité des décompositions obtenues.

2 Initialisations utilisées

2.1 Initialisations monovariables

2.1.1 Modifications de la méthode de Harding automatisée

Comme nous l'avons déjà signalé, la méthode de Harding automatisée (développée par Royer et Mezghache) repose sur une étape de recherche des seuils des classes à partir de la fonction de répartition de la population.

La solution classique consiste à utiliser la fonction de répartition empirique car son calcul est très rapide. Mais on risque d'obtenir un nombre très important de seuils, dont la majorité ne correspondra pas à de réelles modifications au sein de la population, du fait de l'aspect non lissé de cette fonction de répartition (fluctuations d'échantillonnage). Nous avons donc décidé de rajouter deux options dans le calcul de la fonction de répartition monovariable, fournissant une estimation lissée de cette fonction.

La première option consiste à lisser la fonction de répartition empirique par moyennes mobiles : il suffit à l'utilisateur de choisir la taille de la fenêtre de lissage et le pourcentage de recouvrement des fenêtres.

Quant à la seconde option, elle consiste à estimer la densité de la population par la méthode des noyaux avec un noyau du type Epanechnikov (se reporter à l'ouvrage de Silverman en 1986 sur l'utilisation des fonctions noyaux pour l'estimation de la densité), puis à intégrer cette densité afin d'estimer la fonction de répartition. L'utilisateur doit alors fournir la taille du noyau sachant que cette taille contrôle le degré de lissage de l'estimation. Notons que différents travaux sur les méthodes d'estimation de la densité par une fonction noyau proposent une taille optimale de noyau dans le cadre d'hypothèses gaussiennes (Sheather et Jones, 1991).

2.1.2 Seuils initiaux monovariables indiqués par l'utilisateur

L'utilisateur peut avoir une idée de la valeur des seuils d'une variable particulière (par exemple le pourcentage de porosité). Nous avons donc ajouté une option afin que l'utilisateur fournisse lui-même le nombre de seuils et leurs valeurs.

2.1.3 Recherche de classes monovariables de même effectif

Cette option consiste à créer des classes de même effectif sur chaque variable. Mais la solution initiale obtenue en croisant ces classes est très probablement loin de la solution optimale.

2.2 Initialisation multivariable

Nous avons décidé d'utiliser la classification multivariable fournie par la procédure "FASTCLUS" du logiciel SAS comme initialisation à la méthode du maximum de vraisemblance.

Cette initialisation multivariable a l'avantage d'être très rapide. Mais elle présente aussi un inconvénient important : l'utilisateur doit fixer *a priori* le nombre de classes de la partition, contrairement à l'initialisation par la méthode de Harding automatisée. Et il est probable que la décomposition finale de la population dépende du choix de ce nombre de classes.

3 Adjonction d'une contrainte de poids

Nous avons ajouté une contrainte de poids jouant pendant le déroulement complet du programme : si le poids d'une classe devient inférieur à un poids PMIN défini par l'utilisateur ($PMIN \geq 0.01$), cette classe est supprimée. Les probabilités d'appartenance aux classes restantes sont alors renormées.

Cette contrainte influe fortement sur la solution finale puisqu'elle revient implicitement à limiter le nombre de classes. Elle présente les deux avantages suivants.

- Elle limite le nombre de classes en ne recherchant que les composantes majeures de la population. En effet, elle peut permettre de supprimer les classes correspondant à une petite irrégularité locale au sein de la population.
- Elle permet d'obtenir des classes représentatives statistiquement, dont les paramètres (moyenne et matrice de variance-covariance) peuvent être correctement estimés. La contrainte de poids PMIN nous permet ainsi d'adapter le poids minimal des classes de la décomposition en fonction de la taille de la population et du nombre de variables.

4 Qualité de la décomposition

Dans la section précédente, nous avons présenté la méthode de décomposition d'une population multivariable en lois normales. Nous aimerions maintenant avoir un critère permettant de répondre à la question suivante : la décomposition obtenue est-elle de bonne qualité ?

Le problème de décomposition s'énonçant sous la forme :

$$F(x_i) = \sum_{k=1}^{K} p_k F_k(x_i), \quad \forall\, i = 1 \ldots n$$

où on cherche à estimer K, p_k et F_k, pour tout k de 1 à K, nous avons donc implémenté le critère de qualité de la norme L1 :

$$C_1 = \sum_{i=1}^{n} \left| \hat{F}(x_i) - \sum_{k=1}^{K} p_k F_k(x_i) \right|$$

où $\hat{F}$ est la fonction de répartition empirique.

Le second critère implémenté est celui de la norme L2 (critère de l'erreur quadratique MISE) :

$$C_2 = \sum_{i=1}^{n} \left[\hat{F}(x_i) - \sum_{k=1}^{K} p_k F_k(x_i) \right]^2$$

La meilleure décomposition pour une population donnée est celle minimisant le critère C_1 ou le critère C_2. Par rapport au critère C_1, le critère C_2 présente l'inconvénient d'accorder beaucoup d'importance à quelques points très mal ajustés ; ainsi, si notre population comporte une donnée aberrante, le critère C_2 peut amener à choisir une solution pour laquelle l'ajustement est globalement médiocre, et rejeter une solution pour laquelle le point aberrant aurait été très mal traité mais dont l'ajustement aurait été parfait pour le reste de la population. Il peut donc être intéressant de se baser sur ces deux critères pour choisir la meilleure décomposition.

Nous pourrions aussi utiliser le critère de la distance du χ^2 qui s'écrit :

$$C_3 = \sum_{i=1}^{n} \frac{\left[f(x_i) - \sum_{k=1}^{K} p_k f_k(x_i) \right]^2}{f(x_i)}$$

qui accorde plus d'importance à un ajustement médiocre dans les zones de faible densité que dans les zones de forte densité. Mais ce critère nécessite l'estimation de la densité multivariable de la population. Or c'est justement dans les zones de faible densité (queues de distribution) que cette estimation est la moins robuste. Il est donc difficile de mettre en place ce critère de façon fiable.

Par ailleurs, le problème de la qualité de la décomposition peut aussi s'envisager de la façon suivante : puisque nous cherchons à décomposer une population en sous-populations gaussiennes, pourquoi n'appliquerions nous pas différents tests de normalité à chacune des sous-populations obtenues (par exemple le test du χ^2, le critère de Kolmogorov, etc.) ? Ainsi

l'information obtenue ne porterait plus sur la somme des sous-populations gaussiennes mais sur chacune d'entre elles. Cependant, dans le cas de sous-populations chevauchantes, il est difficile de définir les individus devant participer aux tests.

En définitive, seuls les critères C_1 et C_2 ont été implémentés.

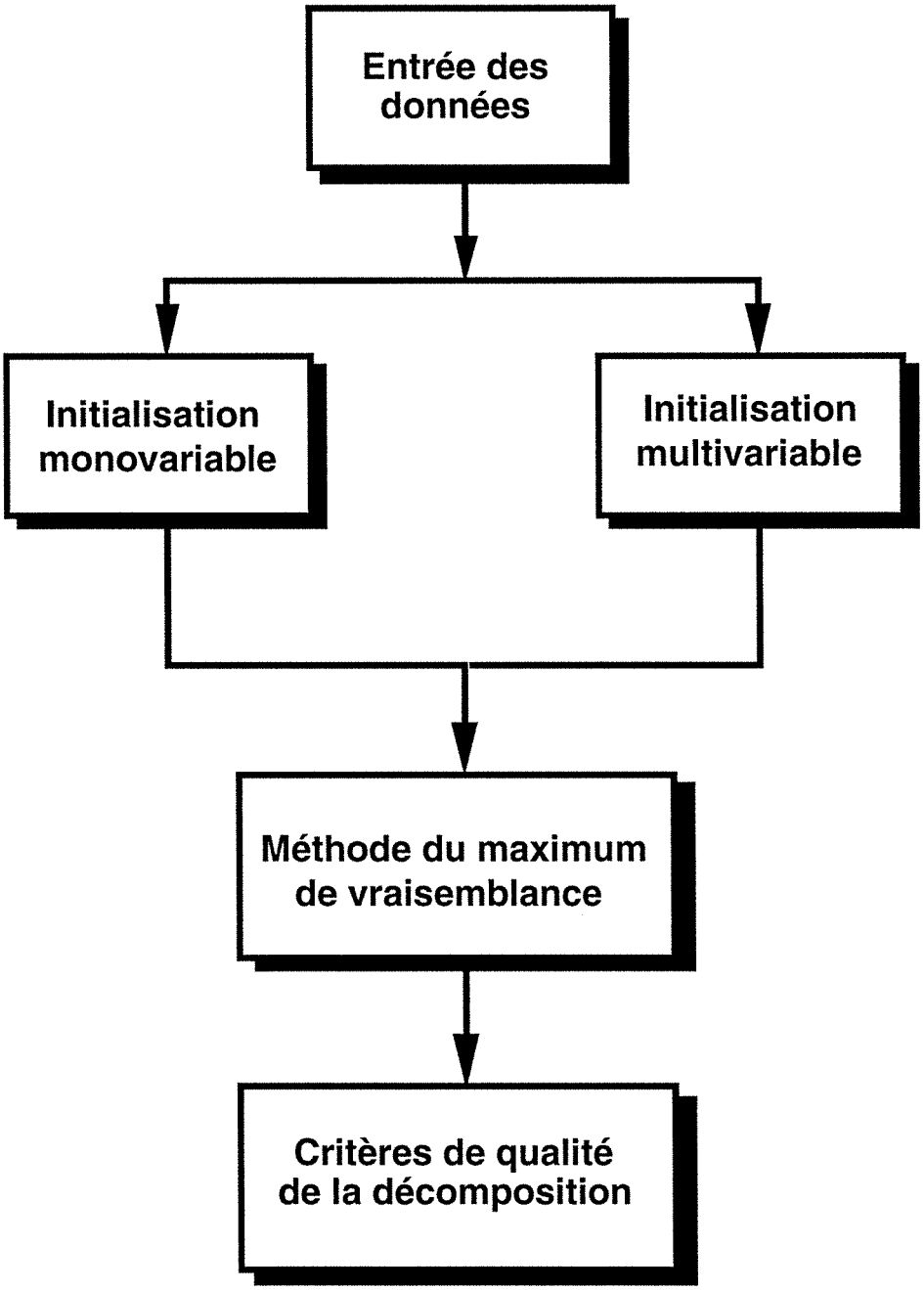
5 Architecture du programme

La méthode de décomposition d'une population en classes gaussiennes et ses diverses extensions forment le corps du programme que nous avons implémenté en Fortran77. Nous présentons dans ce paragraphe l'architecture de ce programme sous forme de trois organigrammes. Le premier organigramme illustre l'architecture générale du programme ; les deux autres développent plus précisément les initialisations implémentées (initialisations monovariables et multivariable).

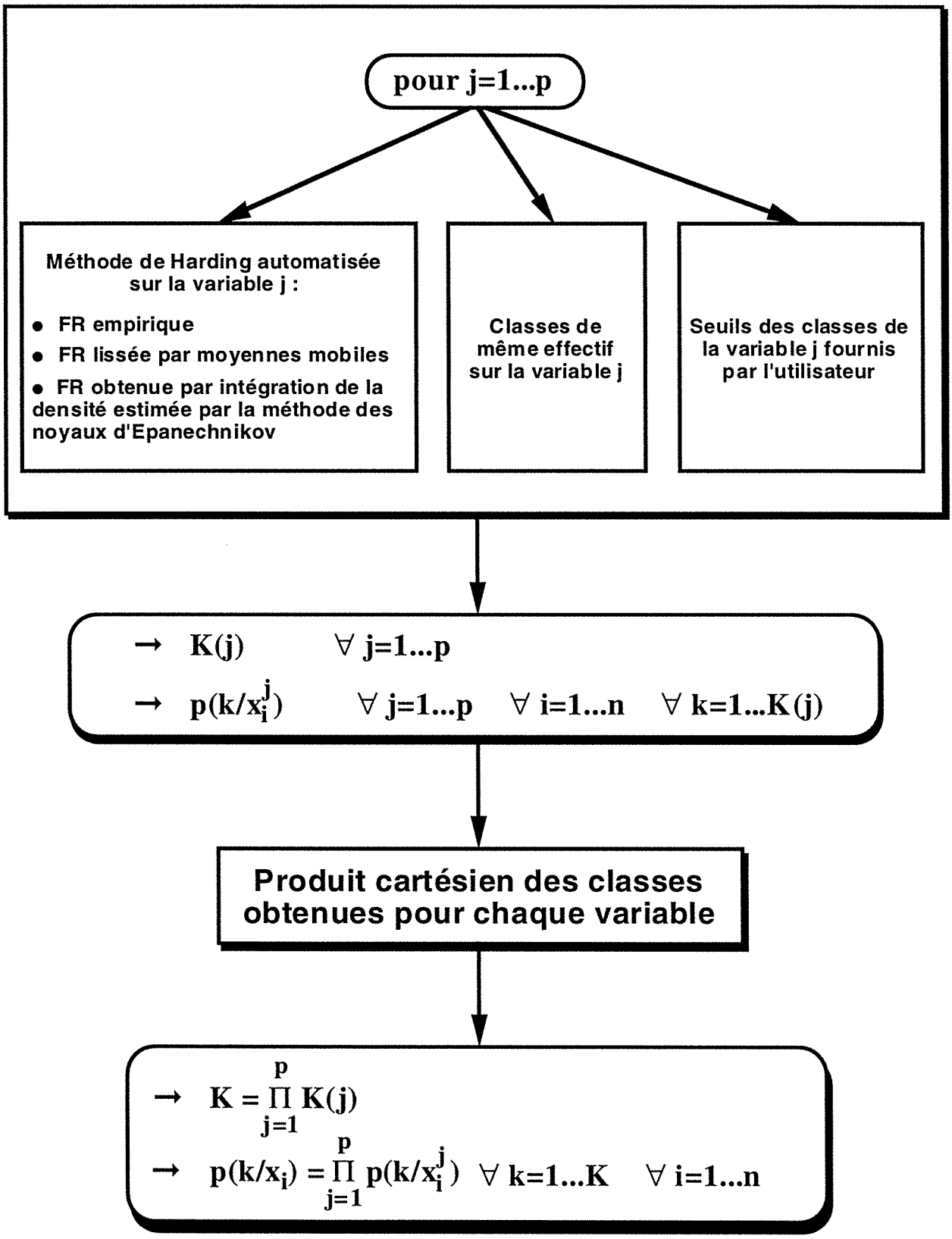
Les notations utilisées sont les suivantes :

- n est le nombre d'individus,
- p est le nombre de variables,
- $x_1 \ldots x_n$ sont les vecteurs des individus,
- $x^1 \ldots x^p$ sont les vecteurs des variables,
- PMIN est le poids minimal pour qu'une classe soit conservée,
- FR est la fonction de répartition,
- $K(j)$ est le nombre de classes pour la variable j, obtenu par initialisation monovariable,
- K est le nombre de classes dans l'espace multivariable, obtenu après initialisation,
- $p\left(k/x_i^j\right)$ est la probabilité *a posteriori* que l'individu x_i appartienne à la classe monovariable k, définie sur la variable j par initialisation monovariable,
- $p(k/x_i)$ est la probabilité *a posteriori* que l'individu x_i appartienne à la classe multivariable k, obtenue après initialisation.

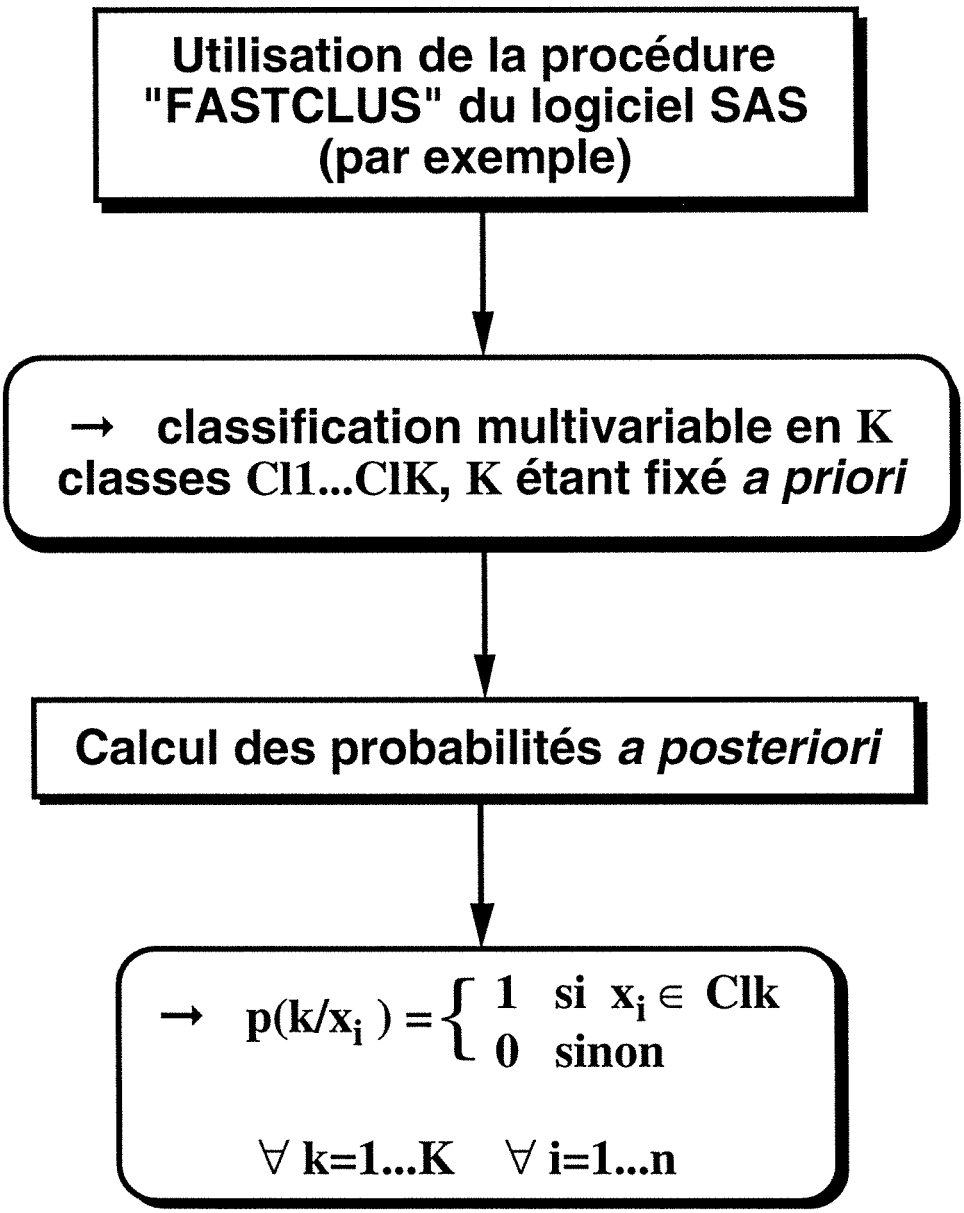
Implémentation de la méthode de décomposition



Initialisation monovariable



Initialisation multivariable



IV - MISE AU POINT SUR DES DONNÉES SYNTHÉTIQUES D'UNE MÉTHODOLOGIE DE DÉCOMPOSITION

Nous avons testé la méthode de décomposition sur plusieurs jeux de données synthétiques que nous avions auparavant simulés. Nous décrivons ci-dessous les fichiers de données. Dans une seconde section, nous analyserons les résultats des tests. Notons que nous utiliserons les critères de qualité C_1 (critère de la norme L1) et C_2 (critère de la norme L2) pour comparer les différentes solutions obtenues. Enfin, nous proposerons une démarche méthodologique pour la segmentation des données en s'appuyant sur les analyses précédentes.

1 Présentation des données

Nous avons généré cinq fichiers de données de différentes tailles, en dimension 2, en fonction de deux variables nommées X^1 et X^2. Pour chaque jeu de données, nous avons estimé la densité multivariable par la méthode des noyaux (avec le noyau d'Epanechnikov). Nous fournirons à chaque fois une figure représentant les données dans R^2 et une figure de la densité estimée. La densité estimée sera représentée par sept courbes de niveaux (les 5^e, 20^e, 35^e, 50^e, 65^e, 80^e et 95^e percentiles de la densité). Cela nous permettra de comparer la décomposition en classes gaussiennes obtenue avec les classes "naturelles" de la population.

Présentons maintenant les données synthétiques simulées.

Fichier 1

Nous avons généré une population de 2500 individus dans R^2 (cf. FIG. 4 et 5), suivant une loi normale de moyenne M et de matrice de variance-covariance Cov avec :

$$M = (1.4\ ,\ 0.4)$$

$$Cov = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$$

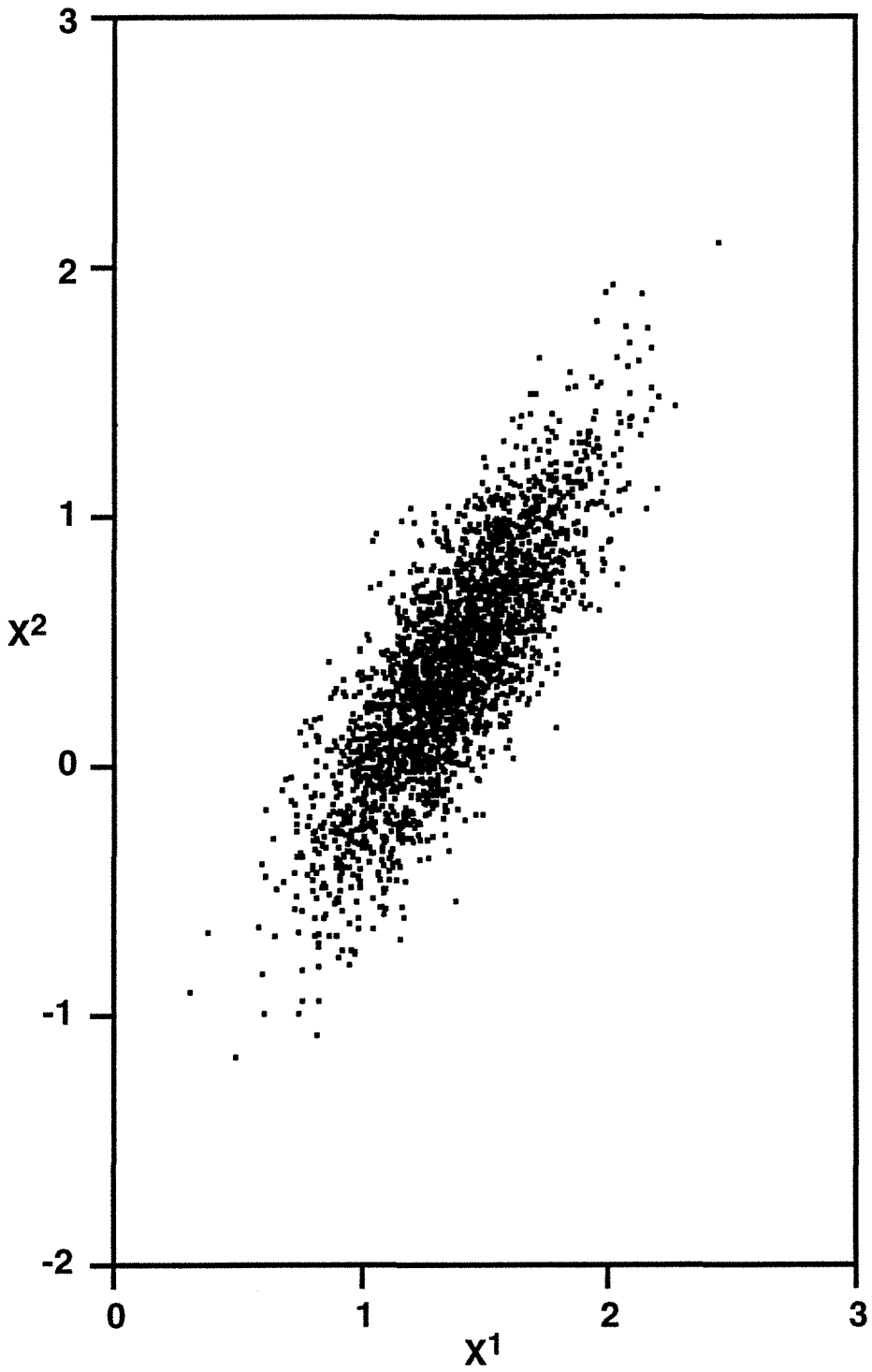


Fig. 4 Représentation de la population du fichier 1

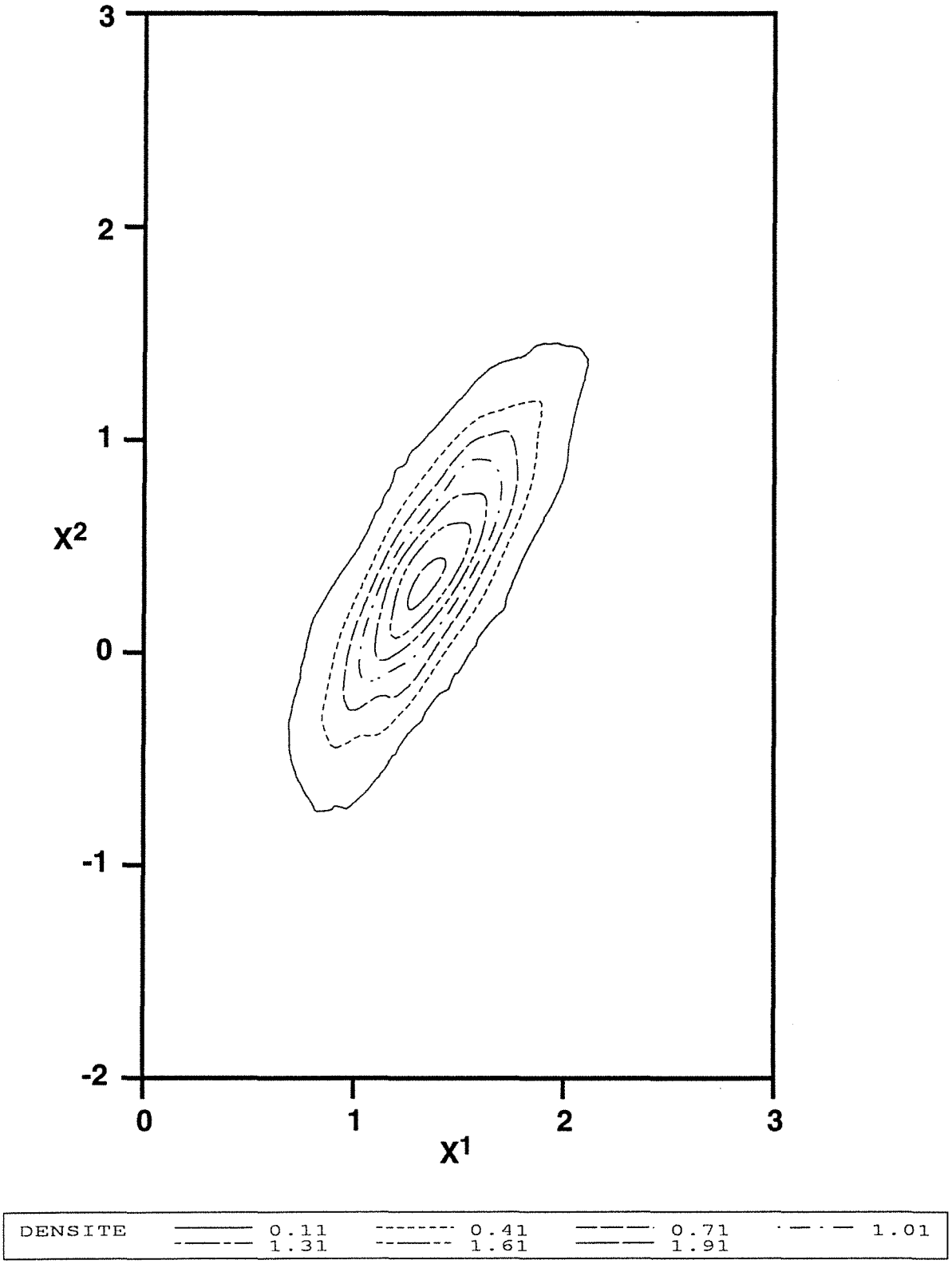


Fig. 5 Courbes de niveau de la densité de la loi empirique
de la population du fichier 1

Fichier 2

Nous avons généré trois gaussiennes non recouvrantes de 500 points chacune (cf. FIG. 6 et 7), de moyennes M_k et de matrices de variance-covariance Cov_k pour k allant de 1 à 3, avec :

$$M_1 = (2, -2)$$
$$M_2 = (0, 2)$$
$$M_3 = (-2, -2)$$
$$Cov_1 = \begin{bmatrix} 0.12 & -0.1 \\ -0.1 & 0.4 \end{bmatrix}$$
$$Cov_2 = \begin{bmatrix} 0.12 & 0.01 \\ 0.01 & 0.4 \end{bmatrix}$$
$$Cov_3 = \begin{bmatrix} 0.12 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}$$

Fichier 3

Nous avons généré trois gaussiennes très recouvrantes de 500 points chacune (cf. FIG. 8 et 9), de moyennes M_k et de matrices de variance-covariance Cov_k pour k allant de 1 à 3, avec :

$$M_1 = (2, -2)$$
$$M_2 = (0, 2)$$
$$M_3 = (-2, -2)$$
$$Cov_1 = \begin{bmatrix} 0.5 & -0.8 \\ -0.8 & 2.7 \end{bmatrix}$$
$$Cov_2 = \begin{bmatrix} 0.5 & 0.05 \\ 0.05 & 2.7 \end{bmatrix}$$
$$Cov_3 = \begin{bmatrix} 0.5 & 0.8 \\ 0.8 & 2.7 \end{bmatrix}$$

Fichier 4

Nous avons généré une population constituée de deux classes de 500 points (cf. FIG. 10 et 11). Chaque classe forme une relation non linéaire entre X^1 et X^2, simulée par le processus suivant :

$$\forall\, i = 1 \ldots 1000$$
$$\begin{cases} x_i^1 = \cos\left(\dfrac{2\pi i}{n}\right) + 1_{i>500} + \dfrac{\varepsilon}{10} \\ x_i^2 = \sin\left(\dfrac{2\pi i}{n}\right) + 1_{i>500} \times 0.3 + \dfrac{\varepsilon}{10} \\ \varepsilon \rightarrow N(0,1) \end{cases}$$

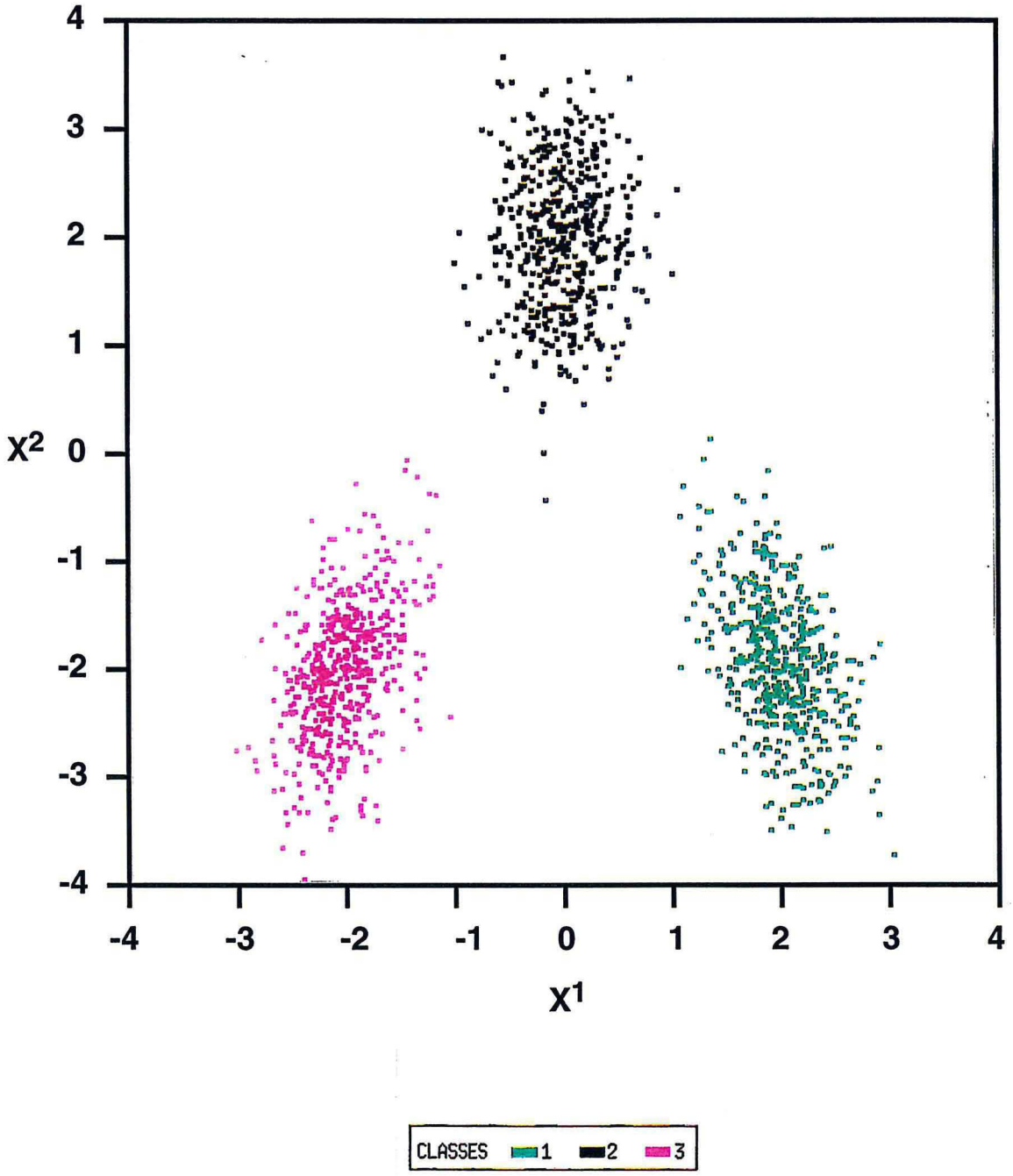


Fig. 6 Représentation de la population du fichier 2

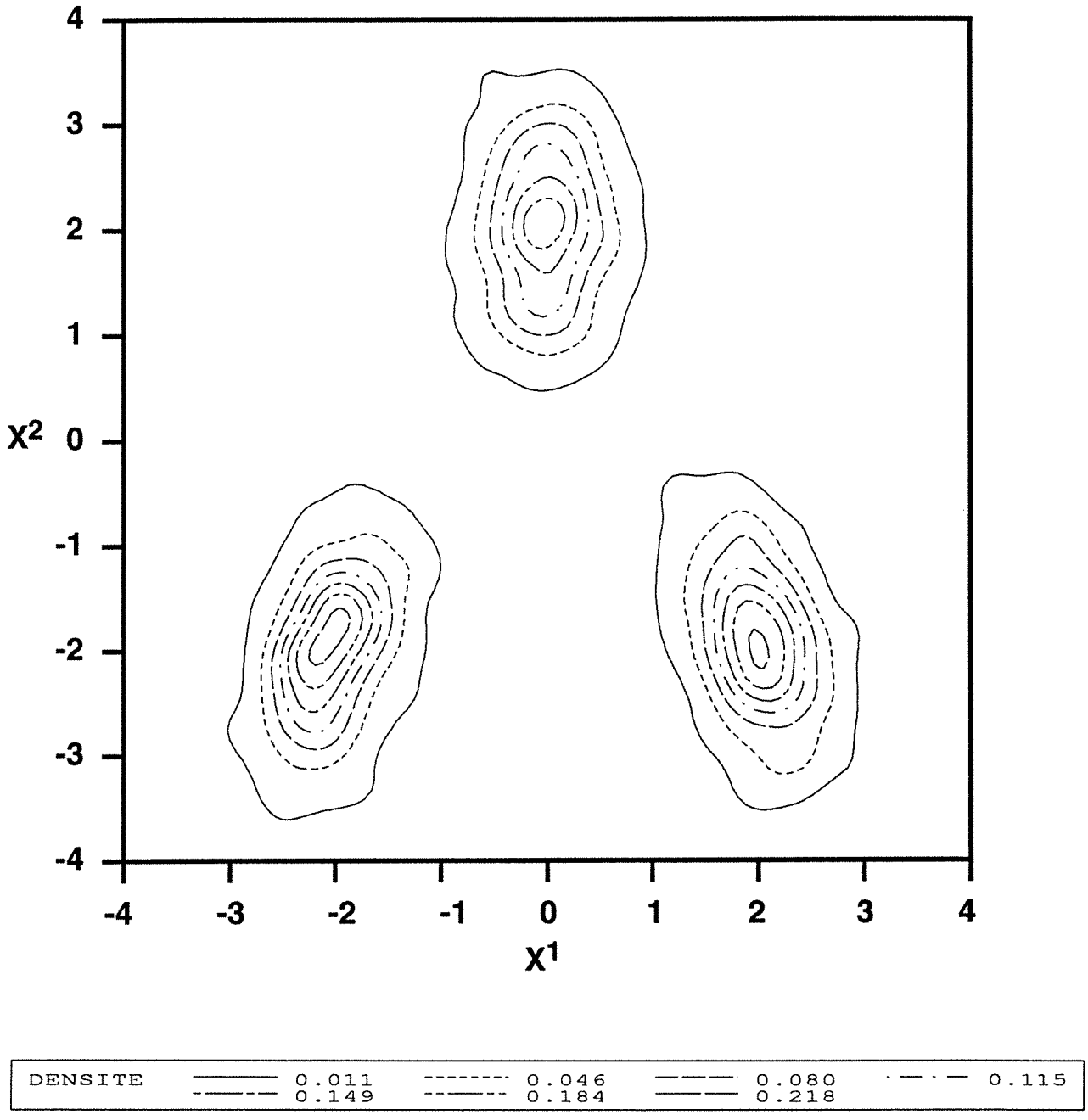


Fig. 7 Courbes de niveau de la densité de la loi empirique
de la population du fichier 2

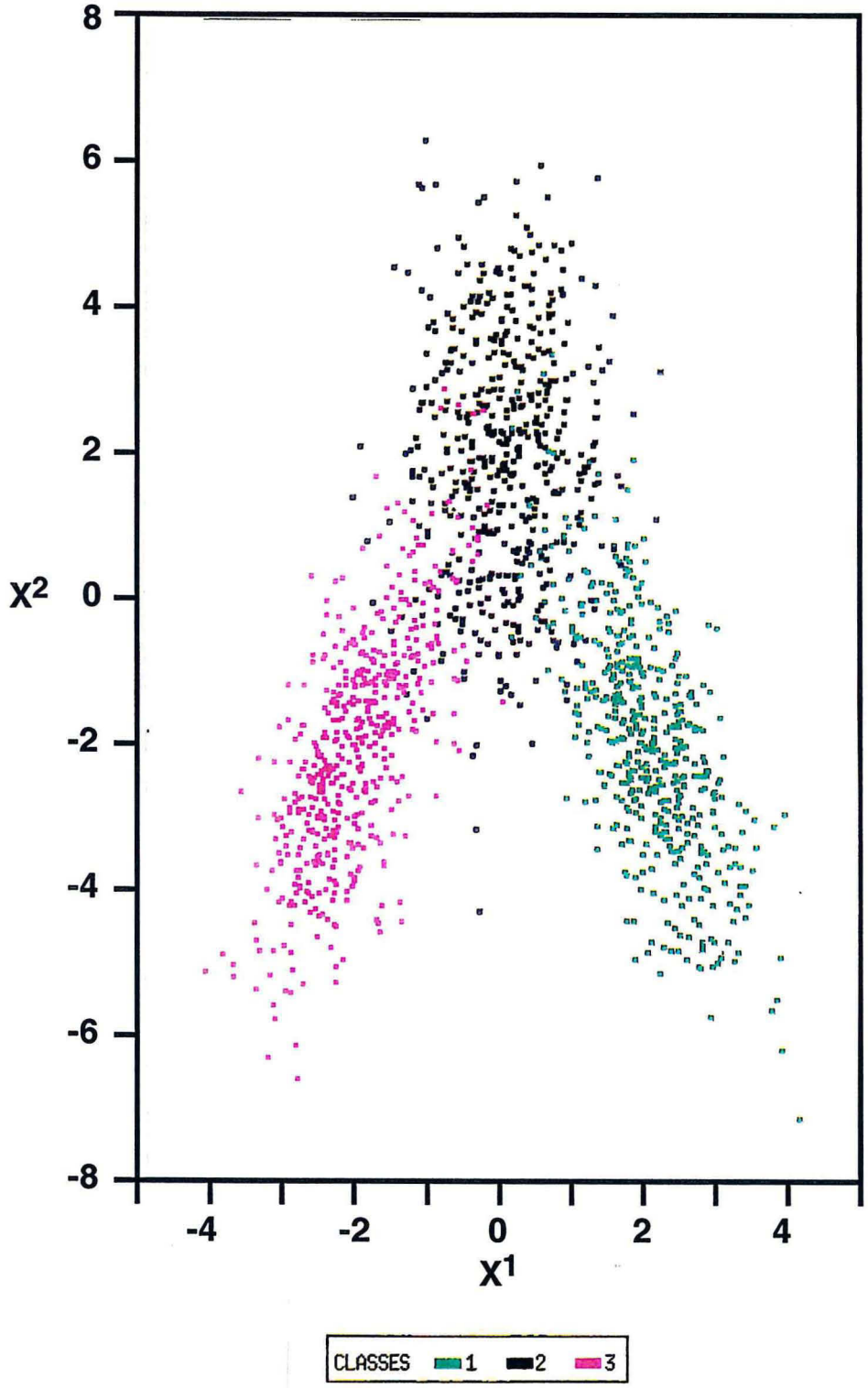


Fig. 8 Représentation de la population du fichier 3

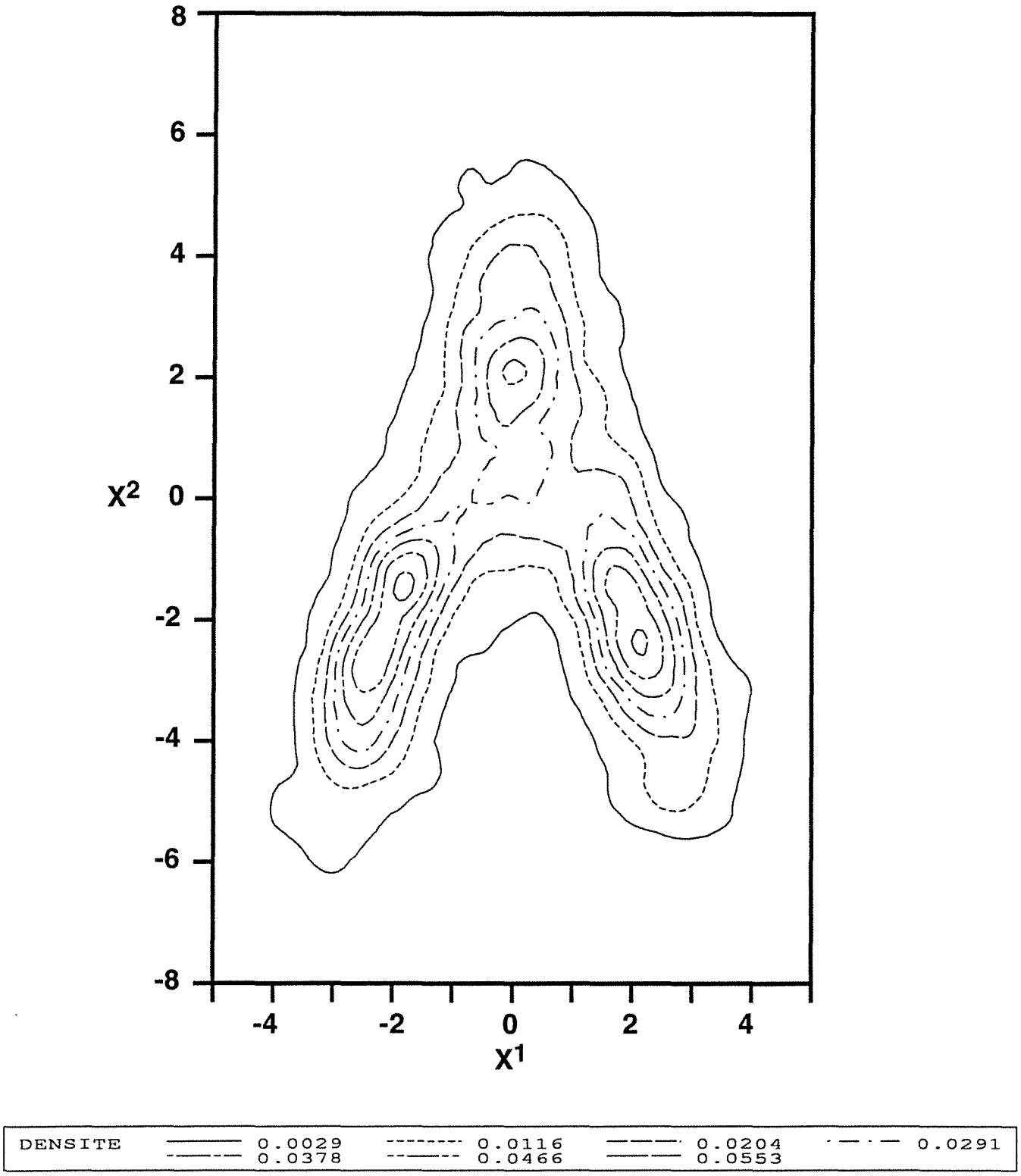


Fig. 9 Courbes de niveau de la densité de la loi empirique
de la population du fichier 3

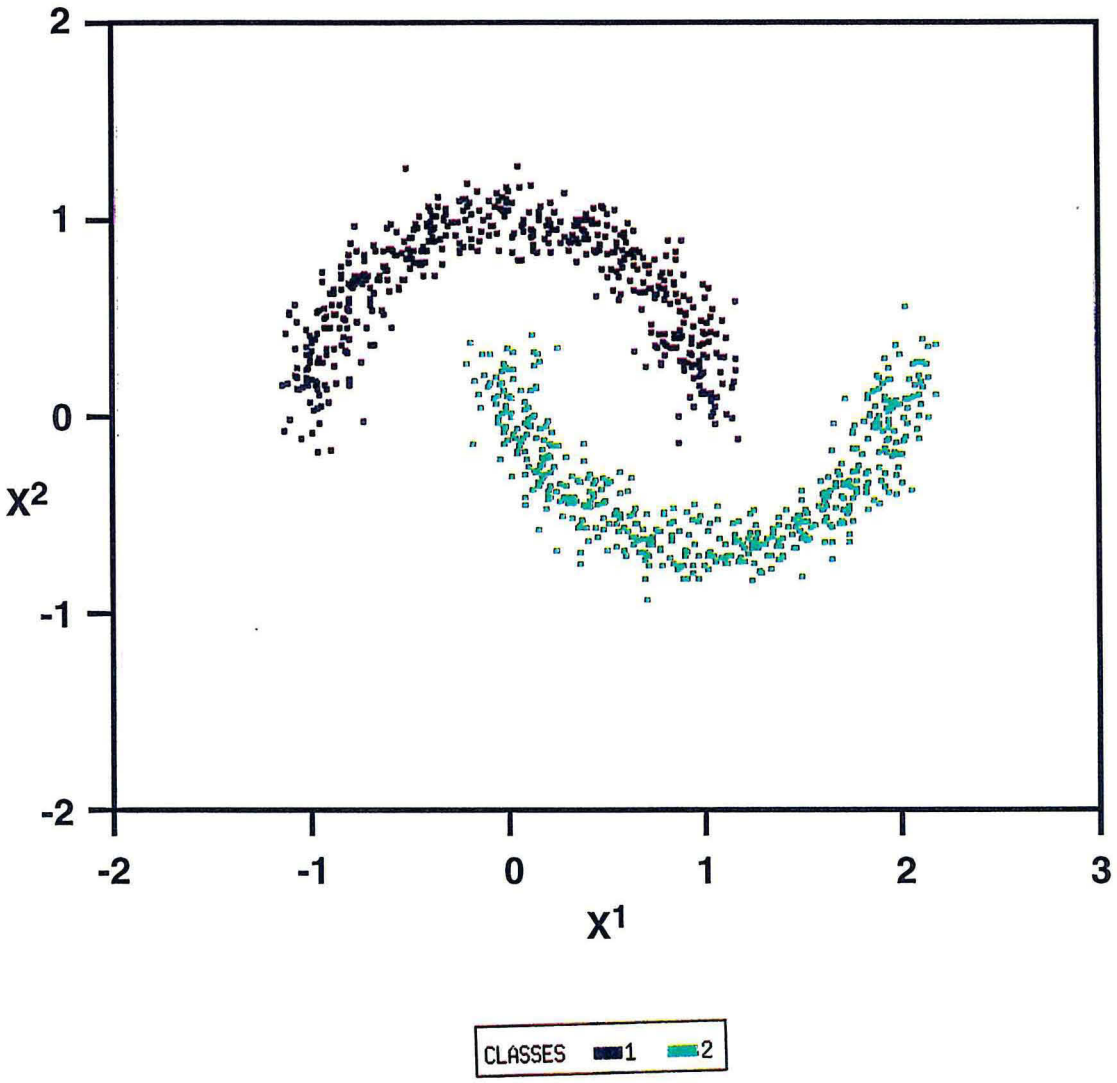


Fig. 10 Représentation de la population du fichier 4

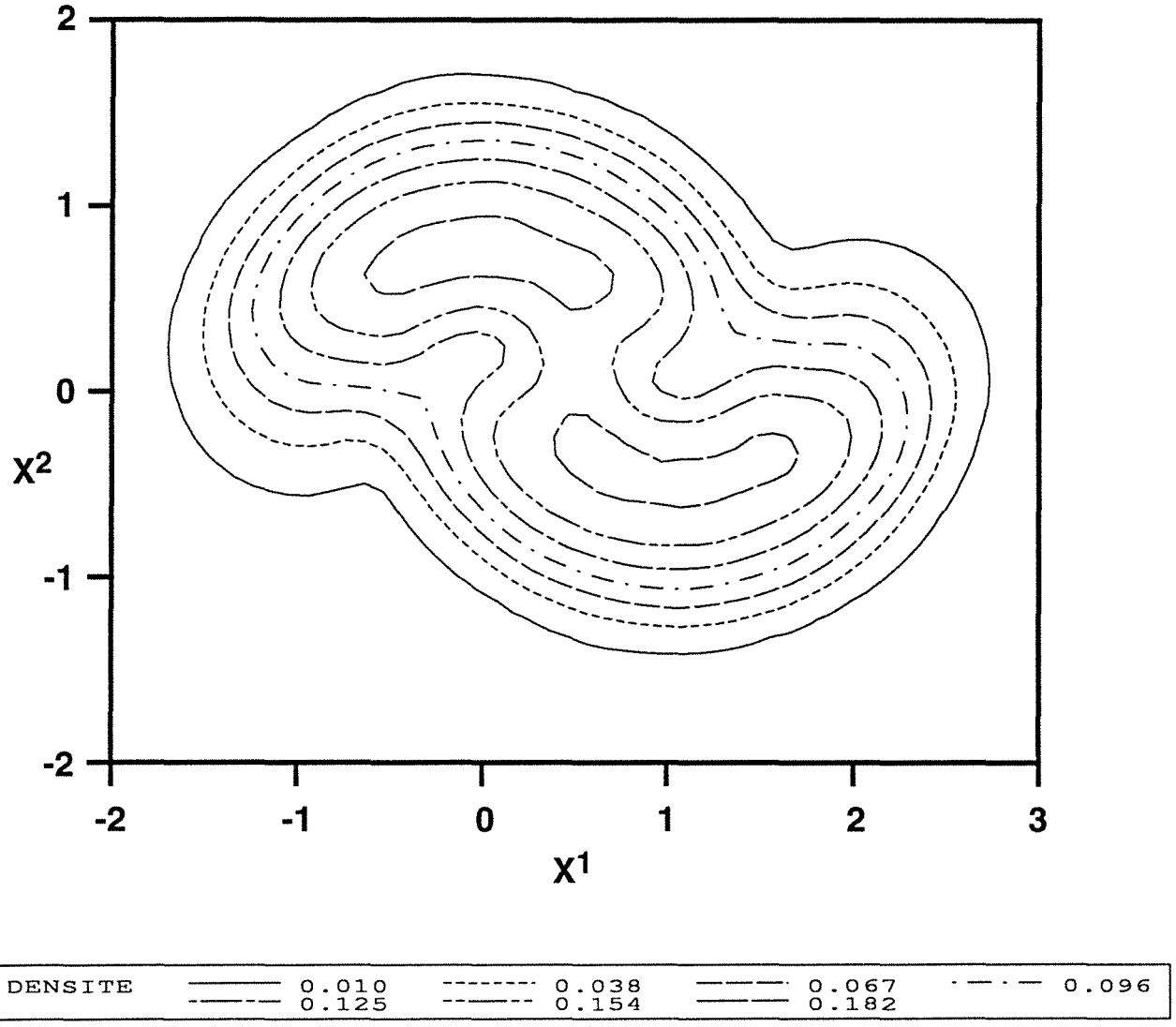


Fig. 11 Courbes de niveau de la densité de la loi empirique
de la population du fichier 4

Fichier 5

Nous avons généré une loi lognormale de 3000 points dans R^2 (cf. FIG. 12 et 13), de moyenne M et de matrice de variance-covariance Cov, avec :

$$M = (2, 0)$$

$$Cov = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.4 \end{bmatrix}$$

2 Analyse des résultats

Sur chacune des populations synthétiques, nous avons testé les algorithmes de décomposition en modifiant les paramètres suivants.

• **Type d'initialisation.**

Les initialisations utilisées sont celles mentionnées dans les organigrammes 2 et 3 présentés précédemment.

—> Initialisation multivariable avec choix de 5, 10, 15 ou 20 groupes *a priori*. Dans les tableaux de résultats figurant par la suite, nous appellerons **Multi** l'initialisation multivariable.

—> Initialisations monovariables suivantes : classes de même effectif sur chaque axe, méthode de Harding automatisée sur la fonction de répartition empirique, méthode de Harding automatisée sur la fonction de répartition lissée par intégration de la densité avec plusieurs valeurs pour le paramètre de lissage h de la fonction d'Epanechnikov, méthode de Harding automatisée sur la fonction de répartition lissée par moyenne mobile avec recouvrement ou non des fenêtres. Dans les tableaux de résultats figurant par la suite, nous utiliserons les notations suivantes pour les initialisations monovariables : **MEff** pour la recherche de classes de même effectif, **FR** pour l'utilisation de la fonction de répartition empirique, **FRE** pour l'utilisation de la fonction de répartition lissée par intégration de la densité, **FRL** pour l'utilisation de la fonction de répartition lissée par moyenne mobile.

• **Poids PMIN**, poids minimal qu'une classe doit avoir pour être conservée. Nous prendrons comme valeurs de PMIN 0.05, 0.1 et 0.15. D'autre part, pour forcer notre méthode de décomposition à trouver comme solution finale celle à une seule classe, nous utiliserons aussi PMIN égal à 0.5 : nous trouverons donc toujours dans les tests présentés ci-dessous la solution à une classe, à titre de référence, même si celle-ci s'ajuste mal aux données.

Analysons maintenant les résultats de ces tests.

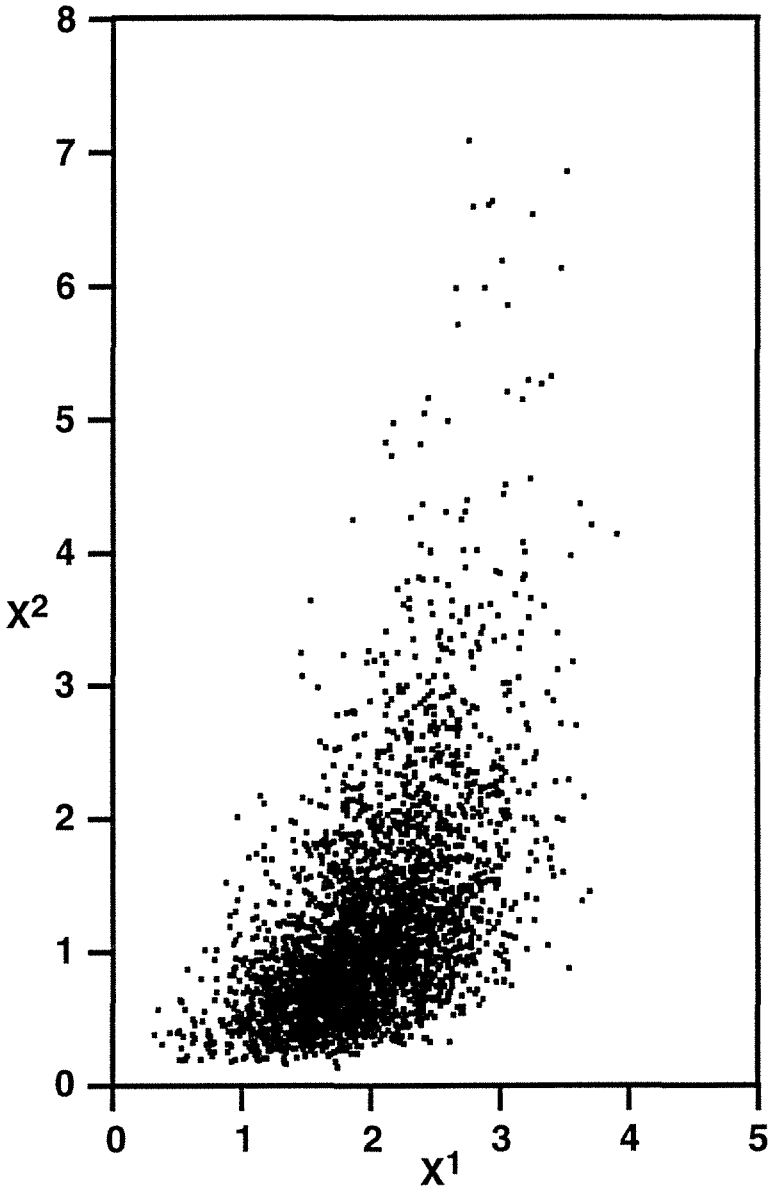
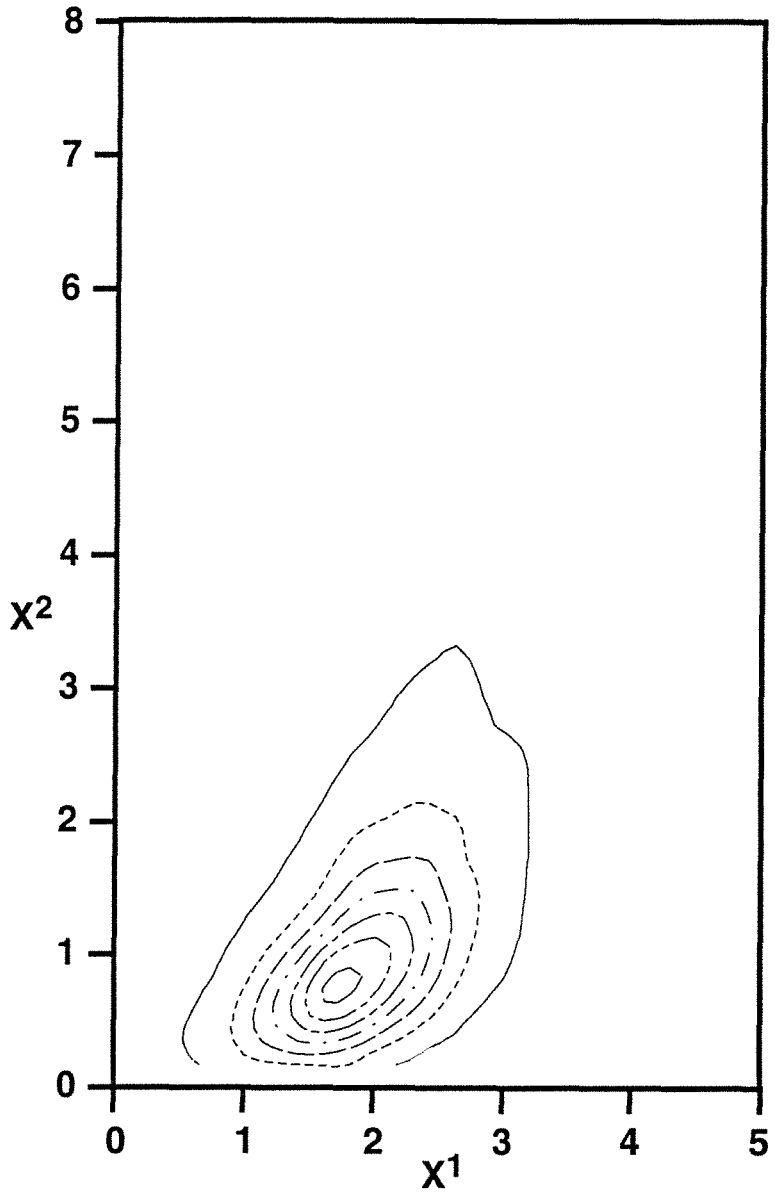


Fig. 12 Représentation de la population du fichier 5



DENSITE 0.033 0.125 0.218 0.310 0.403 0.495 0.588

Fig. 13 Courbes de niveau de la densité de la loi empirique de la population du fichier 5

Fichier 1

Les différents résultats concernant ce fichier se trouvent dans le tableau 1 ci-dessous et en Annexe A (tableaux A-1-1 à A-1-3).

Tableau 1 *Résultats partiels du fichier 1*

	Nb de classes	Critère C1	Critère C2	PMIN	Initialisations
sol0 *	1	131.91	8.82	0.5	toutes
sol1	2	134.41	9.15	0.15, 0.1, 0.05	MEff, FRE, FRL, Multi
sol2	2	132.84	8.95	0.1, 0.05	Multi
sol3	3	133.05	8.93	0.1, 0.05	MEff, FR, FRE, Multi
sol4	3	132.85	9.06	0.05	FRE, FRL, Multi
sol5	5	134.64	9.30	0.05	FRL

(* solution retenue)

Nous obtenons six solutions différentes : une solution à 1 classe (***sol0***), deux solutions à 2 classes (***sol1***, ***sol2***), deux solutions à trois classes (***sol3***, ***sol4***) et une solution à 5 classes (***sol5***). Notons que la solution à 1 classe, qui paraît *a priori* la plus appropriée (cf. FIG. 4 et 5), est obtenue uniquement pour PMIN égal à 0.5.

Pour se faire une idée des autres solutions fournies par la méthode de décomposition, nous joignons les figures 14a et 14b qui décrivent la représentation graphique de la solution ***sol2*** ainsi que sa densité gaussienne.

Nous pouvons déjà tirer quelques conclusions des solutions obtenues. Tout d'abord, la méthode de décomposition ne fournit pas une solution unique quels que soient le poids PMIN ou le type d'initialisation choisis.

Ensuite, nous constatons que le poids PMIN joue un rôle important puisque pour une même initialisation, on obtient des solutions différentes selon la valeur de PMIN. Par exemple, en appliquant la méthode de décomposition sur la fonction de répartition lissée par moyennes mobiles, on obtient soit la solution ***sol1*** (solution à 2 classes) pour PMIN égal à 0.1, soit la solution ***sol4*** (solution à 3 classes) pour PMIN égal à 0.05.

Enfin, à poids PMIN identique, on obtient aussi différentes solutions en fonction du type d'initialisation choisi. Par exemple, pour PMIN égal à 0.05, on obtient cinq solutions différentes (***sol1*** à ***sol5***) selon qu'on utilise une initialisation multivariable ou qu'on travaille sur la fonction de répartition empirique ou lissée.

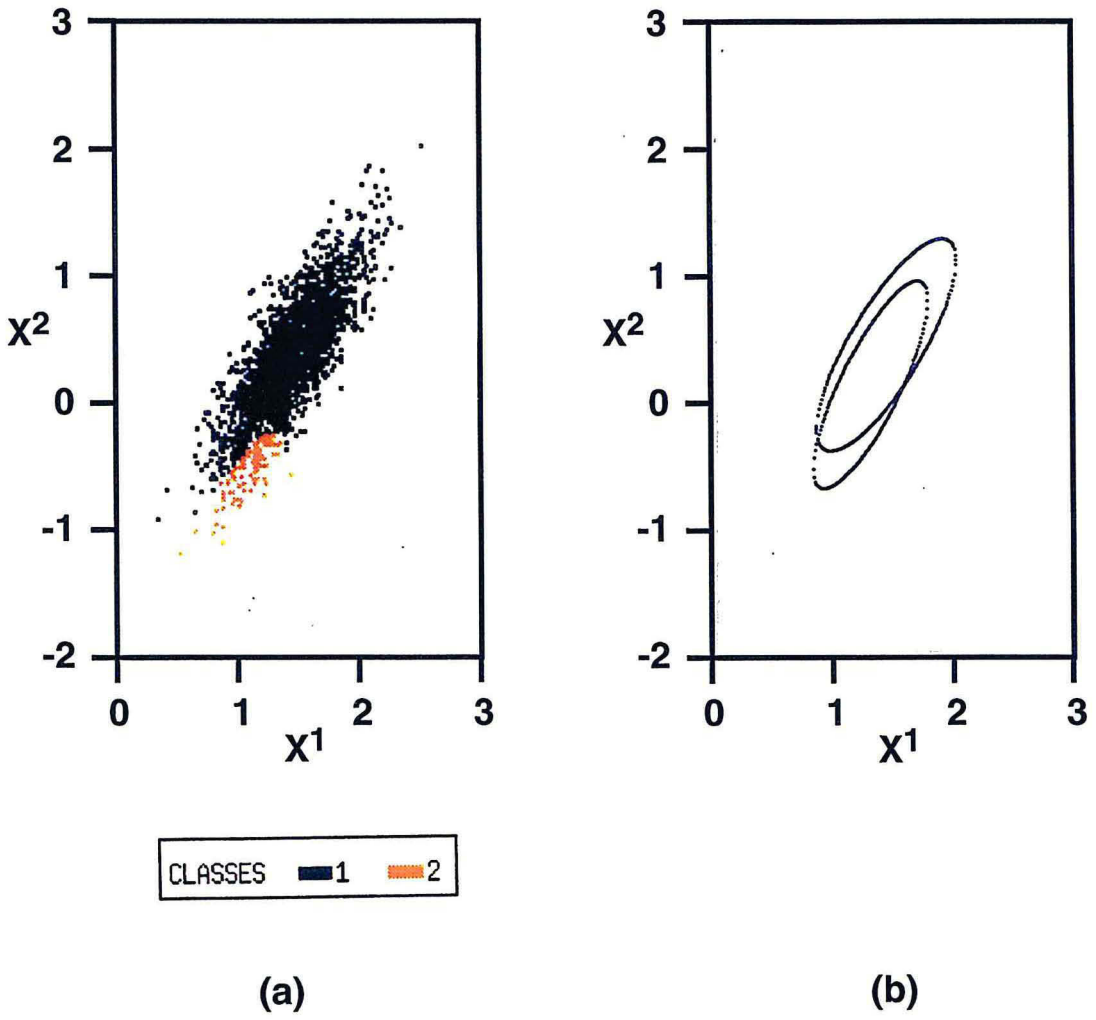


Fig. 14 a. Solution ***sol2*** à deux classes gaussiennes
b. Courbe de niveau à 95% des classes gaussiennes de ***sol2***

Considérons maintenant les critères de qualité de l'ajustement C_1 et C_2. Nous constatons que le critère C_1 (critère de la norme L1) varie peu : il reste compris entre 131.91 et 134.64. De même, le critère C_2 (critère MISE) varie de 8.82 à 9.30. Et même si la solution à 1 classe est celle pour laquelle les deux critères atteignent leur minimum, aucune solution n'est mauvaise en terme de qualité de l'ajustement. Ceci s'explique peut-être par des irrégularités très locales de la population qui seraient plus ou moins prises en compte en fonction de l'initialisation appliquée, d'où différentes solutions mais de qualité assez semblable.

Finalement nous retiendrons comme solution optimale pour cette population celle à 1 classe, qui a été obtenue pour PMIN égal à 0.5 et quelle que soit l'initialisation.

Fichier 2

Les résultats des tests sur ce fichier sont fournis dans le tableau 2 ci-dessous et dans les tableaux A-2-1 à A-2-3 en Annexe A.

Pour ce fichier, nous obtenons une solution à 1 classe ***sol0*** (obtenue pour PMIN égal à 0.5), une solution à 3 classes (***sol1***), deux solutions à 4 classes (***sol2***, ***sol3***), deux solutions à 5 classes (***sol4***, ***sol5***), une solution à 6 classes (***sol6***) et une solution à 7 classes (***sol7***). Nous joignons les figures 15a et 15b représentant graphiquement la solution ***sol4*** et l'estimation de sa densité.

Tableau 2 *Résultats partiels du fichier 2*

	Nb de classes	Critère C1	Critère C2	PMIN	Initialisations
sol0	1	165.26	36.04	0.5	toutes
sol1	3	29.31	0.93	0.15, 0.1	MEff, FRE, FRL, Multi
sol2	4	29.23	0.93	0.1	MEff, FR, FRE
sol3	4	28.71	0.89	0.05	Multi
sol4 *	5	28.64	0.89	0.05	MEff, FRE, Multi
sol5	5	28.66	0.89	0.05	MEff, FRE, Multi
sol6	6	28.64	0.90	0.05	FR, FRE
sol7	7	28.67	0.91	0.05	FRL

(* solution retenue)

Lorsque nous analysons les résultats, nous constatons de nouveau l'importance du poids PMIN. En effet, nous obtenons des solutions comportant 1 à 4 classes pour PMIN supérieur ou

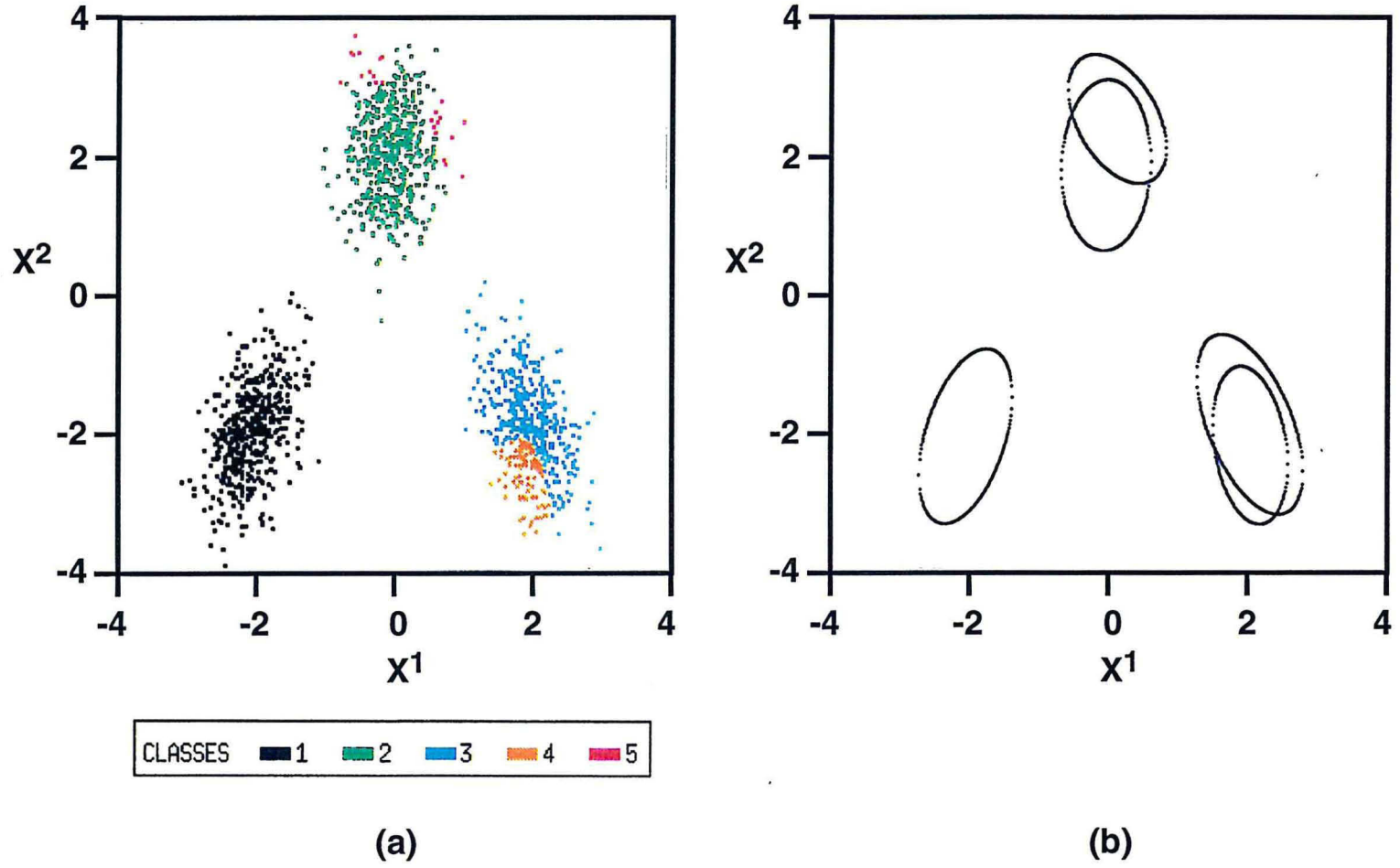


Fig. 15 a. Solution ***sol4*** à cinq classes gaussiennes
b. Courbe de niveau à 95% des classes gaussiennes de ***sol4***

égal à 0.1 (***sol0***, ***sol1*** et ***sol2***), alors que nous obtenons de 4 à 7 classes pour PMIN égal à 0.05 (***sol3*** à ***sol7***).

De même, nous remarquons que le type d'initialisation influe aussi sur la solution. Par exemple, pour PMIN égal à 0.1, nous trouvons 2 solutions différentes :

- ***sol1*** obtenue soit par initialisation multivariable, soit en utilisant la fonction de répartition lissée par moyennes mobiles, soit en calculant la fonction de répartition à partir de la densité estimée par la méthode des noyaux (avec des noyaux de grande taille), soit en recherchant 10 classes de même effectif pour chaque variable ;
- ***sol2*** obtenue soit en utilisant la fonction de répartition empirique, soit en calculant la fonction de répartition à partir de la densité estimée par la méthode des noyaux (avec des noyaux de petite taille), soit en recherchant 5 classes de même effectif pour chaque variable.

D'après les critères de qualité C_1 et C_2, nous pouvons éliminer d'emblée la solution ***sol0*** à 1 classe, qui, compte tenu de la population générée (cf. FIG. 6), était *a priori* irréaliste. Parmi les sept autres solutions, nous pouvons aussi éliminer ***sol1*** et ***sol2*** dont la qualité d'ajustement est légèrement dégradée pour les critères C_1 et C_2. La solution ***sol1*** semblait être *a priori* la décomposition la plus adaptée à la population étudiée puisque celle-ci est composée de trois sous-populations gaussiennes. Son rejet s'explique sans doute par les irrégularités locales visibles sur la figure 7 : ces irrégularités se répercutent sur la fonction de répartition multivariable empirique, et donc sur la valeur des critères C_1 et C_2.

Les cinq solutions restantes (***sol3*** à ***sol7***) sont très similaires (critère C_1 compris entre 28.63 et 28.71, critère C_2 compris entre 0.89 et 0.91). Il apparaît ainsi qu'il n'y a pas une solution optimale mais plusieurs. Nous pouvons cependant faire deux remarques :

- chacune des solutions ***sol4*** à ***sol6*** est issue de ***sol3*** (segmentation d'une des classes gaussiennes de ***sol3*** en 2 ou 3 classes gaussiennes) ;
- il est plus intéressant d'avoir le moins de classes possible, ou tout au moins un nombre suffisant d'individus par classe, afin de pouvoir estimer correctement les paramètres statistiques (moyenne et matrice de variance-covariance) de chaque classe.

Deux solutions peuvent donc être retenues : la solution à 4 classes ***sol3*** si on veut privilégier un faible nombre de classes, ou la solution à 5 classes ***sol4*** si on préfère tenir compte de la valeur des critères. Pour notre part, nous retiendrons la solution ***sol4*** car elle nous semble plus stable que la solution ***sol3*** : une seule initialisation permet d'atteindre la solution ***sol3*** alors que la solution ***sol4*** est obtenue pour cinq initialisations différentes. En effet, la solution ***sol4*** est obtenue pour PMIN égal à 0.05, par initialisation multivariable avec 15 classes *a priori*, ou par initialisation monovariable (recherche de 5 classes de même effectif sur chaque variable, ou

utilisation de la fonction de répartition lissée obtenue par intégration de la densité estimée par la méthode des noyaux).

Fichier 3

Nous pouvons constater que le graphe de l'estimation de la densité (cf. FIG. 9) de ce fichier est très accidenté. Cela va certainement se répercuter sur la fonction de répartition : il est possible qu'on trouve un grand nombre de seuils lors de l'étape d'initialisation, et ainsi un nombre final de classes élevé pour prendre en compte toutes les irrégularités locales. D'autre part, la population générée étant composée de trois sous-populations gaussiennes (cf. FIG. 8), nous espérons trouver des solutions à au moins trois classes par la méthode de décomposition gaussienne.

De fait, nous obtenons neuf solutions différentes : une solution à 1 classe (***sol0***) obtenue pour PMIN égal à 0.5, une solution à 3 classes (***sol1***), deux solutions à 4 classes (***sol2***, ***sol3***), deux solutions à 5 classes (***sol4***, ***sol5***) et trois solutions à 6 classes (***sol6***, ***sol7***, ***sol8***). On trouvera les résultats des tests dans le tableau 3 ci-dessous et dans les tableaux A-3-1 à A-3-3 en Annexe A.

Il est intéressant de noter que si PMIN est égal à 0.1, nous obtenons toujours la solution à 3 classes ***sol1***, alors que pour PMIN égal à 0.05, nous obtenons des solutions comprenant de 4 à 6 classes (***sol2*** à ***sol8***) : la valeur de PMIN permet donc de prendre plus ou moins en compte les irrégularités au sein de la population.

Tableau 3 *Résultats partiels du fichier 3*

	Nb de classes	Critère C1	Critère C2	PMIN	Initialisations
sol0	1	84.27	6.99	0.5	toutes
sol1	3	24.23	0.68	0.15, 0.1	MEff, FR, FRE, FRL, Multi
sol2	4	23.86	0.66	0.05	FRL
sol3	4	23.36	0.64	0.05	Multi
sol4 *	5	22.83	0.60	0.05	MEff, FRE, Multi
sol5	5	23.57	0.66	0.05	FRE
sol6	6	22.85	0.59	0.05	FR, Multi
sol7	6	23.81	0.66	0.05	MEff
sol8	6	23.45	0.65	0.05	FRE

(* solution retenue)

D'après les critères de qualité de l'ajustement, nous pouvons éliminer la solution à 1 classe (***sol0***) qui est de mauvaise qualité. Des huit solutions restantes, nous ne conservons que ***sol4*** (solution à 5 classes) et ***sol6*** (solution à 6 classes) qui sont de qualité équivalente ($C_1 \approx 22.8$ et $C_2 \approx 0.6$). Dans le cadre d'un futur calibrage, nous retiendrons comme solution optimale celle comportant le moins de classes, soit ***sol4***. Cette solution est obtenue pour PMIN égal à 0.05 et pour différentes initialisations, notamment les initialisations multivariables avec 10 ou 20 classes *a priori* et l'initialisation monovariable utilisant une fonction de répartition lissée (obtenue par intégration de la densité estimée par la méthode des noyaux). Nous fournissons ci-joint la représentation graphique de la solution ***sol4*** ainsi que celle de l'estimation de sa densité gaussienne (cf. FIG. 16a et 16b).

Fichier 4

Les résultats sont présentés dans le tableau 4 ci-dessous et dans les tableaux A-4-1 à A-4-4 en Annexe A.

Tableau 4 *Résultats partiels du fichier 4*

	Nb de classes	Critère C1	Critère C2	PMIN	Initialisations
sol0	1	71.02	11.89	0.5	toutes
sol1	4	15.96	0.60	0.1	FRE, Multi
sol2	5	12.14	0.35	0.15	FRE
sol3	5	12.03	0.37	0.1	FR
sol4	5	12.72	0.31	0.1, 0.05	Multi
sol5	6	7.97	0.14	0.1	MEff, FRE, FRL, Multi
sol6	7	7.31	0.13	0.05	MEff
sol7	9	6.03	0.10	0.05	Multi
sol8 *	9	5.11	0.05	0.05	FRE
sol9	9	6.30	0.10	0.05	MEff, Multi
sol10	10	4.98	0.05	0.05	FRE, FRL
sol11	10	5.34	0.05	0.05	FR, FRE
sol12	11	4.93	0.05	0.05	Multi

(* solution retenue)

L'intérêt de ce fichier réside dans le fait que la population générée présente une relation réellement non linéaire entre les deux variables de départ, comme le montrent sa représentation graphique (cf. FIG. 10) et l'estimation de sa densité (cf. FIG. 11). La méthode de décomposition

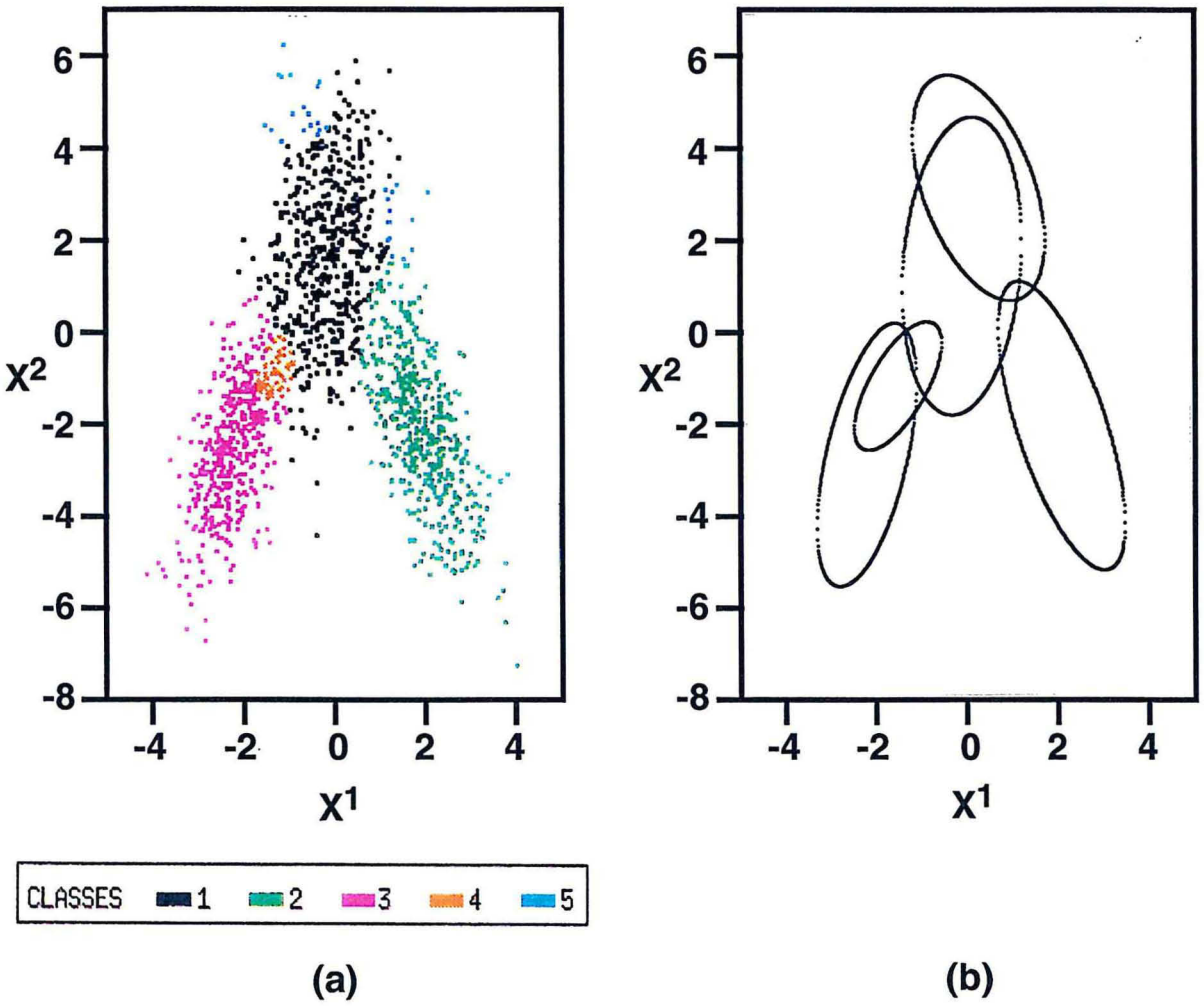


Fig. 16 a. Solution ***sol4*** à cinq classes gaussiennes
b. Courbe de niveau à 95% des classes gaussiennes de ***sol4***

en classes gaussiennes fournit douze solutions différentes (comportant de 1 à 11 classes), suivant les valeurs de PMIN et l'initialisation choisie. Ainsi, nous obtenons des solutions comprenant 1 à 6 classes pour PMIN supérieur ou égal à 0.1 (***sol0*** à ***sol5***), et des solutions comprenant 5 à 11 classes pour PMIN égal à 0.05 (***sol4***, ***sol6*** à ***sol12***). Par ailleurs, pour PMIN égal à 0.1, nous obtenons par exemple la solution ***sol3*** (solution à 5 classes) par initialisation monovariable sur la fonction de répartition empirique, ou la solution ***sol4*** (solution à 5 classes) par initialisation multivariable. La représentation graphique de ***sol3*** et celle de sa densité gaussienne correspondent aux figures 17a et 17b.

Parmi toutes les solutions obtenues, les deux critères C_1 et C_2 permettent de supprimer les solutions à moins de 6 classes (C_1>12.7 et C_2> 0.31). Pour les autres solutions, C_1 est compris entre 4.93 et 7.97 et C_2 est compris entre 0.05 et 0.14. En fait, quatre solutions sont de très bonne qualité : ***sol8***, ***sol10***, ***sol11*** et ***sol12***. Cependant, nous ne conserverons pas la solution ***sol12*** car la plus petite de ses classes a un poids trop faible pour permettre une bonne estimation de ses paramètres. Et nous ne conserverons pas non plus la solution ***sol11*** qui est moins bonne que la solution ***sol10*** avec un même nombre de classes. Deux possibilités s'offrent donc à nous : retenir la solution ***sol8*** qui comporte le moins de classes ou retenir la solution ***sol10*** qui est très légèrement meilleure en terme de critères. Finalement, le faible gain en qualité de la solution ***sol10*** ne nous semble pas suffisant en regard du nombre de classes déjà très important des deux solutions (surtout en vue d'un calibrage). Nous retiendrons donc comme pour les fichiers précédents la solution ***sol8*** comportant le moins de classes (9 classes).

Cette solution est obtenue pour PMIN égal à 0.05 et par initialisation sur la fonction de répartition intégrée à partir de l'estimation de la densité par la méthode des noyaux (avec noyau d'Epanechnikov).

Notons que pour chacune des solutions, le poids des classes est très homogène : ainsi le poids des classes varie de 0.085 à 0.133 pour ***sol8*** et de 0.17 à 0.22 pour ***sol2*** ; d'autre part, les bornes des classes ne sont pas stables d'une solution à l'autre et ne semblent pas associées à des irrégularités au sein de la population (cf. FIG. 11). On peut donc supposer que les classes correspondent à un découpage régulier de la population, le nombre de classes étant contrôlé par PMIN.

Fichier 5

Pour ce fichier (dont les résultats figurent dans le tableau 5 ci-dessous et dans les tableaux A-5-1 et A-5-2 en Annexe A), nous n'obtenons que quatre solutions différentes à 1, 2, 3 et 4 classes.

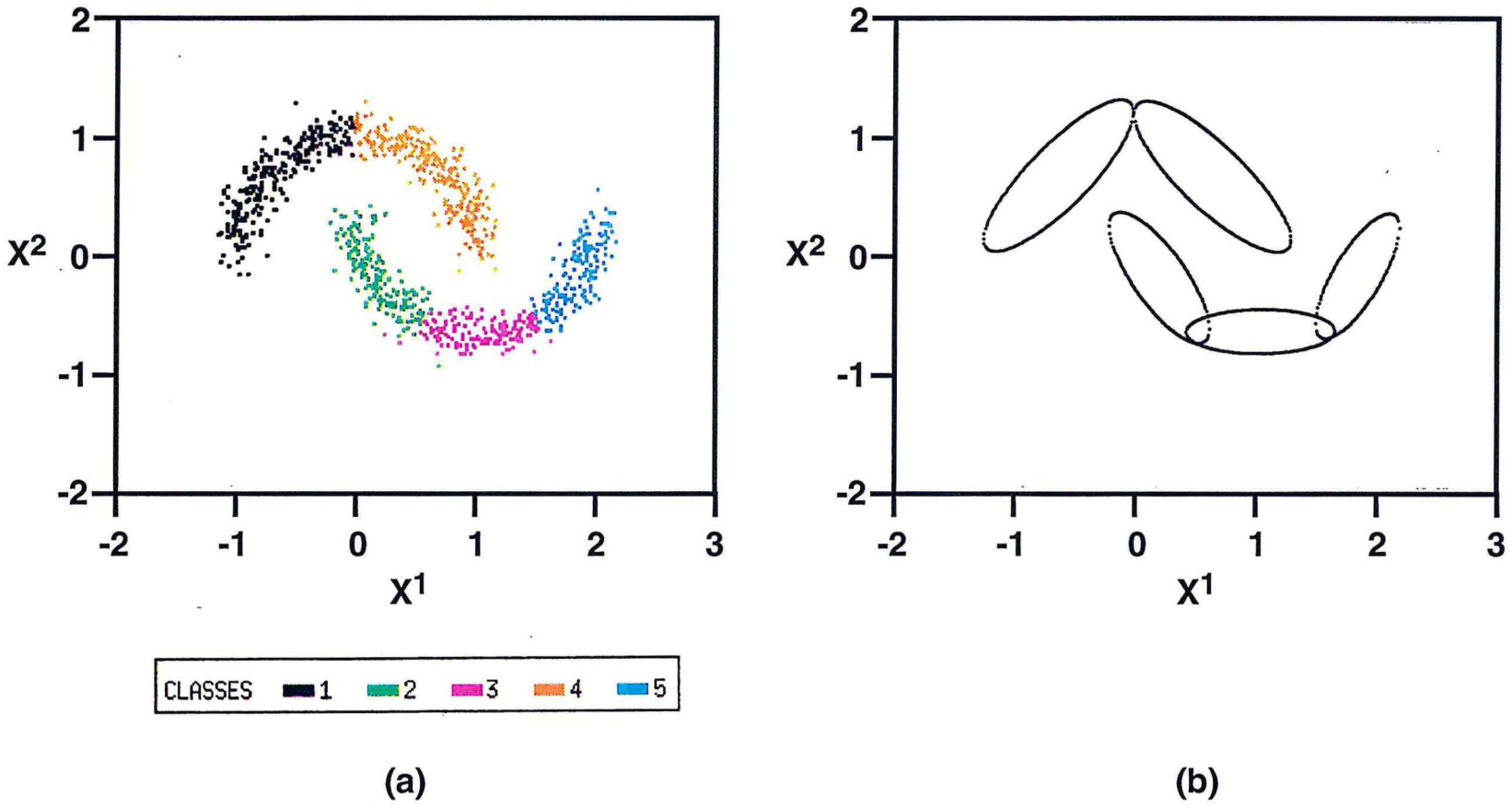


Fig. 17 a. Solution ***sol3*** à cinq classes gaussiennes
b. Courbe de niveau à 95% des classes gaussiennes de ***sol3***

Là encore, nous pouvons constater l'influence de PMIN. En effet, si PMIN est égal à 0.1 ou 0.15, nous obtenons la solution à 2 classes (***sol1***), alors que si PMIN est égal à 0.05, nous obtenons les solutions à 3 et 4 classes (***sol2*** et ***sol3***) ; et la solution à 1 classe (***sol0***) n'est obtenue que pour PMIN égal à 0.5. Par ailleurs, l'influence de l'initialisation choisie apparaît pour PMIN égal à 0.05 : en effet, nous obtenons la solution ***sol2*** (solution à 3 classes) par initialisation multivariable avec 20 classes *a priori* ou en recherchant 10 classes de même effectif sur chaque variable, alors que nous obtenons la solution ***sol3*** (solution à 4 classes) pour tous les autres types d'initialisations.

Nous fournissons ci-joint la représentation graphique de la solution à 4 classes ***sol3*** et celle de sa densité gaussienne (cf. FIG. 18a et 18b).

Tableau 5 *Résultats partiels du fichier 5*

	Nb de classes	Critère C1	Critère C2	PMIN	Initialisations
sol0 *	1	139.15	10.71	0.5	toutes
sol1	2	189.70	19.83	0.15, 0.1	MEff, FR, FRE, FRL, Multi
sol2	3	179.98	16.79	0.05	MEff, Multi
sol3	4	167.98	15.30	0.05	MEff, FR, FRE, FRL, Multi

(* solution retenue)

En ce qui concerne les critères de qualité de l'ajustement C_1 et C_2, ils varient entre 139.15 et 189.7 pour C_1, et entre 10.71 et 19.83 pour C_2. L'ajustement optimal selon ces critères est celui à 1 classe (***sol0***), obtenu uniquement pour PMIN égal à 0.5 ; ceci est surprenant compte tenu de la représentation graphique de la population (cf. FIG. 12), ou de l'estimation de sa densité (cf. FIG. 15) qui n'est pas celle d'une population gaussienne.

En fait, cela peut s'expliquer ; en effet, les critères C_1 et C_2 accordent autant d'importance à un mauvais ajustement dans une zone de faible densité que dans une zone de forte densité. Si on ajuste cette population par une seule classe gaussienne, l'ajustement est mauvais dans les zones de faible densité seulement. Par contre, si on l'ajuste par deux gaussiennes ou plus, on supprime sans doute les problèmes d'ajustement dans les zones de faible densité, mais on obtient certainement de fortes erreurs d'ajustement dans les zones de forte densité (où la densité est très gaussienne). Dans ce cas, la solution appropriée pourrait être d'implémenter le critère C_3 (critère du χ^2) comme autre critère de qualité de l'ajustement : en effet, celui-ci accorde plus d'importance aux erreurs d'ajustement dans les zones de faible densité que dans les zones de forte densité. Ce critère pourrait peut-être permettre de rejeter la solution ***sol0*** à 1 classe, qui n'est pas satisfaisante.

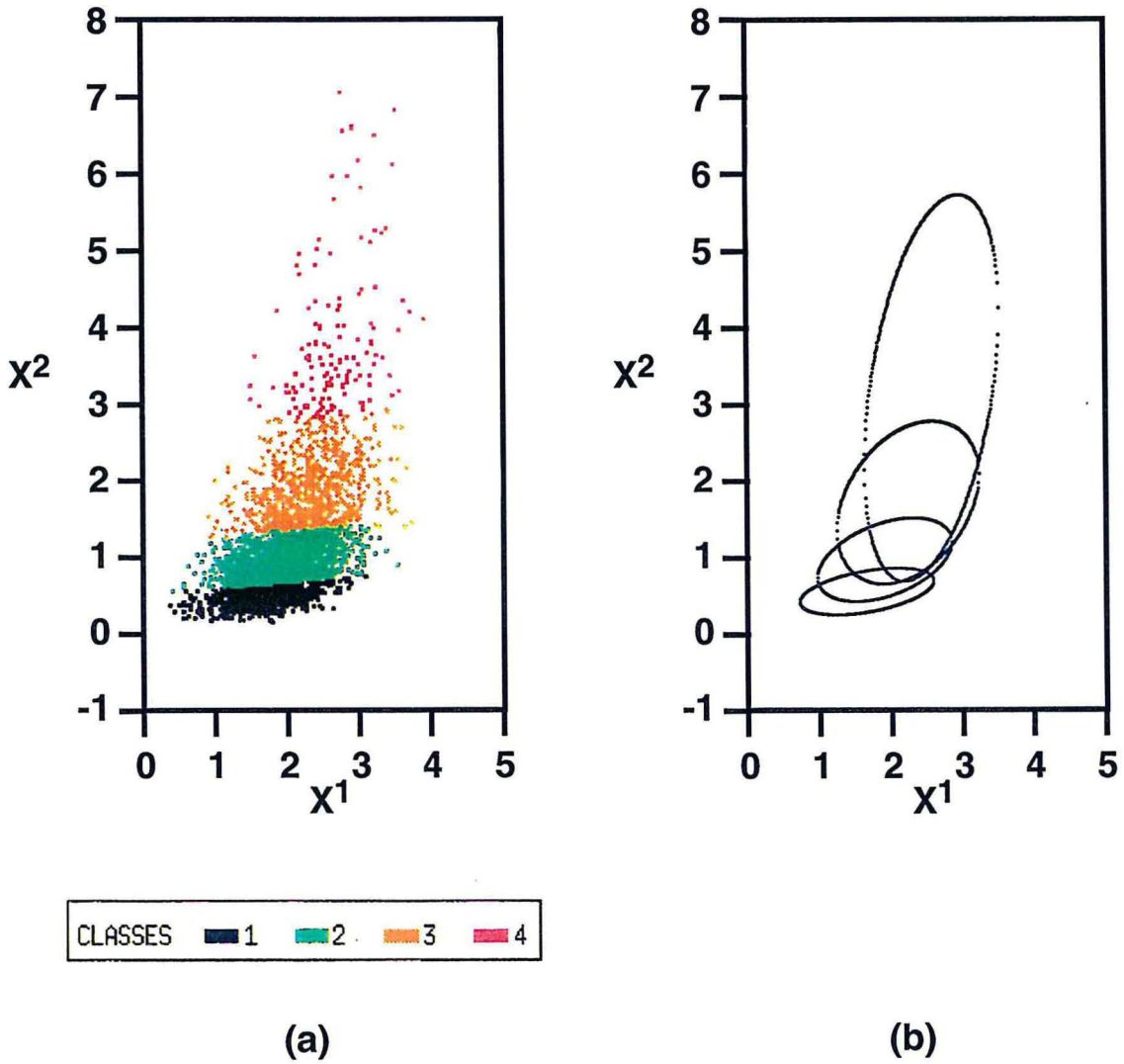


Fig. 18 a. Solution ***sol3*** à quatre classes gaussiennes
b. Courbe de niveau à 95% des classes gaussiennes de ***sol3***

3 Méthodologie de décomposition

L'analyse des résultats des tests effectués permet de dégager des éléments méthodologiques pour la décomposition d'une population. Tout d'abord, nous avons obtenu plusieurs solutions pour chacune des populations synthétiques. Ces solutions dépendent du type d'initialisation choisi, ainsi que de PMIN qui permet de prendre plus ou moins en compte les fluctuations au sein de la population. Ensuite, nous avons constaté que les critères C_1 et C_2 ne sont pas toujours efficaces pour retenir une solution optimale dans l'optique du "calibrage" (cf. fichier 5).

En pratique, est-il quand-même possible de mettre en place une méthodologie permettant d'aboutir à une solution correcte pour toute population ? Supposons que nous ayons plusieurs décompositions pour une population donnée. En utilisant les critères C_1 et C_2, nous pouvons déjà éliminer les solutions de mauvaise qualité (pour lesquelles les valeurs des critères sont les plus élevées) : pour cela, il peut être utile de représenter graphiquement ces critères par valeurs croissantes.

Parmi les solutions restantes, nous supprimons celles ayant un nombre d'individus insuffisant dans une classe pour pouvoir en inférer correctement les paramètres statistiques (moyenne et matrice de variance-covariance). S'il reste plusieurs solutions comparables, nous ne conserverons que celles ayant un nombre de classes compatible avec un calibrage ultérieur : et finalement, nous retiendrons généralement la solution ayant le moins de classes possibles.

Cette méthodologie est celle que nous avons utilisée pour nos tests. Comme nous l'avons constaté, cette méthodologie n'a pas permis de choisir une bonne décomposition pour le fichier 5, puisque nous avons dû retenir la solution à 1 classe. Cependant, nous pouvons faire la remarque suivante : quel que soit le fichier testé, la solution à 1 classe a toujours été obtenue avec PMIN égal à 0.5 (c'est-à-dire en forçant la méthode de décomposition à trouver cette solution), mais elle ne résulte pas d'agrégations successives des classes. Pour ces fichiers, la solution à 1 classe ne correspond donc pas à une décomposition gaussienne "naturelle".

Du coup, que se passe-t-il si nous décidons de ne pas prendre en compte ces solutions dans le choix de la solution optimale ? En appliquant la méthodologie décrite ci-dessus, nous obtenons les mêmes résultats pour les fichiers 2, 3 et 4. Seuls les résultats des fichiers 1 et 5 sont modifiés. Pour le fichier 1, la solution retenue est alors ***sol2*** (solution à 2 classes), dont la qualité d'ajustement est très peu dégradée par rapport à celle de la solution à 1 classe. En ce qui concerne le fichier 5, on retient comme solution optimale ***sol3*** (solution à 4 classes) ; la solution ***sol3*** est de moins bonne qualité que la solution à 1 classe d'après les critères C_1 et C_2, mais elle est plus appropriée pour faire un calibrage.

Ainsi, pour tous les fichiers étudiés, nous sommes parvenus à choisir une décomposition gaussienne de bonne qualité en vue d'un calibrage. Et nous pouvons constater que certaines initialisations permettent d'atteindre à chaque fois la solution optimale. Pour les initialisations monovariables, il semble ainsi préférable d'appliquer la méthode de Harding automatisée sur une fonction de répartition lissée (en utilisant la méthode des noyaux principalement) plutôt que sur la fonction de répartition empirique, ou bien de rechercher des classes de même effectif (avec peu de classes). Il est aussi très intéressant (et très rapide) d'utiliser l'initialisation multivariable en faisant varier le nombre de classes *a priori*. En ce qui concerne PMIN, une valeur de l'ordre de 0.05 semble optimale, mais il faut bien entendu l'adapter à la taille de la population étudiée.

En conclusion, bien qu'il soit toujours nécessaire de relancer plusieurs fois la méthode de décomposition sur une même population, nous sommes à même de trier les solutions et de retenir une décomposition de bonne qualité, avec la démarche méthodologique proposée. Notons que notre objectif n'est pas de rechercher les classes gaussiennes naturelles de la population (si elles existent), mais d'obtenir une décomposition dont les classes sont statistiquement représentatives et telle que la fonction de répartition obtenue approxime au mieux celle de la population.

V - CONCLUSION

Nous avons implémenté une méthode de décomposition d'une population en classes gaussiennes basée sur la méthode du maximum de vraisemblance. Cette méthode itérative nécessite le choix d'une solution initiale. Nous avons donc étudié différentes initialisations possibles et avons notamment retenu celle issue de la méthode de Harding automatisée ainsi qu'une initialisation par classification multivariable.

Lors des tests sur données synthétiques, nous avons constaté que nous obtenions plusieurs décompositions en classes gaussiennes. Nous avons cependant mis en place une méthodologie qui permet le choix d'une décomposition de bonne qualité et optimale en vue d'un calibrage.

Dans le futur, nous utiliserons la décomposition en classes gaussiennes dans le cadre du calibrage géologique d'attributs sismiques. En effet, les techniques statistiques de calibrage (comme l'analyse canonique) ne sont plus utilisables lorsque les relations entre les données étudiées sont non linéaires. Nous espérons résoudre ce problème en appliquant, préalablement à tout calibrage, la méthode de décomposition en classes gaussiennes sur ces données.

Par ailleurs, nous utiliserons aussi cette méthode de décomposition dans le cadre de la régression non paramétrique ; la décomposition nous fournit en effet une approximation de la densité multivariable de la population étudiée.

Dans notre cas, les donnés étudiées sont celles générées aux puits par les propriétés géologiques et les attributs sismiques. Nous serons donc amenés à tester la méthode de décomposition sur des populations comportant un nombre d'individus plus faible pour un nombre de variables plus grand. Il sera intéressant de voir si cela affecte les solutions fournies par la méthode de décomposition, que ce soit en terme de qualité ou de stabilité. Enfin, nous pourrons juger en pratique de l'intérêt de telles décompositions pour l'estimation des propriétés géologiques à partir des données sismiques, en comparant les résultats obtenus à ceux fournis par des méthodologies de calibrage classiques.

CHAPITRE 3

MÉTHODE DE DÉCOMPOSITION EN CLASSES GAUSSIENNES ET ANALYSE CANONIQUE

I - INTRODUCTION

L'analyse canonique est une méthode d'analyse des données développée par Hotelling en 1936. Cette méthode est fondamentale : on peut montrer en effet que des méthodes comme la régression multiple, l'analyse des correspondances ou l'analyse discriminante en sont des cas particuliers (Saporta, 1990, Preisendorfer, 1988).

Le but de l'analyse canonique est de rechercher les liens potentiels existant entre deux groupes de variables ; ainsi, si un ensemble d'individus est décrit par deux groupes de variables différents, l'analyse canonique permet d'examiner les relations linéaires entre ces deux groupes, afin de savoir s'ils mesurent ou non les mêmes propriétés. Par ailleurs, lorsque ces relations sont établies, il est possible de les utiliser à des fins prédictives : connaissant l'expression d'un individu en fonction du premier groupe de variables, on en déduit son expression dans le sous-espace engendré par le second groupe de variables. L'analyse canonique, utilisable pour la description et pour la prédiction, est donc une méthode très intéressante.

Cependant, elle ne permet pas de traiter les relations non linéaires existant entre deux groupes de variables. Pour résoudre ce problème, nous avons décidé de coder les données générées par les deux groupes de variables, avant d'appliquer l'analyse canonique. Le type de codage retenu correspond à une décomposition des données en sous-populations. Afin que chacune de ces sous-populations présente une relation linéaire entre variables, il nous a semblé tout à fait approprié d'utiliser la méthode de décomposition en classes gaussiennes présentée au chapitre précédent.

Dans ce chapitre, nous présenterons tout d'abord la théorie de l'analyse canonique. Puis nous parlerons du type de codage de données retenu, et des conséquences théoriques de ce codage sur l'analyse canonique.

II - ANALYSE CANONIQUE

1 Notations

Soit X le tableau à n lignes et p colonnes $(n \times p)$ des valeurs prises par n individus pour un premier groupe de p variables $X^1 \ldots X^p$.

Soit Y le tableau à n lignes et q colonnes $(n \times q)$ des valeurs prises par n individus pour un second groupe de q variables $Y^1 \ldots Y^q$.

Nous supposerons par la suite que les variables $X^1 \ldots X^p$ et $Y^1 \ldots Y^q$ sont centrées.

Soit $\langle \cdot, \cdot \rangle_D$ et $\| \cdot \|_D$ le produit scalaire et la norme associés à l'espace vectoriel R^n. La métrique D associée à cet espace est la métrique usuelle $\frac{1}{n} I_n$, où I_n est la matrice identité dans R^n.

Alors notons :

$$V_{XX} = \langle X, X \rangle_D = \frac{1}{n} {}^t XX \quad , \text{ matrice } p \times p$$

$$V_{YY} = \langle Y, Y \rangle_D = \frac{1}{n} {}^t YY \quad , \text{ matrice } q \times q$$

$$V_{XY} = \langle X, Y \rangle_D = \frac{1}{n} {}^t XY \quad , \text{ matrice } p \times q$$

$$V_{YX} = {}^t V_{XY}$$

V_{XX} est la matrice de variance-covariance de X. V_{YY} est la matrice de variance-covariance de Y. V_{XY} est la matrice de variance-covariance de X avec Y.

2 Position du problème de l'analyse canonique

Soit W_X et W_Y les deux sous-espaces engendrés respectivement par $X^1 \ldots X^p$ et $Y^1 \ldots Y^q$ dans R^n. Appliquer l'analyse canonique aux tableaux X et Y revient à étudier la position géométrique de W_X et W_Y l'un par rapport à l'autre et à en rechercher les "éléments" les plus proches.

On recherche donc deux nouveaux repères spécifiques $(\varphi^1 \ldots \varphi^r)$ et $(\psi^1 \ldots \psi^r)$, qui sont respectivement des combinaisons linéaires dans W_X de $X^1 \ldots X^p$ et dans W_Y de $Y^1 \ldots Y^q$, r étant la dimension du plus petit des deux espaces W_X et W_Y. Ce problème se résout pas à pas de la façon suivante.

A l'étape 1, on recherche le couple de vecteurs normés (φ^1, ψ^1) de corrélation maximale, φ^1 étant dans W_X et ψ^1 dans W_Y. Ce couple est donc solution du problème suivant :

$$\begin{cases} \text{Max} \quad \rho_1 \\ \|\varphi^1\|_D = \|\psi^1\|_D = 1 \end{cases}$$

où :

$$\begin{aligned} \rho_1 &= \text{Corr}(\varphi^1, \psi^1) \\ &= \frac{\langle \varphi^1, \psi^1 \rangle_D}{\|\varphi^1\|_D \cdot \|\psi^1\|_D} \\ &= \langle \varphi^1, \psi^1 \rangle_D \qquad \text{dans ce cas} \end{aligned}$$

A l'étape 2, connaissant le couple (φ^1, ψ^1), on recherche le couple de vecteurs normés (φ^2, ψ^2) de corrélation maximale, φ^2 étant dans W_X et orthogonal à φ^1, ψ^2 étant dans W_Y et orthogonal à ψ^1. Ce couple est donc solution de :

$$\begin{cases} \text{Max} \quad \rho_2 \\ \|\varphi^2\|_D = \|\psi^2\|_D = 1 \\ \langle \varphi^1, \varphi^2 \rangle_D = \langle \psi^1, \psi^2 \rangle_D = 0 \end{cases}$$

où :

$$\begin{aligned} \rho_2 &= \text{Corr}(\varphi^2, \psi^2) \\ &= \langle \varphi^2, \psi^2 \rangle_D \qquad \text{dans ce cas} \end{aligned}$$

A l'étape 3, connaissant les couples (φ^1, ψ^1) et (φ^2, ψ^2), on recherche le couple de vecteurs normés (φ^3, ψ^3) de corrélation maximale, φ^3 étant dans W_X et orthogonal à φ^1et φ^2, ψ^3 étant dans W_Y et orthogonal à ψ^1 et ψ^2, et ainsi de suite.

De cette façon, on obtient successivement les r couples (φ^1,ψ^1) ... (φ^r,ψ^r), tels que $(\varphi^1 \ldots \varphi^r)$ et $(\psi^1 \ldots \psi^r)$, qui forment respectivement dans W_X et W_Y deux nouveaux repères orthonormés, sont solutions du problème de l'analyse canonique.

Remarquons que, **à une étape i quelconque** ($i \leq r$), φ^i et ψ^i étant respectivement des combinaisons linéaires de X et Y, il existe un vecteur a^i de R^p et un vecteur b^i de R^q, tels que :

$$\begin{cases} \varphi^i = Xa^i \\ \psi^i = Yb^i \end{cases}$$

Il est donc équivalent de rechercher le couple de vecteurs (φ^i,ψ^i), solution du problème :

$$(1) \quad \begin{cases} \text{Max} \quad \rho_i = \langle \varphi^i, \psi^i \rangle_D \\ \|\varphi^i\|_D = \|\psi^i\|_D = 1 \\ \langle \varphi^j, \varphi^i \rangle_D = \langle \psi^j, \psi^i \rangle_D = 0 \quad , \quad \forall\, j \prec i \end{cases}$$

ou le couple de vecteurs (a^i,b^i) solution du problème :

$$(2) \quad \begin{cases} \text{Max} \quad \rho_i = \langle Xa^i, Yb^i \rangle_D \\ \|Xa^i\|_D = \|Yb^i\|_D = 1 \\ \langle Xa^j, Xa^i \rangle_D = \langle Yb^j, Yb^i \rangle_D = 0 \quad , \quad \forall\, j \prec i \end{cases}$$

En fait, la recherche successive des couples (φ^i,ψ^i) ou des couples (a^i,b^i), pour i allant de 1 à r, correspond à la résolution du problème de l'analyse canonique soit dans l'espace des individus soit dans l'espace des variables.

φ^i et ψ^i, pour i allant de 1 à r, sont appelées **variables canoniques** associées respectivement à X et Y.

a^i et b^i, pour i allant de 1 à r, sont appelés **facteurs canoniques** associés respectivement à X et Y.

Et ρ_i, pour i allant de 1 à r, est le **coefficient de corrélation canonique**. Par définition, ρ_i est toujours compris entre -1 et 1 ; mais on le choisira toujours compris entre 0 et 1, en inversant le sens d'un des vecteurs φ^i ou ψ^i si cela est nécessaire.

Par la suite, nous supposerons que les matrices X et Y sont de plein rang, avec $p \leq q$. Nous rechercherons donc p couples canoniques (φ^i,ψ^i) par analyse canonique de X et Y.

3 Résolution géométrique du problème de l'analyse canonique

Il est possible de poser le problème de l'analyse canonique en termes géométriques. Ainsi, à une étape i quelconque, ρ_i, qui s'exprime comme le coefficient de corrélation entre les variables canoniques φ^i et ψ^i, s'exprime aussi comme le cosinus de l'angle formé par ces mêmes variables. On pourra se reporter à la figure 19 pour avoir des exemples d'analyse canonique dans R^3.

Dans ce cadre, recherchons les deux nouveaux repères $(\varphi^1 \ldots \varphi^r)$ de W_X et $(\psi^1 \ldots \psi^r)$ de W_Y, issus de l'analyse canonique de X et Y.

A l'étape 1, le problème associé à la recherche du couple de vecteurs (φ^1, ψ^1), écrit sous forme géométrique, est le suivant :

$$\begin{cases} \text{Max} \quad \rho_1 = \text{Cos}(\varphi^1, \psi^1) \\ \|\varphi^1\|_D = \|\psi^1\|_D = 1 \end{cases}$$

Pour que ρ_1, qui est le cosinus de l'angle formé par φ^1 et ψ^1, soit maximal, il faut que :

- à φ^1 fixé (φ^1 étant un vecteur de W_X), ψ^1 (vecteur de W_Y) soit colinéaire à la projection orthogonale de φ^1 sur W_Y ;
- à ψ^1 fixé, φ^1 soit colinéaire à la projection orthogonale de ψ^1 sur W_X.

Notons Π_X et Π_Y les opérateurs de projection orthogonale respectivement sur W_X et W_Y. On peut établir que :

$$\Pi_X = X \cdot \left({}^tXDX\right)^{-1} \cdot {}^tXD = \frac{1}{n} X \cdot V_{XX}^{-1} \cdot {}^tX$$

$$\Pi_Y = Y \cdot \left({}^tYDY\right)^{-1} \cdot {}^tYD = \frac{1}{n} Y \cdot V_{YY}^{-1} \cdot {}^tY$$

En effet, X et Y étant de plein rang, les matrices V_{XX} et V_{YY} sont inversibles.

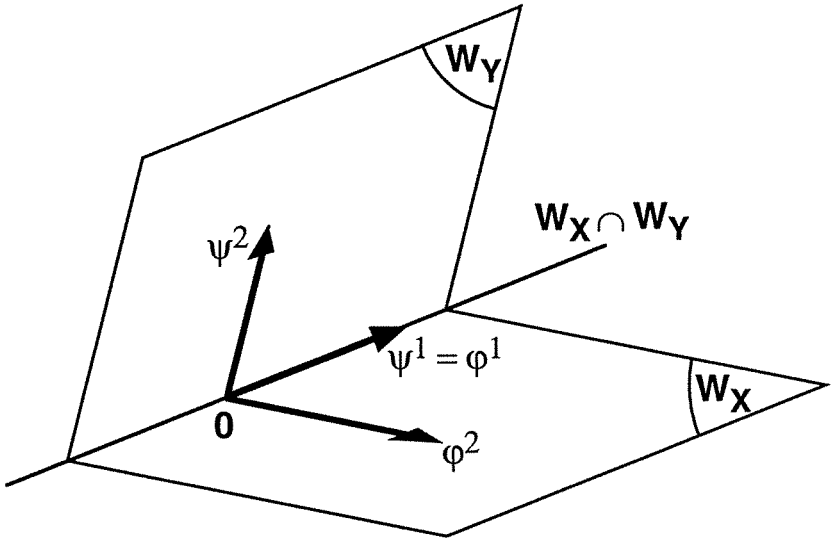
Pour que ρ_1 soit maximal, il faut donc trouver les scalaires α et β, tels que :

$$\begin{cases} \varphi^1 = \alpha \cdot \Pi_X \psi^1 \\ \psi^1 = \beta \cdot \Pi_Y \varphi^1 \end{cases}$$

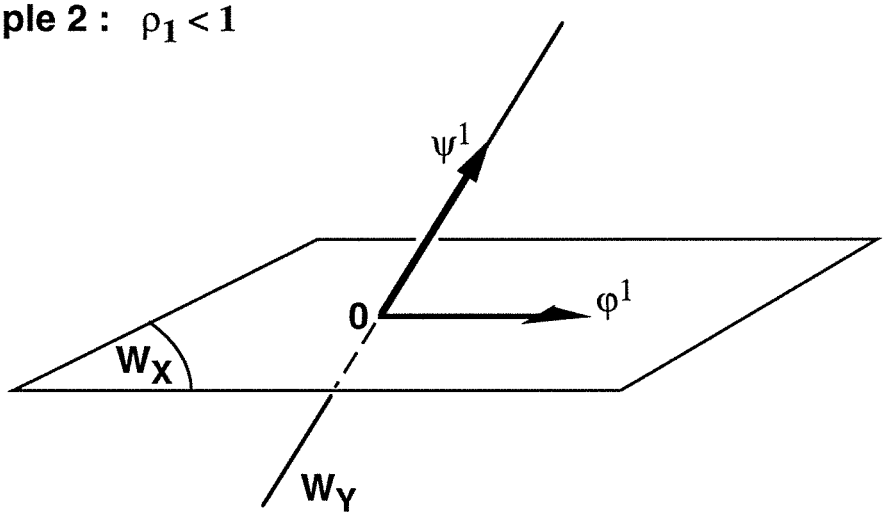
ce qui est équivalent à :

$$(3) \quad \begin{cases} \frac{1}{\alpha} \varphi^1 = \Pi_X \psi^1 \\ \frac{1}{\beta} \psi^1 = \Pi_Y \varphi^1 \end{cases}$$

Exemple 1 : $\rho_1 = 1,\ \rho_2 < 1$



Exemple 2 : $\rho_1 < 1$



Exemple 3 : $\rho_1 = 0$

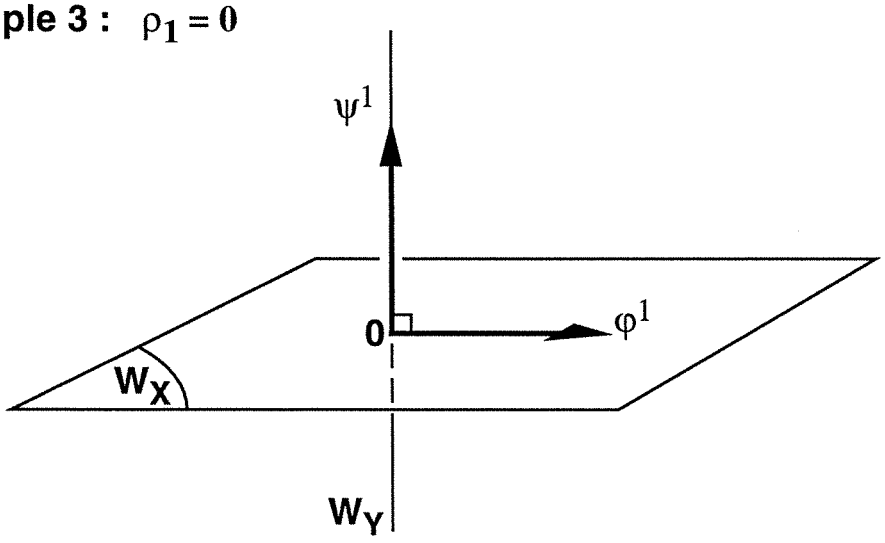


Fig. 19 Exemples d'analyse canonique dans R^3

Par définition,

$$\rho_1 = \mathrm{Cos}\left(\varphi^1, \psi^1\right)$$

$$= \frac{\left\| \Pi_X \psi^1 \right\|_D}{\left\| \varphi^1 \right\|_D}$$

$$= \frac{\left\| \frac{1}{\alpha} \varphi^1 \right\|_D}{\left\| \varphi^1 \right\|_D} \qquad \text{d'après (3)}$$

$$= \frac{1}{\alpha}$$

Et de même,

$$\rho_1 = \mathrm{Cos}\left(\psi^1, \varphi^1\right)$$

$$= \frac{\left\| \Pi_Y \varphi^1 \right\|_D}{\left\| \psi^1 \right\|_D}$$

$$= \frac{1}{\beta}$$

On a donc :

$$\rho_1 = \frac{1}{\alpha} = \frac{1}{\beta}$$

En utilisant ce résultat dans le système (3), on obtient le système suivant :

$$\begin{cases} \rho_1 \varphi^1 = \Pi_X \psi^1 \\ \rho_1 \psi^1 = \Pi_Y \varphi^1 \end{cases}$$

avec ρ_1 maximal (cf. FIG. 20).

En appliquant le projecteur Π_Y à la première équation de ce système et en utilisant la seconde, puis en appliquant le projecteur Π_X à la seconde équation de ce système et en utilisant la première, il vient :

$$(4) \quad \begin{cases} \rho_1^2 \psi^1 = \Pi_Y \Pi_X \psi^1 \\ \rho_1^2 \varphi^1 = \Pi_X \Pi_Y \varphi^1 \end{cases}$$

d'où, en utilisant l'expression des opérateurs de projection Π_X et Π_Y :

$$\begin{cases} \rho_1^2 \psi^1 = \frac{1}{n} Y \cdot V_{YY}^{-1} \cdot V_{YX} \cdot V_{XX}^{-1} \cdot {}^t X \cdot \psi^1 \\ \rho_1^2 \varphi^1 = \frac{1}{n} X \cdot V_{XX}^{-1} \cdot V_{XY} \cdot V_{YY}^{-1} \cdot {}^t Y \cdot \varphi^1 \end{cases}$$

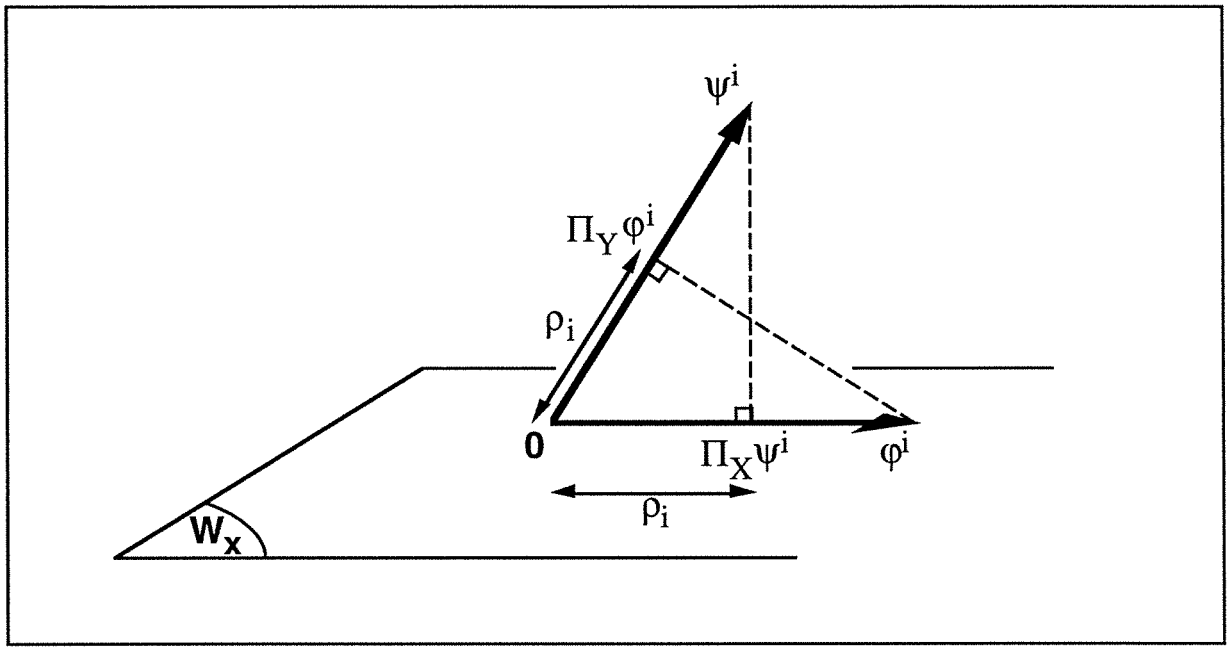


Fig. 20 Propriétés géométriques des variables canoniques

Les équations du système (4) sont appelées **équations canoniques associées aux variables canoniques** φ^1 et ψ^1.

Résoudre l'étape 1 de l'analyse canonique revient à résoudre le système (4) en maximisant ρ_1, c'est-à-dire à trouver les variables canoniques φ^1 et ψ^1 qui sont respectivement vecteurs propres des matrices $(\Pi_X\Pi_Y)$ et $(\Pi_Y\Pi_X)$ associés à la plus grande et même valeur propre ρ_1^2.

Connaissant le couple (φ^1,ψ^1), on passe **à l'étape 2** de l'analyse canonique, qui consiste à rechercher le couple de vecteurs (φ^2,ψ^2). Par un raisonnement géométrique similaire à celui de l'étape 1, on peut montrer que le couple (φ^2,ψ^2) est solution de :

$$\begin{cases}\rho_2^2\,\psi^2 = \Pi_Y\Pi_X\,\psi^2\\ \rho_2^2\,\varphi^2 = \Pi_X\Pi_Y\,\varphi^2\end{cases}$$

avec ρ_2 maximal. Les variables canoniques φ^2 et ψ^2 sont donc respectivement les vecteurs propres des matrices $(\Pi_X\Pi_Y)$ et $(\Pi_Y\Pi_X)$ associés à la seconde plus grande et même valeur propre ρ_2^2.

De façon générale, on peut résoudre de façon géométrique **l'étape i** de l'analyse canonique. Cette étape consiste à rechercher le couple de vecteurs $\left(\varphi^i, \psi^i\right)$, pour i allant de 1 à p, qui est solution du problème suivant :

$$\begin{cases} \text{Max} \quad \rho_i = \text{Cos}\left(\varphi^i, \psi^i\right) \\ \left\|\varphi^i\right\|_D = \left\|\psi^i\right\|_D = 1 \\ \left\langle \varphi^j, \varphi^i \right\rangle_D = \left\langle \psi^j, \psi^i \right\rangle_D = 0 \quad , \quad \forall\, j \prec i \end{cases}$$

On aboutit au système d'équations :

$$\text{(4bis)} \qquad \begin{cases} \rho_i^2\, \psi^i = \Pi_Y \Pi_X\, \psi^i \\ \rho_i^2\, \varphi^i = \Pi_X \Pi_Y\, \varphi^i \end{cases}$$

qui sont les équations canoniques associées aux variables canoniques φ^i et ψ^i. Et les variables canoniques φ^i et ψ^i sont les vecteurs propres des matrices $\left(\Pi_X \Pi_Y\right)$ et $\left(\Pi_Y \Pi_X\right)$ associés à la ième plus grande et même valeur propre ρ_i^2.

Comme on peut montrer que les matrices $\left(\Pi_X \Pi_Y\right)$ et $\left(\Pi_Y \Pi_X\right)$ ont p valeurs propres positives ou nulles, les équations canoniques ont toujours une solution.

Récapitulons toutes les relations associées aux variables canoniques φ^i et ψ^i ainsi qu'au coefficient de corrélation canonique ρ_i.

(5) • $\rho_i = \left\langle \varphi^i, \psi^i \right\rangle_D$

(6) • $\left\|\varphi^i\right\|_D = \left\|\psi^i\right\|_D = 1$

(7) • $\left\langle \varphi^i, \varphi^j \right\rangle_D = \left\langle \psi^i, \psi^j \right\rangle_D = 0 \quad , \quad \text{si } i \neq j$

(8) • $\rho_i\, \varphi^i = \Pi_X\, \psi^i$

$\rho_i\, \psi^i = \Pi_Y\, \varphi^i$

(9) • $\rho_i^2 \varphi^i = \Pi_X \Pi_Y \varphi^i$

$\rho_i^2 \psi^i = \Pi_Y \Pi_X \psi^i$

De plus, on peut en déduire que :

(10) $\left\langle \varphi^i, \psi^j \right\rangle_D = 0 \quad , \quad \text{si } i \neq j$

En effet,

$$\begin{aligned}
\left\langle \varphi^i, \psi^j \right\rangle_D &= \frac{1}{\rho_j} \cdot \left\langle \varphi^i, \rho_j \psi^j \right\rangle_D \\
&= \frac{1}{\rho_j} \cdot \left\langle \varphi^i, \Pi_Y \varphi^j \right\rangle_D && \text{d'après (8)} \\
&= \frac{1}{\rho_j} \cdot \left\langle \Pi_X \varphi^i, \Pi_Y \varphi^j \right\rangle_D && \text{car } \varphi^i \in W_X \\
&= \frac{1}{\rho_j} \cdot \left\langle \varphi^i, {}^t\Pi_X \, \Pi_Y \varphi^j \right\rangle_D \\
&= \frac{1}{\rho_j} \cdot \left\langle \varphi^i, \Pi_X \, \Pi_Y \varphi^j \right\rangle_D && \text{car } {}^t\Pi_X = \Pi_X \\
&= \frac{1}{\rho_j} \cdot \left\langle \varphi^i, \rho_j^2 \varphi^j \right\rangle_D && \text{d'après (9)} \\
&= \rho_j \cdot \left\langle \varphi^i, \varphi^j \right\rangle_D \\
&= 0 \quad , \quad \text{si } i \neq j && \text{d'après (7)}
\end{aligned}$$

4 Résolution algébrique du problème de l'analyse canonique

Il est possible de résoudre les p étapes du problème de l'analyse canonique de façon algébrique, en se plaçant dans l'espace des variables. Nous recherchons donc cette fois-ci les vecteurs $a^1 \ldots a^p$ de R^p et les vecteurs $b^1 \ldots b^p$ de R^q, appelés facteurs canoniques.

A l'étape 1, le couple $\left(a^1, b^1\right)$ est solution du problème :

$$\begin{cases} \text{Max} \quad \rho_1 = \left\langle Xa^1, Yb^1 \right\rangle_D \\ \left\| Xa^1 \right\|_D = \left\| Yb^1 \right\|_D = 1 \end{cases}$$

Ce problème d'optimisation peut se résoudre par la méthode du Lagrangien, qui s'écrit :

$$L(\lambda,\mu) = \left\langle Xa^1, Yb^1 \right\rangle_D - \lambda \cdot \left(\left\langle Xa^1, Xa^1 \right\rangle_D - 1 \right) - \mu \cdot \left(\left\langle Yb^1, Yb^1 \right\rangle_D - 1 \right)$$

On aboutit au système (11) suivant :

$$(11) \quad \begin{cases} V_{XY} \, b^1 - 2\lambda \cdot V_{XX} \, a^1 = 0 \\ V_{YX} \, a^1 - 2\mu \cdot V_{YY} \, b^1 = 0 \end{cases}$$

En multipliant à gauche la première équation de (11) par ${}^t a^1$ et la seconde équation par ${}^t b^1$, on obtient le système :

$$\begin{cases} {}^t a^1 V_{XY} b^1 = 2\lambda \cdot {}^t a^1 V_{XX} a^1 \\ {}^t b^1 V_{YX} a^1 = 2\mu \cdot {}^t b^1 V_{YY} b^1 \end{cases}$$

Sachant que

$$\begin{aligned} \left\| Xa^1 \right\|_D &= 1 \\ &= \left\langle Xa^1, Xa^1 \right\rangle_D \\ &= {}^t a^1 V_{XX} a^1 \end{aligned}$$

et que

$$\begin{aligned} \left\| Yb^1 \right\|_D &= 1 \\ &= \left\langle Yb^1, Yb^1 \right\rangle_D \\ &= {}^t b^1 V_{YY} b^1 \end{aligned}$$

ce système est donc équivalent à :

$$(12) \quad \begin{cases} {}^t a^1 V_{XY} b^1 = 2\lambda \\ {}^t b^1 V_{YX} a^1 = 2\mu \end{cases}$$

Comme ${}^t a^1 V_{XY} b^1$ est un scalaire, il est égal à son transposé, c'est-à-dire égal à ${}^t b^1 V_{YX} a^1$. On déduit donc de (12) que :

$$\lambda = \mu$$

De plus, d'après nos notations,

$$\begin{aligned} \rho_1 &= \left\langle Xa^1, Yb^1 \right\rangle_D \\ &= {}^t a^1 V_{XY} b^1 \end{aligned}$$

Donc,

$$\lambda = \mu = \frac{1}{2}\rho_1$$

En appliquant ce résultat au système (11), il vient :

$$\begin{cases} V_{XY} b^1 - \rho_1 \cdot V_{XX} a^1 = 0 \\ V_{YX} a^1 - \rho_1 \cdot V_{YY} b^1 = 0 \end{cases}$$

ou bien

$$\begin{cases} \rho_1 a^1 = V_{XX}^{-1} V_{XY} b^1 \\ \rho_1 b^1 = V_{YY}^{-1} V_{YX} a^1 \end{cases}$$

En utilisant la première équation de ce système dans la seconde et réciproquement, on obtient le système suivant :

$$(13)\quad \begin{cases} \rho_1^2 a^1 = V_{XX}^{-1} V_{XY} V_{YY}^{-1} V_{YX} a^1 \\ \rho_1^2 b^1 = V_{YY}^{-1} V_{YX} V_{XX}^{-1} V_{XY} b^1 \end{cases}$$

Les équations de ce système sont appelées **équations canoniques associées aux facteurs canoniques** a^1 et b^1.

Résoudre l'étape 1 de l'analyse canonique revient à résoudre le système (13) en maximisant ρ_1, c'est-à-dire à trouver les facteurs canoniques a^1 et b^1 qui sont respectivement vecteurs propres des matrices $\left(V_{XX}^{-1} V_{XY} V_{YY}^{-1} V_{YX}\right)$ et $\left(V_{YY}^{-1} V_{YX} V_{XX}^{-1} V_{XY}\right)$ associés à la plus grande et même valeur propre ρ_1^2.

Et de façon générale, **à l'étape i**, connaissant les couples $\left(a^1, b^1\right) \ldots \left(a^{i-1}, b^{i-1}\right)$, on peut rechercher de façon similaire les facteurs canoniques a^i et b^i. Comme précédemment, on aboutit à la résolution du système :

$$(13\text{bis})\quad \begin{cases} \rho_i^2 a^i = V_{XX}^{-1} V_{XY} V_{YY}^{-1} V_{YX} a^i \\ \rho_i^2 b^i = V_{YY}^{-1} V_{YX} V_{XX}^{-1} V_{XY} b^i \end{cases}$$

a^i et b^i sont donc les vecteurs propres respectifs des matrices $\left(V_{XX}^{-1} V_{XY} V_{YY}^{-1} V_{YX}\right)$ et $\left(V_{YY}^{-1} V_{YX} V_{XX}^{-1} V_{XY}\right)$ associés à la ième plus grande et même valeur propre ρ_i^2. Comme on peut montrer que ces deux matrices ont p valeurs propres positives ou nulles, il est donc toujours possible de calculer de proche en proche les couples $\left(a^1, b^1\right) \ldots \left(a^p, b^p\right)$.

Remarquons qu'il est identique, à une étape i donnée, de résoudre le système (4bis) obtenu par raisonnement géométrique et le système (13bis) obtenu algébriquement. Rappelons les équations du système (4bis) :

$$\begin{cases} \rho_i^2 \psi^i = \Pi_Y \Pi_X \psi^i \\ \rho_i^2 \varphi^i = \Pi_X \Pi_Y \varphi^i \end{cases}$$

En remplaçant les projecteurs Π_X et Π_Y par leur expression, et sachant que

$$\begin{cases} \varphi^i = X a^i \\ \psi^i = Y b^i \end{cases}$$

on a :

$$\begin{cases} \rho_i^2 \, Yb^i = Y V_{YY}^{-1} V_{YX} V_{XX}^{-1} \underbrace{{}^tXDY} b^i \\ \rho_i^2 \, Xa^i = X V_{XX}^{-1} V_{XY} V_{YY}^{-1} \underbrace{{}^tYDX} a^i \end{cases}$$

En multipliant à gauche la première équation par $\left[\left({}^tYDY\right)^{-1} \cdot {}^tYD\right]$ et la seconde équation par $\left[\left({}^tXDX\right)^{-1} \cdot {}^tXD\right]$, on obtient le système (13bis) :

$$\begin{cases} \rho_i^2 \, b^i = V_{YY}^{-1} V_{YX} V_{XX}^{-1} V_{XY} \, b^i \\ \rho_i^2 \, a^i = V_{XX}^{-1} V_{XY} V_{YY}^{-1} V_{YX} \, a^i \end{cases}$$

En fait, la résolution du problème de l'analyse canonique par le système (4bis) se fait dans l'espace des individus R^n (avec la recherche des variables canoniques). Par contre, avec le système (13bis), cette résolution se fait dans l'espace des variables R^{p+q} (avec la recherche des facteurs canoniques).

5 Formule de reconstitution des données

Comme nous l'avions indiqué au début de ce chapitre, l'analyse canonique peut être utilisée dans un but descriptif ou dans un but prédictif. Dans les paragraphes précédents, nous avons abordé l'**aspect descriptif** : en effet, rechercher les variables ou les facteurs canoniques permet de décrire la position respective des sous-espaces W_X et W_Y engendrés par les deux groupes de variables $X^1 \ldots X^p$ et $Y^1 \ldots Y^q$.

Dans ce paragraphe, nous nous attacherons à l'**aspect prédictif** de l'analyse canonique (Preisendorfer, 1988). Dans ce cadre, nous souhaitons prédire la position dans le sous-espace W_Y de tout individu à partir de sa position (connue) dans W_X. Pour cela, il nous faut établir une formule (dite formule de reconstitution des données) fournissant l'expression (ou "prédiction") de Y en fonction de X, en supposant connus les résultats d'une analyse canonique entre les deux groupes de variables $X^1 \ldots X^p$ et $Y^1 \ldots Y^q$.

Pour simplifier les écritures, nous travaillerons sous forme matricielle. Notons :

$\varphi = \left[\varphi^1 \ldots \varphi^p\right]$ matrice $n \times p$

$\psi = \left[\psi^1 \ldots \psi^p\right]$ matrice $n \times p$

$A = \left[a^1 \ldots a^p\right]$ matrice $p \times p$

$$B = \left[b^1 \ldots b^p \right] \qquad \text{matrice } q \times p$$

$$\Theta = \begin{bmatrix} \rho_1 & & 0 \\ & \ddots & \\ 0 & & \rho_p \end{bmatrix} \qquad \text{matrice } p \times p$$

$$I_r = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \qquad \text{matrice identité dans } R^r$$

Les équations du problème de l'analyse canonique sous forme matricielle sont :

(14) $\varphi = XA$

(15) $\psi = YB$

(16) ${}^t\psi D \varphi = \Theta$

(17) ${}^t\varphi D \varphi = I_p$ et ${}^t\psi D \psi = I_p$

et l'équation suivante doit, entre autres, être vérifiée :

(18) $V_{YY}^{-1} V_{YX} A = B\Theta$

Cherchons à l'aide de ces équations à exprimer Y en fonction de X.

• Essayons tout d'abord d'exprimer Y en fonction de ψ sous la forme :

(19) $Y = \psi C$

où C est une matrice de taille $p \times q$ à définir. Nous sommes assurés qu'une telle relation linéaire existe entre Y et ψ : en effet, d'après la relation (15), Y est de la forme ψB^{-}.

Cherchons à expliciter la matrice C. En multipliant l'équation (19) à gauche par $\left({}^t\psi D \right)$, il vient :

$${}^t\psi D Y = \underbrace{{}^t\psi D \psi} C$$

D'après (17), ${}^t\psi D \psi = I_p$, donc :

(20) ${}^t\psi D Y = C$

Par ailleurs, en multipliant chaque terme de l'équation (15) à gauche par $\left({}^tY D \right)$, on obtient :

$$\begin{aligned} {}^tY D \psi &= {}^tY D Y B \\ &= V_{YY} B \end{aligned}$$

En transposant cette équation, et puisque ${}^tD = D$,

$$\begin{aligned} {}^t\psi D Y &= {}^tB V_{YY} \\ &= C \qquad \text{d'après (20)} \end{aligned}$$

L'équation (19) s'écrit donc :

$$Y = \Psi\, {}^{t}B V_{YY}$$

• Exprimons maintenant Ψ en fonction de φ. En transposant les deux membres de l'équation (16), il vient :

$${}^{t}\varphi D \Psi = {}^{t}\Theta$$

d'où :

$$(21) \quad \Psi = D^{-1}\left({}^{t}\varphi\right)^{-} {}^{t}\Theta$$

$$= n\left({}^{t}\varphi\right)^{-} {}^{t}\Theta$$

• Exprimons à présent Y en fonction de X. De (19) et (21), il vient :

$$Y = n\left({}^{t}\varphi\right)^{-} {}^{t}\Theta\, {}^{t}B V_{YY}$$

$$= n\left({}^{t}\varphi\right)^{-} {}^{t}(B\Theta) V_{YY}$$

En utilisant (18),

$$Y = n\left({}^{t}\varphi\right)^{-} {}^{t}A V_{XY}$$

D'après (14), on obtient donc comme expression de Y en fonction de X :

$$\boxed{Y = n\left({}^{t}A\,{}^{t}X\right)^{-} {}^{t}A V_{XY}}$$

qui est la formule de reconstitution des données dans sa forme générale.

En fait, d'après (17), une des inverses généralisées de ${}^{t}\varphi$ étant :

$$\left({}^{t}\varphi\right)^{-} = D\varphi$$

une des relations reliant Y et X s'écrit :

$$Y = X A\, {}^{t}A V_{XY}$$

Théoriquement, cette formule pourrait donc permettre de prédire la position d'un individu quelconque dans W_Y, connaissant celle qu'il a dans W_X.

Mais du fait de la non unicité de cette formule de prédiction, elle ne peut être utilisée en pratique. L'analyse canonique ne peut donc pas être mise en pratique dans un but prédictif.

6 Problème du centrage des données

6.1 Analyse canonique sur données non centrées

Jusqu'à présent, tous les développements ont été effectués en supposant les données centrées. Or dans certains cas, centrer les données peut faire perdre une de leurs spécificités (par exemple, données de type entier exclusivement...), et se révéler inapproprié. Nous allons donc considérer l'analyse canonique sans centrage préalable des données.

Les matrices V_{XX}, V_{YY}, V_{XY} et V_{YX} sont des matrices de variance-covariance si les tableaux X et Y sont centrés. Ce n'est plus le cas si on ne centre pas ces tableaux.

Si les variables $X^1 \ldots X^p$ et $Y^1 \ldots Y^q$ sont non centrées, φ^i et ψ^i qui en sont des combinaisons linéaires ne le sont généralement pas. De ce fait, ρ_i associé au couple canonique (φ^i, ψ^i) ne peut plus s'interpréter comme un coefficient de corrélation : on ne parlera donc plus de corrélation canonique. En outre, ρ_i est compris entre 0 et 1 lorsque les tableaux X et Y sont centrés, ce qui n'est plus vrai si ces tableaux sont non centrés.

Mis à part ces points, les résultats de l'analyse canonique (et développements mathématiques associés) restent valables, mais leur interprétation est plus ardue.

6.2 Cas particulier : $1_n \in (W_X \cap W_Y)$

Notons 1_n le vecteur $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ de R_n. Nous allons démontrer que si $1_n \in (W_X \cap W_Y)$, il est équivalent de travailler sur les données centrées ou non centrées (Mallet, 1988). Dans ce cas, le problème du centrage des données ne se pose donc plus.

Notons X et Y les tableaux non centrés de données initiales, X^c et Y^c les tableaux centrés :

$$X^c = X - 1_n\, {}^t m_X \qquad \text{matrice } n \times p$$

$$Y^c = Y - 1_n\, {}^t m_Y \qquad \text{matrice } n \times q$$

où :

$$(22) \quad \begin{cases} m_X = \dfrac{1}{n}\, {}^t X 1_n = \langle X, 1_n \rangle_D & \text{vecteur de } R^p \\ m_Y = \dfrac{1}{n}\, {}^t Y 1_n = \langle Y, 1_n \rangle_D & \text{vecteur de } R^q \end{cases}$$

• L'analyse canonique de X et Y fournit comme maxima à chaque étape du problème de l'analyse canonique $\rho_1 \ldots \rho_p$ (ordonnés par ordre décroissant), associés aux facteurs canoniques $\left(a^1 \ldots a^p\right)$ pour le tableau X et $\left(b^1 \ldots b^p\right)$ pour le tableau Y.

Par hypothèse, $1_n \in \left(W_X \cap W_Y\right)$. Nous sommes donc assurer de trouver après analyse canonique de X et Y un indice i_0 tel que :

$$Xa^{i_0} = Yb^{i_0} = 1_n \quad \text{et} \quad \rho_{i_0} = 1$$

Affectons la valeur 0 à cet indice i_0, et réordonnons par ordre décroissant les (p-1) autre valeurs propres ρ_i et leurs facteurs canoniques associés. Nous avons donc :

$$\begin{cases} \rho_0 = 1 \\ \rho_0 \geq \rho_1 \geq \ldots \geq \rho_{p-1} \end{cases}$$

Par ailleurs, pour tout i allant de 0 à (p-1), nous avons le résultat suivant :

$$(23) \quad \rho_i = \left\langle Xa^i, Yb^i \right\rangle_D$$

• En ce qui concerne les tableaux de données centrés, montrons tout d'abord que le rang des matrices X^c et Y^c sont respectivement (p-1) et (q-1), et non p et q comme les matrices X et Y.

Nous savons qu'il existe un vecteur a^0 de R^p tel que :

$$Xa^0 = 1_n \quad , \text{ car } \quad 1_n \in W_X$$

En remplaçant X par son expression en fonction de X^c, il vient :

$$\left(X^c + 1_n \, {}^t m_X\right) a^0 = 1_n$$

d'où :

$$\begin{aligned} X^c a^0 &= 1_n - 1_n \, {}^t m_X \, a^0 \\ &= 1_n - 1_n \, {}^t 1_n \, D \underbrace{X a^0}_{} && \text{d'après (22)} \\ &= 1_n - 1_n \underbrace{{}^t 1_n D 1_n}_{} && \text{car } Xa^0 = 1_n \\ &= 1_n - 1_n && \text{car } {}^t 1_n D 1_n = 1 \\ &= 0_n \end{aligned}$$

où $0_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ est un vecteur de R_n.

De ce fait, l'un des vecteurs de la matrice X^c peut s'exprimer comme une combinaison linéaire des autres vecteurs de cette matrice. La matrice X^c est donc de rang (p-1), et non de rang p comme la matrice X.

De la même façon, comme $1_n \in W_Y$, on peut montrer que la matrice Y^c est de rang (q-1), et non de rang q comme la matrice Y.

• Le rang des matrices X^c et Y^c étant respectivement (p-1) et (q-1) avec $p \leq q$, l'analyse canonique de X^c et Y^c fournit (p-1) corrélations canoniques $\mu_1 \ldots \mu_{p-1}$ (ordonnées par ordre décroissant), associées aux facteurs canoniques $\left(c^1 \ldots c^{p-1}\right)$ pour X^c et $\left(d^1 \ldots d^{p-1}\right)$ pour Y^c.

Par ailleurs, pour tout i allant de 1 à (p-1), nous avons le résultat suivant :

(24) $\mu_i = \left\langle X^c c^i, Y^c d^i \right\rangle_D$

Cherchons donc les relations existant entre ρ_i et μ_i, a^i et c^i, b^i et d^i.

• Tout d'abord, il est possible de démontrer que :

$$\boxed{\rho_i = \mu_i \quad , \forall\, i = 1 \ldots (p-1)}$$

$\rho_1 \ldots \rho_{p-1}$, qui sont calculés sur données non centrées, s'interprètent donc aussi comme des coefficients de corrélation.

• Que peut-on en déduire pour les facteurs canoniques ?

Dans l'équation (23), remplaçons X et Y par leur expression en fonction de X^c et Y^c. On obtient, pour tout i allant de 0 à (p-1) :

$$(25) \quad \rho_i = \left\langle \left(X^c + 1_n\, {}^t m_X\right) a^i, \left(Y^c + 1_n\, {}^t m_Y\right) b^i \right\rangle_D$$

$$= \left\langle X^c a^i, Y^c b^i \right\rangle_D + \left\langle 1_n\, {}^t m_X a^i, 1_n\, {}^t m_Y b^i \right\rangle_D$$
$$+ \left\langle X^c a^i, 1_n\, {}^t m_Y b^i \right\rangle_D + \left\langle 1_n\, {}^t m_X a^i, Y^c b^i \right\rangle_D$$

Nous savons que :

$$\rho_0 = 1 \quad \text{et} \quad X a^0 = Y b^0 = 1_n$$

et que, par hypothèse du problème de l'analyse canonique :

$$\left.\begin{array}{l} \left\langle Xa^i, Xa^j \right\rangle_D = 0 \\ \left\langle Yb^i, Yb^j \right\rangle_D = 0 \end{array}\right\} \forall\, i \neq j$$

Donc, pour tout i allant de 1 à (p-1),

$$\begin{aligned} \left\langle Xa^0, Xa^i \right\rangle_D &= 0 \\ &= \left\langle 1_n, Xa^i \right\rangle_D \\ &= {}^t 1_n D X a^i \end{aligned}$$

D'après la première équation du système (22),

$${}^t 1_n D X = {}^t m_X$$

On a donc, pour tout i allant de 1 à (p-1) :

$$\left\langle Xa^0, Xa^i \right\rangle_D = {}^t m_X a^i = 0$$

Et de même, on trouverait :

$$\left\langle Yb^0, Yb^i \right\rangle_D = {}^t m_Y b^i = 0 \quad , \forall\, i = 1 \ldots (p-1)$$

En reportant ces deux résultats dans l'équation (25), il vient :

$$\rho_i = \left\langle X^c a^i, Y^c b^i \right\rangle_D \quad , \forall\, i = 1 \ldots (p-1)$$

Or, d'après (24),

$$\mu_i = \left\langle X^c c^i, Y^c d^i \right\rangle_D \quad , \forall\, i = 1 \ldots (p-1)$$

Comme $\rho_i = \mu_i$ pour tout i allant de 1 à (p-1), a^i et b^i apparaissent aussi comme les facteurs canoniques associés respectivement à X^c et Y^c. Par unicité des facteurs canoniques, on en déduit donc que :

$$\boxed{\begin{cases} a^i = c^i \\ b^i = d^i \end{cases} \quad \forall\, i = 1 \ldots (p-1)}$$

Les **facteurs canoniques** associés aux tableaux de données centrés et non centrés sont donc identiques.

• Que peut-on en déduire pour les variables canoniques ?

Notons φ^i et ψ^i les variables canoniques associées aux tableaux X et Y, et $\tilde{\varphi}^i$ et $\tilde{\psi}^i$ les variables canoniques associées aux tableaux centrés X^c et Y^c.

Remarquons tout d'abord que :

$$\begin{cases} \varphi^0 = Xa^0 = 1_n \\ \psi^0 = Yb^0 = 1_n \end{cases}$$

Par ailleurs, pour tout i allant de 1 à (p-1), on a :

$$\begin{aligned} \tilde{\varphi}^i &= X^c a^i \\ &= \left(X - 1_n\, {}^t m_X\right) a^i \\ &= Xa^i - 1_n\, {}^t m_X\, a^i \end{aligned}$$

Comme nous avons vu précédemment que :

$${}^t m_X\, a^i = 0 \quad , \forall\, i = 1 \ldots (p-1)$$

il vient :

$$\boxed{\tilde{\varphi}^i = Xa^i = \varphi^i \quad , \forall\, i = 1 \ldots (p-1)}$$

Et de même, on montrerait que :

$$\boxed{\tilde{\psi}^i = \psi^i \quad , \forall\, i = 1 \ldots (p-1)}$$

Les **variables canoniques** obtenues sur tableaux de données centrés ou non centrés sont donc identiques.

De plus, pour tout i allant de 1 à (p-1),

$$\begin{aligned} m_{\tilde{\varphi}^i} &= m_{\varphi^i} \\ &= \left\langle \tilde{\varphi}^i, 1_n \right\rangle_D \\ &= \left\langle X^c a^i, 1_n \right\rangle_D \\ &= {}^t a^i\, \underbrace{{}^t X^c D 1_n} \\ &= {}^t a^i\, m_{X^c} && \text{d'après (22)} \\ &= 0 && \text{car } X^c \text{ est centré} \end{aligned}$$

Donc,

$$\boxed{m_{\tilde{\varphi}^i} = m_{\varphi^i} = 0 \quad , \forall\, i = 1 \ldots (p-1)}$$

Et de même, on montrerait que :

$$\boxed{m_{\tilde{\psi}^i} = m_{\psi^i} = 0 \quad , \forall\, i = 1 \ldots (p-1)}$$

Donc les variables canoniques (associées aux tableaux de données centrés ou non centrés) sont centrées.

En conclusion, si $1_n \in (W_X \cap W_Y)$, les coefficients canoniques ρ_i obtenus sur données non centrées sont égaux aux corrélations canoniques μ_i obtenues sur données centrées, pour tout i allant de 1 à (p-1).

De plus, il y a égalité des facteurs canoniques correspondants, ainsi que des variables canoniques correspondantes. Enfin, ces variables canoniques sont centrées.

On constate donc qu'il n'est **pas nécessaire de centrer les données avant d'appliquer l'analyse canonique** dans ce cas particulier.

III - CODAGE DES DONNÉES ET ANALYSE CANONIQUE

1 Introduction

L'analyse canonique est une méthode recherchant des relations linéaires entre deux groupes de variables $\left(X^1 \ldots X^p\right)$ et $\left(Y^1 \ldots Y^q\right)$. Si des relations non linéaires existent entre ces deux groupes, elles ne peuvent être mises en évidence par analyse canonique. Pour pouvoir établir de telles relations, nous avons donc décidé de coder les tableaux de données X et Y (ce qui revient en fait à transformer les variables initiales) préalablement à l'analyse canonique (Epitalon, 1985, Mallet *et al.*, 1985).

Deux codages ont été utilisés : le codage probabiliste et le codage disjonctif. Ces codages sont liés à la définition de classes sur la population générée par chacun des tableaux de données. Nous avons donc fait intervenir la méthode de décomposition en classes gaussiennes pour définir les classes nécessaires au codage.

Dans cette partie, nous présentons les deux types de codage envisagés. Puis, nous montrons les conséquences de ces codages sur la méthode d'analyse canonique.

2 Codage d'un tableau de données

Soit X un tableau de données à n individus et p variables $X^1 \ldots X^p$, que nous souhaitons coder. La méthode de décomposition en classes gaussiennes, appliquée à la population définie par le tableau X, fournit K_X classes notées $C_1 \ldots C_{K_X}$.

Nous décrivons ci-dessous les codages probabiliste et disjonctif appliqués au tableau de données X, connaissant cette démonstration.

2.1 Codage probabiliste

Dans le codage probabiliste, ce sont directement les probabilités d'appartenance aux différentes classes qui remplacent les classes initiales. Notons I_X le tableau de données après codage probabiliste. I_X est un tableau de taille $n \times K_X$, tel que :

$$\begin{aligned} I_X(i,k) &= \text{valeur de l'individu i pour la variable k} \\ &= p\left(k/X_i\right) \end{aligned}$$

si nous reprenons la notation du chapitre 2 sur la décomposition en classes gaussiennes.

Notons $I^1 \ldots I^{K_X}$ les vecteurs des variables du tableau I_X. Par définition,

$$\sum_{k=1}^{K_X} I^k = 1_n$$

ou bien

$$\sum_{k=1}^{K_X} I_X(i,k) = 1 \qquad ,\forall\, i = 1 \ldots n$$

Donc, quel que soit le tableau initial X, le vecteur 1_n appartient au sous-espace engendré par I_X après codage. Nous noterons ce sous-espace W_{I_X}.

Finalement, les caractéristiques du codage probabiliste peuvent se résumer dans le tableau suivant.

Avant codage	**Après codage**
• tableau X	• tableau I_X
• p variables $X^1 \ldots X^p$	• K_X variables $I^1 \ldots I^{K_X}$
• X(i,j) quelconque	• $I_X(i,k) \in [0,1]$
• pas de propriétés particulières sur W_X	• $1_n \in W_{I_X}$

2.2 Codage disjonctif

Le codage disjonctif consiste à attribuer chaque individu à la classe la plus probabiliste ; les "indicatrices" ainsi formées prennent donc les valeurs 0 ou 1. Notons J_X le tableau de données après codage disjonctif. J_X est un tableau de taille $n \times K_X$ dont le ième individu prend la valeur suivante pour la variable k :

$$J_X(i,k) = \begin{cases} 1 & \text{si } X_i \in C_k \\ 0 & \text{sinon} \end{cases}$$

Les classes obtenues par la méthode de décomposition en classes gaussiennes étant généralement recouvrantes, nous dirons que l'individu X_i appartient à la classe j qui maximise sa probabilité d'appartenance. On peut donc écrire :

$$J_X(i,j) = 1 \qquad \text{si} \quad p(j/X_i) = \underset{k=1 \ldots K_X}{\text{Max}}\ p(k/X_i)$$

pour faire apparaître les probabilités *a posteriori* fournies par la décomposition en classes gaussiennes.

De ce fait, comme pour le codage probabiliste, nous avons :

$$\sum_{k=1}^{K_X} J^k = 1_n$$

ou

$$\sum_{k=1}^{K_X} J_X(i,k) = 1 \qquad ,\forall\, i = 1...n$$

si nous notons J^k la kième variable du tableau J_X. Le vecteur 1_n appartient donc au sous-espace engendré par J_X, sous-espace que nous noterons W_{J_X}.

En résumé, le codage disjonctif possède donc les caractéristiques suivantes.

Avant codage	Après codage
• tableau X	• tableau J_X
• p variables $X^1...X^p$	• K_X variables $J^1...J^{K_X}$
• $X(i,j)$ quelconque	• $J_X(i,k) \in \{0,1\}$
• pas de propriétés particulières sur W_X	• $1_n \in W_{J_X}$

3 Analyse canonique sur données codées

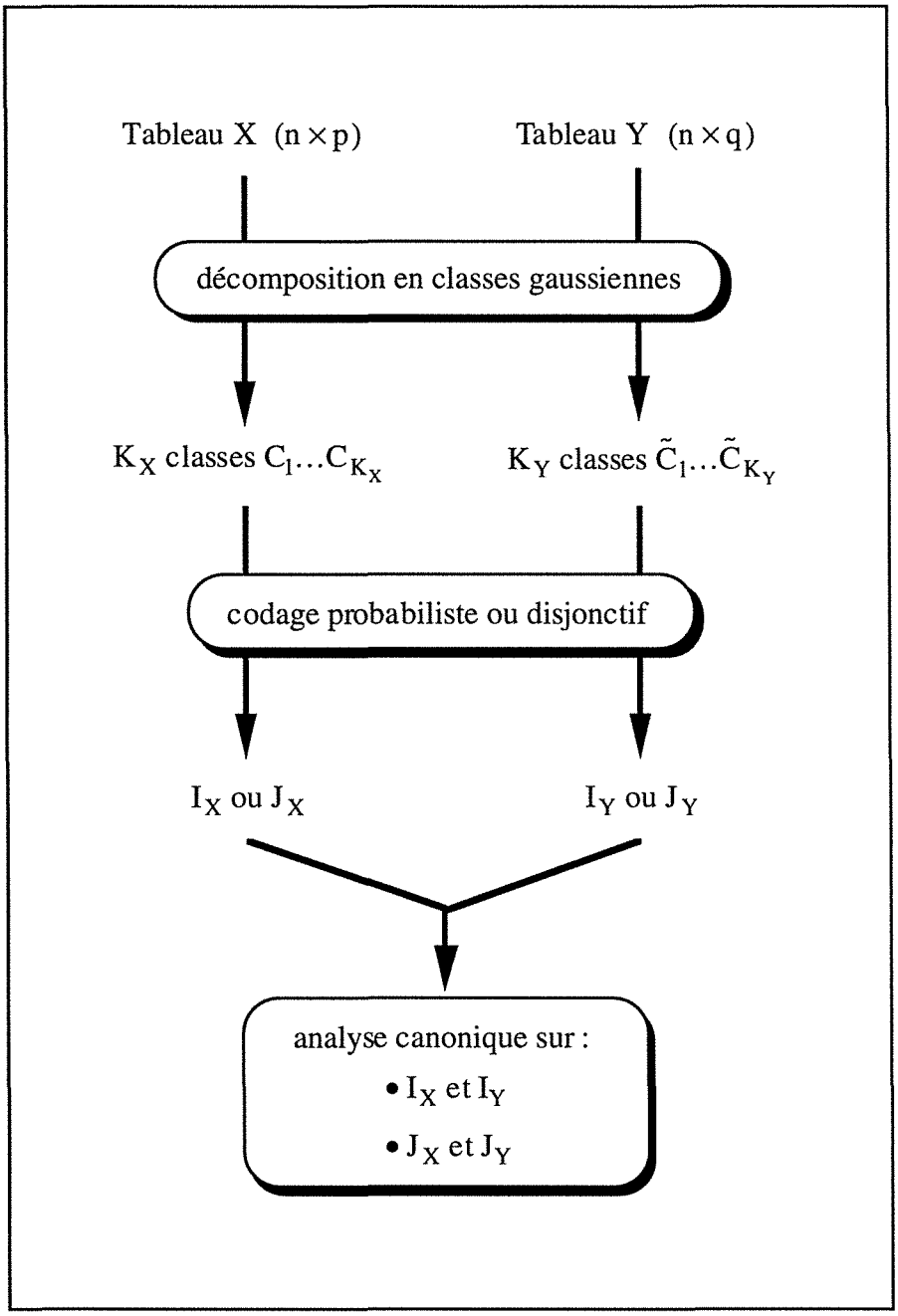
Nous présentons dans ce paragraphe les principes et les conséquences de l'analyse canonique sur données codées, par codage disjonctif ou probabiliste.

3.1 Méthodologie

Soit X et Y les tableaux de données initiaux, de tailles respectives $n \times p$ et $n \times q$. La méthode de décomposition en classes gaussiennes appliquée au tableau de données X fournit K_X classes notées $C_1...C_{K_X}$. De même, nous obtenons, pour le tableau de données Y, K_Y classes notées $\tilde{C}_1...\tilde{C}_{K_Y}$.

Les codages probabiliste et disjonctif du tableau de données X fournissent respectivement les tableaux I_X et J_X, tous deux de taille $n \times K_X$. De même, nous obtenons les tableaux I_Y et J_Y (de taille $n \times K_Y$) par codage du tableau de données Y.

Notre but est de faire l'analyse canonique des tableaux I_X et I_Y d'une part, J_X et J_Y d'autre part, au lieu de faire l'analyse canonique de X et Y. Cette méthodologie peut donc se schématiser de la façon suivante.



3.2 Conséquences des deux codages sur l'analyse canonique

Nous avons vu que le vecteur 1_n appartient à W_{I_X}, où I_X est le tableau de données obtenu après codage probabiliste de X. De même, 1_n appartient au sous-espace W_{I_Y} issu du codage probabiliste de Y. On a donc :

$$1_n \in \left(W_{I_X} \cap W_{I_Y}\right)$$

On en déduit qu'il n'y a **pas nécessité de centrer les tableaux codés** I_X et I_Y avant de leur appliquer l'analyse canonique.

Ce résultat est aussi valable pour l'analyse canonique de J_X et J_Y (tableaux issus du codage disjonctif de X et Y), car on pourrait montrer que :

$$1_n \in \left(W_{I_X} \cap W_{I_Y}\right)$$

Nous allons donc analyser ci-dessous les conséquences théoriques de chaque codage pour l'analyse canonique.

3.2.1 Codage probabiliste

Rappelons que l'analyse canonique de deux tableaux X et Y quelconques consiste à résoudre successivement p problèmes. Et à une étape i quelconque (pour i allant de 1 à p), le problème posé est le suivant :

$$\begin{cases} \text{Max} \quad \rho_i = \left\langle Xa^i, Yb^i \right\rangle_D \\ \left\| Xa^i \right\|_D = \left\| Yb^i \right\|_D = 1 \\ \left\langle Xa^i, Xa^j \right\rangle_D = \left\langle Yb^i, Yb^j \right\rangle_D = 0 \quad , \quad \forall\, j \prec i \end{cases}$$

les inconnues étant les vecteurs a^i et b^i (respectivement de R^p et R^q) appelés facteurs canoniques, et le coefficient de corrélation canonique ρ_i. Ce problème revient à résoudre le système d'équations canoniques suivant :

$$\begin{cases} \rho_i^2\, a^i = V_{XX}^{-1} V_{XY} V_{YY}^{-1} V_{YX}\, a^i \\ \rho_i^2\, b^i = V_{YY}^{-1} V_{YX} V_{XX}^{-1} V_{XY}\, b^i \end{cases}$$

Après codage probabiliste de X et Y, ce système s'écrit :

$$\begin{cases} \rho_i^2\, a^i = V_{I_XI_X}^{-1} V_{I_XI_Y} V_{I_YI_Y}^{-1} V_{I_YI_X}\, a^i \\ \rho_i^2\, b^i = V_{I_YI_Y}^{-1} V_{I_YI_X} V_{I_XI_X}^{-1} V_{I_XI_Y}\, b^i \end{cases}$$

où a^i et b^i sont respectivement des vecteurs de R^{K_X} et R^{K_Y}.

Cherchons l'expression des matrices $V_{I_XI_X}$, $V_{I_YI_Y}$, $V_{I_XI_Y}$ et $V_{I_YI_X}$, qui ont pour tailles respectives $K_X \times K_X$, $K_Y \times K_Y$, $K_X \times K_Y$ et $K_Y \times K_X$.

- $V_{I_X I_X} = {}^t I_X \, D \, I_X$

$$= \begin{bmatrix} & s_{kl} & \\ & \vdots & \end{bmatrix} \text{ ligne k}$$

colonne l

où $s_{kl} = \frac{1}{n} \sum_{i=1}^{n} \left[p(k/X_i) \cdot p(l/X_i) \right]$

$$= \frac{1}{n} P\left\{ X_i \in (C_k \cap C_l), \forall i \right\}$$

s_{kl} s'interprète comme la moyenne des probabilités d'appartenance à l'intersection des classes C_k et C_l obtenues sur le tableau X.

- $V_{I_Y I_Y} = {}^t I_Y \, D \, I_Y$

$$= \begin{bmatrix} & t_{kl} & \\ & \vdots & \end{bmatrix} \text{ ligne k}$$

colonne l

où $t_{kl} = \frac{1}{n} \sum_{i=1}^{n} \left[p(k/Y_i) \cdot p(l/Y_i) \right]$

$$= \frac{1}{n} P\left\{ Y_i \in (\tilde{C}_k \cap \tilde{C}_l), \forall i \right\}$$

t_{kl} s'interprète comme la moyenne des probabilités d'appartenance à l'intersection des classes $\tilde{C}_k$ et $\tilde{C}_l$ obtenues sur le tableau Y.

- $V_{I_X I_Y} = {}^t I_X \, D \, I_Y$

$$= \begin{bmatrix} & u_{kl} & \\ & \vdots & \end{bmatrix} \text{ ligne k}$$

colonne l

où $u_{kl} = \frac{1}{n}\sum_{i=1}^{n}\left[p(k/X_i)\cdot p(l/Y_i)\right]$

$= \frac{1}{n} P\left\{X_i \in C_k \quad \text{et} \quad Y_i \in \tilde{C}_l, \forall i\right\}$

u_{kl} s'interprète comme la moyenne des probabilités d'appartenance à la classe C_k obtenue sur le tableau X et à la classe $\tilde{C}_l$ obtenue sur le tableau Y.

- $V_{I_Y I_X} = {}^t V_{I_X I_Y}$

Nous n'avons pas détaillé plus les développements théoriques, et ne pouvons donc pas fournir d'interprétation probabiliste des équations canoniques.

3.2.2 Codage disjonctif

A une étape i quelconque (pour i allant de 1 à p), le système des équations canoniques, après codage disjonctif des tableaux de données X et Y, s'écrit :

$$\begin{cases} \rho_i^2 a^i = V_{J_X J_X}^{-1} V_{J_X J_Y} V_{J_Y J_Y}^{-1} V_{J_Y J_X} a^i \\ \rho_i^2 b^i = V_{J_Y J_Y}^{-1} V_{J_Y J_X} V_{J_X J_X}^{-1} V_{J_X J_Y} b^i \end{cases}$$

où les inconnues sont les vecteurs a^i et b^i (respectivement de R^{K_X} et R^{K_Y}) et ρ_i.

Cherchons à expliciter les matrices $V_{J_X J_X}$, $V_{J_Y J_Y}$, $V_{J_X J_Y}$ et $V_{J_Y J_X}$, de tailles respectives $K_X \times K_X$, $K_Y \times K_Y$, $K_X \times K_Y$ et $K_Y \times K_X$.

- $V_{J_X J_X} = {}^t J_X D J_X$

$$= \frac{1}{n}\begin{bmatrix} n_{1.} & & 0 \\ & \ddots & \\ 0 & & n_{K_X.} \end{bmatrix} = \begin{bmatrix} p_{1.} & & 0 \\ & \ddots & \\ 0 & & p_{K_X.} \end{bmatrix}$$

où $n_{k.}$, appelé effectif de la classe C_k, est le nombre d'individus i tels que X_i appartienne à la classe C_k, et où $p_{k.} = \frac{n_{k.}}{n}$ est le pourcentage d'individus appartenant à la classe C_k.

- $V_{J_Y J_Y} = {}^t J_Y D J_Y$

$$= \frac{1}{n} \begin{bmatrix} n_{.1} & & 0 \\ & \ddots & \\ 0 & & n_{.K_Y} \end{bmatrix} = \begin{bmatrix} p_{.1} & & 0 \\ & \ddots & \\ 0 & & p_{.K_Y} \end{bmatrix}$$

où $n_{.k}$, appelé effectif de la classe $\tilde{C}_k$, est le nombre d'individus i tels que Y_i appartienne à la classe $\tilde{C}_k$, et où $p_{.k} = \frac{n_{.k}}{n}$ est le pourcentage d'individus appartenant à la classe $\tilde{C}_k$.

- $V_{J_X J_Y} = {}^t J_X D J_Y$

$$= \frac{1}{n} \begin{bmatrix} & n_{kl} & \text{---- ligne k} \\ & \vdots & \\ & \text{colonne l} & \end{bmatrix}$$

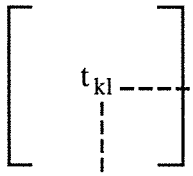
où n_{kl} est le nombre d'individus i tels que $\left\{X_i \in C_k \quad \text{et} \quad Y_i \in \tilde{C}_l\right\}$, et où $p_{kl} = \frac{n_{kl}}{n}$ est le pourcentage de ces individus.

- $V_{J_Y J_X} = {}^t V_{J_X J_Y}$

Développons maintenant les matrices $\left(V_{J_X J_X}^{-1} V_{J_X J_Y}\right)$ et $\left(V_{J_Y J_Y}^{-1} V_{J_Y J_X}\right)$, qui sont de tailles respectives $K_X \times K_Y$ et $K_Y \times K_X$.

- $V_{J_X J_X}^{-1} V_{J_X J_Y} = \begin{bmatrix} & s_{kl} & \text{---- ligne k} \\ & \vdots & \\ & \text{colonne l} & \end{bmatrix}$

où $s_{kl} = \frac{p_{kl}}{p_{k.}} = \frac{n_{kl}}{n_{k.}}$ s'interprète comme le pourcentage d'individus i réalisant $\left\{Y_i \in \tilde{C}_l\right\}$ parmi ceux réalisant $\left\{X_i \in C_k\right\}$.

- $V_{J_Y J_Y}^{-1} V_{J_Y J_X} =$  ligne k

colonne l

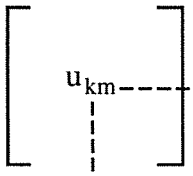
où $t_{kl} = \frac{p_{lk}}{p_{.k}} = \frac{n_{lk}}{n_{.k}}$ s'interprète comme le pourcentage d'individus i réalisant $\{X_i \in C_l\}$ parmi ceux réalisant $\{Y_i \in \tilde{C}_k\}$.

On peut alors expliciter les matrices :

$$M_1 = V_{J_X J_X}^{-1} V_{J_X J_Y} V_{J_Y J_Y}^{-1} V_{J_Y J_X}$$

$$M_2 = V_{J_Y J_Y}^{-1} V_{J_Y J_X} V_{J_X J_X}^{-1} V_{J_X J_Y}$$

qui sont de tailles respectives $K_X \times K_X$ et $K_Y \times K_Y$, et qui interviennent dans les équations canoniques. Par exemple,

$M_1 =$  ligne k

colonne m

où :

$$u_{km} = \sum_{l=1}^{K_Y} \left(\frac{n_{kl}}{n_{k.}} \times \frac{n_{ml}}{n_{.l}} \right)$$

Donc dans le cadre d'un codage disjonctif, il nous est possible de calculer directement la matrice M_1 à partir de l'expression précédente. Et nous savons que le facteur canonique a^i (vecteur de R^{K_X}) est le vecteur propre de cette matrice associé à la ième plus grande valeur propre ρ_i^2.

De même, nous pouvons expliciter la matrice M_2 pour laquelle b^i (vecteur de R^{K_Y}) est vecteur propre associé à la ième plus grande valeur propre ρ_i^2. Nous pouvons aussi directement calculer b^i à l'aide de l'équation :

$$b^i = \frac{1}{\rho_i} V_{J_Y J_Y}^{-1} V_{J_Y J_X} a^i$$

puisque nous avons pu expliciter la matrice $\left(V_{J_Y J_Y}^{-1} V_{J_Y J_X} \right)$.

Pour conclure, nous pourrions démontrer qu'il est équivalent de pratiquer :

- l'analyse canonique de J_X et J_Y,
- l'analyse factorielle des correspondances du tableau $J = \left[J_X \middle| J_Y \right]$,
- l'analyse factorielle des correspondances du tableau de contingence $N = {}^t J_X \cdot J_Y$,
- l'analyse en composantes principales du tableau des profils-lignes (qui est en fait la matrice $V_{J_X J_X}^{-1} V_{J_X J_Y}$), ou du tableau des profils-colonnes (matrice $V_{J_Y J_Y}^{-1} V_{J_Y J_X}$), issus du tableau de contingence N.

On pourra se reporter à l'ouvrage de Saporta (1990) pour établir ces démonstrations.

4 Conclusion

Pour pouvoir traiter les relations de type non linéaire existant entre deux groupes de variables, nous avons décidé de coder les tableaux de données générés par ces deux groupes, avant d'appliquer l'analyse canonique.

Nous avons retenu deux codages, l'un probabiliste, l'autre disjonctif, faisant tous deux intervenir les résultats de la méthode de décomposition en classes gaussiennes. Nous avons montré que ces deux codages présentent des propriétés intéressantes, et qu'entre autres, il n'est pas nécessaire de centrer les données codées pour l'analyse canonique.

Nous désirons ultérieurement appliquer cette méthodologie dans le cadre du calibrage géologique des données sismiques, pour pouvoir traiter les cas où des relations non linéaires existent entre les propriétés géologiques aux puits et les attributs sismiques des traces adjacentes aux puits.

CHAPITRE 4

MÉTHODE DE DÉCOMPOSITION EN CLASSES GAUSSIENNES ET RÉGRESSION NON PARAMÉTRIQUE

I - THÉORIE DE LA RÉGRESSION

1 Cas général

Supposons que nous cherchions à prédire une variable Y en fonction d'un ensemble X de p variables $X^1 \ldots X^p$. Nous connaissons une population de calibrage, c'est-à-dire un ensemble d'individus pour lesquels les variables $X^1 \ldots X^p$ et Y ont été observées. Par ailleurs, pour d'autres individus, seules les variables $X^1 \ldots X^p$ ont été observées : nous souhaitons donc prédire pour ces individus la variable Y. Supposons que X et Y ne soient pas indépendants. Alors régresser Y en X consiste à prédire Y à partir des observations faites sur X, à l'aide d'une formule de prévision du type :

$$\hat{Y} = R(X)$$

où R, qui est appelée fonction de régression de Y en X, est une fonction de X à définir.

Notons ε l'erreur de prévision faite sur Y en utilisant cette formule :

$$\varepsilon = Y - \hat{Y} = Y - R(X)$$

ε est un résidu aléatoire qui peut être important. De ce fait, on recherche la fonction $R(X)$ pour laquelle l'espérance de l'erreur de prévision ε est nulle et sa variance minimale, c'est-à-dire telle que :

$$\begin{cases} E(\varepsilon) = 0 = E[Y - R(X)] \\ Var(\varepsilon) = Var[Y - R(X)] \quad \text{est minimale} \end{cases}$$

On peut alors montrer que cette fonction $R(X)$ est en fait l'espérance de Y conditionnellement en X, d'où la **formule de prévision** :

$$\boxed{\hat{Y} = R(X) = E(Y/X)}$$

De plus, l'erreur de prévision ε a la propriété d'être non corrélée avec X, c'est-à-dire que $E(\varepsilon X) = 0$.

Supposons connue une observation $x = \left(x^1, \ldots, x^p\right)$ faite sur X. La fonction de régression de Y en $\{X = x\}$ s'écrit donc :

$$\begin{aligned} R(X = x) &= E(Y/X = x) \\ &= \int_R y \, dP(y/X = x) \end{aligned}$$

où y est une réalisation possible de Y, et où $P(y/X=x)$ est la probabilité d'obtenir la réalisation $\{Y=y\}$ sachant que $\{X=x\}$.

Si de plus, dP admet une densité de probabilité notée f, alors :

$$dP(y/X=x) = f(y/X=x)\,dy$$

On obtient donc comme fonction de régression la fonction :

$$\begin{aligned} R(X=x) &= E(Y/X=x) \\ &= \int_R y\, f(y/X=x)\,dy \\ &= \int_R y\, \frac{f(x,y)}{f(x,.)}\,dy \end{aligned}$$

qui peut être utilisée pour prédire Y connaissant toute observation x de X.

On pourra se reporter aux ouvrages de Rao (1965) ou Kendall et Stuart (1979) qui présentent, entre autres, la théorie de la régression.

2 Cas particulier : la régression linéaire

La régression linéaire est le cas le plus important en pratique. Nous allons présenter séparément le cas p=1 (on parle alors de **régression simple**) et le cas p>1 (**régression multiple**).

2.1 Régression simple

Soit X et Y deux variables observées sur une population de calibrage. Pour d'autres individus, seule la variable X a été observée. Alors, la régression linéaire simple consiste à prédire Y à partir de X par une formule de régression du type :

$$\hat{Y} = R(X) = aX + b$$

Nous allons démontrer que, dans ce cas, les paramètres a et b prennent les valeurs suivantes :

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var(X)} \\ b = E(Y) - \dfrac{Cov(X,Y)}{Var(X)} \cdot E(X) \end{cases}$$

L'erreur de prévision ε étant d'espérance nulle, nous avons :

$$E(Y) = E[R(X)]$$

Dans le cas particulier de la régression linéaire, nous obtenons donc :

$$\begin{aligned} E(Y) &= E(aX+b) \\ &= a \cdot E(X) + b \end{aligned}$$

d'où :

$$\boxed{b = E(Y) - a \cdot E(X)}$$

Par ailleurs, si ε est l'erreur de prévision faite sur Y en utilisant la formule de régression, nous obtenons :

$$\begin{aligned} Y &= R(X) + \varepsilon \\ &= aX + b + \varepsilon \\ &= aX + [E(Y) - a \cdot E(X)] + \varepsilon \end{aligned}$$

en utilisant l'expression de b précédente.

Il vient donc :

$$Y - E(Y) = a[X - E(X)] + \varepsilon$$

Multiplions chaque membre de cette expression par $[X - E(X)]$, et calculons son espérance. On obtient :

$$E[(Y - E(Y))(X - E(X))] = a \cdot E[(X - E(X))^2] + E[\varepsilon(X - E(X))]$$

soit :

$$Cov(X,Y) = a \cdot Var(X) + E(\varepsilon X) - E(\varepsilon) \cdot E(X)$$

L'erreur de prévision ε est d'espérance nulle, donc $E(\varepsilon) = 0$. De plus, elle est non corrélée avec X, donc $E(\varepsilon X) = 0$. De ce fait :

$$\boxed{Cov(X,Y) = a \cdot Var(X)}$$

En conclusion, dans le cadre de la **régression simple**, la formule de **régression de Y en X** s'écrit :

$$\boxed{\hat{Y} = aX + b}$$

avec :

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var(X)} \\ b = E(Y) - \dfrac{Cov(X,Y)}{Var(X)} \cdot E(X) \end{cases}$$

2.2 Régression multiple

Soit $X^1,\ldots,X^p$ et Y un ensemble de variables observées sur une population de calibrage. Pour d'autres individus, seules les variables $X^1,\ldots,X^p$ ont été observées. Pour simplifier les écritures, nous supposerons que les variables $X^1,\ldots,X^p$ et Y sont centrées. Alors, la régression multiple consiste à prédire Y à partir des variables $X^1,\ldots,X^p$ à l'aide de la formule de régression suivante :

$$\hat{Y} = R\left(X^1 \ldots X^p\right) = a_1 X^1 + \ldots + a_p X^p$$

ce qui peut s'écrire :

$$\hat{Y} = R(X) = Xa$$

avec :

- $X = \left[X^1 \ldots X^p\right]$
- $a = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$

Nous allons démontrer que dans ce cas :

$$a = \left[E\left({}^tXX\right)\right]^{-1} \cdot E\left({}^tXY\right)$$

Notons ε l'erreur de prévision faite sur Y en utilisant la formule de régression. Par hypothèse, nous devons avoir :

$$\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y - Xa) \quad \text{minimale}$$

$$\begin{aligned} \mathrm{Var}(Y - Xa) &= E\left[{}^t(Y - Xa)(Y - Xa)\right] \\ &= E\left({}^tYY\right) - E\left({}^tYXa\right) - E\left[{}^t(Xa)Y\right] + E\left[{}^t(Xa)Xa\right] \\ &= E\left({}^tYY\right) - 2 \cdot \left[E\left({}^tYX\right)\right] \cdot a + {}^ta \cdot E\left({}^tXX\right) \cdot a \end{aligned}$$

Dérivons cette formule par rapport à a pour déterminer le paramètre â qui réalise l'optimum de $\mathrm{Var}(Y - Xa)$. Il vient :

$$2 \cdot \left[E\left({}^tXX\right)\right] \cdot a - 2 \cdot E\left({}^tXY\right) = 0$$

d'où l'optimum :

$$\boxed{\hat{a} = \left[E\left({}^tXX\right)\right]^{-1} \cdot E\left({}^tXY\right)}$$

Montrons maintenant que l'optimum $\hat{a}$ réalise effectivement le minimum de $\mathrm{Var}(Y - Xa)$. Notons δa un vecteur quelconque de R^p. Il nous faut démontrer que :

$$\mathrm{Var}\left[Y - X(\hat{a} + \delta a)\right] \geq \mathrm{Var}(Y - X\hat{a}) \quad , \forall\, \delta a \in R^p$$

A l'optimum $\hat{a}$, nous savons d'une part que :

$$(1) \quad \mathrm{Var}(Y - X\hat{a}) = E\left({}^tYY\right) - 2 \cdot \left[E\left({}^tYX\right)\right] \cdot \hat{a} + {}^t\hat{a} \cdot E\left({}^tXX\right) \cdot \hat{a}$$

et d'autre part que :

$$(2) \quad \left(E\left({}^tXX\right)\right) \cdot \hat{a} - E\left({}^tXY\right) = 0$$

d'après la formule dérivée permettant son calcul.

Alors :

$$\begin{aligned}
\mathrm{Var}\left[Y - X(\hat{a} + \delta a)\right] &= E\left[{}^t(Y - X\hat{a} - X\delta a)(Y - X\hat{a} - X\delta a)\right] \\
&= E\left({}^tYY\right) - 2 \cdot \left[E\left({}^tYX\right)\right] \cdot \hat{a} + {}^t\hat{a} \cdot E\left({}^tXX\right) \cdot \hat{a} \\
&\quad + {}^t\delta a \cdot E\left({}^tXX\right) \cdot \delta a + 2 \cdot {}^t\delta a \cdot \left[E\left({}^tXX\right)\right] \cdot \hat{a} - 2 \cdot {}^t\delta a \cdot \left[E\left({}^tXY\right)\right] \\
&= \mathrm{Var}(Y - X\hat{a}) \\
&\quad + {}^t\delta a \cdot E\left({}^tXX\right) \cdot \delta a + 2 \cdot {}^t\delta a \cdot \left[E\left({}^tXX\right)\right] \cdot \hat{a} - 2 \cdot {}^t\delta a \cdot \left[E\left({}^tXY\right)\right] \qquad \text{d'après (1)} \\
&= \mathrm{Var}(Y - X\hat{a}) \\
&\quad + {}^t\delta a \cdot E\left({}^tXX\right) \cdot \delta a + 2 \cdot {}^t\delta a \cdot \left[\left(E\left({}^tXX\right)\right) \cdot \hat{a} - E\left({}^tXY\right)\right] \\
&= \mathrm{Var}(Y - X\hat{a}) + {}^t\delta a \cdot E\left({}^tXX\right) \cdot \delta a \qquad \text{d'après (2)} \\
&= \mathrm{Var}(Y - X\hat{a}) + \underbrace{E\left[{}^t(X\delta a)(X\delta a)\right]}_{\geq 0 \quad , \forall\, \delta a}
\end{aligned}$$

Donc :

$$\boxed{\mathrm{Var}\left[Y - X(\hat{a} + \delta a)\right] \geq \mathrm{Var}(Y - X\hat{a}) \quad , \forall\, \delta a \in R^p}$$

En conclusion, le vecteur $\hat{a}$ minimise $\mathrm{Var}(Y - Xa)$ qui est la variance de l'erreur de prévision. Dans le cadre de la régression multiple, la formule théorique de régression de Y en X s'écrit donc :

$$\boxed{\hat{Y} = X\hat{a} = X \cdot \left[E\left({}^tXX\right)\right]^{-1} \cdot E\left({}^tXY\right)}$$

En pratique, cette formule théorique peut être utilisée de la façon suivante. Supposons connues n réalisations $\left\{\left(x_i^1,\ldots,x_i^p,y_i\right), i=1\ldots n\right\}$ des variables $X^1,\ldots,X^p$ et Y. Notons :

$$x=\begin{bmatrix} x_1^1 & \cdots & x_1^p \\ \vdots & & \vdots \\ x_n^1 & \cdots & x_n^p \end{bmatrix}$$

$$y=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Nous supposerons que le tableau x et le vecteur y sont centrés. Alors, une estimation de $\hat{a}$ est :

$$\hat{a}=\left({}^t xx\right)^{-1}.{}^t xy$$

Soit $\tilde{x}=\left(\tilde{x}^1\ldots\tilde{x}^p\right)$ une nouvelle réalisation des variables $X^1,\ldots,X^p$ pour laquelle la variable Y n'a pas été observée. Alors, la prédiction $\tilde{y}$ de Y associée à $\tilde{x}$ par régression multiple est :

$$\tilde{y}=\tilde{x}\hat{a}=\tilde{x}\cdot\left({}^t xx\right)^{-1}.{}^t xy$$

Notons que, si les variables $X^1,\ldots,X^p$ et Y sont telles que la loi de $\left(X^1,\ldots,X^p,Y\right)$ est une loi normale dans $R^{(p+1)}$, la régression sans modèle *a priori* (cas général) de Y en $\left(X^1\ldots X^p\right)$ correspond à la régression linéaire. Nous en fournissons la démonstration en Annexe B dans le cas p=1.

II - RÉGRESSION NON PARAMÉTRIQUE

1 Régression non paramétrique

Supposons que nous cherchions à prédire une variable Y en fonction d'un ensemble X de variables $X^1 \ldots X^p$. Alors, régresser Y en X nécessite :

- soit d'utiliser un modèle de régression (par exemple, le modèle linéaire), ce qui correspond le plus souvent à faire des hypothèses sur la loi des variables aléatoires,
- soit d'estimer la fonction de densité conditionnelle $f(Y/X)$ pour rester dans le cadre de la théorie générale de la régression, ce qui évite de faire de telles hypothèses ; on parle dans ce cas de **régression non paramétrique**.

Le plus souvent, dans le cadre de la régression non paramétrique, la méthode d'estimation de la fonction de densité conditionnelle est la méthode des noyaux (Silverman, 1986). Cette méthode permet en fait d'estimer une fonction de densité quelconque. Dans le cadre de la régression, elle permet donc d'approximer, pour tout Y, $f(X = x, Y)$ et $f(X = x, \cdot)$, et donc $E(Y/X = x)$. Cette méthode a été utilisée, entre autres, par Collomb (1977) et Härdle et Marron (1985). Cependant, cette méthode suppose le choix de la taille du noyau utilisé, qui contrôle le degré de lissage pour le calcul de la fonction de densité. Ce paramètre est donc très important, et en pratique difficile à choisir (Sheather et Jones, 1991).

D'autres méthodes, telle la méthode des k plus proches voisins, pourraient aussi être utilisées pour l'estimation des fonctions de densité $f(X = x, Y)$ et $f(X = x, \cdot)$, et donc servir dans le cadre de la régression non paramétrique.

2 Méthodologie de régression non paramétrique développée

Supposons que nous cherchions à prédire un ensemble de q variables $Y^1 \ldots Y^q$ en fonction d'un ensemble de p variables $X^1 \ldots X^p$. Nous connaissons une population de calibrage, c'est-à-dire un ensemble d'individus pour lesquels les variables $X^1 \ldots X^p$ et $Y^1 \ldots Y^q$ ont été observées. Par ailleurs, pour d'autres individus, seules les variables $X^1 \ldots X^p$ ont été observées : nous souhaitons donc prédire pour ces individus les variables $Y^1 \ldots Y^q$.

La méthodologie de régression non paramétrique que nous avons développée est la suivante. Sur la population de calibrage, nous estimons la fonction de densité multivariable associée aux variables $X^1 \ldots X^p, Y^1 \ldots Y^q$ en utilisant les résultats fournis par la méthode de décomposition en classes gaussiennes. Ainsi :

$$\hat{f}\left(X^1 \ldots X^p, Y^1 \ldots Y^q\right) = \sum_{k=1}^{K} p_k \cdot f_k\left(X^1 \ldots X^p, Y^1 \ldots Y^q\right)$$

où :

- $\hat{f}$ est la fonction de densité estimée,
- K est le nombre de classes gaussiennes obtenues sur la population de calibrage,
- p_k est le poids de la kième classe gaussienne,
- f_k est la fonction de densité gaussienne multivariable associée à la kième classe gaussienne.

Nous pouvons alors calculer une estimation de la fonction de densité conditionnelle des variables $Y^1 \ldots Y^q$ connaissant une réalisation $x^1 \ldots x^p$ des variables $X^1 \ldots X^p$:

$$\hat{f}\left(Y^1 \ldots Y^q / x^1 \ldots x^p\right) = \frac{\sum_{k=1}^{K} p_k \cdot f_k\left(x^1 \ldots x^p, Y^1 \ldots Y^q\right)}{\sum_{k=1}^{K} p_k \cdot f_k\left(x^1 \ldots x^p\right)}$$

En pratique, nous échantillonnons à pas constant les variables à prédire, afin d'obtenir une estimation de la fonction de densité conditionnelle sur une grille régulière. Cette fonction de densité conditionnelle, à $x^1 \ldots x^p$ fixés, est en fait une surface de dimension q, qu'on peut caractériser par différents paramètres :

- son espérance mathématique, ce qui revient à faire de la régression,
- son mode, qui correspond dans ce cas à l'individu $\left(y^1 \ldots y^q\right)$ maximisant la probabilité $P\left(Y^1 \ldots Y^q / x^1 \ldots x^p\right)$ (les variables $Y^1 \ldots Y^q$ étant échantillonnées à pas constant, les probabilités sont proportionnelles aux densités conditionnelles),
- si q=1, les quantiles, intervalles interquartiles... qui fournissent des indications sur la dispersion de la fonction de densité conditionnelle.

Connaissant un individu $\left(x^1 \ldots x^p\right)$, on peut donc caractériser la distribution des variables $Y^1 \ldots Y^q$, et donc obtenir une prédiction de $\left(Y^1 \ldots Y^q\right)$.

Cette méthodologie est celle que nous avons utilisée pour prédire une ou plusieurs propriétés géologiques connaissant les valeurs d'attributs sismiques, et ayant à notre disposition une

population de calibrage composée des valeurs des propriétés géologiques aux puits et des valeurs des attributs sismiques calculées sur les traces adjacentes aux puits. Et du fait de sa forte couverture spatiale, l'information géologique ainsi extraite des traces sismiques est particulièrement intéressante.

CHAPITRE 5

CAS PRATIQUE

La méthodologie de régression non paramétrique, une des deux méthodologies statistiques présentées précédemment, a été appliquée sur les données d'un champ pétrolier, pour calibrer géologiquement les données sismiques aux puits, puis pour prédire les propriétés géologiques connaissant les données sismiques entre les puits.

Dans une première partie, nous présentons donc le champ étudié, ainsi que les données disponibles sur ce champ. Des méthodes statistiques ayant déjà été appliquées sur ce champ pour calibrer géologiquement les données sismiques, nous en fournissons les principales conclusions.

Enfin, dans une seconde partie, nous présentons les résultats de la régression non paramétrique, utilisée pour prédire séparément puis conjointement deux propriétés géologiques.

I - PRÉSENTATION DES DONNÉES

1 Présentation du champ étudié

Le gisement étudié est une structure productrice d'huile, dans des niveaux carbonatés de l'Albien supérieur. Le piégeage est mixte : structural et sédimentaire. La structure correspond à un anticlinal d'origine halocinétique affecté par deux failles majeures respectivement Est/Ouest et Nord-Est/Sud-Ouest (cf. FIG. 21). La faille Est/Ouest délimite l'extension Nord du réservoir qui disparaît par biseau stratigraphique.

L'ensemble réservoir présente une épaisseur d'environ 70 à 100 mètres, les plus fortes épaisseurs étant localisées dans le compartiment affaissé de la faille Nord-Est/Sud-Ouest, suite au jeu synsédimentaire de cet accident.

Ce réservoir, déposé dans un environnement tidal à supratidal, correspond à des intercalations plurimétriques de dolomies, dolomies vacuolaires et grès producteurs d'huile, et de niveaux argilo-dolomitiques qui jouent un rôle de couverture. Il s'agit donc d'un réservoir multicouche.

La répartition tant verticale qu'horizontale des lithofaciès est extrêmement variable : elle est essentiellement contrôlée par le jeu synsédimentaire des deux grands accidents. Le réservoir étudié se caractérise donc par une extrême hétérogénéité qui se répercute sur la distribution des propriétés pétrophysiques (porosités, en particulier). Cependant, des zones homogènes apparaissent sur le champ (cf. FIG. 21). En effet, les plus fortes porosités sont localisées au sommet de la structure, où le réservoir est à dominante de grès et de dolomies vacuolaires associés à des couvertures plutôt argileuses. Au Nord et à l'Est du champ, les porosités sont plus faibles, le réservoir comprenant des réservoirs dolomitiques et des couvertures dolomicritiques et anhydritiques.

La grande hétérogénéité de cet ensemble réservoir explique certaines difficultés apparues au cours du développement du champ : implantation des puits problématique, productions mal

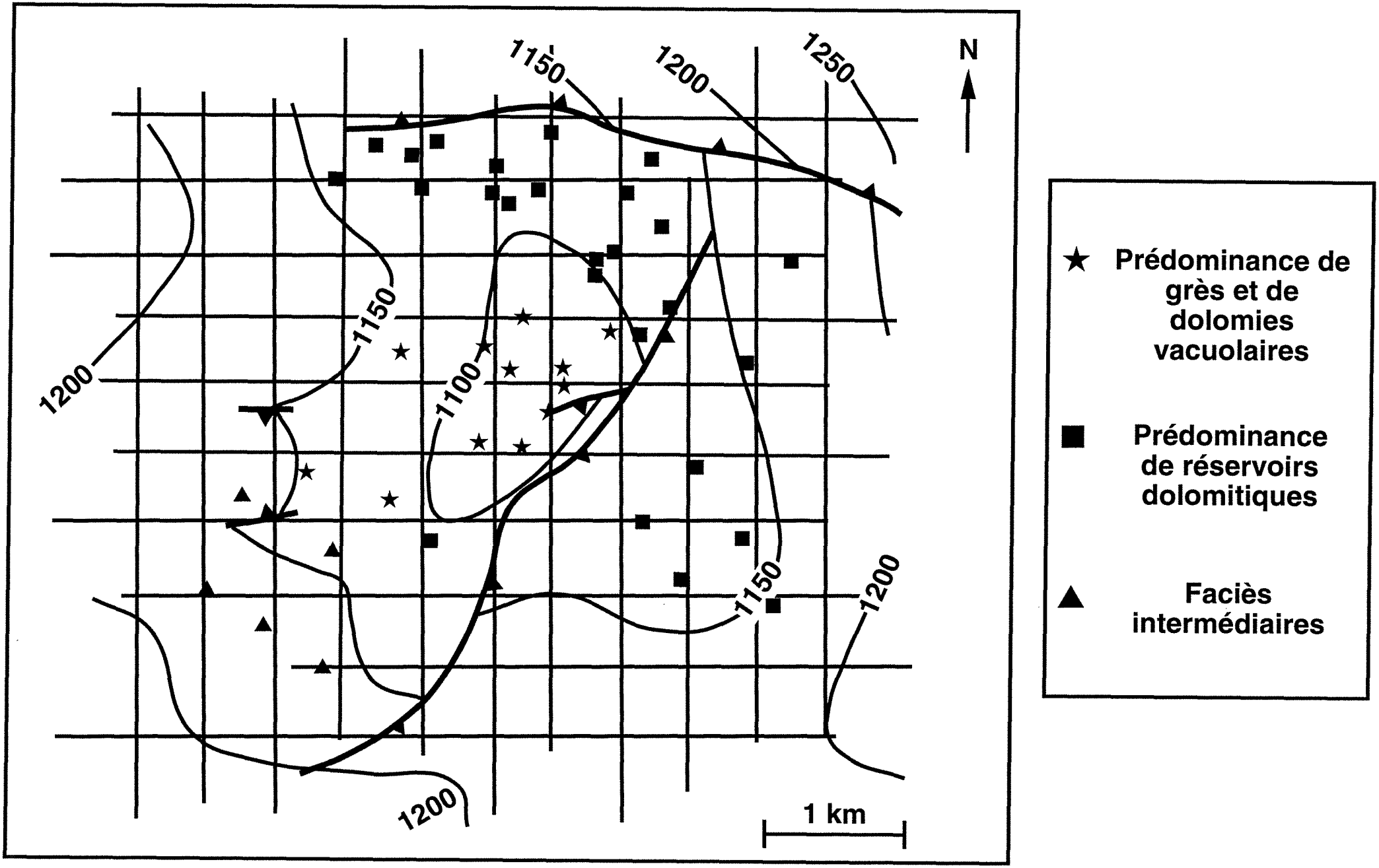


Fig. 21 Carte structurale et principales variations lithologiques du champ étudié

estimées. Toute information géologique en plus de celle connue aux puits (en particulier issue de la sismique) est donc très intéressante, car elle peut permettre de mieux comprendre les variations des caractéristiques du réservoir entre les puits. Il nous semble donc approprié d'appliquer sur ce champ les techniques de calibrage géologique des données sismiques que nous avons développées.

2 Présentation des données disponibles

Quarante-quatre puits ont été forés sur ce champ (cf. FIG. 21), et un jeu complet de mesures diagraphiques y est disponible.

Par ailleurs, trois de ces puits ont été carottés, ce qui a permis de caler l'interprétation des diagraphies, tant du point de vue lithologique que des propriétés pétrophysiques du réservoir. Ainsi, sur l'ensemble des puits, l'interprétation lithologique (basée essentiellement sur les outils gamma-ray, sonic, densité et porosité-neutron) a permis de définir six lithofaciès (cf. FIG. 22) :

- anhydrite,
- couvertures argileuses,
- couvertures dolomitiques (dolomicrites),
- dolomies vacuolaires,
- grès,
- réservoirs dolomitiques.

La distribution spatiale de ces lithofaciès contrôlant les propriétés pétrophysiques du réservoir, c'est une information géologique de ce type qu'il serait intéressant de prédire à partir des données sismiques.

Ces données sismiques correspondent à une acquisition marine 2D à mailles denses, de 22 lignes (cf. FIG. 21), avec un espacement entre lignes d'environ 500 mètres, un intertrace de 17 mètres et un pas d'échantillonnage de 4 ms temps double. La fréquence dominante au niveau du réservoir est d'environ 30 Hz. Le rapport signal/bruit est supérieur à 6dB dans la plage 6-52 Hz.

Le calage entre les traces sismiques synthétiques aux puits et les traces réelles adjacentes a permis l'estimation du signal sismique et l'identification des horizons associés au toit et à la base de l'ensemble réservoir (cf. FIG. 23). Leur pointé a ensuite été effectué sur l'ensemble des profils sismiques. La fenêtre temporelle ainsi définie au niveau du réservoir varie de 28 à 36 ms. Cette

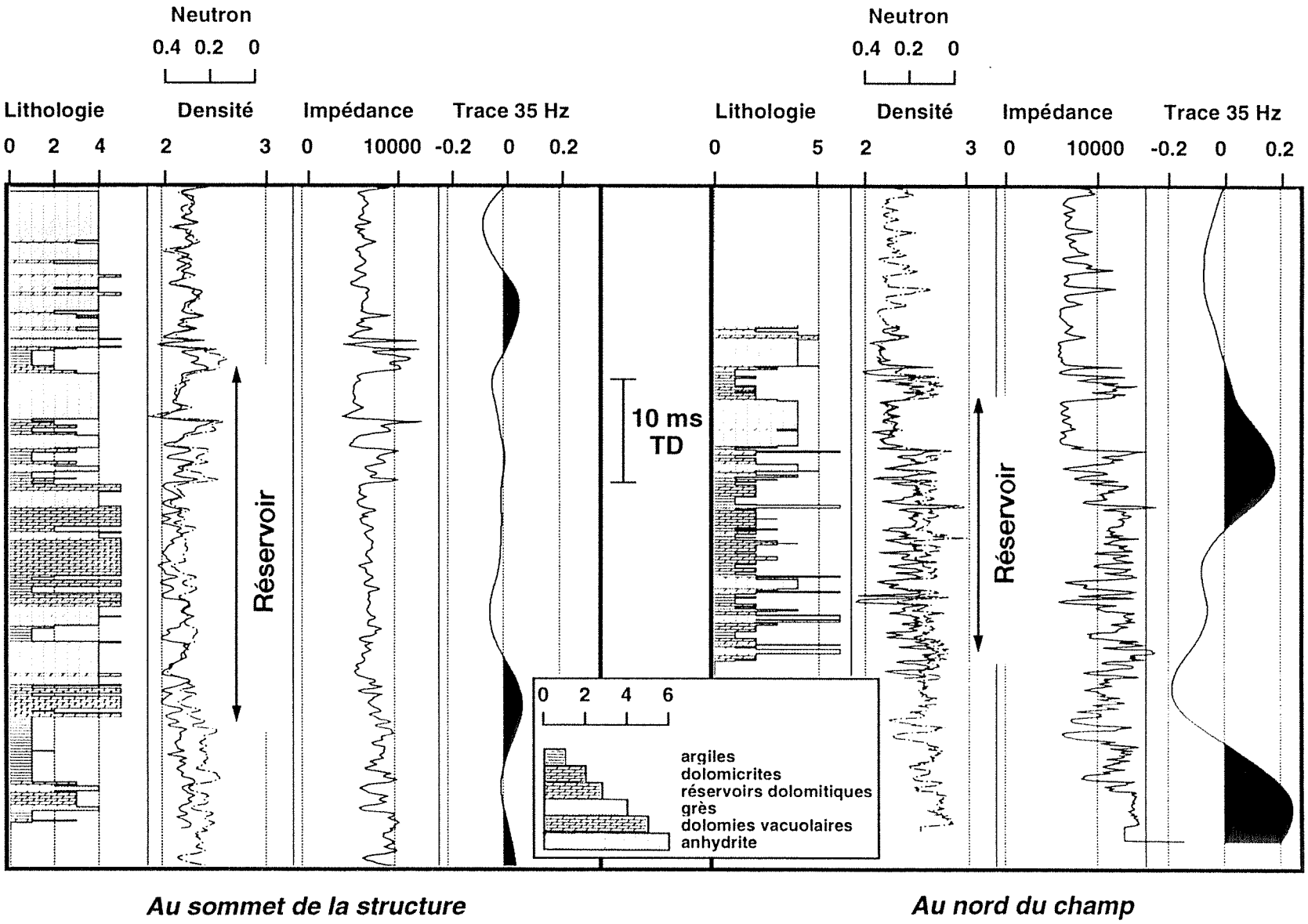


Fig. 22 Lithologies, logs et traces synthétiques associés à deux puits

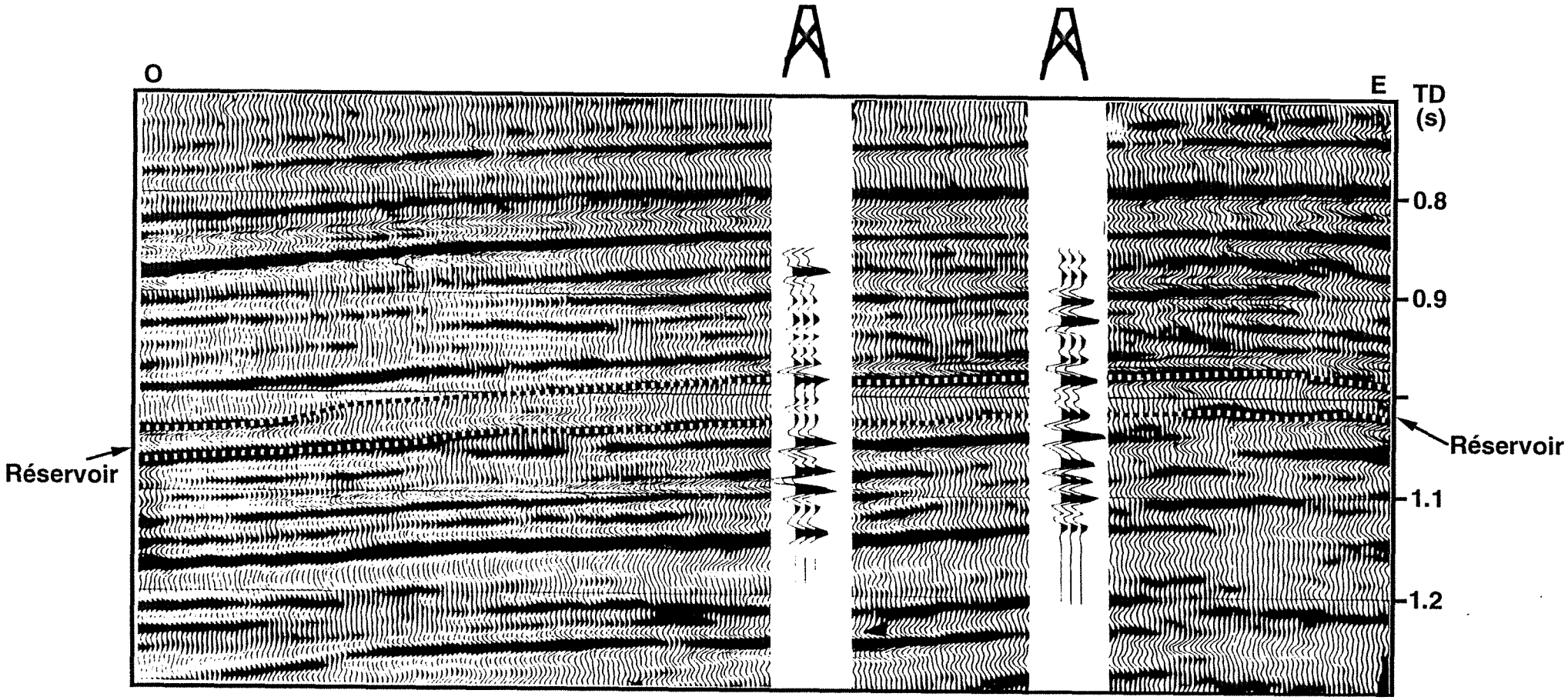


Fig. 23 Définition de l'intervalle réservoir

variation étant de faible amplitude, ce sont les caractéristiques sismiques d'une fenêtre constante de 32 ms, définie à partir du toit du réservoir, que nous avons analysées.

3 Résultats antérieurs

A partir des données sismiques et lithologiques disponibles sur ce champ, une étude de calibrage géologique des données sismiques a déjà été effectuée, la méthode de calibrage retenue étant l'analyse canonique (Fournier et Derain, 1992-a).

Les propriétés géologiques prédites à partir de l'information sismique sont les épaisseurs des six lithofaciès, cumulées sur l'ensemble réservoir. Les attributs sismiques, représentant l'information sismique à calibrer sont les quatre premières composantes d'une analyse en composantes principales, appliquée sur les valeurs des neuf amplitudes décrivant les traces sismiques au niveau de la fenêtre temporelle constante de 32 ms précédemment définie.

La population de calibrage (ensemble des individus pour lesquels sont connues les données géologiques et les données sismiques) a été générée par l'association des puits et de traces adjacentes à ces puits. Sur cette population, une très forte non linéarité entre variables géologiques et sismiques a été mise en évidence. Cette non linéarité est liée à la présence de zones géologiquement différentes sur le champ : ainsi, la zone à dominante gréseuse au sommet de la structure ne présente pas les mêmes relations entre propriétés géologiques et attributs sismiques que la zone à dominante dolomitique située au Nord et à l'Est du champ. Du fait de cette non linéarité, le calibrage a été réalisé indépendamment sur des sous-domaines du champ : la définition spatiale de ces sous-domaines s'appuie sur une reconnaissance des faciès sismiques, puis sur l'association de ces faciès sismiques aux différentes zones géologiques. Cette méthodologie de calibrage a permis, à partir des données sismiques, l'estimation d'épaisseurs de lithofaciès, cumulées sur l'intervalle réservoir, aux puits puis entre les puits.

Dans un second temps, ces estimations et les incertitudes associées ont été intégrées dans des simulations géostatistiques de la distribution spatiale des lithologies (Fournier et Derain, 1992-b) ce qui a mis en évidence l'intérêt de données supplémentaires (même "floues" mais de forte densité spatiale) pour contraindre ces réalisations.

Cependant, ces travaux ont soulevé deux problèmes. D'une part, la recherche préliminaire de sous-domaines permet de contourner le problème de la non linéarité entre propriétés géologiques et

attributs sismiques ; mais certains sous-domaines ne comprennent pas un nombre de puits suffisants pour assurer un calibrage multivariable fiable. D'autre part, les incertitudes associées aux valeurs estimées sont souvent grandes.

Il serait donc intéressant de voir si la méthodologie de calibrage par régression non paramétrique que nous avons développée dans ce travail peut supprimer ces deux problèmes et améliorer les résultats de calibrage. En vue d'une comparaison, nous avons donc appliqué cette méthodologie aux mêmes propriétés géologiques et aux mêmes attributs sismiques que ci-dessus.

II - APPLICATION DE LA MÉTHODOLOGIE DE RÉGRESSION NON PARAMÉTRIQUE

La méthodologie de régression non paramétrique a été appliquée pour le calibrage géologique des données sismiques sur le champ que nous venons de présenter.

Deux propriétés géologiques nous ont paru particulièrement intéressantes à étudier, car elles contrôlent la distribution des porosités sur ce champ. Il s'agit de l'épaisseur cumulée de grès (que nous noterons G_1) et de l'épaisseur cumulée de dolomies vacuolaires (notée G_2).

Quant aux attributs sismiques utilisés, il s'agit des quatre composantes principales ayant déjà servi aux travaux de calibrage antérieurs sur ce champ. Nous noterons ces attributs S_1 à S_4.

La population de calibrage a été définie par l'association de chacun des quarante-quatre puits aux trois traces sismiques les plus proches, d'où une population de calibrage comprenant 132 individus.

La méthodologie de régression non paramétrique (avec utilisation de la décomposition en classes gaussiennes) a été appliquée, afin d'estimer séparément puis conjointement les valeurs des deux propriétés géologiques G_1 et G_2, connaissant les valeurs des attributs sismiques S_1 à S_4.

Dans une **première étape**, nous travaillons sur la population de calibrage. Nous recherchons sur cette population une décomposition en classes gaussiennes de bonne qualité, afin de pouvoir estimer, pour chaque individu de la population, la fonction de densité conditionnelle de la (ou des) propriété(s) géologique(s) considérée(s) connaissant les attributs sismiques S_1 à S_4.

L'espérance mathématique, le mode et l'intervalle interquartile sont calculés sur les fonctions de densité conditionnelle, afin de caractériser la (ou les) propriété(s) géologique(s) pour chaque individu. Les résultats sont analysés et comparés (lorsque c'est possible) à ceux obtenus antérieurement par calibrage sur ce champ (Fournier et Derain, 1992-a, Fournier, 1992). Des tests effectués pour valider la méthodologie sont aussi présentés.

Dans une **deuxième étape**, nous calculons la fonction de densité conditionnelle en chaque trace sismique du champ étudié, et fournissons donc entre les puits une prédiction de la (ou des) propriété(s) géologique(s) étudiée(s). Présentons maintenant ces résultats.

1 Prédiction de l'épaisseur cumulée de grès

Les résultats de cette partie ont fait l'objet d'une communication au congrès mondial de la Society of Exploration Geophysicists (Joseph *et al.*, 1993). Nous en fournissons l'abstract en Annexe C.

Dans cette partie, nous avons cherché à prédire l'épaisseur cumulée de grès (propriété géologique G_1) en fonction des quatre attributs sismiques S_1 à S_4. La population de calibrage étudiée comporte donc 132 individus définis dans un espace de dimension 5.

Sur cette population de calibrage, les caractéristiques de l'épaisseur cumulée de grès sont les suivantes. L'épaisseur moyenne de grès est de 9 mètres, les épaisseurs minimale et maximale étant respectivement de 1.8 et 19.7 mètres. Enfin, 25% des puits présentent une épaisseur cumulée de grès inférieure à 5.7 mètres ; et pour 75% des puits, l'épaisseur cumulée de grès est inférieure à 11.8 mètres.

Nous allons maintenant présenter les résultats fournis par la méthodologie de régression non paramétrique, utilisée pour prédire l'épaisseur cumulée de grès au niveau du réservoir.

1.1 Choix d'une décomposition en classes gaussiennes

Nous avons appliqué la méthodologie de décomposition en classes gaussiennes sur la population de calibrage, en faisant varier le paramètre PMIN et en choisissant différentes initialisations. Nous avons obtenu 29 solutions différentes (sur 44 essais), comportant de 1 à 7 classes : la stabilité de la méthode de décomposition est donc fortement affectée par la diminution du nombre d'individus de la population, conjointe à l'augmentation de la dimension de l'espace, par rapport aux tests effectués sur données synthétiques.

Les critères de qualité de ces 29 solutions varient de 0.52 à 1.65 pour le critère de l'erreur quadratique C_2 (avec une moyenne de 1.2), et de 4.11 à 6.21 pour le critère de l'erreur en valeur absolue C_1 (avec une moyenne de 4.98). Nous fournissons dans le tableau 6 ci-dessous les caractéristiques (nombre de classes, poids de la plus grande et de la plus petite classe, valeurs des critères C_1 et C_2) des 7 meilleures solutions notées ***Sol1*** à ***Sol7***, de la moins bonne solution notée ***Sol29*** et d'une solution intermédiaire notée ***Sol10***. Les résultats complets sont fournis en Annexe D (tableaux D-1 à D-4).

Compte tenu de la dimension de l'espace de calibrage, nous ne conservons pas les solutions ***Sol1*** à ***Sol5*** : le poids de la plus petite classe de chacune de ces solutions nous semble trop faible pour permettre une estimation correcte de ses paramètres (moyenne et matrice de variance-covariance). En ce qui concerne la solution ***Sol7***, elle est de meilleure qualité que la solution ***Sol6*** en terme de critère C_1, mais par contre sa valeur du critère C_2 est très dégradée par

rapport à celle de ***Sol6***. Nous rejetons donc la solution ***Sol7***. De toutes les solutions restantes (par exemple ***Sol10***), la solution ***Sol6*** présente les meilleures valeurs pour les critères C_1 et C_2. Nous retenons finalement cette solution à 5 classes gaussiennes, dont nous fournissons en Annexe D les paramètres (poids, moyenne et écart-type des classes).

Tableau 6 *Caractéristiques des décompositions obtenues*

	Nombre de classes	Poids des classes minimal	Poids des classes maximal	Critère C_1	Critère C_2
Sol1	7	0.11	0.17	4.24	0.52
Sol2	6	0.13	0.23	4.18	0.60
Sol3	6	0.11	0.22	4.12	0.86
Sol4	6	0.14	0.24	4.35	0.69
Sol5	4	0.13	0.39	4.21	0.90
Sol6	5	0.16	0.27	4.39	0.76
Sol7	5	0.16	0.25	4.32	0.96
Sol10	4	0.18	0.32	4.91	0.89
Sol29	4	0.16	0.36	6.21	1.65

La figure 24 représente la décomposition ***Sol6*** dans certains plans de l'espace de calibrage. Les individus sont attribués à la classe gaussienne qui maximise leur probabilité d'appartenance, et codés en couleur en fonction de cette appartenance.

Une autre façon de représenter la décomposition ***Sol6*** consiste, pour chacune des variables S_1, S_2, S_3, S_4 et G_1, à estimer sa fonction de densité marginale à partir de la décomposition en classes gaussiennes, puis à la comparer à son histogramme. Ainsi, les figures 25a à 25e présentent la superposition de la fonction de densité marginale (approximée par la décomposition gaussienne) et de l'histogramme pour chacune des variables. Les fonctions de densité marginale des cinq classes gaussiennes de ***Sol6***, pondérées par la proportion d'individus de la classe, y sont aussi représentées. On constate pour la plupart des variables qu'il y a une très bonne adéquation entre l'histogramme et la fonction de densité marginale approximée : entre autres, les modes de ces deux fonctions sont identiques.

La décomposition en classes gaussiennes ***Sol6*** étant de qualité satisfaisante, nous l'avons donc utilisée dans le cadre de la régression non paramétrique.

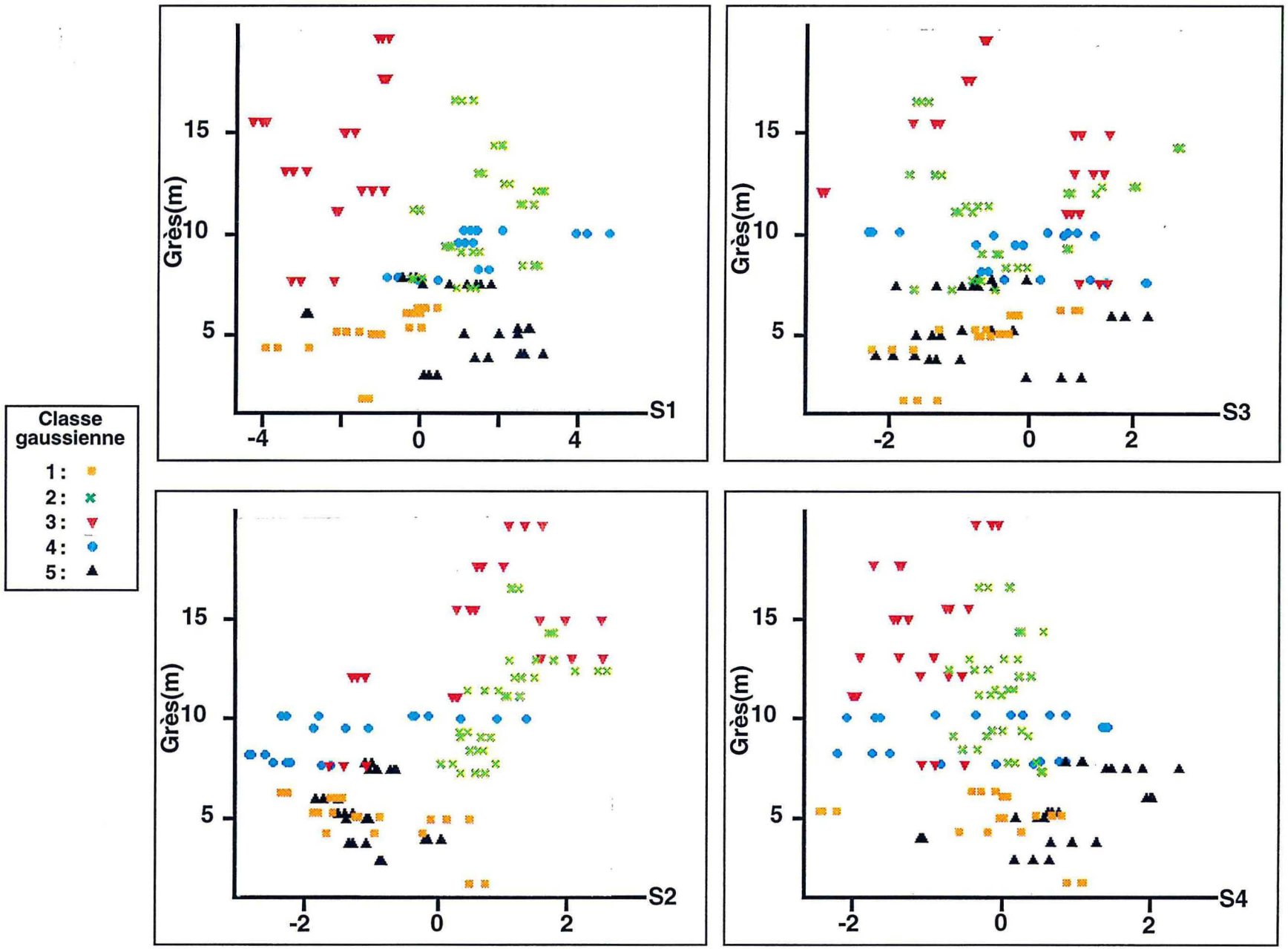


Fig. 24 Représentation de la décomposition Sol6 dans l'espace de calibrage

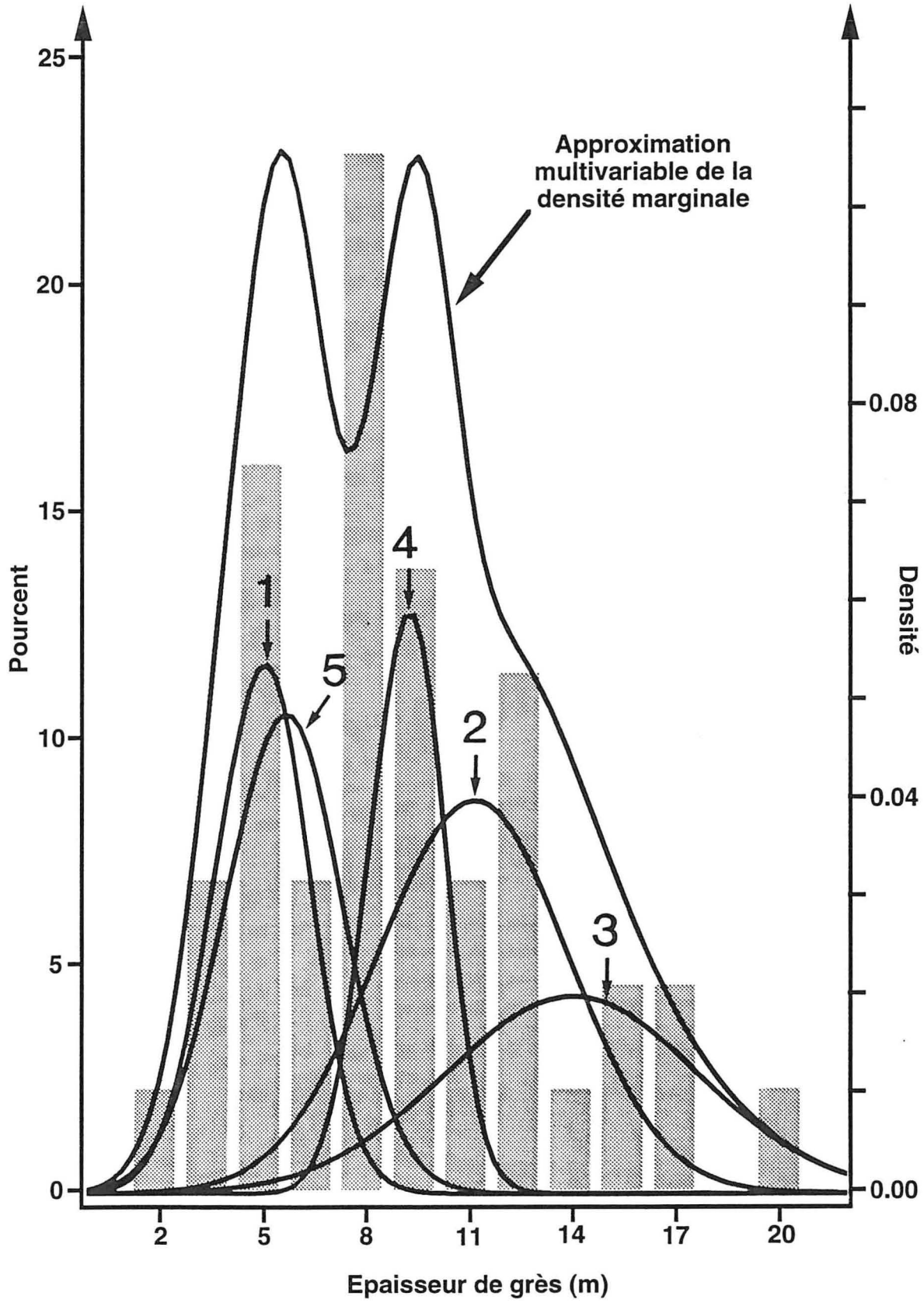


Fig. 25a Fonction de densité marginale de l'épaisseur de grès

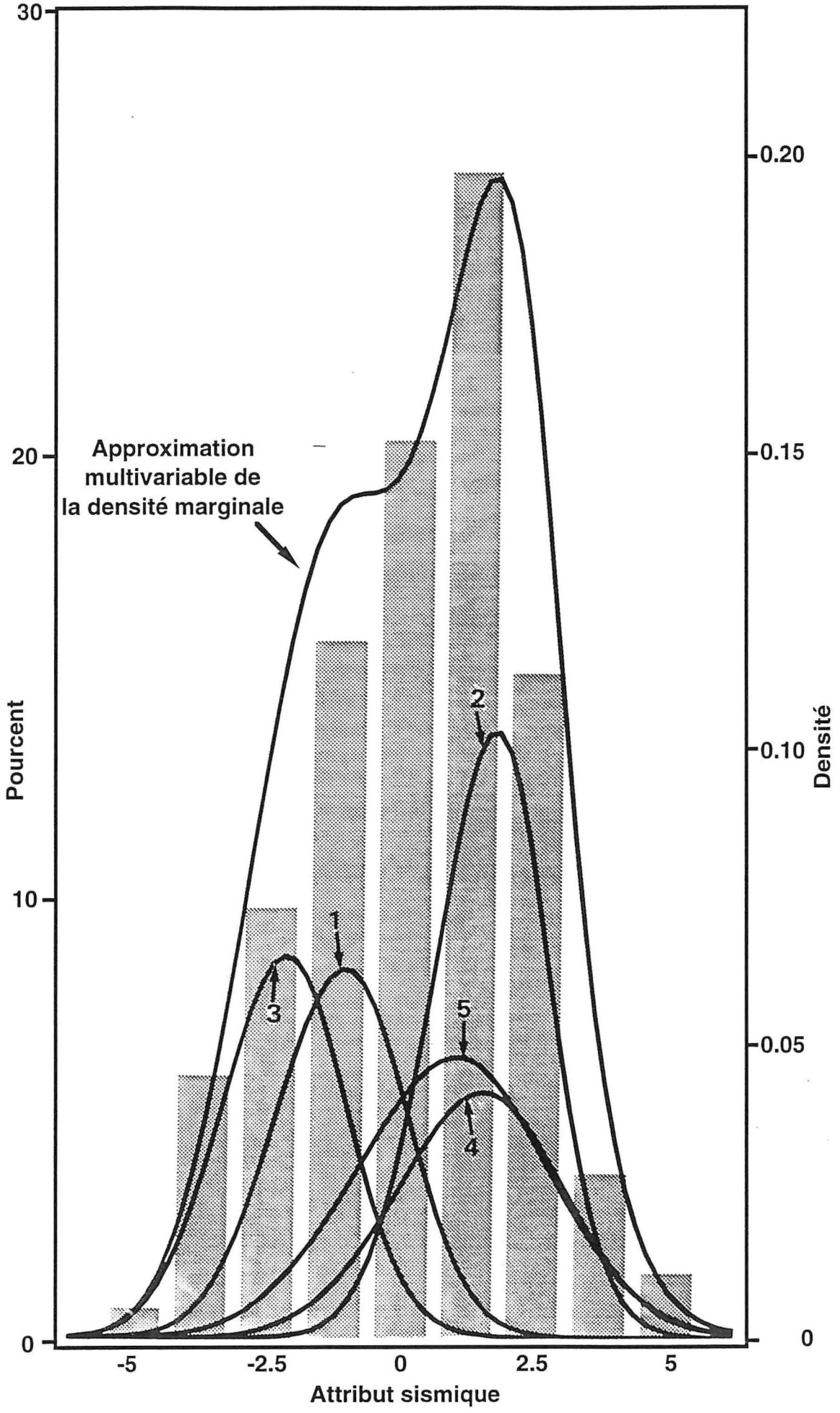


Fig. 25b Fonction de densité marginale du premier attribut sismique

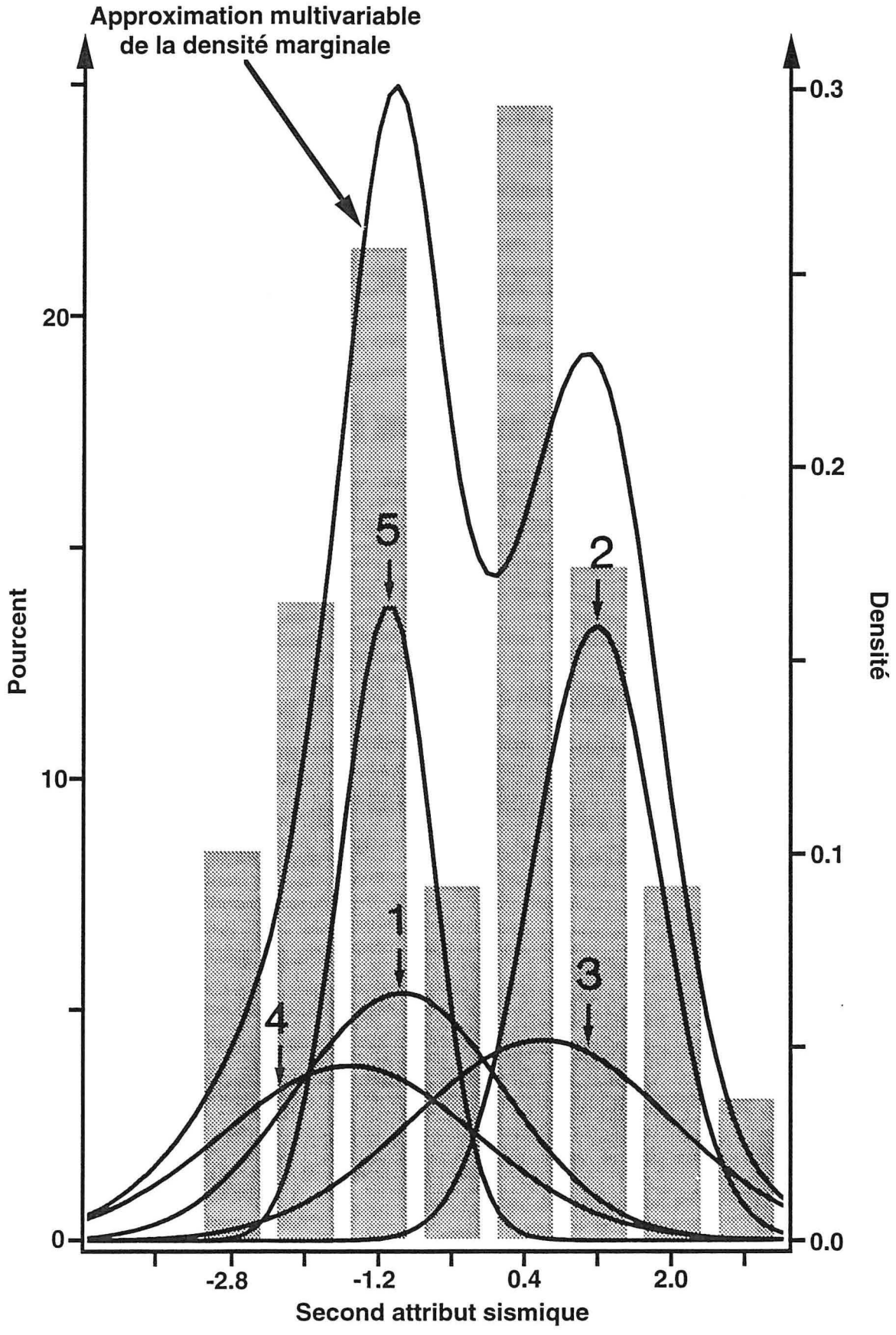


Fig. 25c Fonction de densité marginale du second attribut sismique

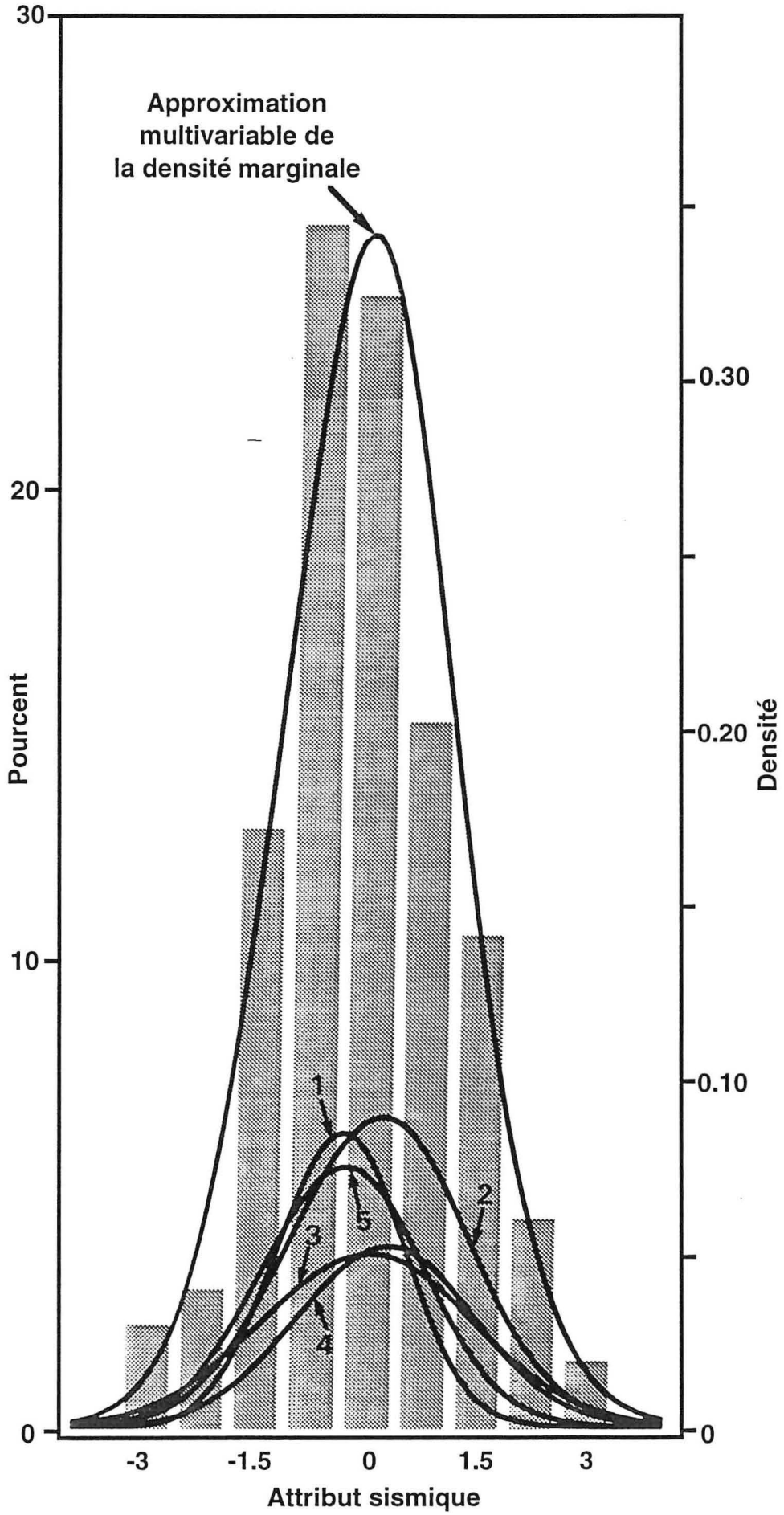


Fig. 25d Fonction de densité marginale du troisième attribut sismique

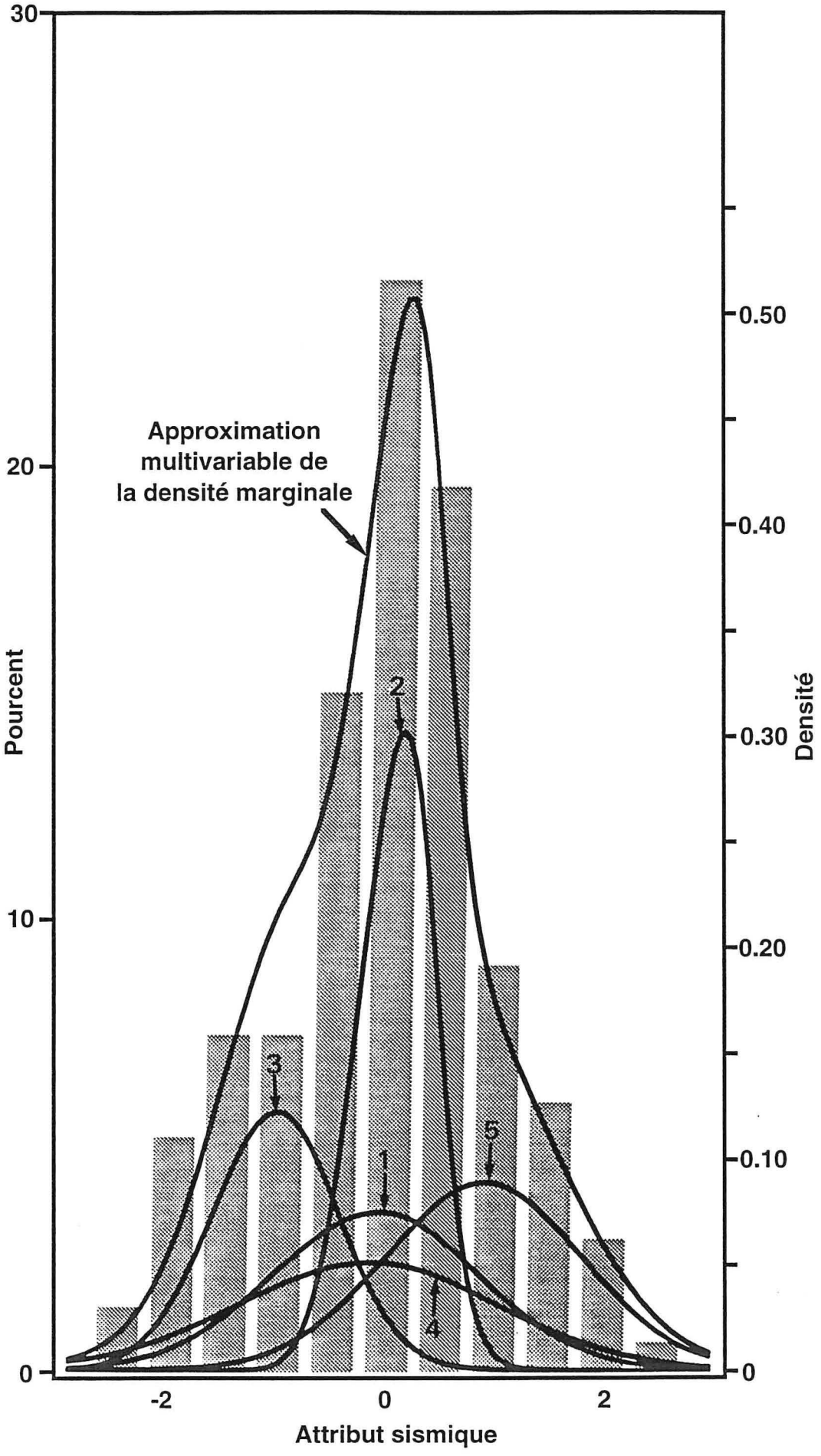


Fig. 25e Fonction de densité marginale du quatrième attribut sismique

1.2 Application de la régression non paramétrique à la décomposition retenue

1.2.1 Prédiction de l'épaisseur cumulée de grès aux puits

Nous avons calculé la fonction de densité conditionnelle de l'épaisseur cumulée de grès pour chaque individu de la population de calibrage (donc en chaque puits et pour chacune de ses trois traces sismiques adjacentes).

Les figures 26, 27 et 28 fournissent les fonctions de densité conditionnelle associées respectivement à trois puits nommés A, B et C. L'épaisseur réelle de grès à ces puits y est aussi indiquée.

Pour le puits A (cf. FIG. 26), les fonctions de densité conditionnelle associées à chacune de ses trois traces sismiques adjacentes sont presque identiques et très fortement unimodales. De plus, le mode de ces fonctions est très proche de l'épaisseur réelle de grès à ce puits : la différence est d'environ 0.3 mètre. Il en va de même pour l'espérance mathématique. Considérons enfin l'intervalle interquartile qui est la différence entre le troisième et le premier quartile : il permet de quantifier la dispersion de la distribution d'une variable quelconque. Pour le puits A, on constate que cette dispersion est très faible (inférieure au mètre).

Pour le puits B (cf. FIG. 27), les fonctions de densité conditionnelle sont aussi unimodales. Le mode et l'espérance mathématique de ces fonctions sont encore proches de l'épaisseur réelle de grès à ce puits (différence inférieure à 0.5 mètre). Par contre, la dispersion est beaucoup plus forte : en effet, l'intervalle interquartile est d'environ 3 mètres.

Pour le puits C (cf. FIG. 28), les fonctions de densité conditionnelle sont nettement bimodales, le second mode étant plus ou moins marqué suivant les traces. De ce fait, l'espérance mathématique de ces fonctions est très éloignée de l'épaisseur réelle de grès à ce puits : la différence est comprise entre 2.5 et 5 mètres. En ce qui concerne le mode, il est presque identique à l'épaisseur réelle de grès pour les deux traces les plus proches du puits. Par contre, le mode fourni par la troisième trace est à plus de 6.5 mètres de l'épaisseur réelle de grès. Enfin, la dispersion (moyennée sur les trois traces) est très forte, d'environ 7 mètres.

En définitive, nous constatons que les fonctions de densité conditionnelle de l'épaisseur de grès sont très variables d'un puits à l'autre.

Les fonctions de densité conditionnelle ayant été calculées pour l'ensemble des 132 individus de la population de calibrage, nous avons décidé de prédire l'épaisseur cumulée de grès, soit par le mode, soit par l'espérance mathématique de ces fonctions.

La figure 29 présente le graphe des prédictions par le mode en fonction des épaisseurs réelles de grès, pour tous les individus. Cette figure permet de constater que la prédiction par le mode est

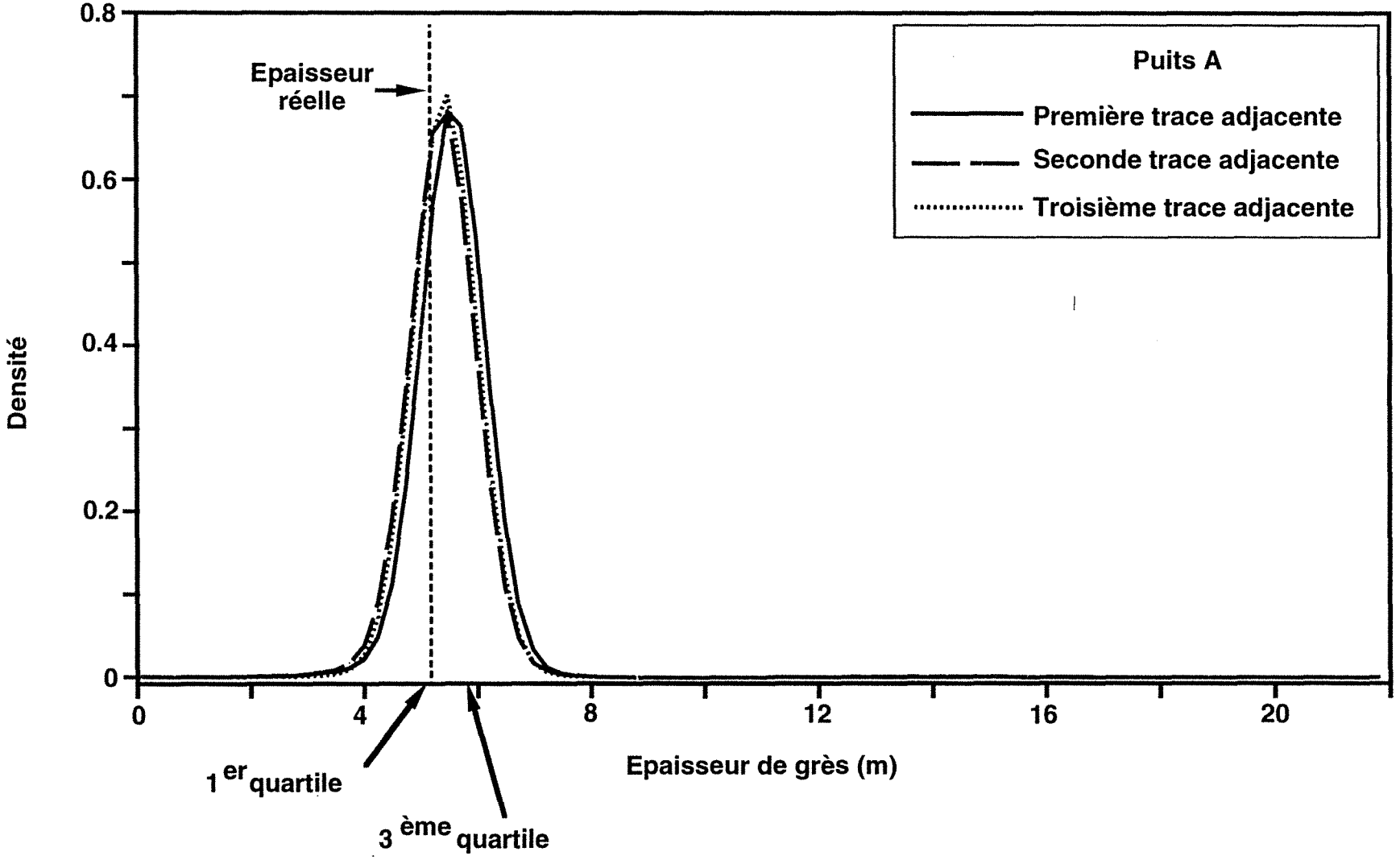


Fig. 26 Fonction de densité conditionnelle de l'épaisseur de grès au puits A

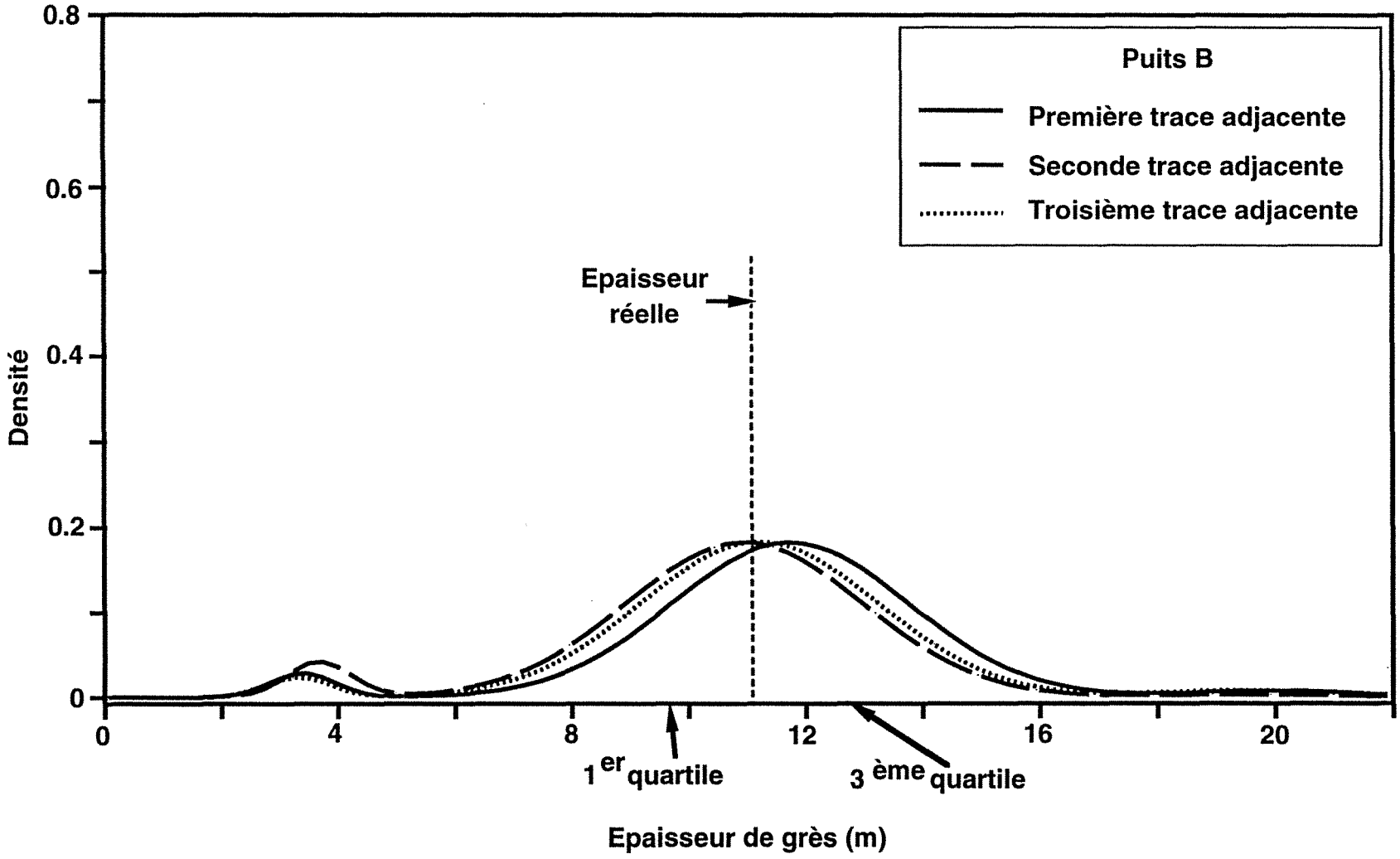


Fig. 27 Fonction de densité conditionnelle de l'épaisseur de grès au puits B

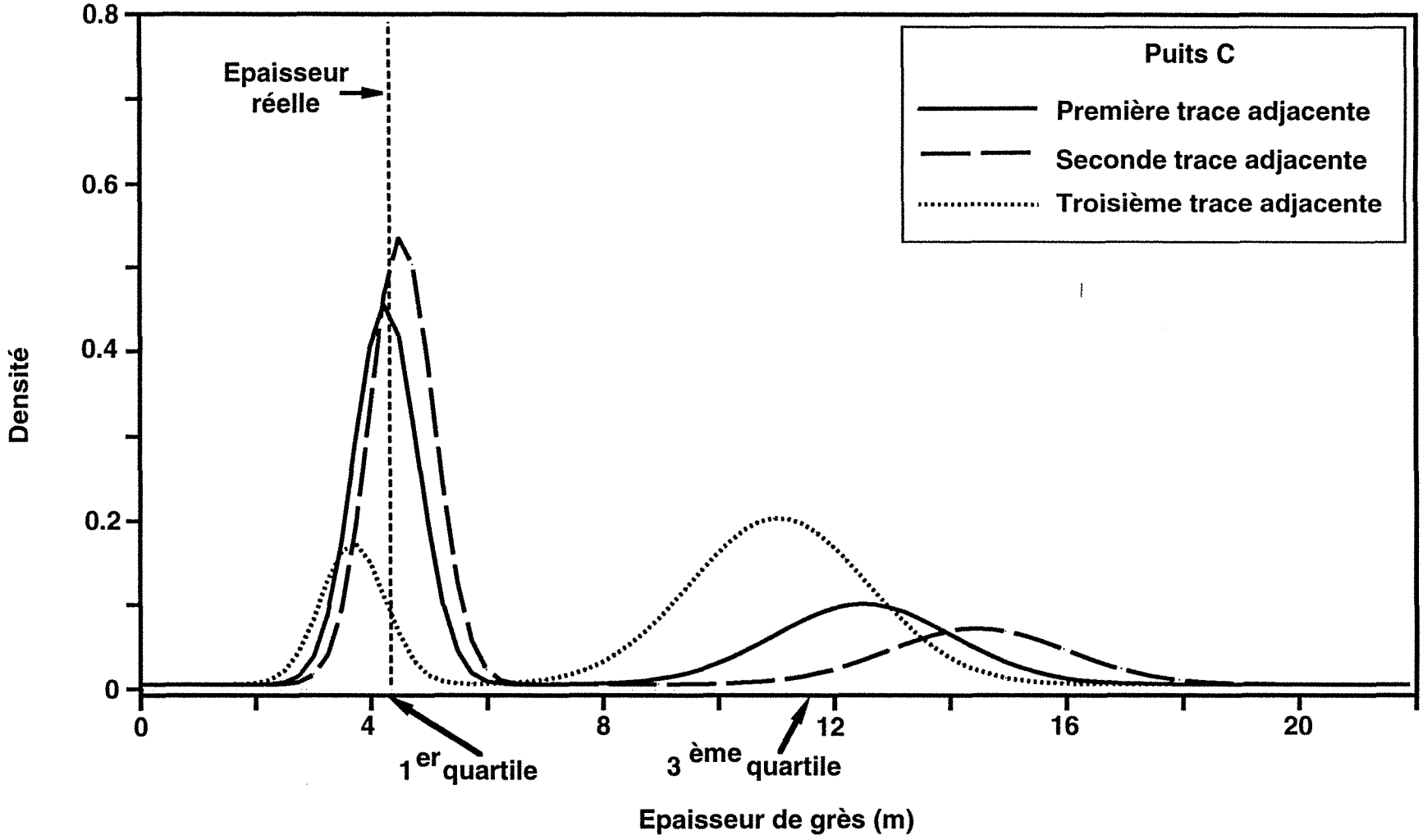


Fig. 28 Fonction de densité conditionnelle de l'épaisseur de grès au puits C

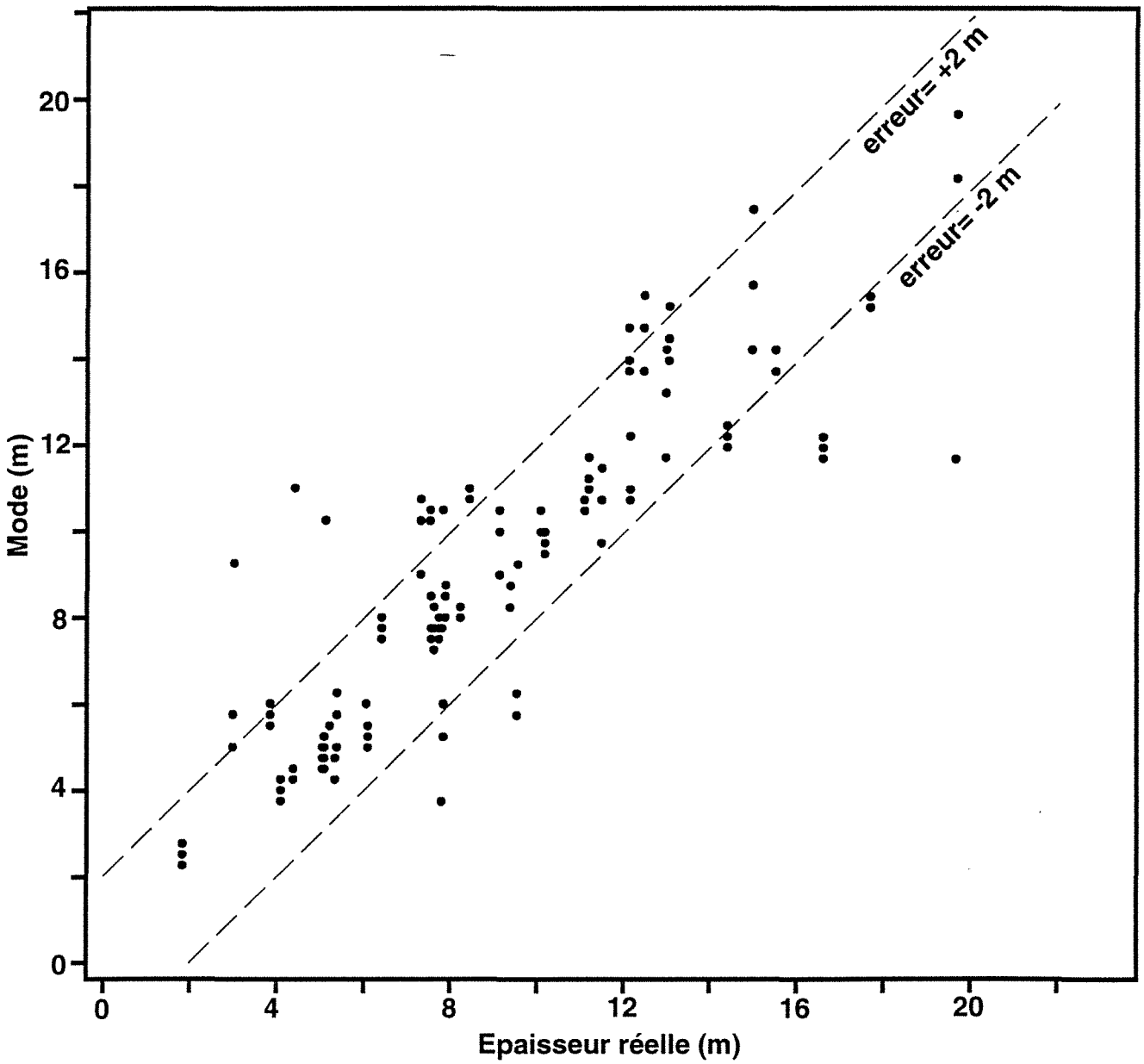


Fig. 29 Prédiction par le mode de l'épaisseur de grès aux puits

globalement satisfaisante, la majorité des modes se trouvant dans un intervalle du type [épaisseur réelle - 2 mètres , épaisseur réelle + 2 mètres]. Ceci est confirmé par la figure 30, qui représente l'histogramme des erreurs de prédiction faites en prenant le mode, codé par la valeur de la densité conditionnelle au mode. On constate que pour environ 75% des individus, l'erreur de prédiction est inférieure à 2 mètres. D'autre part, il est intéressant de constater que les modes associés à de fortes valeurs de la densité conditionnelle (≥ 0.7) correspondent toujours à des erreurs de prédiction inférieures à 2 mètres.

Considérons maintenant les résultats de la prédiction de l'épaisseur de grès par l'espérance mathématique des fonctions de densité conditionnelle. D'après la figure 31 qui fournit le graphe des valeurs prédites en fonction des épaisseurs réelles de grès, nous constatons que les résultats sont assez similaires à ceux obtenus en considérant le mode. Toutefois, d'après l'histogramme des erreurs de prédiction représenté sur la figure 32, nous constatons que les erreurs de prédiction sont moins fortes. En effet, elles n'excèdent pas 6 mètres, alors qu'en considérant le mode, les erreurs de prédiction pouvaient atteindre près de 8 mètres (cf. Tableau 7).

Tableau 7 *Comparaison des erreurs de prédiction en prenant le mode ou l'espérance mathématique*

	Erreur		Erreur maximale		% d'individus avec une erreur ≥ 2 m
	moyenne	quadratique	négative	positive	
Mode	1.33	3.84	-6.65	7.90	25%
Espérance mathématique	1.34	3.37	-5.00	5.90	27%

En fait, il nous paraît plus approprié d'utiliser le mode comme valeur prédite, en lui associant la probabilité correspondante. En effet, si l'espérance mathématique prend en compte la globalité de la distribution des épaisseurs prédites, le mode correspond à une valeur précise de cette distribution dont on connaît la probabilité associée, ce qui est très important. Par la suite, nous considérerons donc toujours les prédictions obtenues en prenant le mode de la fonction de densité conditionnelle.

Afin de quantifier la dispersion de la distribution de l'épaisseur de grès, nous avons représenté l'histogramme des intervalles interquartiles aux puits (cf. FIG. 33). Nous constatons que pour plus de 80% des puits, cet intervalle est inférieur à 3 mètres ce qui est faible.

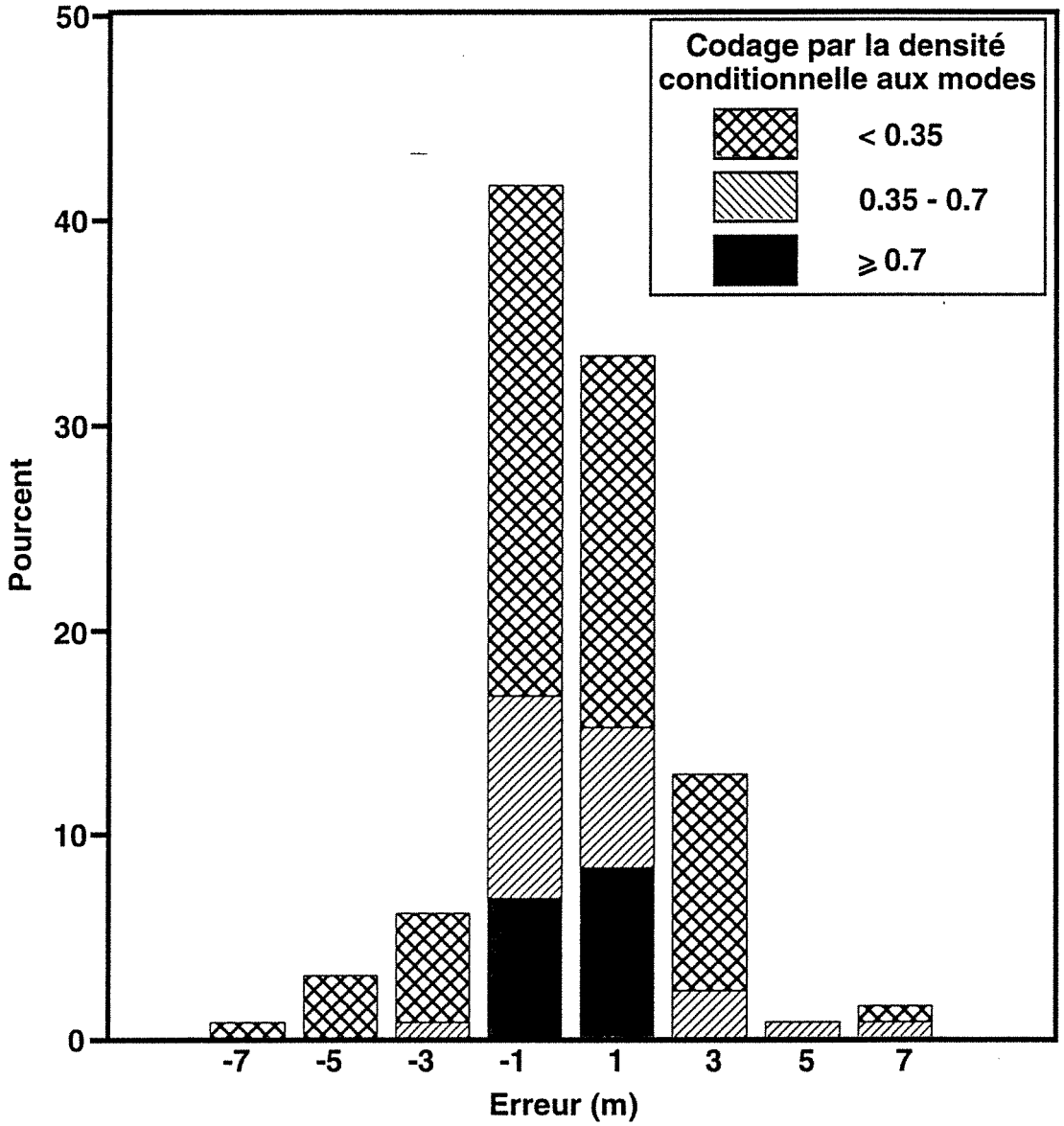


Fig. 30 Histogramme des erreurs de prédiction de l'épaisseur de grès aux puits par le mode

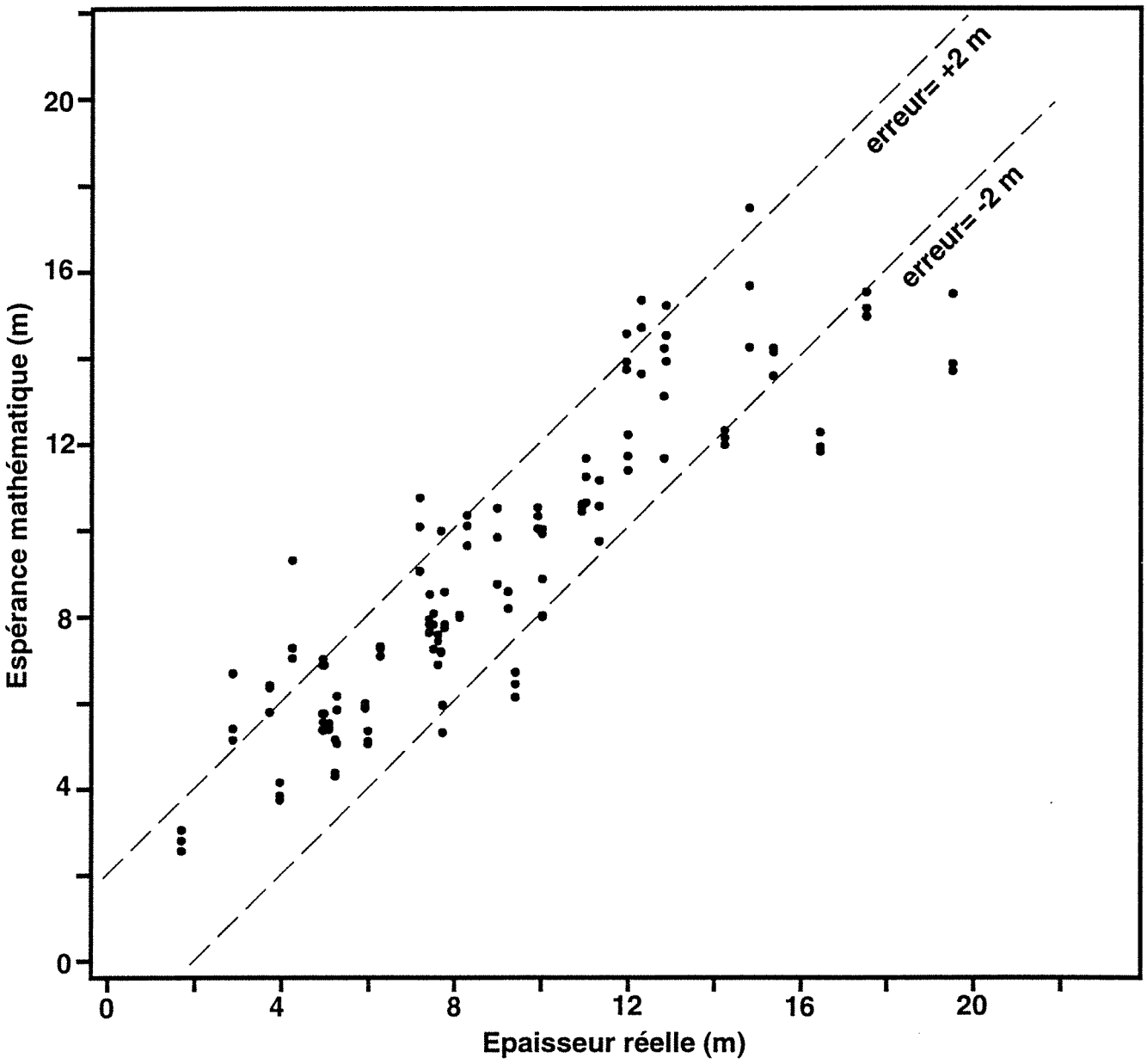


Fig. 31 Prédiction par l'espérance mathématique de l'épaisseur de grès aux puits

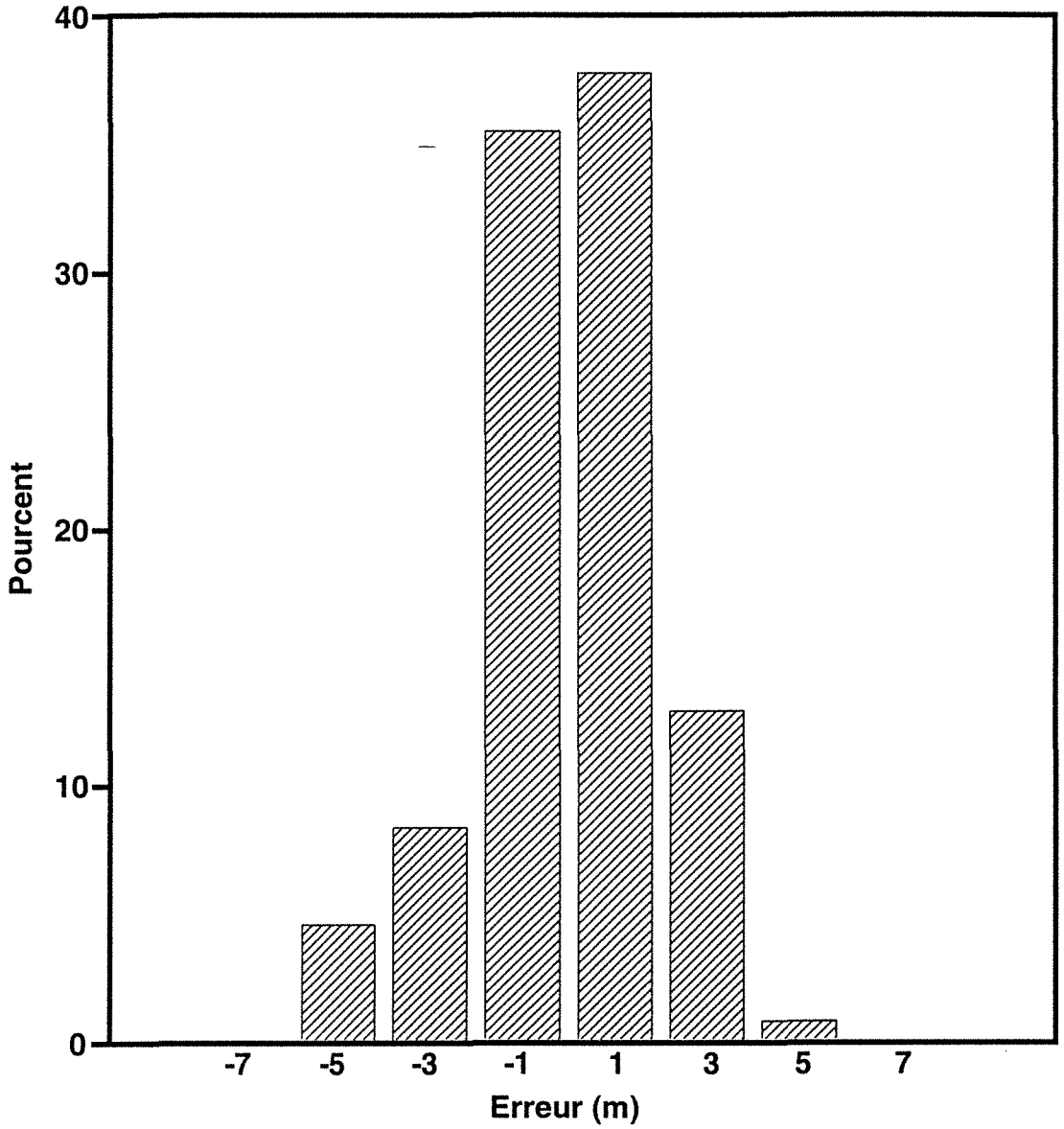


Fig. 32 Histogramme des erreurs de prédiction de l'épaisseur de grès aux puits par l'espérance mathématique

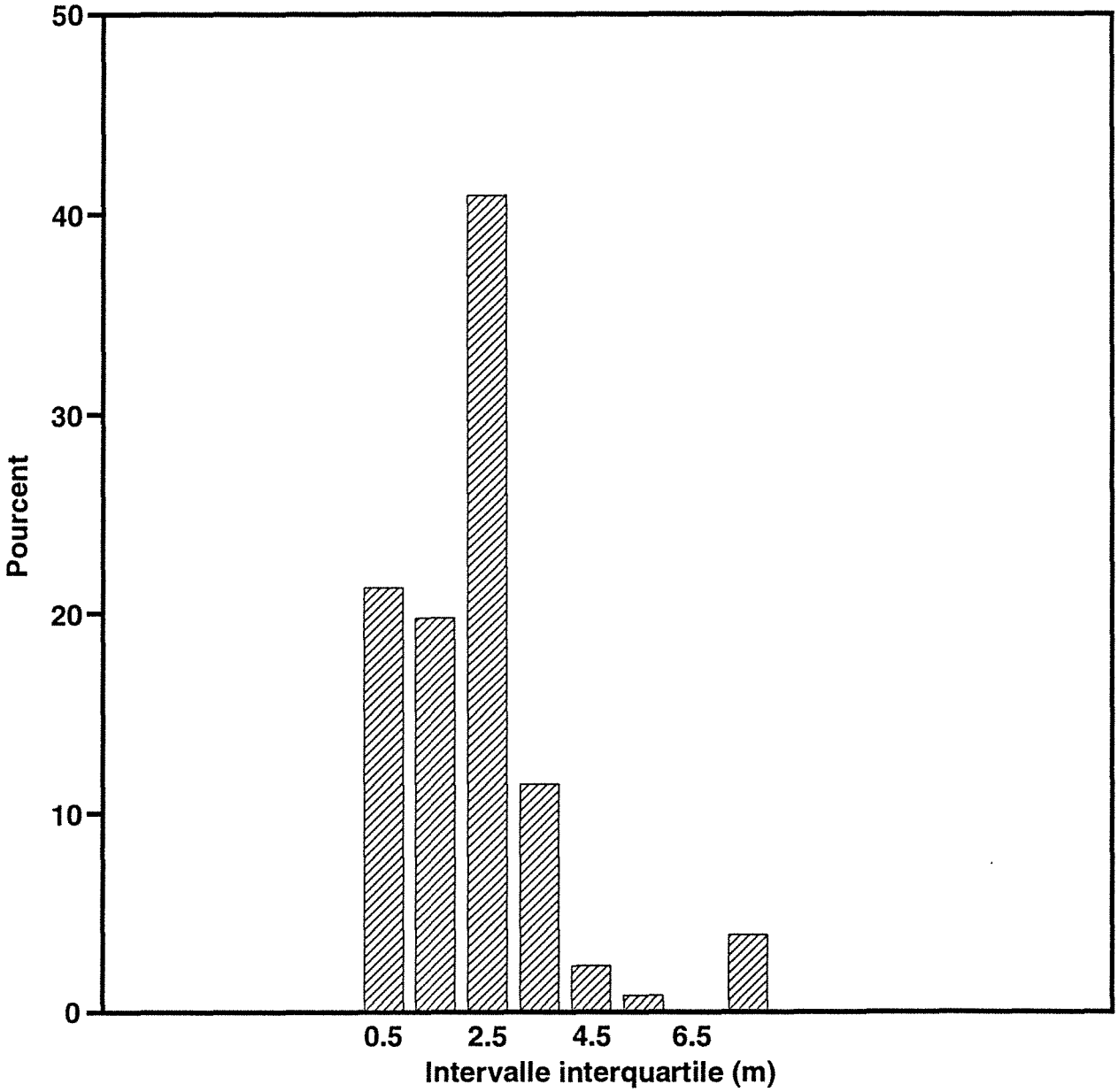


Fig. 33 Histogramme des intervalles interquartiles aux puits

En conclusion, nous avons pu, à partir d'une décomposition en classes gaussiennes de bonne qualité (***Sol6***), prédire de façon satisfaisante les épaisseurs de grès aux puits par régression non paramétrique. Toutefois, afin de valider cette méthodologie, nous avons effectué des blind-tests dont nous présentons ci-dessous les résultats.

1.2.2 Validation de la méthodologie par la méthode des blind-tests

La méthode des blind-tests consiste à supprimer un ou plusieurs puits de la population de calibrage (et donc du processus de calibrage), puis à prédire l'épaisseur de grès par régression non paramétrique connaissant les traces sismiques associées à ce (ou ces) puits. On peut alors comparer les valeurs ainsi prédites aux épaisseurs réelles de grès aux puits.

Nous avons effectué trois tests. Pour le premier test, nous avons retiré trois puits, dont deux puits situés au sommet de la structure mais présentant de faibles épaisseurs cumulées de grès. La population de calibrage restante comprend donc 123 individus. Pour les deuxième et troisième tests, nous avons retiré quatre puits représentatifs des différentes zones géologiques du champ. Les populations de calibrage restantes comprennent donc 120 individus.

Nous synthétisons les résultats obtenus en présentant le graphe des épaisseurs de grès prédites par le mode en fonction des épaisseurs réelles de grès, pour les puits des trois blind-tests (cf. FIG. 34). Et nous détaillons les résultats en fournissant les erreurs de prédiction dans le tableau 8 ci-dessous.

Tableau 8 *Erreurs de prédiction pour les puits des blind-tests*

	Erreur		Erreur maximale	
	moyenne	quadratique	négative	positive
Blind-test 1	1.75	4.66	-3.40	0.37
Blind-test 2	1.48	2.67	-2.55	0.00
Blind-test 3	2.33	6.67	-0.69	4.33

Nous constatons que, sur l'ensemble des blind-tests, les erreurs de prédiction sont inférieures à 2 mètres pour huit des onze puits, et n'excèdent jamais 4.5 mètres.

Par ailleurs, nous avons vérifié que ces erreurs de prédiction sont du même ordre de grandeur que celles obtenues lorsque ces puits font partie du processus de calibrage : le fait de ne pas faire intervenir ces puits pendant le processus de calibrage n'a que peu modifié les valeurs de

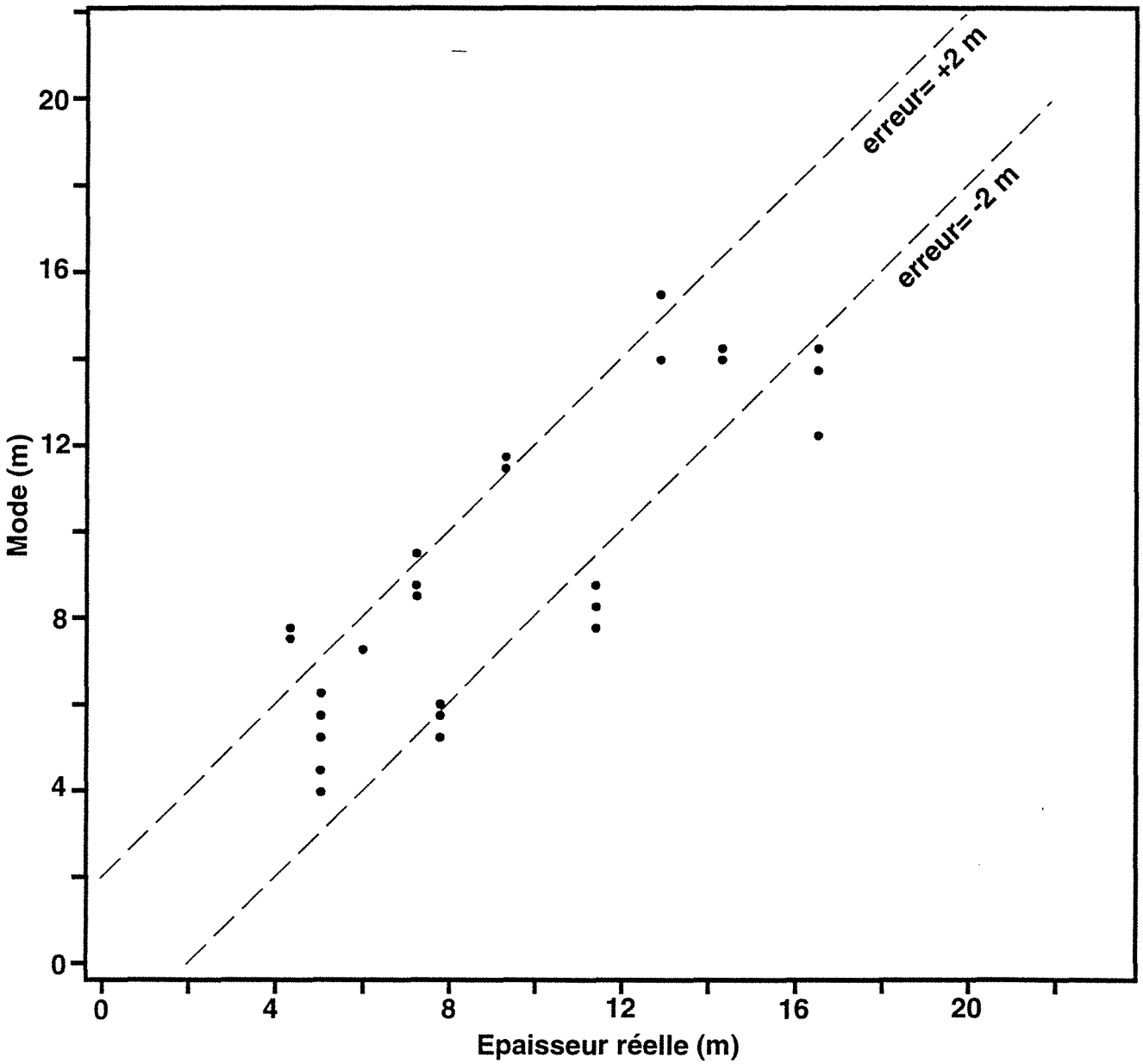


Fig. 34 Prédiction par le mode de l'épaisseur de grès aux puits
Puits des blind-tests

grès prédites. Donc, bien que les puits participant aux blind-tests soient représentatifs de zones géologiques diverses, aucun n'influence trop la stabilité de la décomposition, au risque d'entraîner une instabilité s'il est supprimé.

En conclusion, le calibrage proposé nous semble fiable pour prédire l'épaisseur cumulée de grès à partir de l'information sismique.

1.2.3 Prédiction de l'épaisseur cumulée de grès entre les puits

Le calibrage ayant été validé par des blind-tests, nous avons décidé de l'appliquer en toute trace sismique du champ étudié. Nous avons calculé en chaque trace les valeurs des attributs sismiques S_1 à S_4, et avons estimé la fonction de densité conditionnelle de l'épaisseur de grès connaissant ces valeurs des attributs.

La figure 35 fournit les épaisseurs de grès prédites par le mode, représentées sur une carte spatiale du champ. Nous constatons que les fortes épaisseurs de grès ($\geq$ 10 mètres), codées en rouge et en orange, apparaissent au sommet de la structure mais aussi dans la zone Sud-Est du champ ; en fait, dans la zone Sud-Est, le réservoir a une épaisseur d'environ 100 mètres contre 70 mètres sur le reste du champ, ce qui explique ces fortes épaisseurs de grès prédites. En ce qui concerne les très faibles épaisseurs de grès ($<$ 6 mètres) codées en noir, elles sont situées dans la zone Ouest du champ. Et les zones Nord, Nord-Est et Sud-Ouest présentent des épaisseurs cumulées de grès intermédiaires.

Les intervalles interquartiles ont aussi été calculés sur les fonctions de densité conditionnelle. Rappelons que les intervalles interquartiles décrivent la dispersion de la distribution des épaisseurs de grès, et donc l'incertitude associée à la prédiction des épaisseurs de grès par le mode. D'après la figure 36 fournissant la distribution spatiale de l'intervalle interquartile, on constate que celui-ci est majoritairement inférieur à 2.5 mètres (codage en bleu). D'autre part, s'il est possible d'identifier la zone Sud-Est du champ pour laquelle l'intervalle interquartile est plutôt compris entre 2.5 et 5 mètres (codage en orange), les fortes valeurs de l'intervalle interquartile ($\geq$ 5 mètres, codées en rouge) sont quant à elles réparties de façon aléatoire sur l'ensemble du champ, ce qui est intéressant.

En conclusion, en utilisant une décomposition en classes gaussiennes obtenue sur la population de calibrage, il nous a été possible de prédire en toute trace sismique l'épaisseur cumulée de grès, et d'estimer ainsi la distribution spatiale de cette épaisseur. Nous avons donc pu obtenir de l'information géologique avec une forte couverture spatiale, contrairement à l'information fournie par les données de puits.

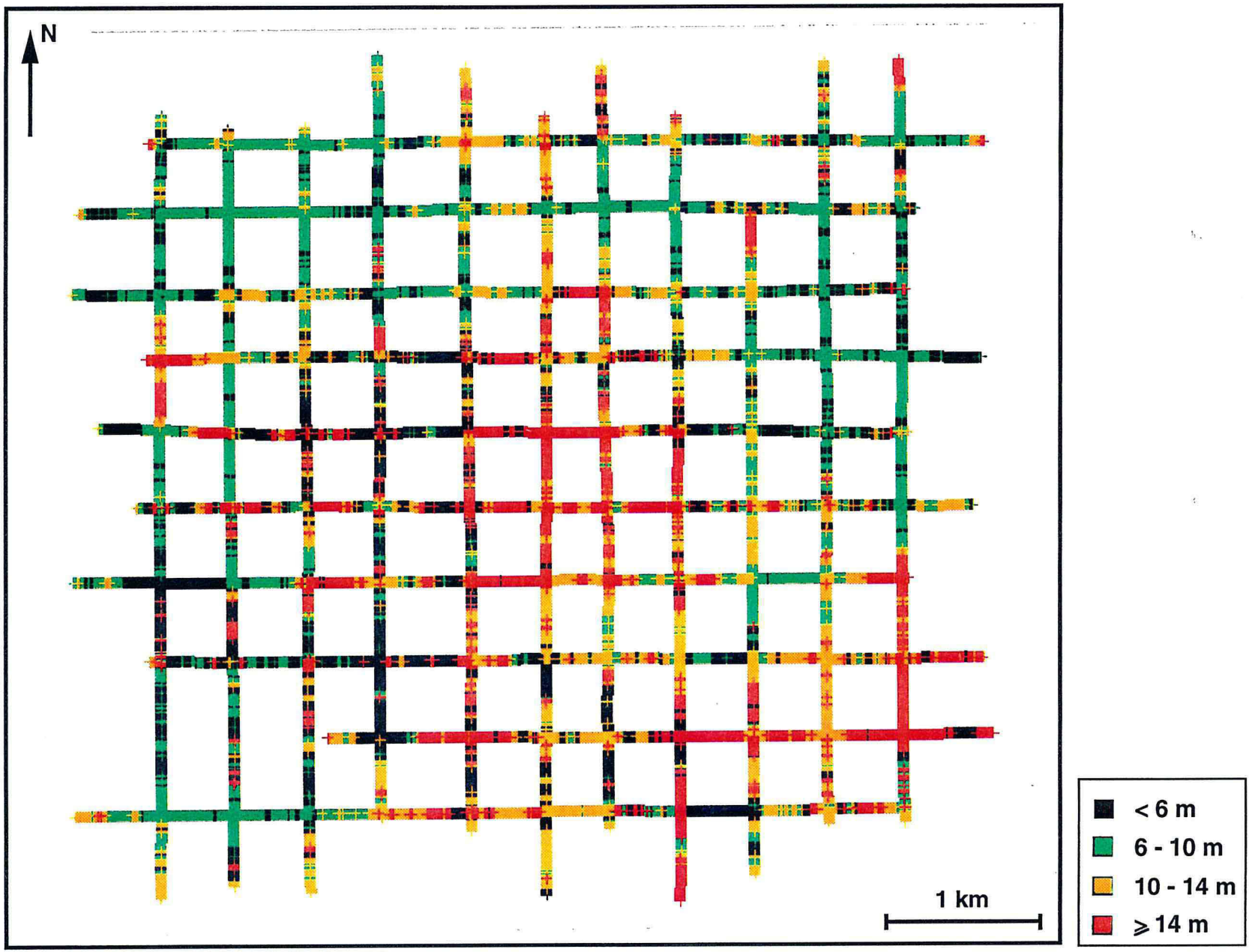


Fig. 35 Distribution spatiale de l'épaisseur de grès prédite par le mode

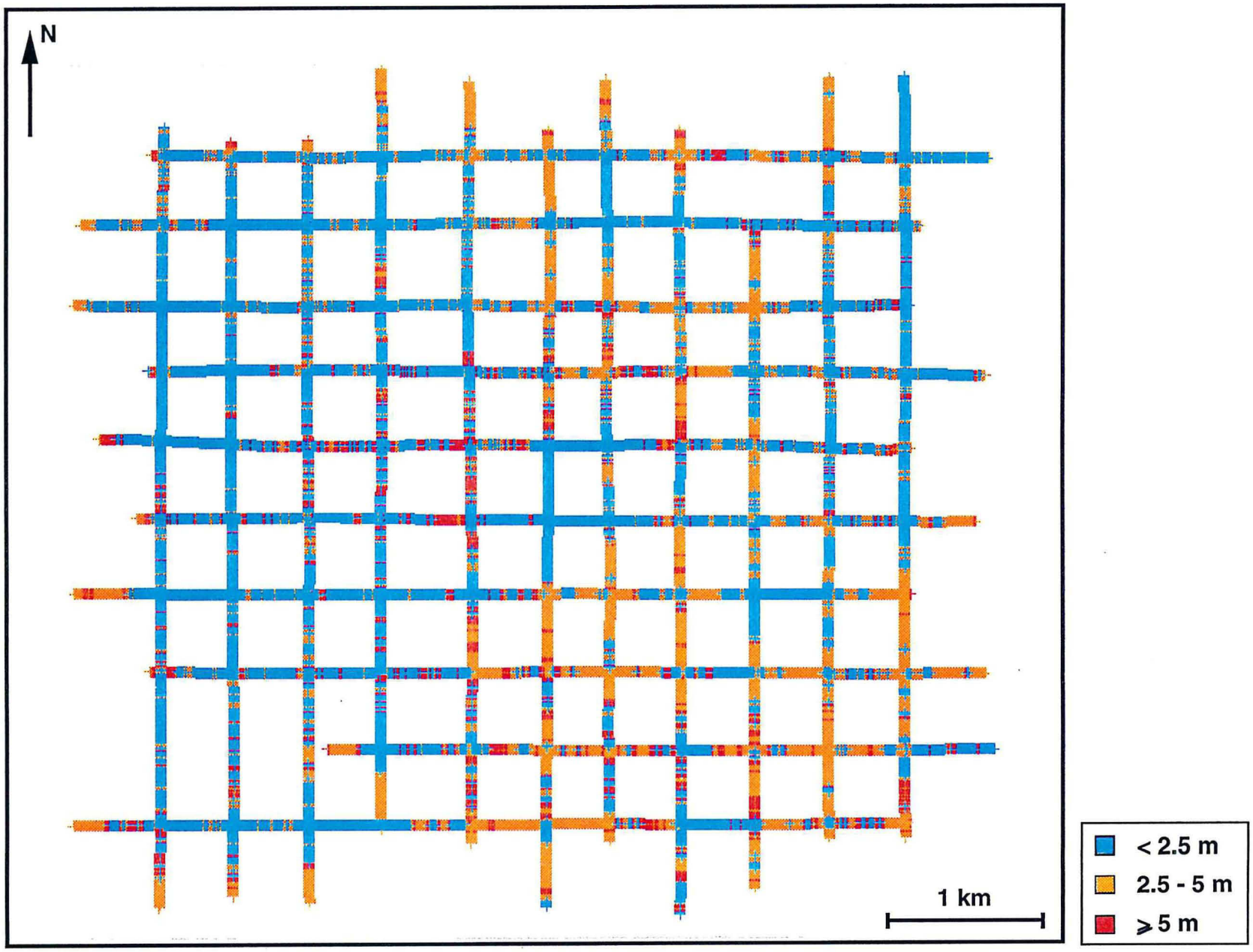


Fig. 36 Distribution spatiale de l'intervalle interquartile

1.3 Influence de la décomposition en classes gaussiennes retenue sur la qualité de la prédiction

Nous avons constaté qu'il était possible d'obtenir des prédictions de l'épaisseur de grès par la méthodologie de régression non paramétrique, et que les prédictions aux puits étaient satisfaisantes. Or, le calcul de la fonction de densité conditionnelle reposant sur la connaissance d'une décomposition en classes gaussiennes, nous avons voulu vérifier l'influence de la décomposition retenue sur la qualité des prédictions, et avons donc comparé les résultats de prédiction obtenus avec ***Sol6*** à ceux obtenus avec des décompositions de moins bonne qualité que ***Sol6***.

Nous présentons les résultats de prédiction obtenus pour l'une des décompositions en classes gaussiennes considérées. Cette décomposition, notée ***Sol10*** (cf. tableau 6), est une décomposition en 4 classes gaussiennes. Les critères de qualité ont pour valeur 4.91 pour C_1 (contre 4.39 pour ***Sol6***), et 0.89 pour C_2 (contre 0.76 pour ***Sol6***).

La méthodologie de régression non paramétrique a donc été appliquée sur la population de calibrage en utilisant la décomposition ***Sol10***. La figure 37 présente le graphe des prédictions par le mode en fonction des épaisseurs réelles de grès, pour les individus de la population de calibrage. Nous constatons qu'il y a de fortes erreurs de prédiction, ce que confirme l'histogramme des erreurs de prédiction (cf. FIG. 38) ; en fait, certaines erreurs atteignent plus de 13 mètres, et seulement 58% de la population de calibrage est prédite avec moins de deux mètres d'erreur.

Comparons ces résultats avec ceux obtenus en utilisant la décomposition ***Sol6***, pour laquelle l'ajustement des classes gaussiennes est de bonne qualité. D'après le tableau 9 ci-dessous, l'erreur de prédiction moyenne faite en utilisant ***Sol10*** est de 2.52 mètres (contre 1.33 mètres pour ***Sol6***), et l'erreur quadratique atteint 13 mètres (contre 3.85 mètres pour ***Sol6***). Les résultats obtenus avec ***Sol10*** sont donc fortement dégradés par rapport à ceux obtenus avec ***Sol6***.

Tableau 9 *Comparaison des erreurs de prédiction obtenues avec **Sol6** et **Sol10***

	Erreur		Erreur maximale		% d'individus avec une erreur
	moyenne	quadratique	négative	positive	≥ 2 m
Sol6	1.33	3.84	-6.65	7.90	25%
Sol10	2.52	13.03	-13.41	7.65	42%

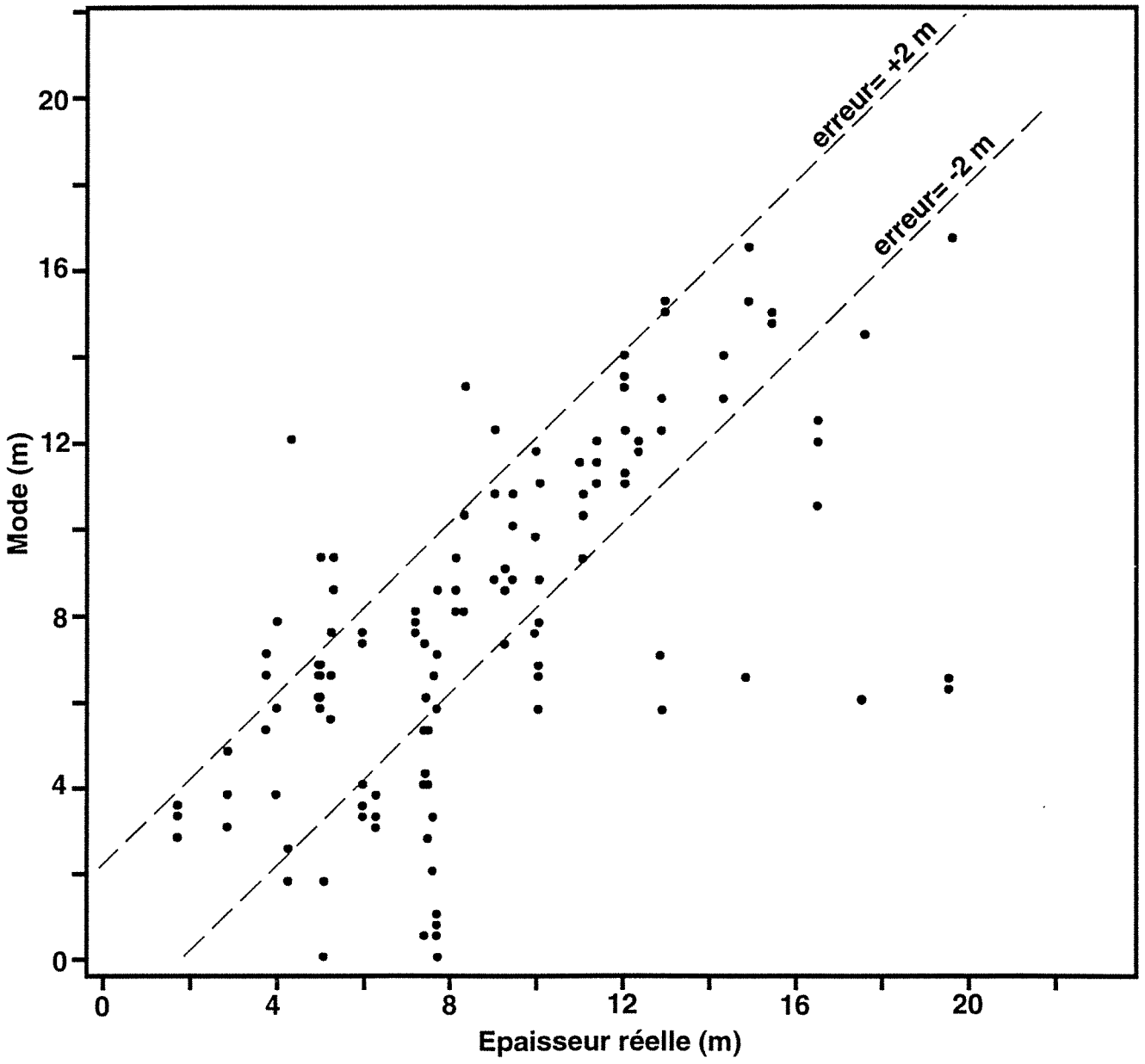


Fig. 37 Prédiction par le mode de l'épaisseur de grès aux puits (Sol10)

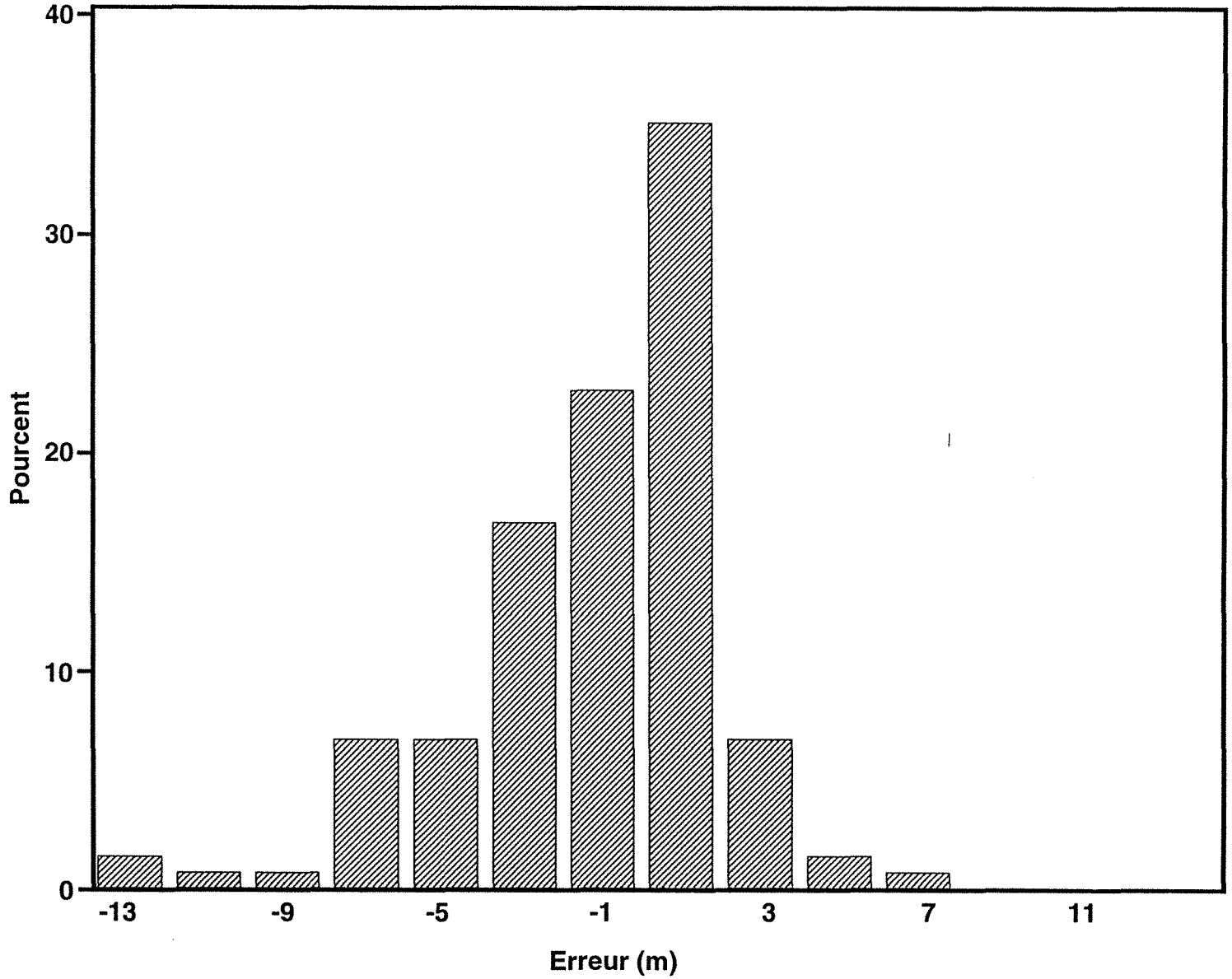


Fig. 38 Histogramme des erreurs de prédiction de l'épaisseur de grès aux puits par le mode (Sol10)

Par la suite, nous avons considéré d'autres décompositions en classes gaussiennes de moins bonne qualité que ***Sol6*** pour tester leur influence sur le calibrage. Pour chacune d'elles, les prédictions obtenues aux puits sont toujours de moins bonne qualité que celles obtenues à partir de ***Sol6***. Il apparaît donc que la qualité de la décomposition en classes gaussiennes influence la qualité des prédictions. Il est donc nécessaire de choisir avec attention la décomposition en classes gaussiennes devant être utilisée dans la méthodologie de calibrage.

1.4 Comparaison des résultats de prédiction avec ceux obtenus antérieurement sur ce champ

Des méthodes statistiques ont été appliquées sur le champ étudié pour le calibrage géologique des données sismiques, antérieurement à ce travail (Fournier et Derain, 1992-a, Fournier, 1992).

Tout d'abord, une analyse canonique effectuée sur l'ensemble des puits du champ a permis d'étalonner à partir de la sismique l'épaisseur cumulée de grès aux puits, pour la prédire ensuite entre les puits. Enfin, une analyse détaillée des données sismiques et géologiques a permis de déterminer des zones sismiquement et géologiquement homogènes sur le champ. Une analyse canonique a été appliquée sur une de ces zones (zone comprenant 22 puits dans la partie Nord/Nord-Est du champ), afin de prédire l'épaisseur cumulée de grès sur cette zone.

Nous présentons dans le tableau 10 ci-dessous les erreurs de prédiction aux puits obtenues par ces méthodes, ainsi que les erreurs de prédiction obtenues par régression non paramétrique calculées soit sur tous les puits, soit sur les 22 puits de la zone Nord/Nord-Est.

Tableau 10 *Comparaison des erreurs de prédiction de l'épaisseur de grès pour différentes méthodologies de calibrage*

		Erreur		Erreur maximale		% d'individus avec une erreur ≥ 2 m
		moyenne	quadratique	négative	positive	
Analyse canonique	sur tous les puits	2.39	9.96	-7.58	6.58	54%
	sur 22 puits	1.35	3.65	-5.38	4.13	14%
Régression non paramétrique	sur tous les puits	1.33	3.84	-6.65	7.90	25%
	sur 22 puits	1.26	3.21	-3.78	6.28	24%

Nous constatons tout d'abord que la régression non paramétrique fournit de meilleurs résultats que l'analyse canonique appliquée sur tous les puits : l'erreur de prédiction moyenne est de 1.3 mètre pour la régression non paramétrique et de 2.4 mètres pour l'analyse canonique. En fait, l'analyse canonique ne s'applique pas dans le cas où des relations non linéaires existent entre les données à calibrer, contrairement à la régression non paramétrique : or, d'après la figure 24, de telles relations non linéaires existent sur ce champ entre l'épaisseur cumulée de grès et les quatre attributs sismiques.

En ce qui concerne la zone Nord/Nord-Est du champ, les résultats de la régression non paramétrique sont légèrement meilleurs que ceux de l'analyse canonique (erreur moyenne de 1.26 contre 1.35 mètre), mais la différence n'est pas très significative. Cependant, l'obtention de prédictions satisfaisantes par analyse canonique fait suite au découpage du champ en plusieurs zones. Or, mise à part la zone Nord/Nord-Est, les autres zones n'ont pu être calibrées par analyse canonique car elles ne comportent pas assez de puits. Ce problème ne se pose pas pour la méthodologie de régression non paramétrique. Et d'ailleurs, la qualité des prédictions obtenues est aussi bonne pour l'ensemble des puits (représentatifs de zones géologiques différentes) que pour les 22 puits de la zone Nord/Nord-Est (géologiquement homogène).

En conclusion, la méthodologie de régression non paramétrique semble bien adaptée pour le calibrage géologique des données sismiques. Les résultats obtenus sont globalement meilleurs que ceux fournis par d'autres méthodes de calibrage, déjà appliquées sur les données étudiées. Et cette méthodologie permet manifestement de traiter le cas de relations non linéaires existant entre les données géologiques et les données sismiques, contrairement à l'analyse canonique par exemple.

2 Prédiction de l'épaisseur cumulée de dolomies vacuolaires

Nous avons appliqué la méthodologie de régression non paramétrique pour prédire l'épaisseur cumulée de dolomies vacuolaires en fonction des quatre attributs sismiques S_1 à S_4. Comme dans le cas précédent, la population de calibrage comporte donc 132 individus définis dans un espace de dimension 5.

Pour cette population de calibrage, les caractéristiques de l'épaisseur cumulée de dolomies vacuolaires sont les suivantes. L'épaisseur moyenne de dolomies vacuolaires est de 7,8 mètres, les épaisseurs minimale et maximale étant respectivement de 0.1 et 22,1 mètres. De plus, 25% des puits présentent une épaisseur de dolomies vacuolaires inférieures à 2,9 mètres ; et pour 75% des puits cette épaisseur est inférieure à 13 mètres.

Pour prédire l'épaisseur cumulée de dolomies vacuolaires, nous avons appliqué la même démarche que celle utilisée pour la prédiction de l'épaisseur cumulée de grès. Nous ne présentons donc que les principaux résultats.

2.1 Choix d'une décomposition en classes gaussiennes

Nous avons appliqué la méthode de décomposition en classes gaussiennes sur la population de calibrage, en faisant varier les paramètres. Nous avons obtenu 31 solutions comportant de 1 à 6 classes. Les critères de qualité de ces solutions sont compris entre 0.49 et 3.84 pour le critère C_2 (avec une moyenne de 1.68), et entre 4.61 et 12.23 pour le critère C_1 (avec une moyenne de 7.83).

Nous fournissons dans le tableau 11 ci-dessous les caractéristiques des 6 meilleures solutions, notées ***Sol1*** à ***Sol6***, ainsi que celles de la plus mauvaise solution notée ***Sol31***. Les résultats complets sont fournis en Annexe E (tableaux E-1 à E-4).

Tableau 11 *Caractéristiques des décompositions obtenues*

	Nombre de classes	Poids des classes		Critère C_1	Critère C_2
		minimal	maximal		
Sol1	6	0.11	0.25	4.61	0.49
Sol2	6	0.14	0.19	5.26	0.51
Sol3	4	0.18	0.40	5.48	0.58
Sol4	5	0.15	0.30	5.57	0.61
Sol5	5	0.15	0.28	5.68	0.58
Sol6	6	0.14	0.20	5.80	0.72
Sol31	2	0.47	0.53	12.23	3.84

Compte tenu de la dimension de l'espace de calibrage, nous ne conservons pas la solution ***Sol1*** dont la plus petite classe a un poids trop faible pour permettre une estimation correcte de ses paramètres (moyenne et matrice de variance-covariance). Des solutions restantes, la décomposition ***Sol2*** est celle pour laquelle les critères C_1 et C_2 sont minimaux. Nous avons donc retenu la décomposition ***Sol2*** à 6 classes, car elle présente le meilleur compromis entre "minimisation des critères" et "poids des classes suffisamment important pour inférer de façon fiable les moyennes et matrices de variance-covariance des classes". Nous fournissons en Annexe E les paramètres (poids, moyenne et écart-type) des classes gaussiennes de la solution ***Sol2***.

La figure 39 correspond à la représentation de la décomposition ***Sol2*** dans certains plans de l'espace de calibrage, les individus étant codés en fonction de leur appartenance aux classes. Nous constatons que les classes présentent de fortes variances pour la plupart des variables, notamment pour la variable géologique. En outre, ce phénomène est aussi vrai pour les autres décompositions obtenues sur cette population.

2.2 Application de la régression non paramétrique à la décomposition retenue

2.2.1 Prédiction de l'épaisseur cumulée de dolomies vacuolaires aux puits

En utilisant la décomposition ***Sol2***, nous avons estimé les fonctions de densité conditionnelle de l'épaisseur cumulée de dolomies vacuolaires pour les trois traces sismiques associées à chaque puits du champ.

Le mode des fonctions de densité conditionnelle a été considéré pour prédire l'épaisseur cumulée de dolomies vacuolaires. Sur la figure 40, les valeurs prédites par le mode sont représentées en fonction des épaisseurs réelles de dolomies vacuolaires aux puits. Globalement, les prédictions obtenues sont satisfaisantes ; cependant, pour cinq puits (et chacune de leurs trois traces sismiques adjacentes), les prédictions présentent de fortes erreurs.

Ces résultats sont confirmés par l'histogramme des erreurs de prédiction (cf. FIG. 41), qui est unimodal et symétrique. En effet, l'erreur moyenne de prédiction est de 1.6 mètre, et 74% des individus de la population de calibrage présentent des erreurs de prédiction inférieures à 2 mètres. Mais ces erreurs peuvent atteindre 6.5 mètres. Sur cet histogramme, nous constatons aussi que, mis à part pour un individu, les prédictions par le mode associées aux fortes valeurs de la densité conditionnelle correspondent à des erreurs de prédiction de moins de 2 mètres : ceci concerne environ un tiers de la population de calibrage. Toutefois, les valeurs des fonctions de densité conditionnelle aux modes n'excèdent pas 0.45, ce qui est très faible, mais qui est à relier aux variances très importantes des classes gaussiennes de la décomposition ***Sol2***. Pour comparaison, les valeurs des fonctions de densité conditionnelle associées à l'épaisseur cumulée de grès atteignaient 1.14.

Enfin, si nous regardons la figure 42, qui présente l'histogramme des intervalles interquartiles, nous constatons que ceux-ci sont inférieurs à 5 mètres pour 90% de la population de calibrage, ce qui est correct.

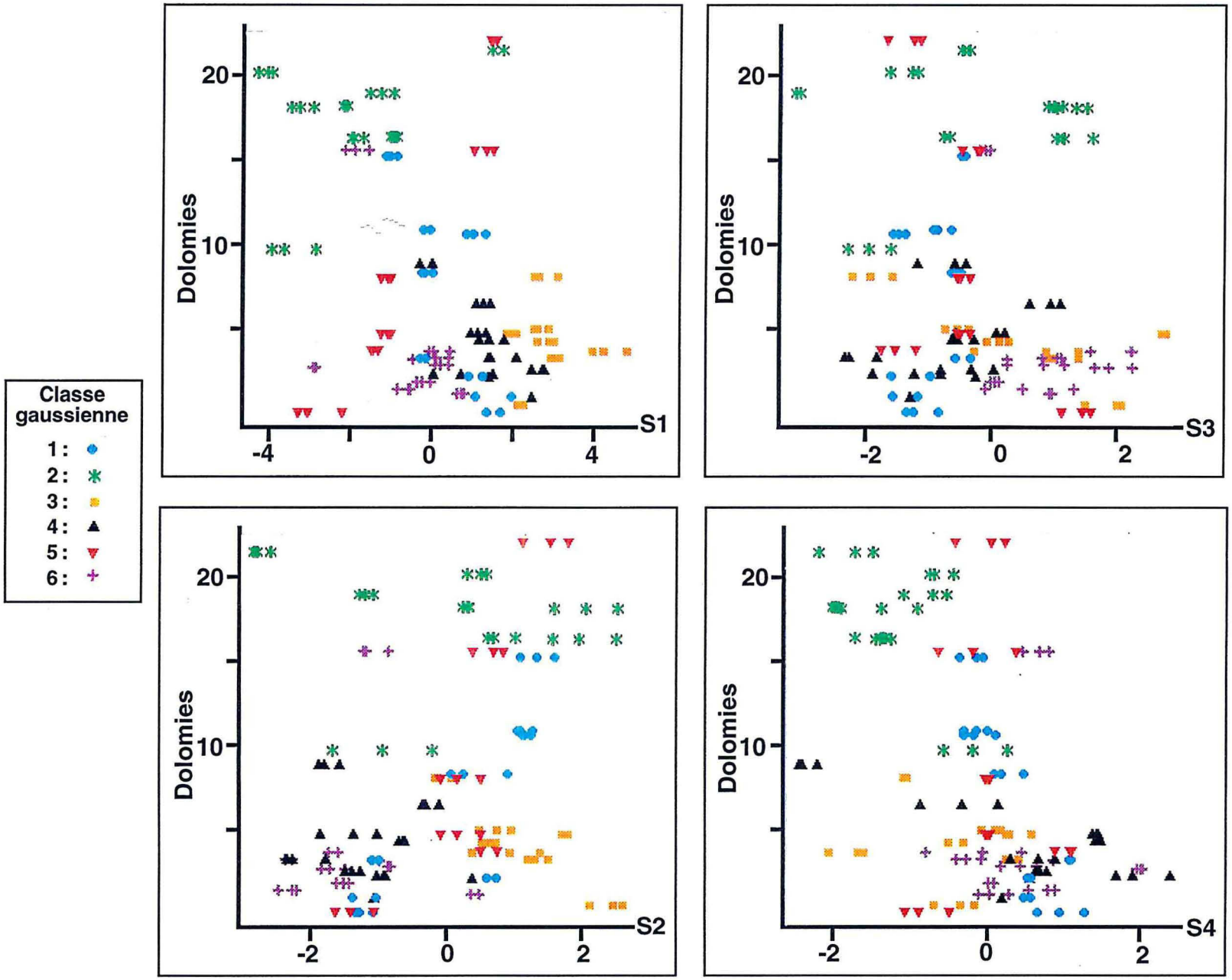


Fig. 39 Représentation de la décomposition Sol2 dans l'espace de calibrage

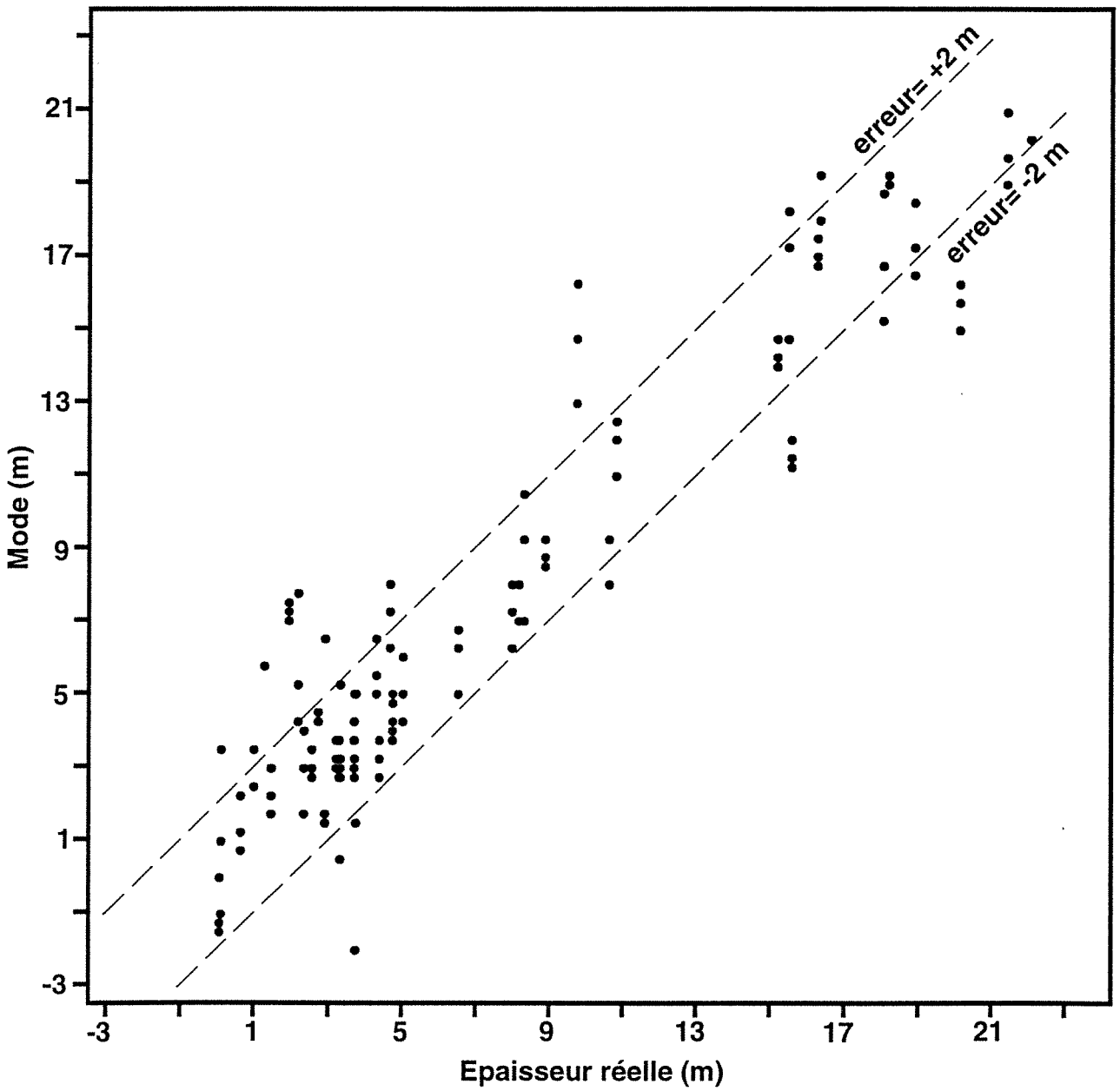


Fig. 40 Prédiction par le mode de l'épaisseur de dolomies vacuolaires aux puits

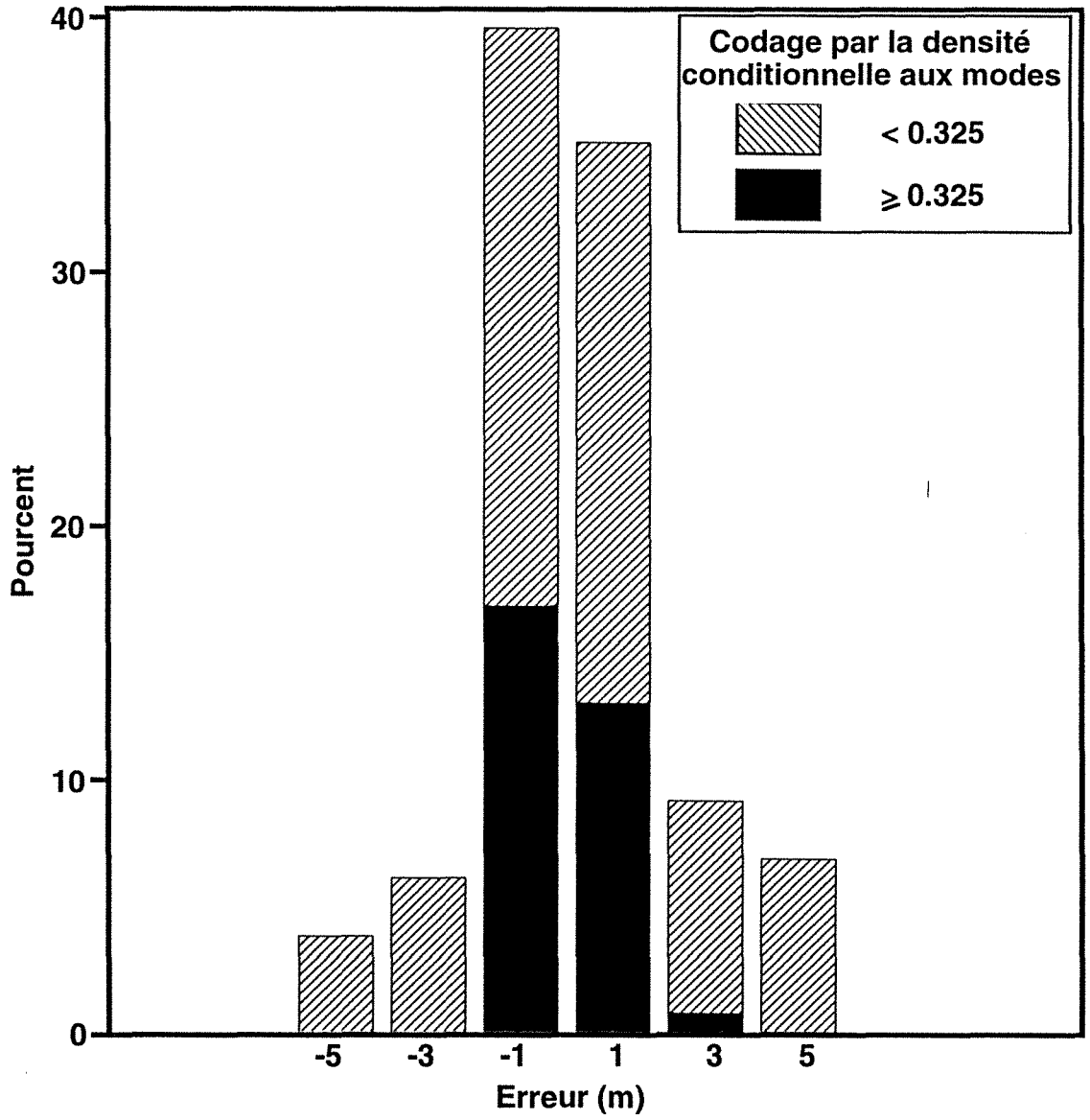


Fig. 41 Histogramme des erreurs de prédiction de l'épaisseur de dolomies vacuolaires aux puits par le mode

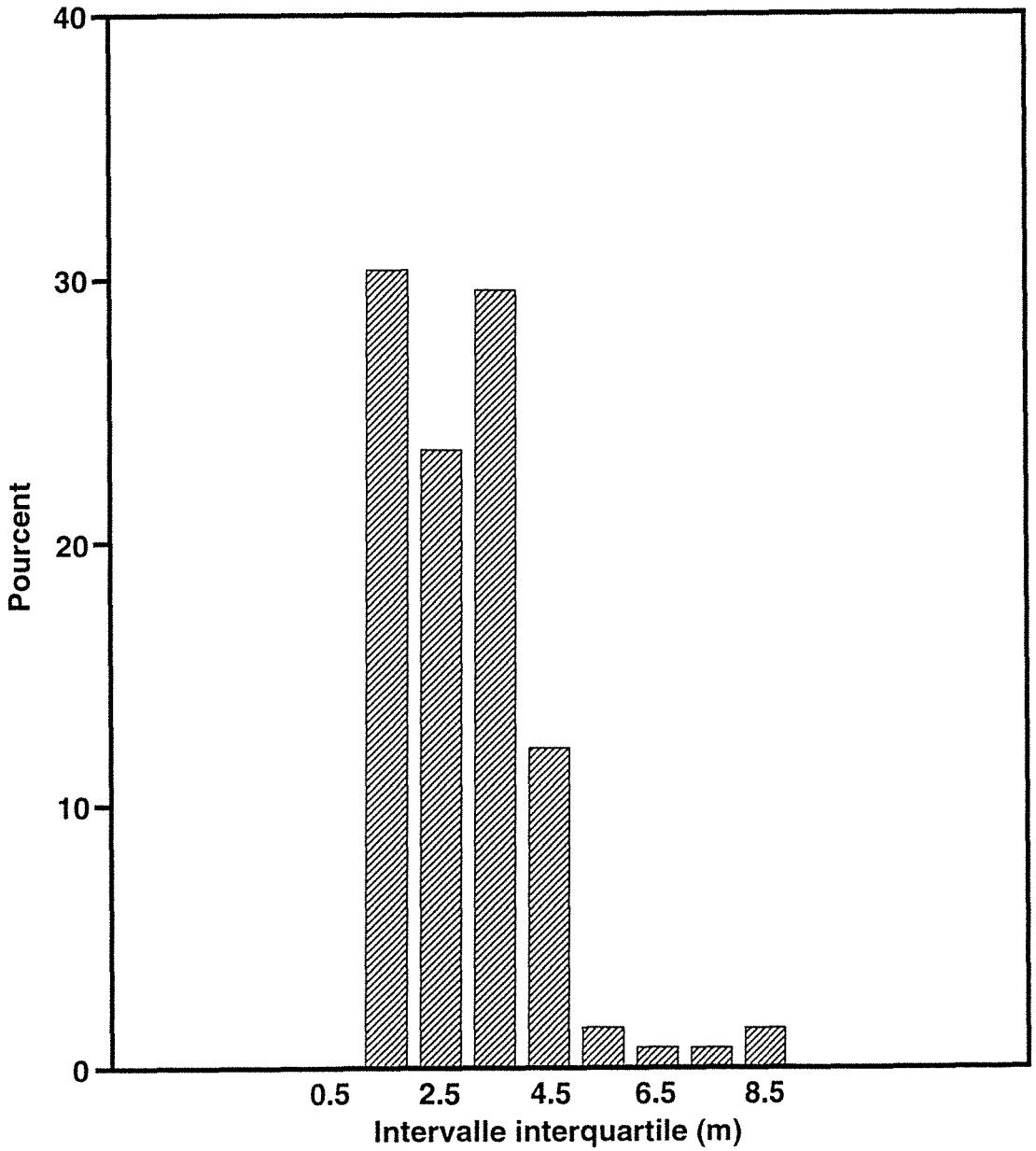


Fig. 42 Histogramme des intervalles interquartiles aux puits

2.2.2 Prédiction de l'épaisseur cumulée de dolomies vacuolaires entre les puits

Nous avons appliqué la méthodologie de régression non paramétrique pour prédire l'épaisseur cumulée de dolomies vacuolaires entre les puits, c'est-à-dire en toute trace sismique du champ. Le mode des fonctions de densité conditionnelle a été utilisé pour prédire cette épaisseur.

La figure 43 présente la carte spatiale des prédictions ainsi obtenues. Les fortes épaisseurs de dolomies vacuolaires (≥ 12 mètres), codées en rouge et en orange, se situent au niveau du sommet de la structure et sur une petite zone à l'Ouest du champ. Les zones Nord et Est sont associées principalement à des épaisseurs de dolomies vacuolaires inférieures à 6 mètres, et codées en noir. La répartition spatiale de l'épaisseur cumulée de dolomies vacuolaires n'est donc pas la même que celle de l'épaisseur cumulée de grès (sauf au sommet de la structure).

En ce qui concerne les incertitudes associées à ces prédictions, nous constatons sur l'histogramme des intervalles interquartiles (cf. FIG. 44) qu'elles sont inférieures à 5 mètres pour 89% des traces sismiques traitées. Mais elles atteignent 23 mètres pour une centaine de traces, ce qui laisse supposer que la fonction de densité conditionnelle associée à ces traces est très aplatie, et donc que la prédiction est peu sûre.

2.3 Comparaison des résultats avec ceux obtenus antérieurement sur ce champ

Antérieurement, l'analyse canonique a été appliquée sur l'ensemble des puits du champ pour prédire l'épaisseur cumulée de dolomies vacuolaires aux puits, puis entre les puits (Fournier, 1992). Par contre, il n'a pas été possible d'appliquer le calibrage indépendamment sur les zones sismiquement et géologiquement homogènes définies sur le champ (cf. chapitre 5 II-1.4 : prédiction de l'épaisseur cumulée de grès sur la zone Nord/Nord-Est) ; en effet, soit les zones ne comportent pas assez de puits pour permettre un calibrage fiable, soit les zones ne présentent pas de relations entre l'épaisseur de dolomies vacuolaires et les attributs sismiques S_1 à S_4.

Nous comparons donc les résultats obtenus par analyse canonique appliquée sur l'ensemble des puits, à ceux obtenus par régression non paramétrique. Le tableau 12 ci-dessous récapitule en terme d'erreurs de prédiction les résultats fournis par ces deux méthodologies.

Nous constatons que l'erreur moyenne de prédiction est deux fois plus forte par analyse canonique que par régression non paramétrique (3.19 contre 1.63). De plus, le pourcentage d'individus présentant une erreur de prédiction supérieure à 2 mètres est beaucoup plus fort pour l'analyse canonique ; et ces erreurs peuvent être beaucoup plus importantes par analyse canonique que par régression non paramétrique, comme l'indique le critère de l'erreur quadratique. En fait,

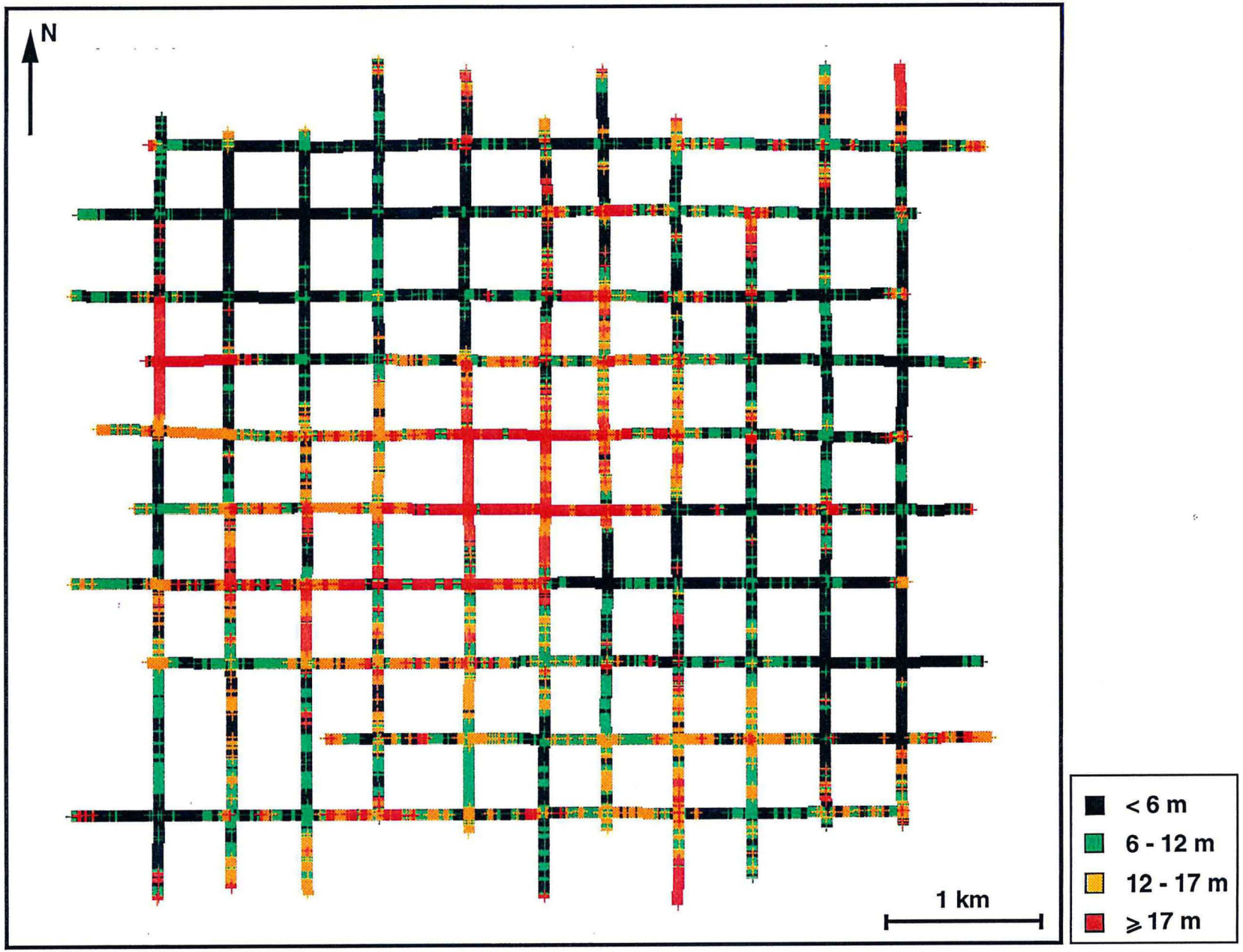


Fig. 43 Distribution spatiale de l'épaisseur de dolomies vacuolaires prédite par le mode

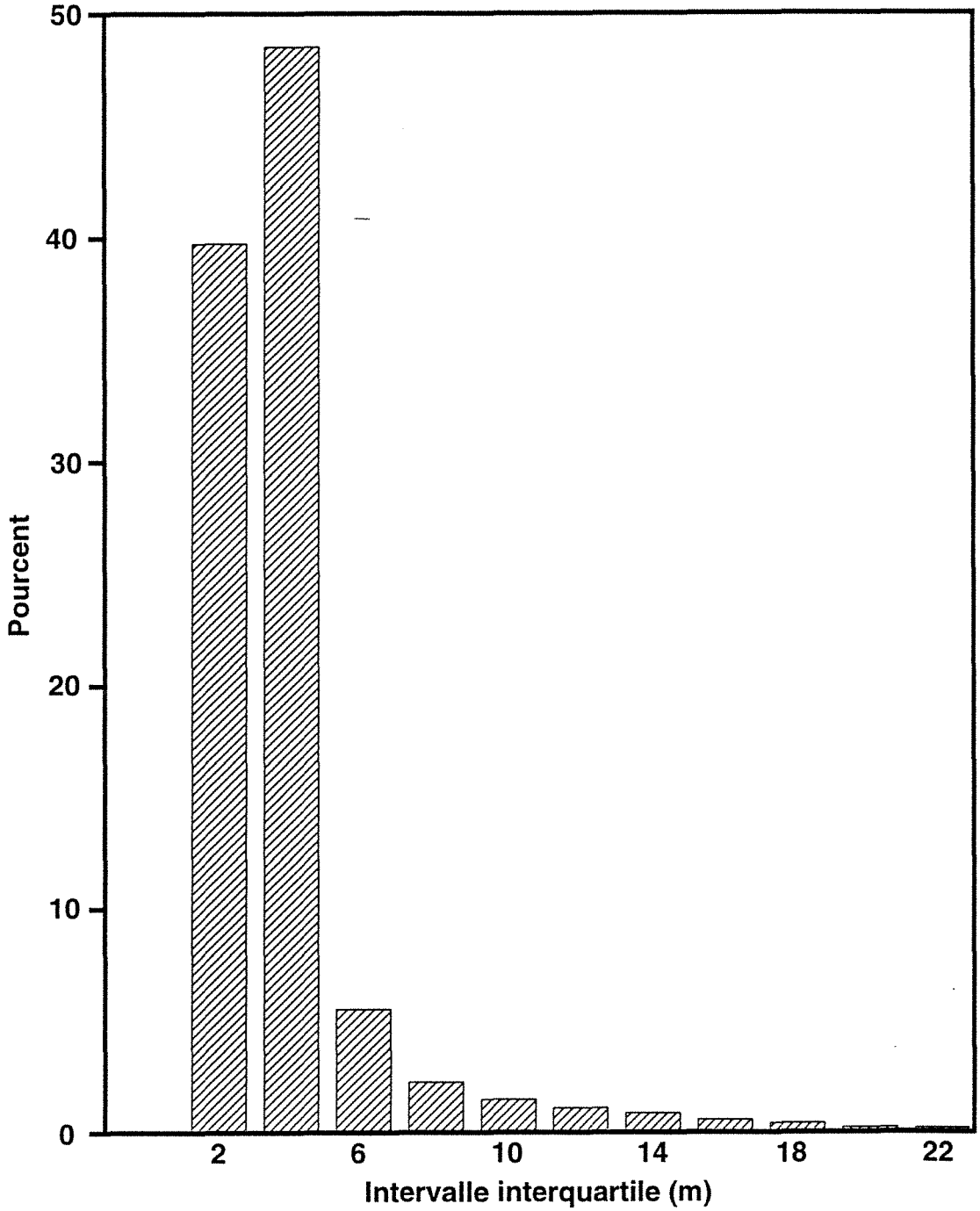


Fig. 44 Histogramme des intervalles interquartiles pour toutes les traces sismiques

les erreurs de prédiction peuvent atteindre plus de 14 mètres par analyse canonique contre 6.5 mètres par régression non paramétrique.

Tableau 12 *Comparaison des erreurs de prédiction de l'épaisseur de dolomies vacuolaires pour deux méthodologies de calibrage*

	Erreur		Erreur maximale		% d'individus avec une erreur ≥ 2 m
	moyenne	quadratique	négative	positive	
Analyse canonique	3.19	19.14	-14.38	9.88	56%
Régression non paramétrique	1.63	4.78	-5.80	6.48	26%

La méthodologie de régression non paramétrique permet donc d'estimer de façon beaucoup plus fiable que précédemment l'épaisseur de dolomies vacuolaires aux puits, comme cela avait été le cas pour la prédiction de l'épaisseur cumulée de grès. Ceci n'est pas surprenant compte tenu des relations très non linéaires existant entre l'épaisseur de dolomies vacuolaires et les quatre attributs sismiques (cf. FIG. 39).

3 Prédiction conjointe des épaisseurs cumulées de grès et de dolomies vacuolaires

3.1 Introduction

Jusqu'à présent, nous avons utilisé la méthodologie de régression non paramétrique pour prédire une unique propriété géologique en fonction d'attributs sismiques. Y-a-t-il un intérêt à prédire conjointement deux propriétés géologiques par cette méthodologie ?

En théorie, si G_1 et G_2 sont les deux propriétés géologiques à prédire connaissant les valeurs $s=(s_1,s_2\ldots)$ d'un ensemble d'attributs $S=(S_1,S_2,\ldots)$, on a :

$$f(G_1,\bullet/S=s)=\frac{f(G_1,\bullet,S=s)}{f(S=s)}$$

$$=\frac{f(G_1,S=s)}{f(S=s)}$$

$$=f(G_1/S=s)$$

et de même :

$$f(\bullet, G_2/S=s)=f(G_2/S=s)$$

Donc, si on considère la fonction de densité conditionnelle du couple (G_1,G_2) sachant $S=s$, ses fonctions de densité marginale sont égales aux fonctions de densité conditionnelle obtenues par régression non paramétrique sur chacune des propriétés G_1 et G_2.

Toutefois, ceci ne signifie pas que la prédiction conjointe du couple (G_1,G_2) sachant $S=s$ soit identique aux prédictions de G_1 sachant $S=s$ et de G_2 sachant $S=s$; en effet, si on considère le mode d'une fonction de densité conditionnelle comme valeur prédite, le mode de la fonction de densité conditionnelle du couple (G_1,G_2) sachant $S=s$ n'est pas toujours égal au couple des modes de ses fonctions de densité marginale (cf. FIG. 45).

De ce fait, il nous a semblé intéressant d'appliquer la méthodologie de régression non paramétrique pour prédire conjointement les épaisseurs cumulées de grès (propriété G1) et de dolomies vacuolaires (propriété G2). Nous comparons donc les prédictions obtenues à celles fournies par régression non paramétrique sur chacune des épaisseurs cumulées séparément (cf. chapitre 5 II-1 et II-2). Enfin, nous comparons les avantages ou inconvénients de ces deux approches.

3.2 Choix d'une décomposition en classes gaussiennes

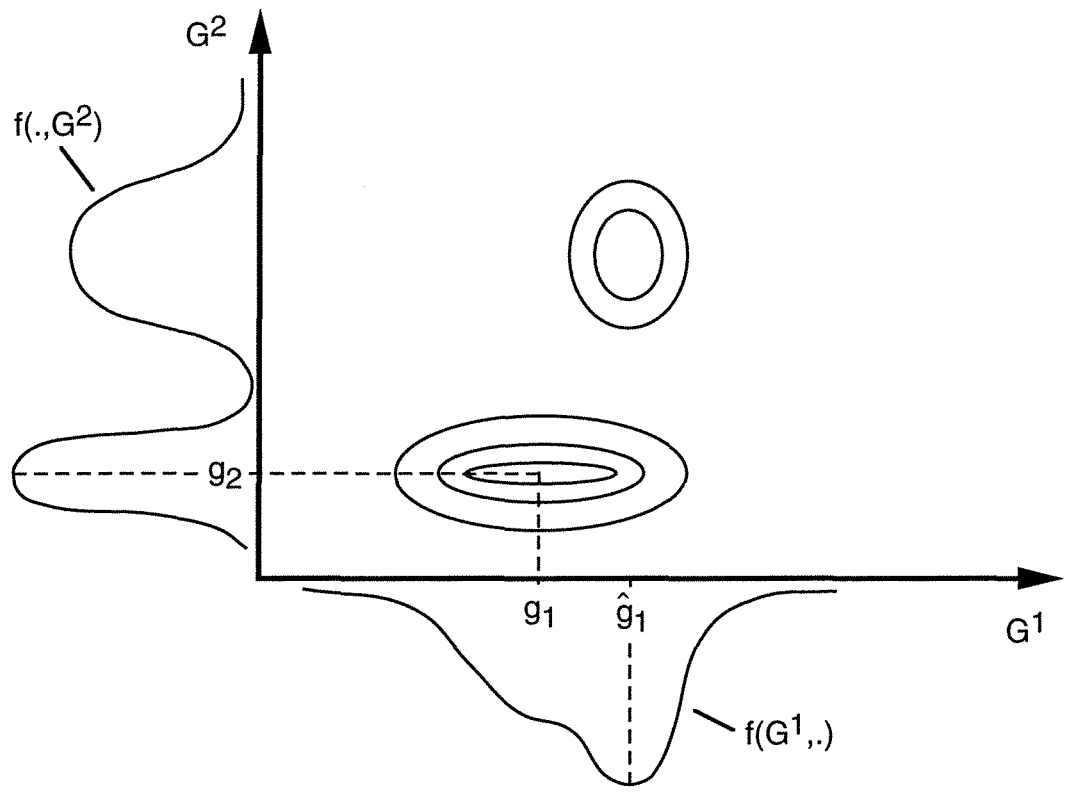
La population de calibrage comporte dans ce cas 132 individus définis dans un espace de dimension 6 (deux propriétés géologiques G_1 et G_2, quatre attributs sismiques S_1 à S_4).

Nous avons appliqué sur cette population la méthode de décomposition en classes gaussiennes. Nous avons obtenu diverses décompositions comportant de 1 à 6 classes, et dont les critères de qualité varient de 0.61 à 2.93 pour C_2 (avec une moyenne de 1.56), et de 4.66 à 9.84 pour C_1 (avec une moyenne de 7.31). La décomposition retenue est une solution à 4 classes dont le critère C_2 vaut 0.83 et le critère C_1 vaut 5.96.

3.3 Application de la régression non paramétrique à la décomposition retenue

3.3.1 Prédiction des épaisseurs cumulées de grès et de dolomies vacuolaires aux puits

Nous avons appliqué la méthodologie de régression non paramétrique sur la population de calibrage, afin de prédire aux puits les épaisseurs cumulées de grès (propriété G_1) et de dolomies vacuolaires (propriété G_2).



La fonction de densité multivariable $f(G^1,G^2)$, représentée par courbes de niveau, a pour mode le couple (g_1,g_2).

Les fonctions de densité marginales de $f(G^1,G^2)$, notées $f(G^1,.)$ et $f(.,G^2)$, ont pour modes respectifs $\hat{g}_1$ et $\hat{g}_2$.

On constate que $\hat{g}_2 = g_2$ mais $\hat{g}_1 \neq g_1$.

Fig. 45 Cas de non égalité entre le mode d'une fonction de densité multivariable et les modes de ses fonctions de densité marginales

Nous avons calculé les fonctions de densité conditionnelle du couple (G_1,G_2). Cette fois-ci, les fonctions de densité conditionnelle sont donc des surfaces 2D, définies en fonction de G1 et G_2. Les figures 46a, 47a, 48a représentent les fonctions de densité conditionnelle obtenues aux trois puits A, B et C (pour leur trace sismique la plus proche), ces puits ayant déjà été présentés lors de la prédiction de l'épaisseur de grès. Les figures 46b, 47b et 48b représentent ces mêmes fonctions par courbes de niveau.

D'un puits à l'autre, nous constatons que les fonctions de densité conditionnelle sont assez différentes, présentant des modes plus ou moins accentués. Les fonctions de densité marginale de ces fonctions de densité conditionnelle 2D devant théoriquement être identiques aux fonctions de densité conditionnelle issues des prédictions séparées de G_1 et G_2, nous fournissons les figures 46c, 47c et 48c et les figures 46d, 47d et 48d qui sont respectivement les fonctions de densité conditionnelle de G_1 sachant $S=(S_1,S_2,S_3,S_4)$ (cf. chapitre 5 II-1) et de G_2 sachant $S=(S_1,S_2,S_3,S_4)$ (cf. chapitre 5 II-2).

Pour le puits B, on constate que l'adéquation entre les "fonctions de densité marginale de la fonction de densité conditionnelle 2D" et les "fonctions de densité conditionnelle respectivement de G_1 et de G_2" est bonne. Il en va de même pour le puits C. Par contre, pour le puits A, on constate un décalage important entre le mode de la fonction de densité marginale de G_2 calculée à partir de la densité conditionnelle 2D et le mode de la fonction de densité conditionnelle de G_2. Nous constatons en fait que, suivant que l'on prédise séparément ou conjointement les deux propriétés géologiques G_1 et G_2, les modes ne sont pas toujours identiques.

Après avoir calculé les fonctions de densité conditionnelles associées au couple (G_1,G_2) sur la population de calibrage, nous avons prédit les épaisseurs cumulées de grès et de dolomies vacuolaires par le mode de ces fonctions. Les intervalles interquartiles n'étant pas définis dans le cas multivariable, nous n'avons pas pu calculer de critère de dispersion autour des valeurs prédites. Les figures 49 et 50 présentent respectivement les épaisseurs cumulées de grès et de dolomies vacuolaires prédites, en fonction des épaisseurs réelles aux puits. On constate que la prédiction de l'épaisseur de grès est satisfaisante aux puits. Par contre, la prédiction de l'épaisseur de dolomies vacuolaires est de moins bonne qualité que celle issue de la régression non paramétrique appliquée uniquement sur l'épaisseur de dolomies vacuolaires.

En fait, si on considère les figures 51 et 52, on constate que les valeurs prédites pour l'épaisseur de grès G_1 sont peu différentes, qu'on utilise la régression non paramétrique sur G_1 ou la régression non paramétrique sur le couple (G_1,G_2). Mais, par contre, il y a une forte différence de prédiction de l'épaisseur de dolomies vacuolaires par ces deux méthodes.

Nous comparons dans le tableau 13 ci-dessous la qualité des prédictions obtenues en tenant compte soit de chaque propriété géologique, soit du couple de propriétés géologiques.

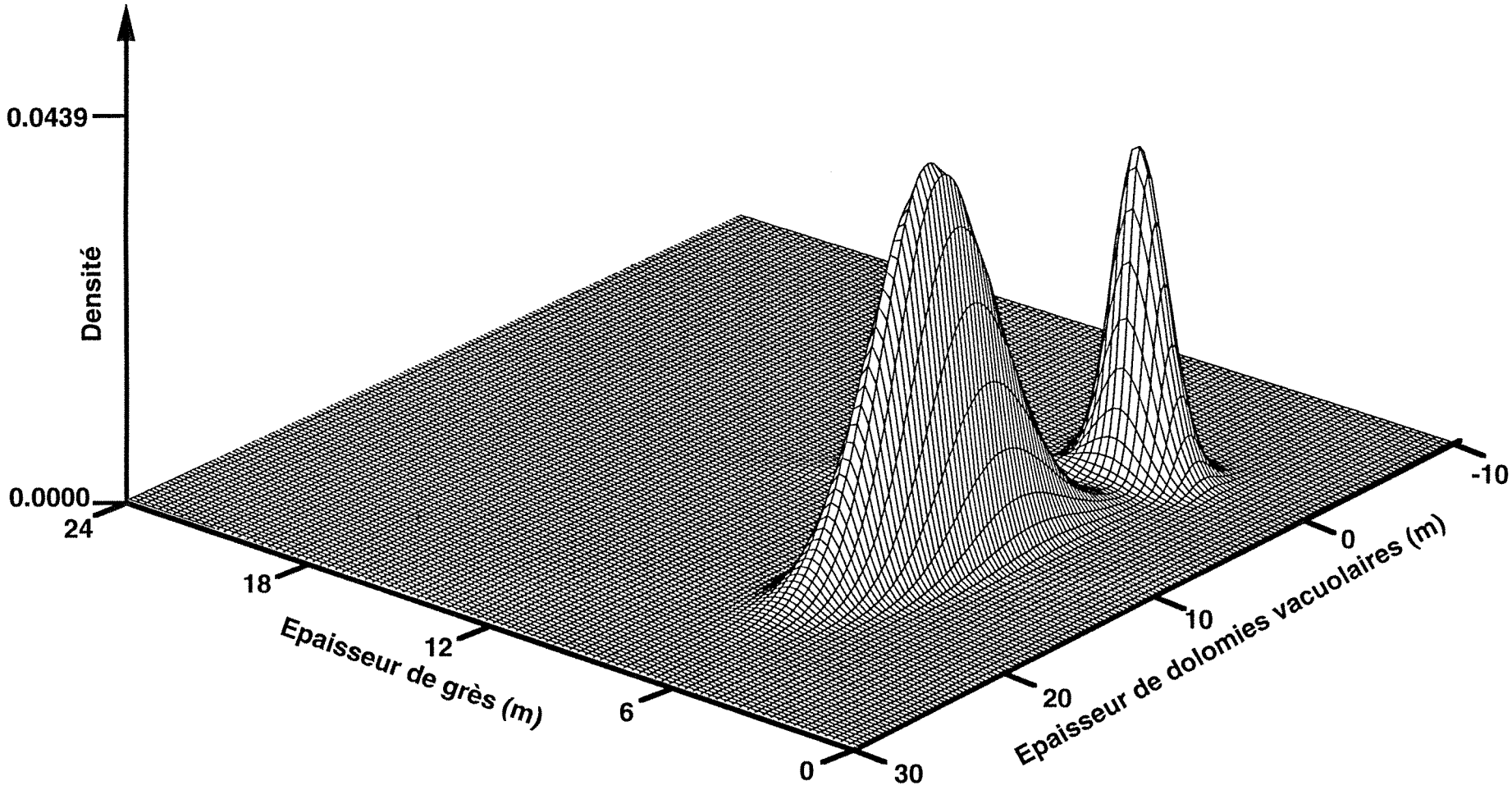


Fig. 46a Fonction de densité conditionnelle des épaisseurs de grès et de dolomies vacuolaires au puits A
Première trace adjacente

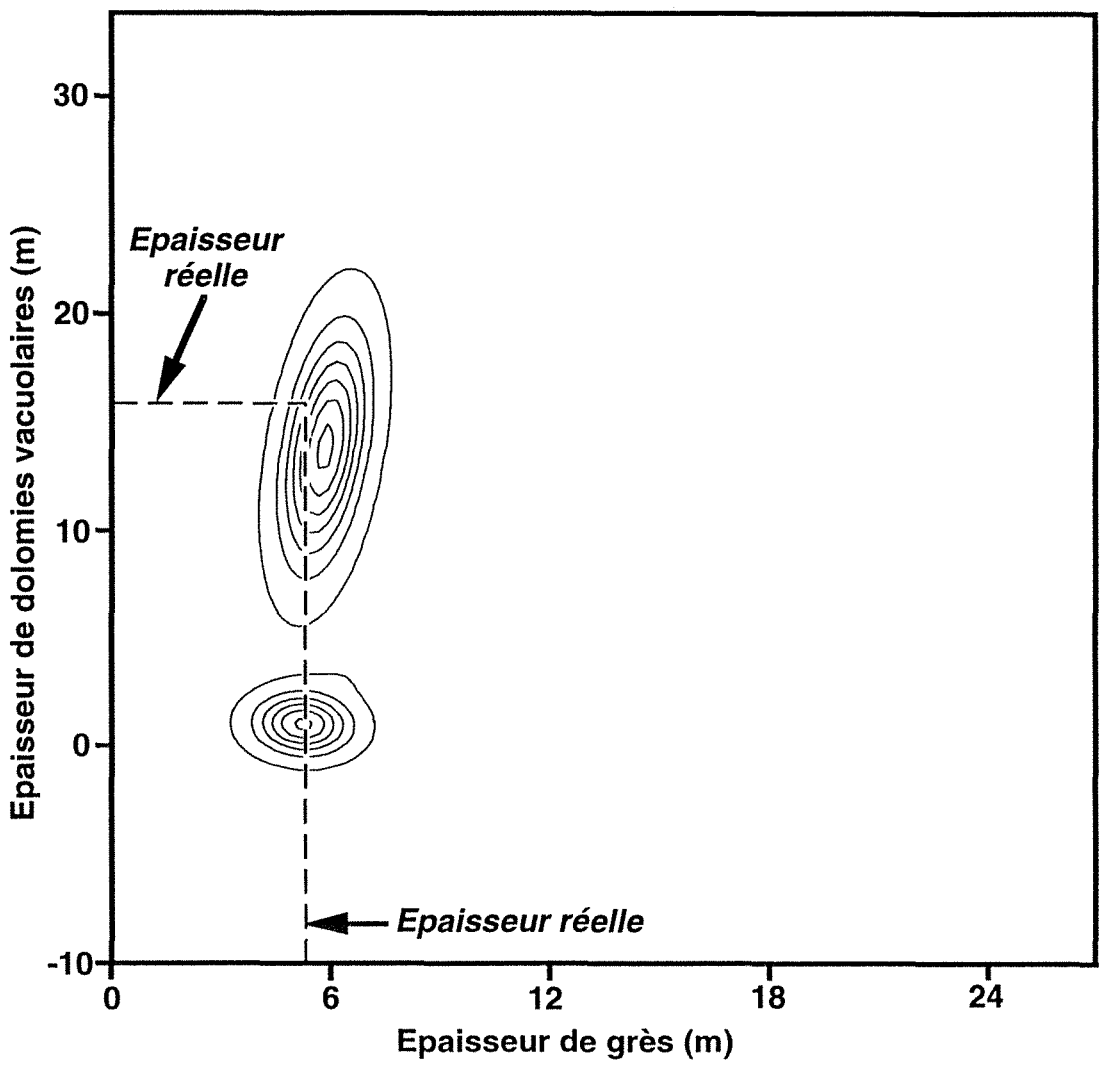


Fig. 46b Courbes de niveau de la fonction de densité conditionnelle des épaisseurs de grès et de dolomies vacuolaires au puits A. *Première trace adjacente*

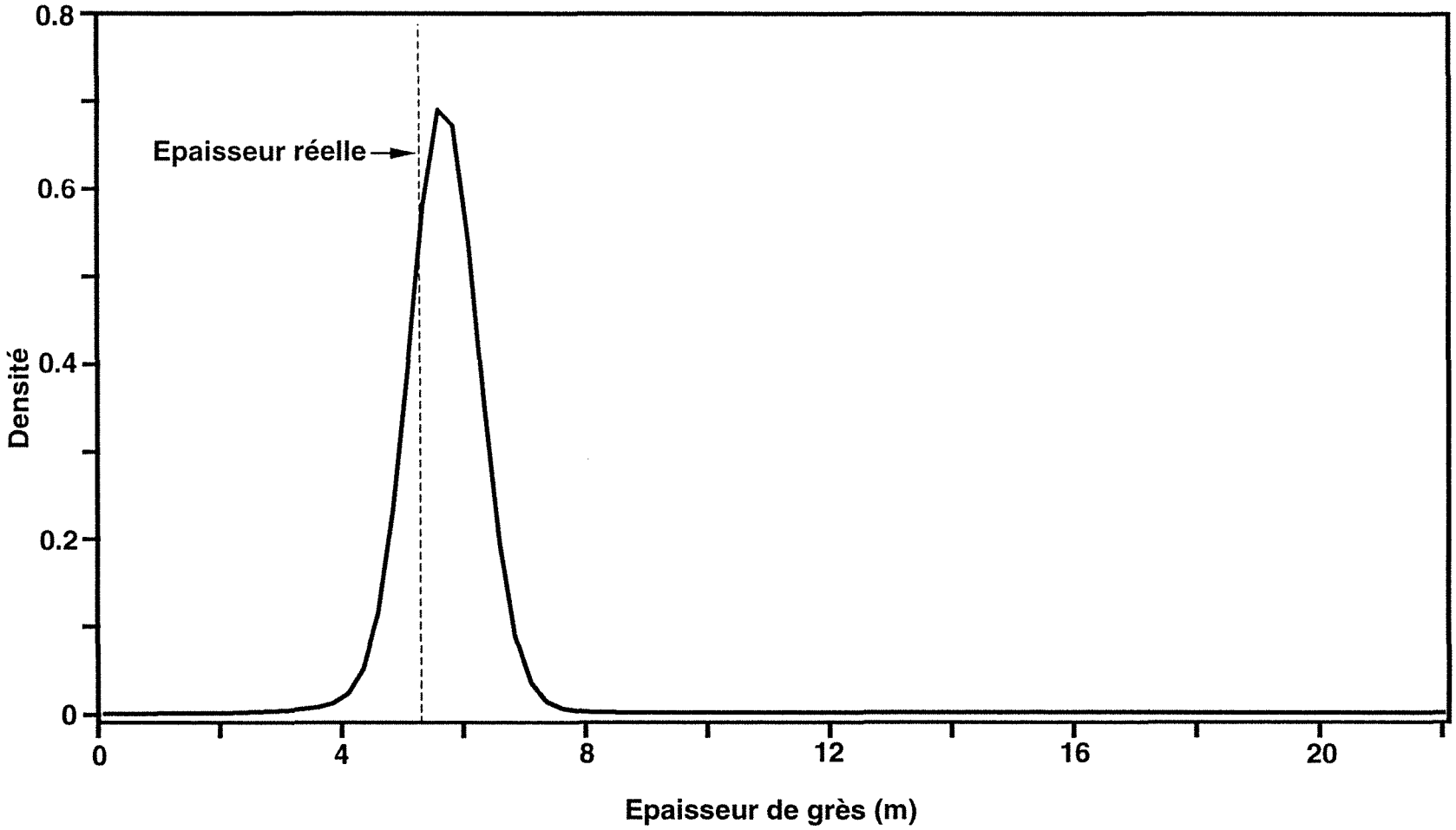


Fig. 46c Fonction de densité conditionnelle de l'épaisseur de grès au puits A
Première trace adjacente

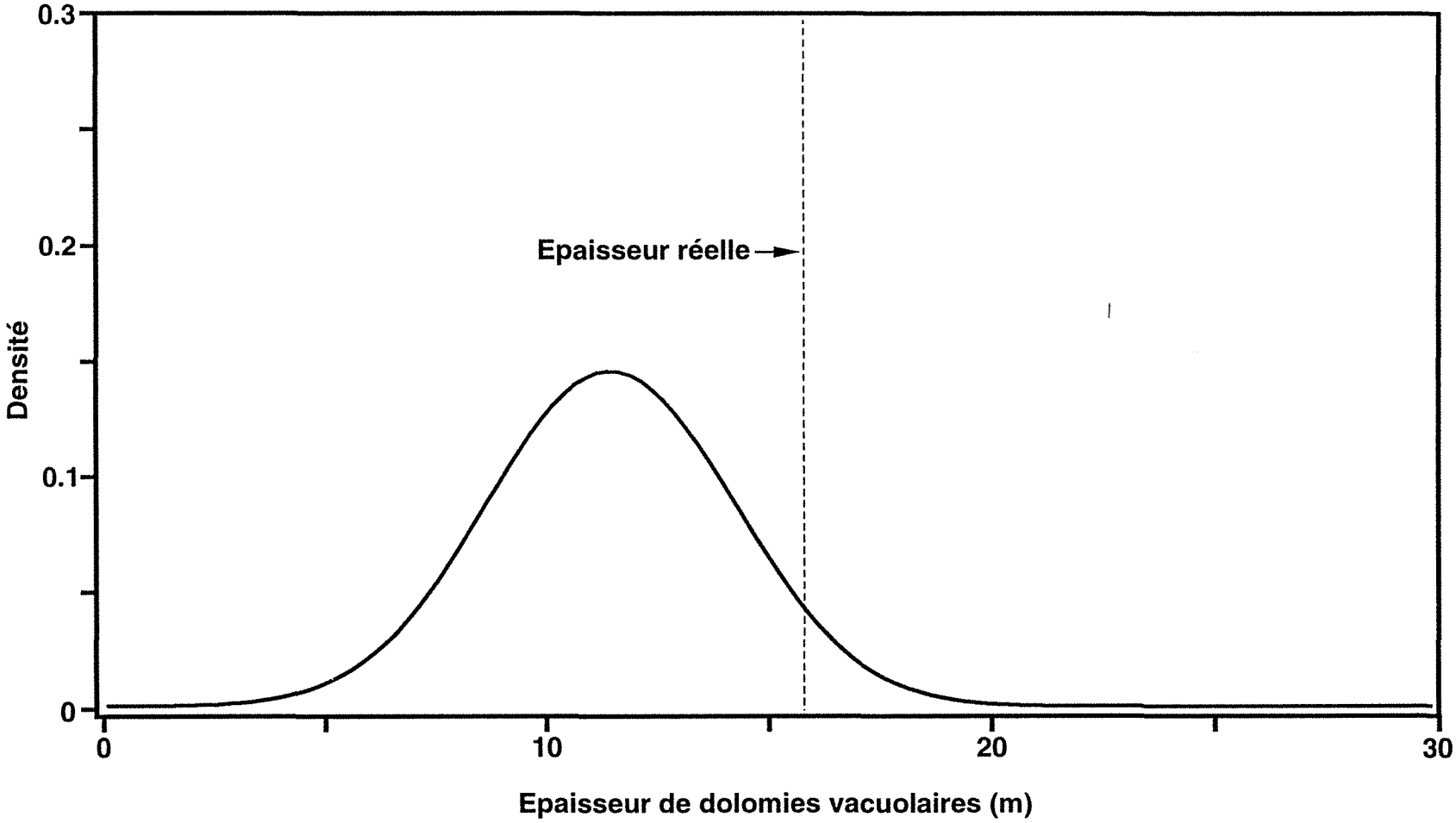


Fig. 46d Fonction de densité conditionnelle de l'épaisseur de dolomies vacuolaires au puits A
Première trace adjacente

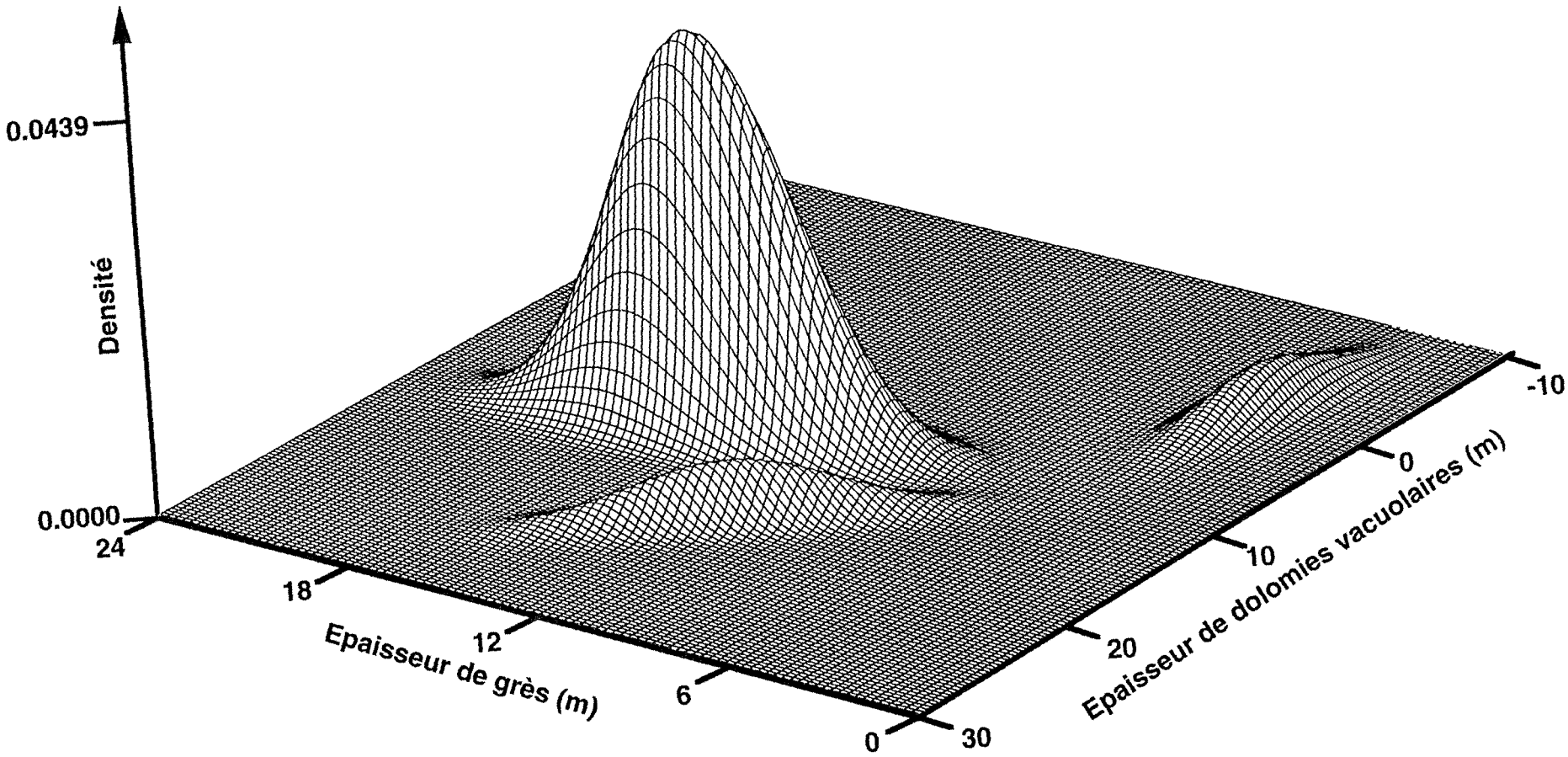


Fig. 47a Fonction de densité conditionnelle des épaisseurs de grès et de dolomies vacuolaires au puits B
Première trace adjacente

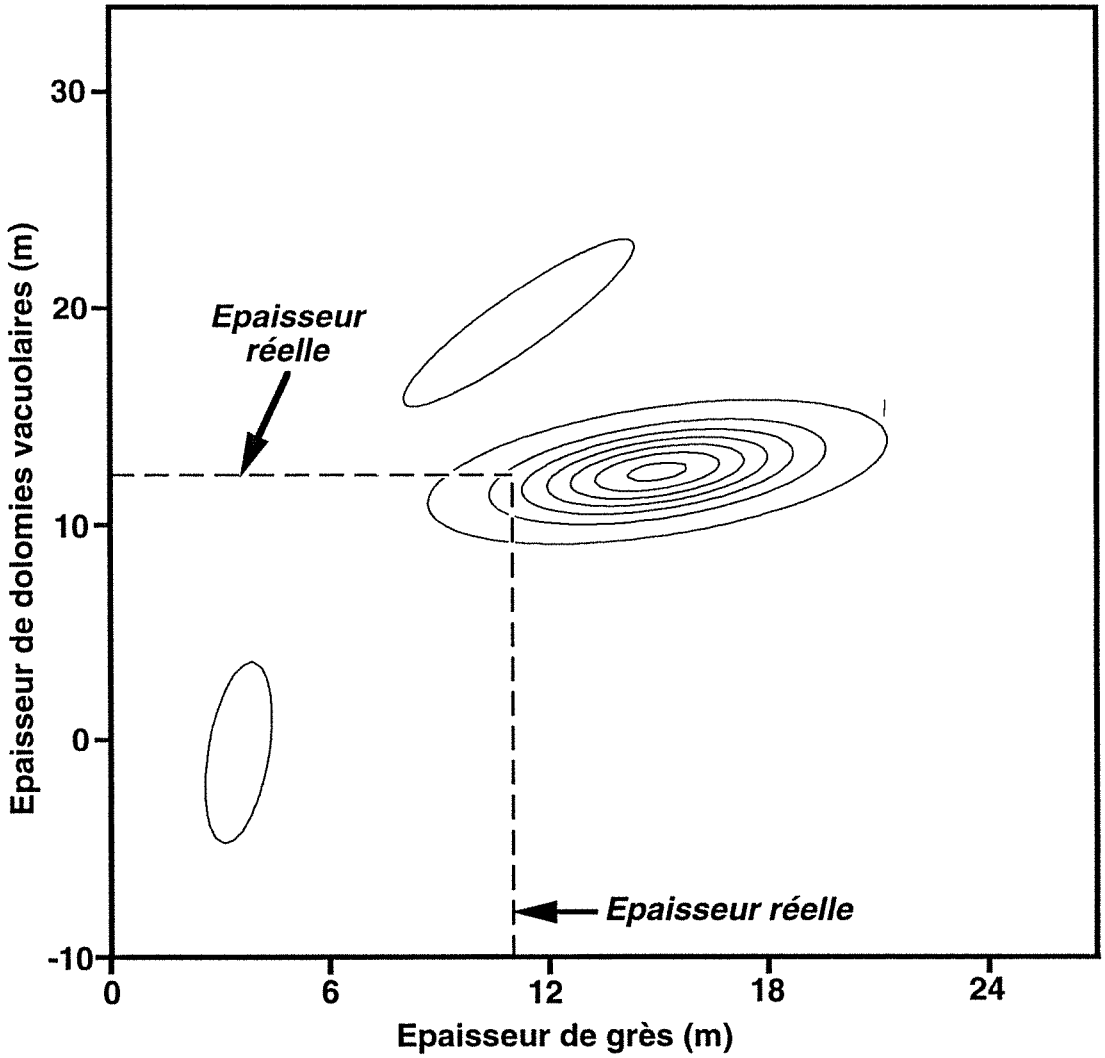


Fig. 47b Courbes de niveau de la fonction de densité conditionnelle des épaisseurs de grès et de dolomies vacuolaires au puits B. *Première trace adjacente*

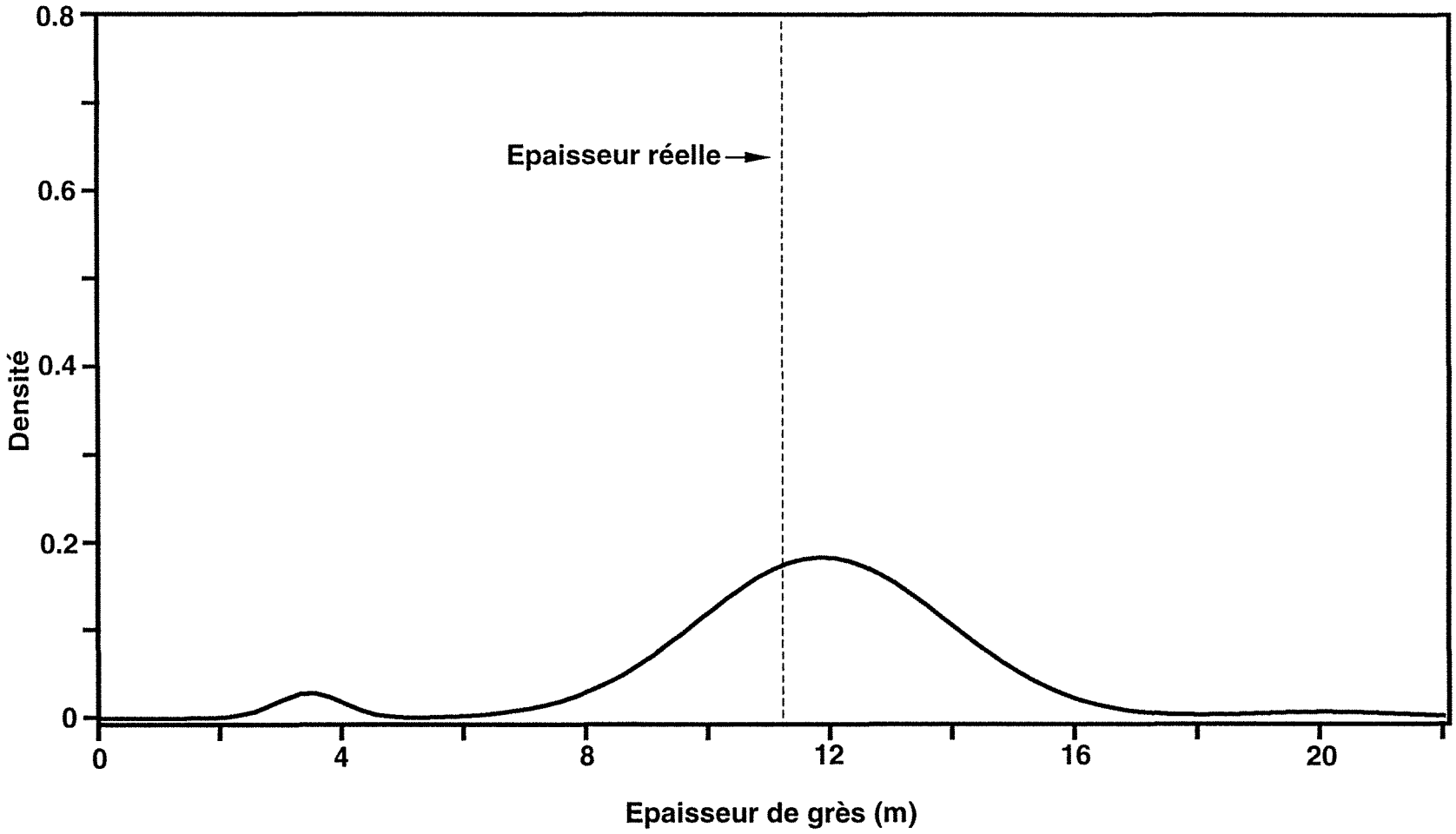


Fig. 47c Fonction de densité conditionnelle de l'épaisseur de grès au puits B
Première trace adjacente

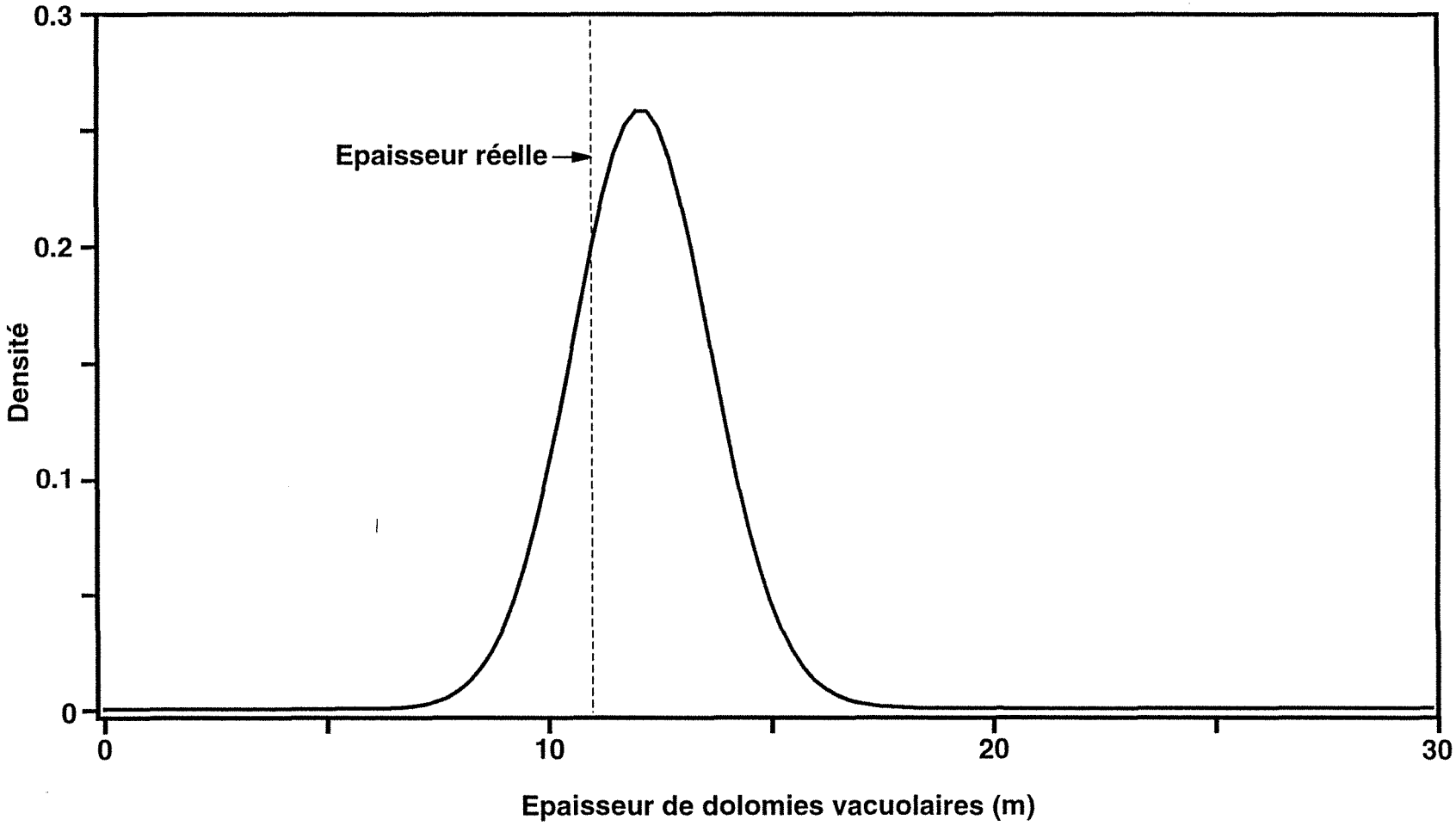


Fig. 47d Fonction de densité conditionnelle de l'épaisseur de dolomies vacuolaires au puits B
Première trace adjacente

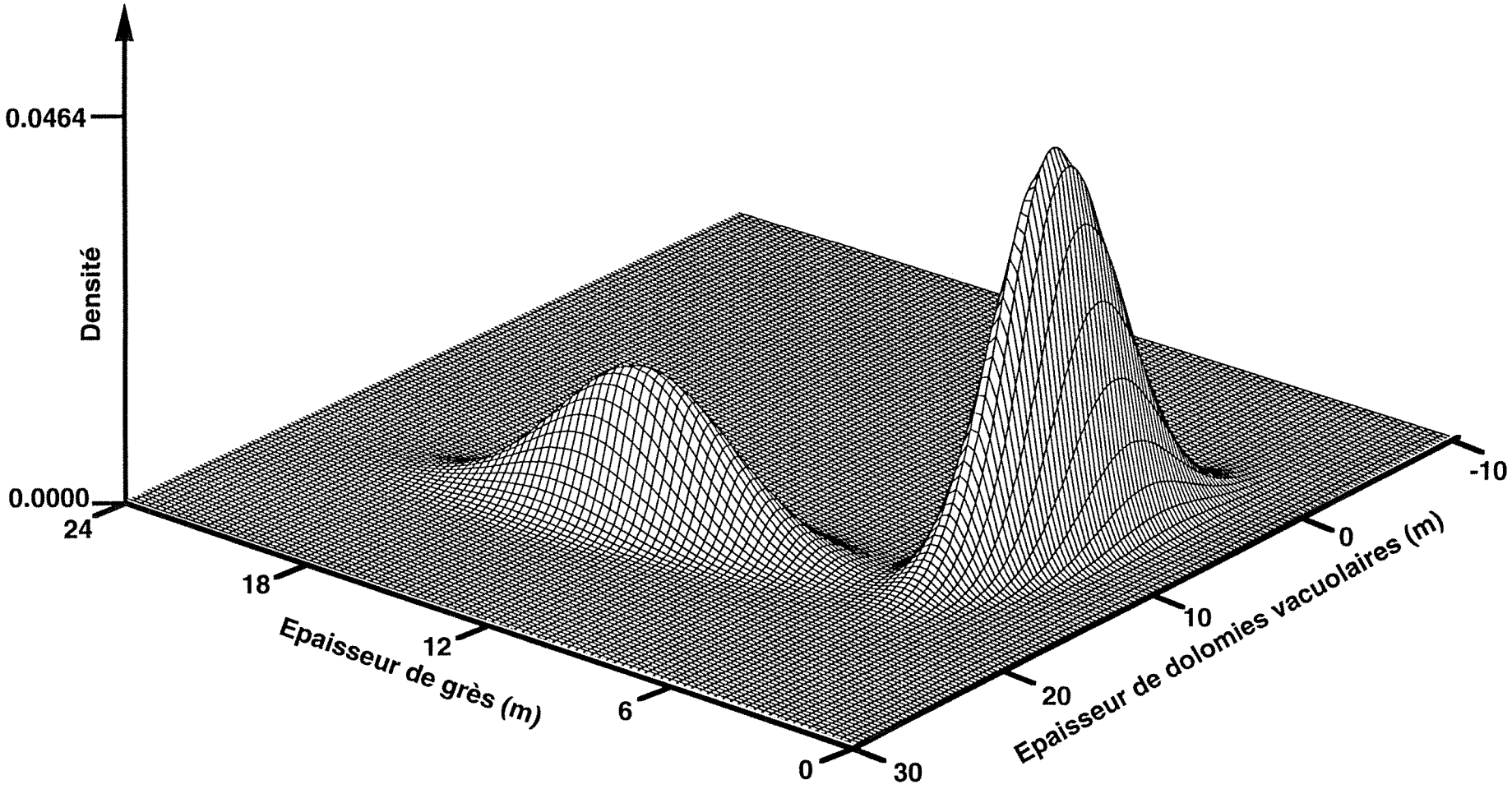


Fig. 48a Fonction de densité conditionnelle des épaisseurs de grès et de dolomies vacuolaires au puits C
Première trace adjacente

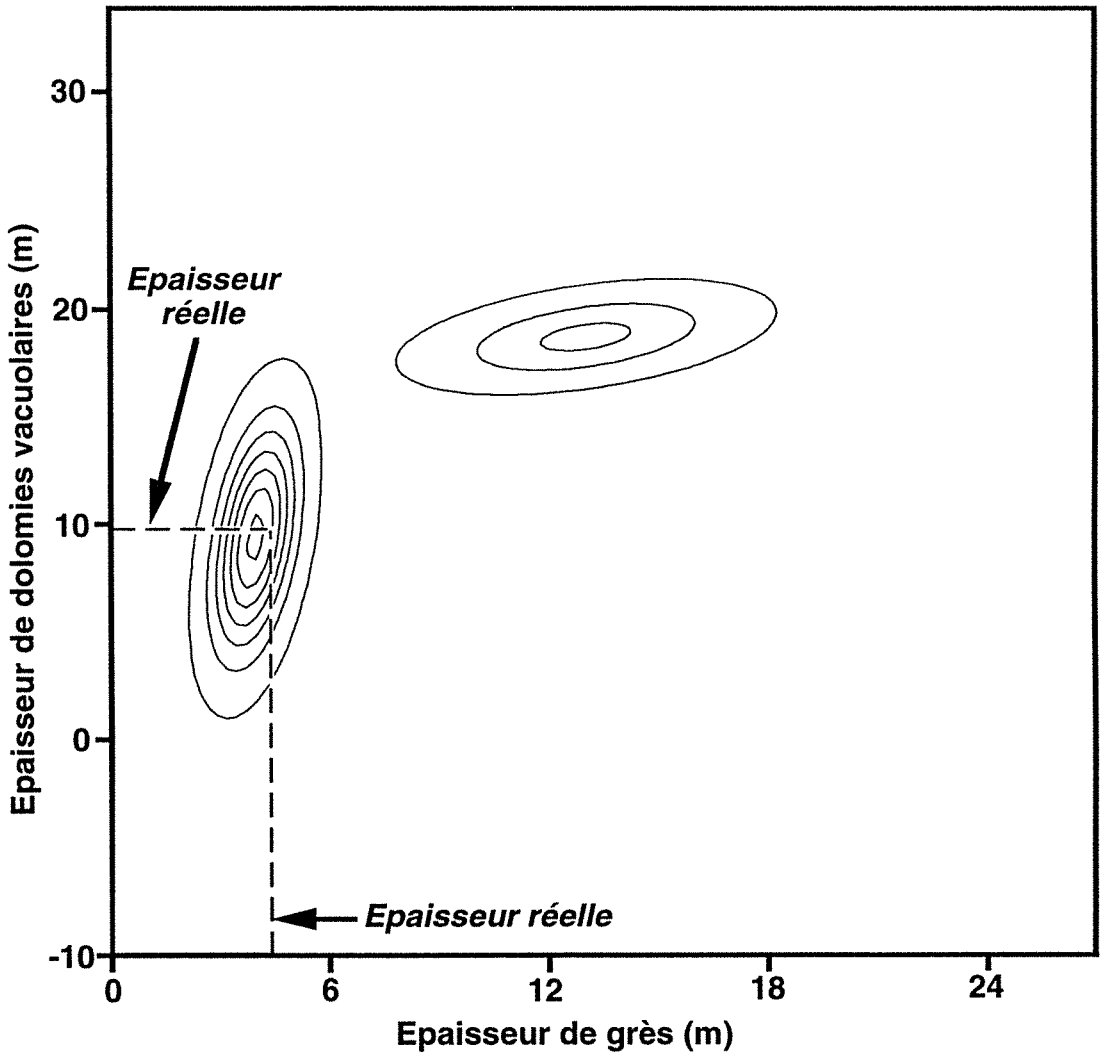


Fig. 48b Courbes de niveau de la fonction de densité conditionnelle des épaisseurs de grès et de dolomies vacuolaires au puits C. *Première trace adjacente*

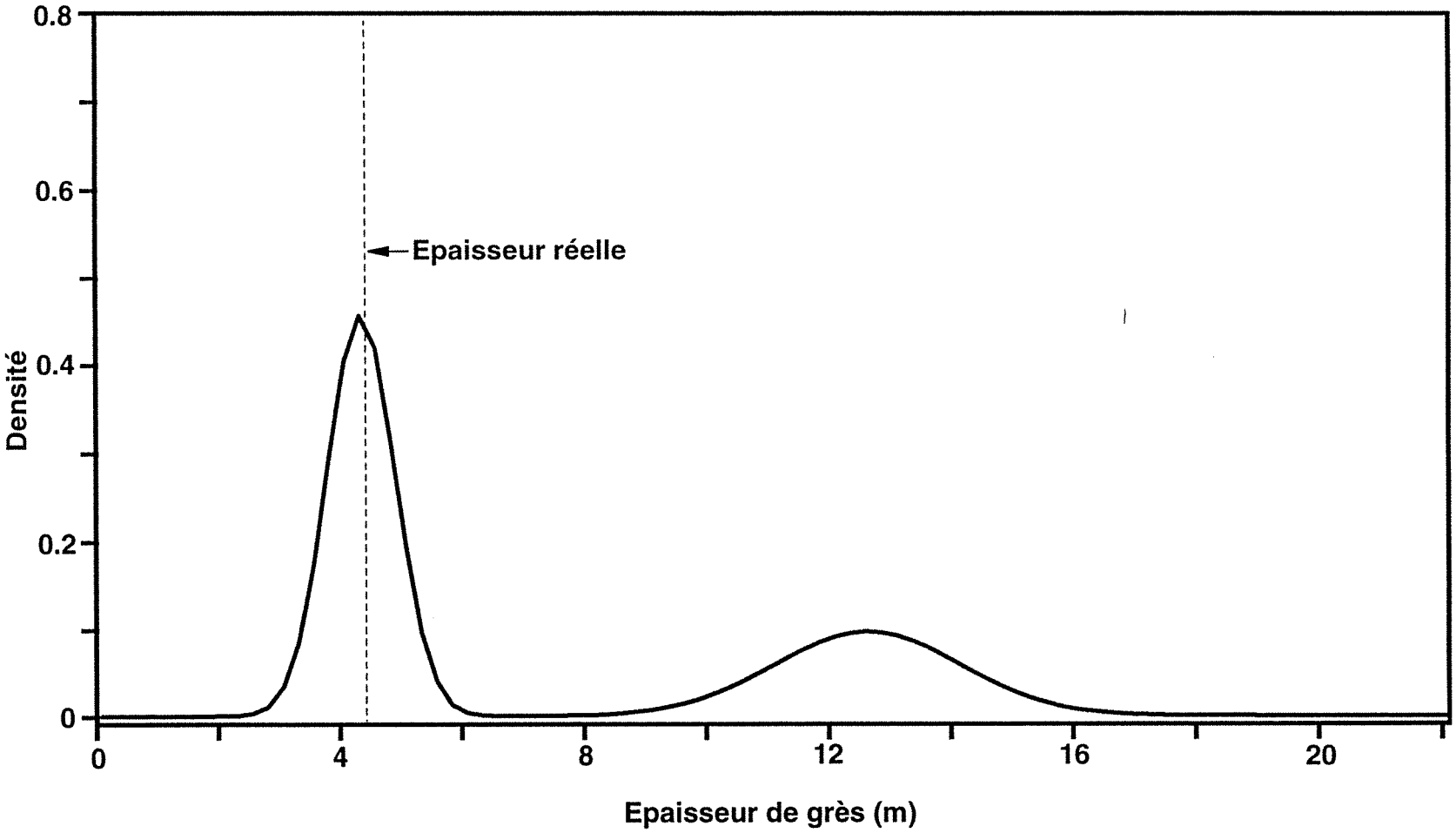


Fig. 48c Fonction de densité conditionnelle de l'épaisseur de grès au puits C
Première trace adjacente

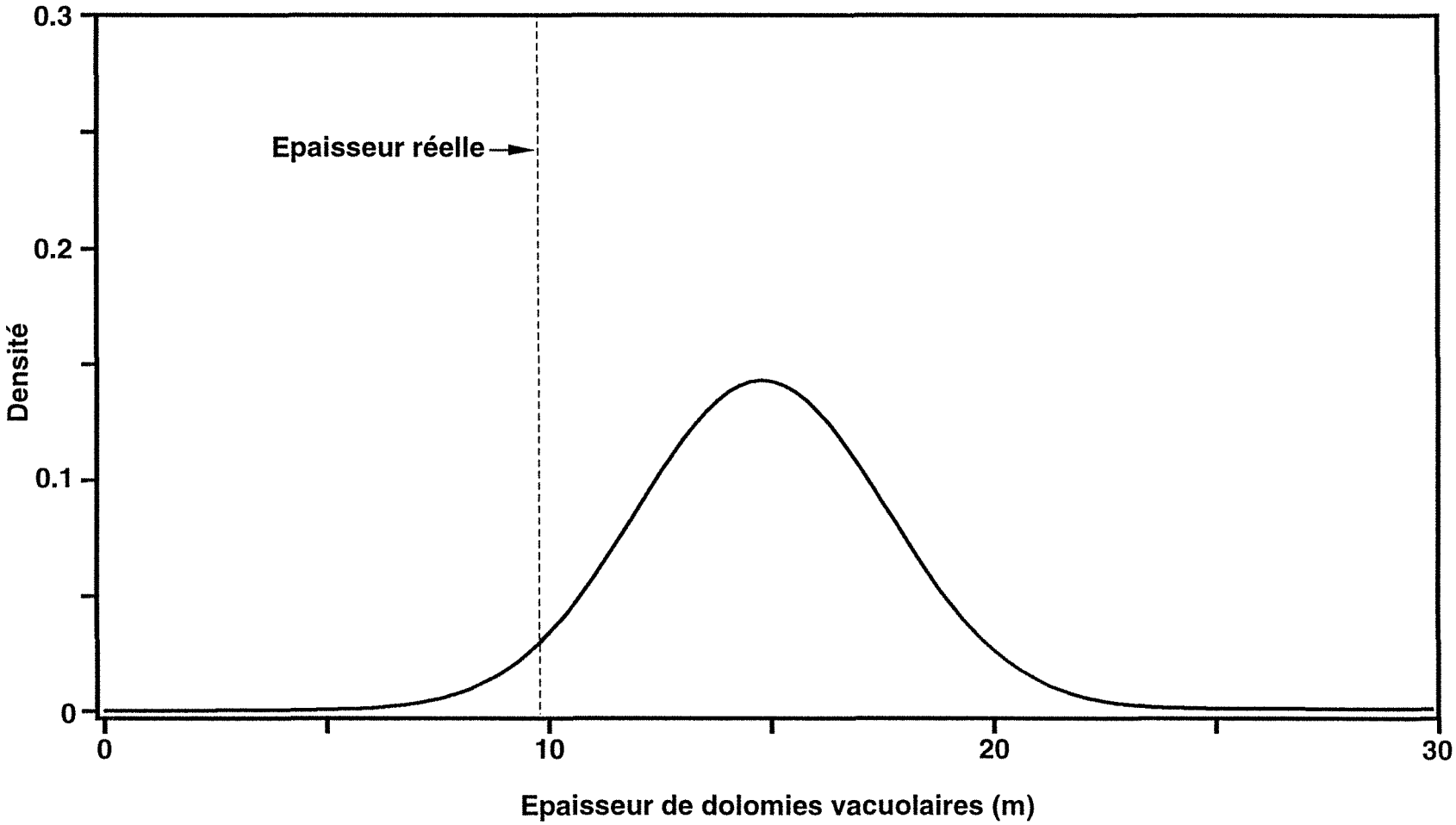


Fig. 48d Fonction de densité conditionnelle de l'épaisseur de dolomies vacuolaires au puits C
Première trace adjacente

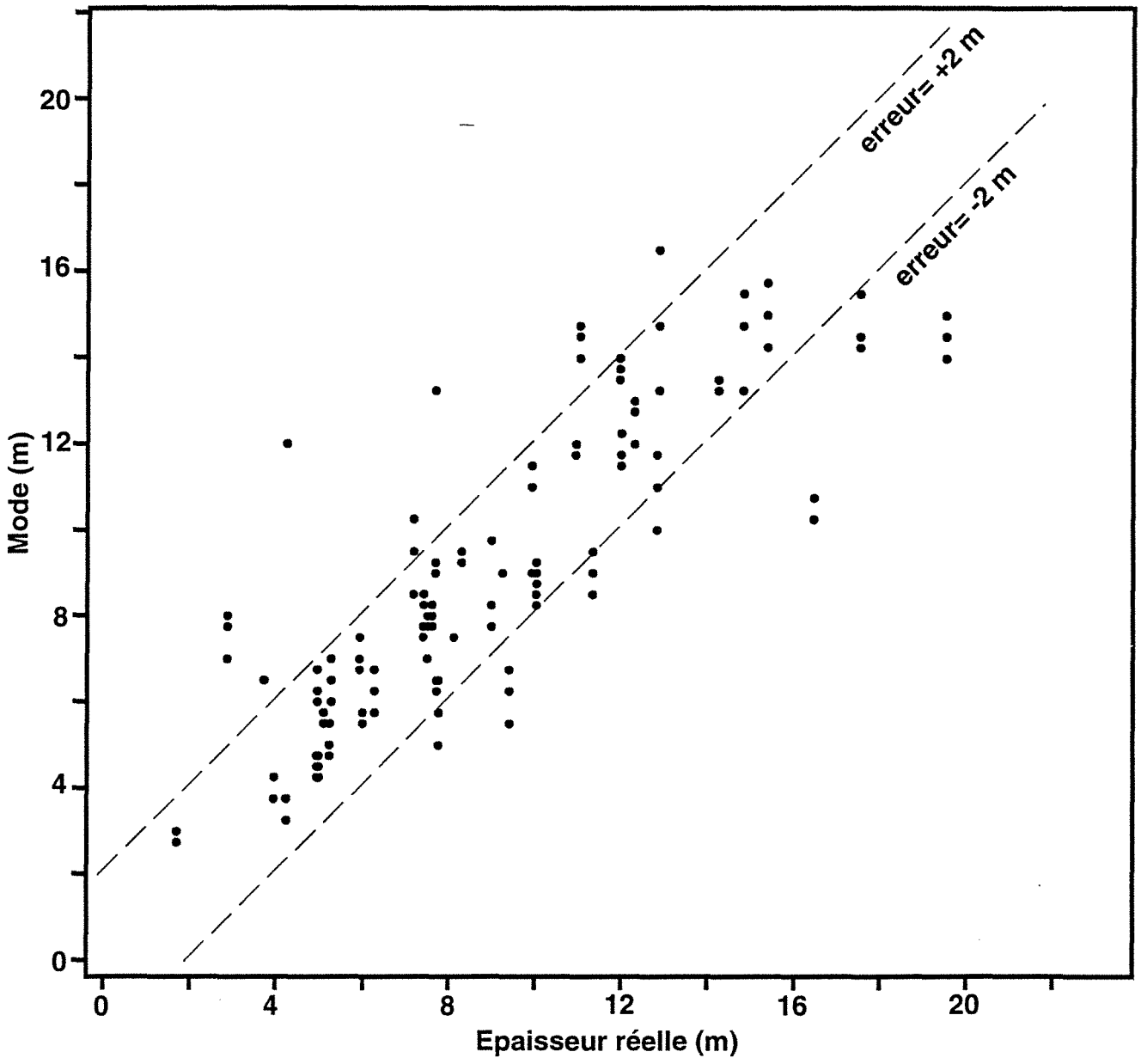


Fig. 49 Prédiction par le mode de l'épaisseur de grès aux puits
Prédiction conjointe des épaisseurs de grès et de dolomies vacuolaires

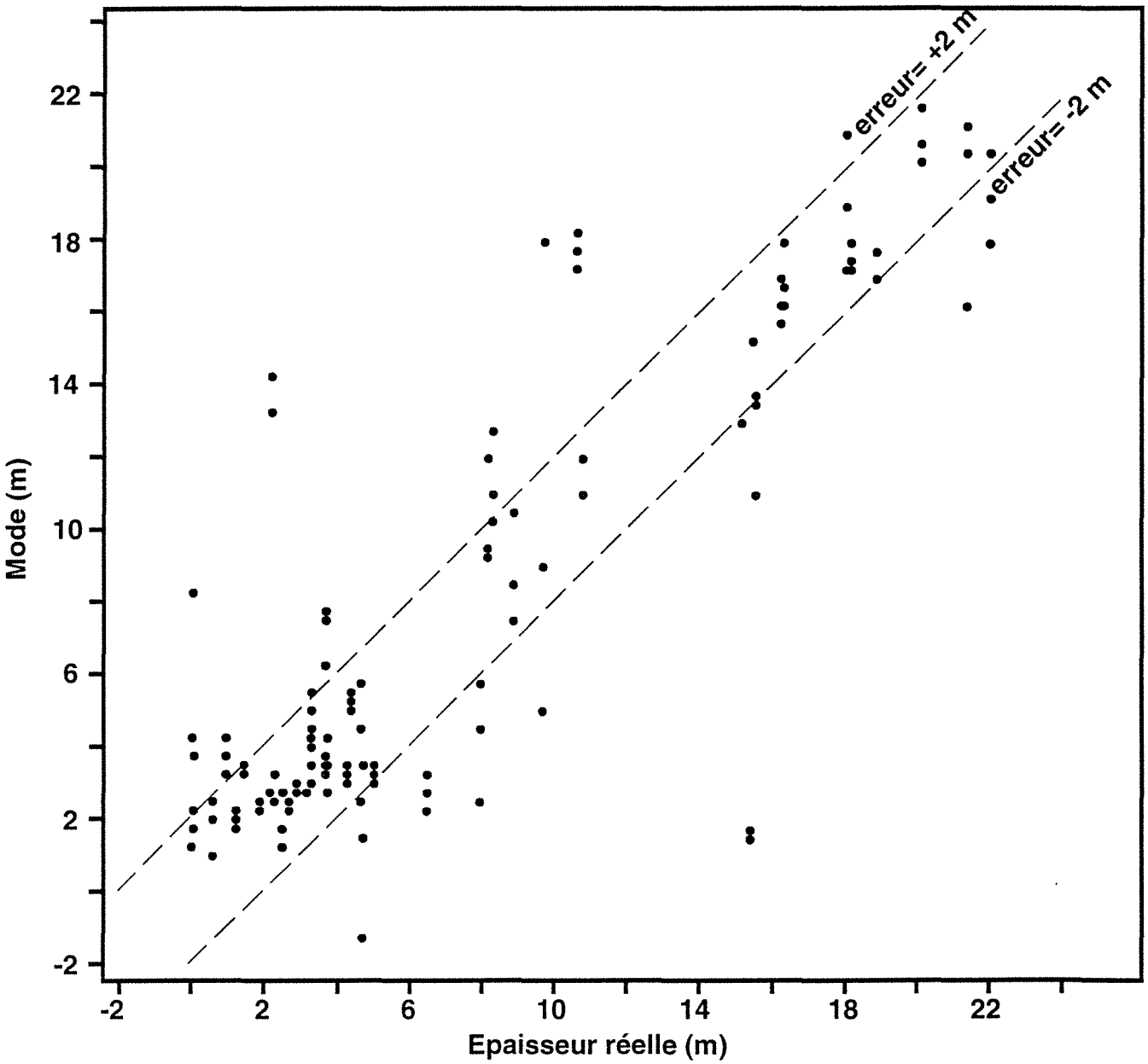


Fig. 50 Prédiction par le mode de l'épaisseur de dolomies vacuolaires aux puits
Prédiction conjointe des épaisseurs de grès et de dolomies vacuolaires

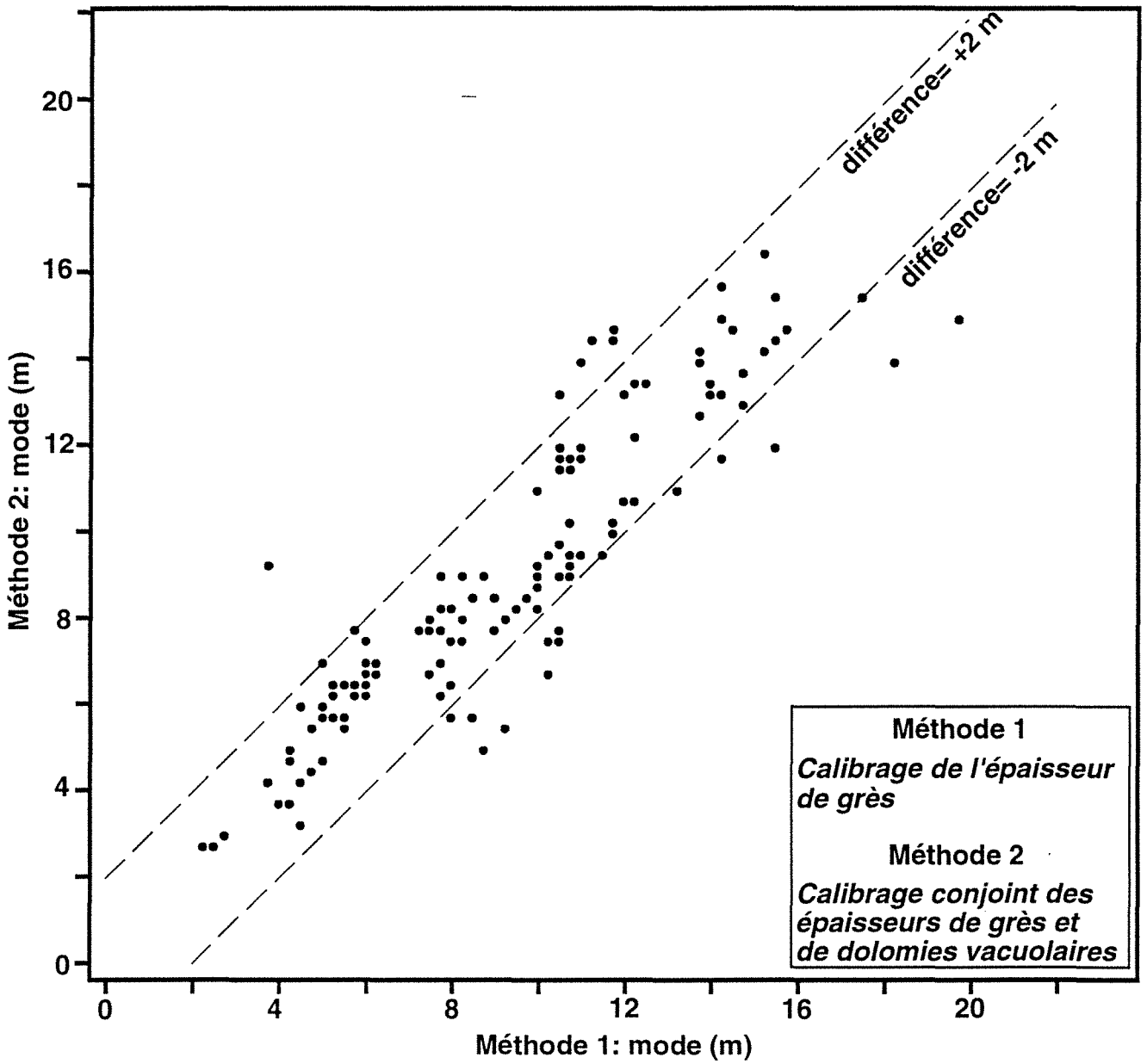


Fig. 51 Comparaison des prédictions de l'épaisseur de grès aux puits

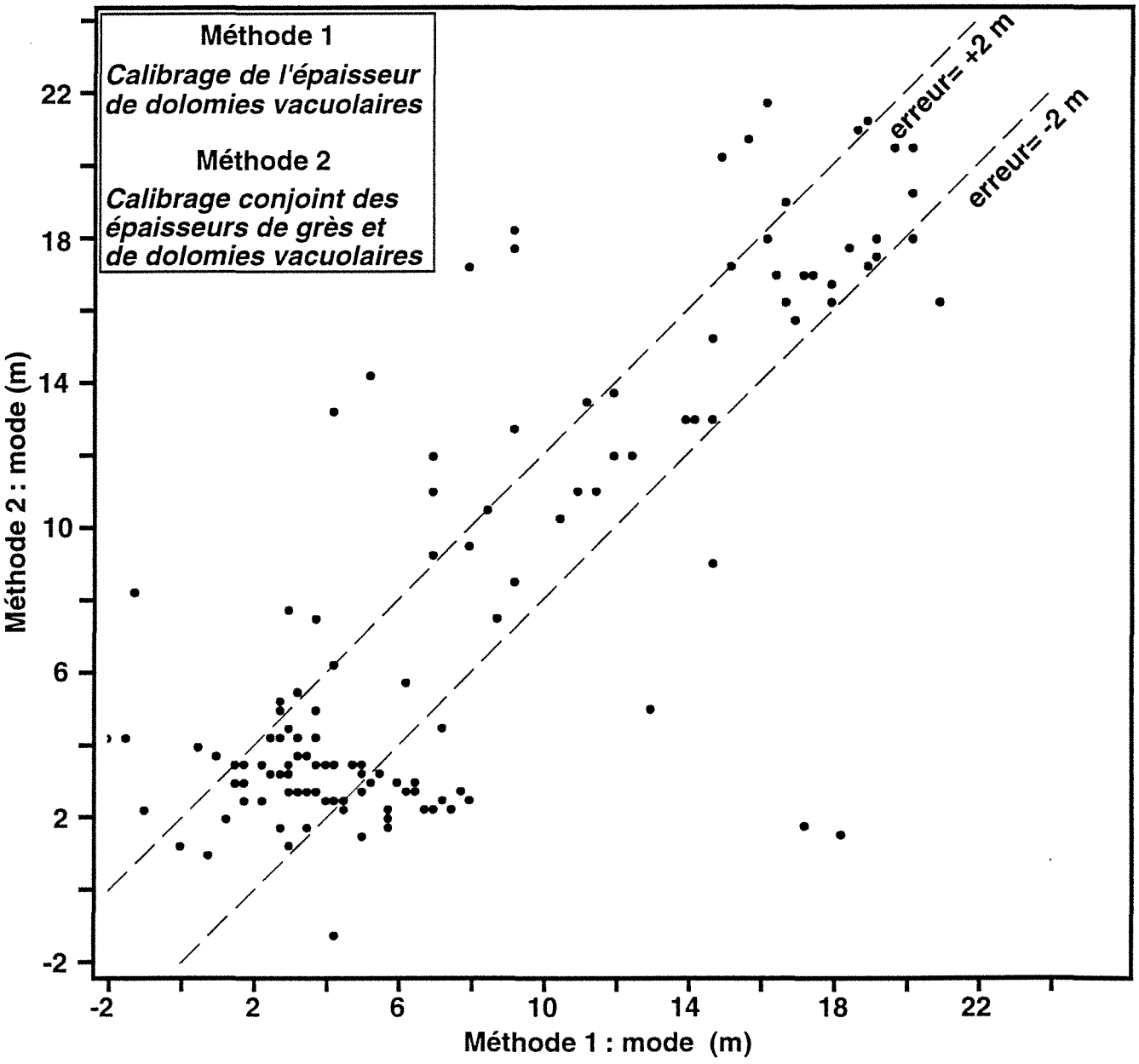


Fig. 52 Comparaison des prédiction de l'épaisseur de dolomies vacuolaires aux puits

Tableau 13 *Comparaison des erreurs de prédiction des propriétés géologiques G_1 et G_2, soit par régression non paramétrique sur G_1 ou G_2, soit par régression non paramétrique sur le couple (G_1,G_2)*

	Régression non paramétrique sur	Erreur moyenne	Erreur quadratique	Erreur maximale négative	Erreur maximale positive	% d'individus avec une erreur ≥ 2 m
Prédiction de G_1	G_1	1.33	3.84	-6.65	7.90	25%
	(G_1,G_2)	1.53	4.64	-7.65	6.33	23%
Prédiction de G_2	G_2	1.63	4.78	-5.80	6.48	26%
	(G_1,G_2)	2.09	10.94	-14.03	12.02	32%

Nous constatons ainsi que la régression non paramétrique appliquée soit pour prédire G1, soit pour prédire le couple (G_1,G_2), fournit des résultats de qualité semblable. Par contre, il est clair que, pour la prédiction de la propriété G2, la régression non paramétrique appliquée au couple (G_1,G_2) fournit des résultats très dégradés, les erreurs de prédiction étant très fortes : elles atteignent 14 mètres, alors que la régression non paramétrique appliquée à la seule prédiction de G_2 fournit des erreurs inférieures à 6.5 mètres.

En fait, en ce qui concerne G_2, nous avions déjà constaté que, par régression non paramétrique appliquée sur G_2, seules des décompositions en classes gaussiennes à au moins 6 classes permettaient une prédiction de très bonne qualité, les décompositions en 4 classes gaussiennes fournissant des résultats très dégradés. Or, dans le cas de la régression non paramétrique appliquée au couple (G_1,G_2), nous avons retenu une décomposition en 4 classes gaussiennes : en effet, les décompositions à plus de 4 classes gaussiennes présentent pour leur plus petite classe un poids vraiment trop faible pour permettre une estimation fiable des paramètres de cette classe en dimension 6. Il semble donc que 4 classes gaussiennes ne soient pas suffisantes pour expliquer la distribution de l'épaisseur de dolomies vacuolaires aux puits, ce qui entraîne une dégradation de la qualité des prédictions.

3.3.2 Prédiction des épaisseurs cumulées de grès et de dolomies vacuolaires entre les puits

Nous avons appliqué la régression non paramétrique sur toutes les traces sismiques du champ, pour prédire conjointement les épaisseurs cumulées de grès et de dolomies vacuolaires. Les figures 53 et 54 fournissent respectivement pour G_1 et G_2 une carte spatiale des prédictions obtenues.

Si on compare la carte correspondant à l'épaisseur cumulée de grès (cf. FIG. 53) avec celle obtenue par régression non paramétrique appliquée uniquement sur G_1 (cf. FIG. 35), nous constatons qu'elles sont semblables. De même, la distribution de l'épaisseur de dolomies vacuolaires (cf. FIG. 54) est similaire à celle obtenue par régression non paramétrique appliquée uniquement sur G_2 (cf. FIG. 43). On constate cependant que la régression non paramétrique appliquée au couple de propriétés géologiques semble fournir des prédictions d'épaisseurs cumulées légèrement moins fortes (moins de traces sismiques du sommet de la structure codées en rouge par exemple).

3.4 Avantage et limitations de la prédiction conjointe d'un couple de propriétés géologiques

Théoriquement, la notion de régression fait intervenir une seule variable à prédire, et est donc définie comme $f(G/s_1,s_2,...)$. Il semble cependant plus avantageux d'appliquer la méthodologie de régression non paramétrique sur un couple de propriétés géologiques : en effet, cela permet de prendre en compte les corrélations pouvant exister entre ces propriétés.

Mais cette approche pose des problèmes sur le plan pratique. En effet, il devient très difficile (sinon impossible) de quantifier les incertitudes associées aux prédictions. De plus, l'espace de calibrage étant de dimension supérieure, cela entraîne des problèmes d'inférence des paramètres des classes gaussiennes au niveau de la décomposition. Comme nous avons vu que la qualité des prédictions est fortement influencée par le choix de la décomposition en classes gaussiennes, ceci se révèle être une limitation importante.

Finalement, compte tenu que la méthodologie de régression non paramétrique appliquée à une seule propriété géologique nous fournit de bons résultats aux puits (confirmés par des blind-tests), il nous semble préférable de prédire séparément plutôt que conjointement plusieurs propriétés géologiques à partir d'attributs sismiques.

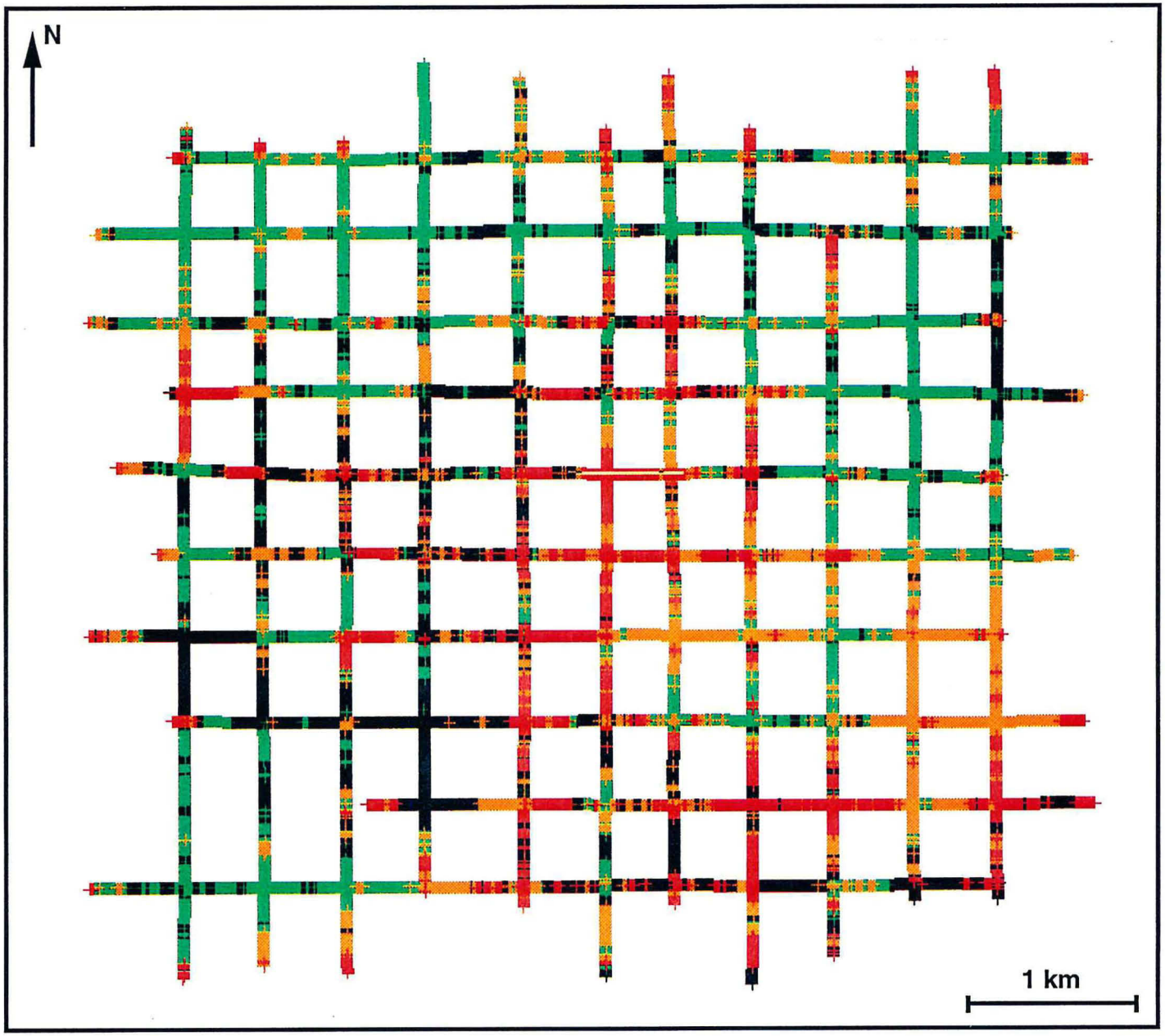


Fig. 53 Distribution spatiale de l'épaisseur de grès
Prédiction conjointe des épaisseurs de grès et de dolomies vacuolaires

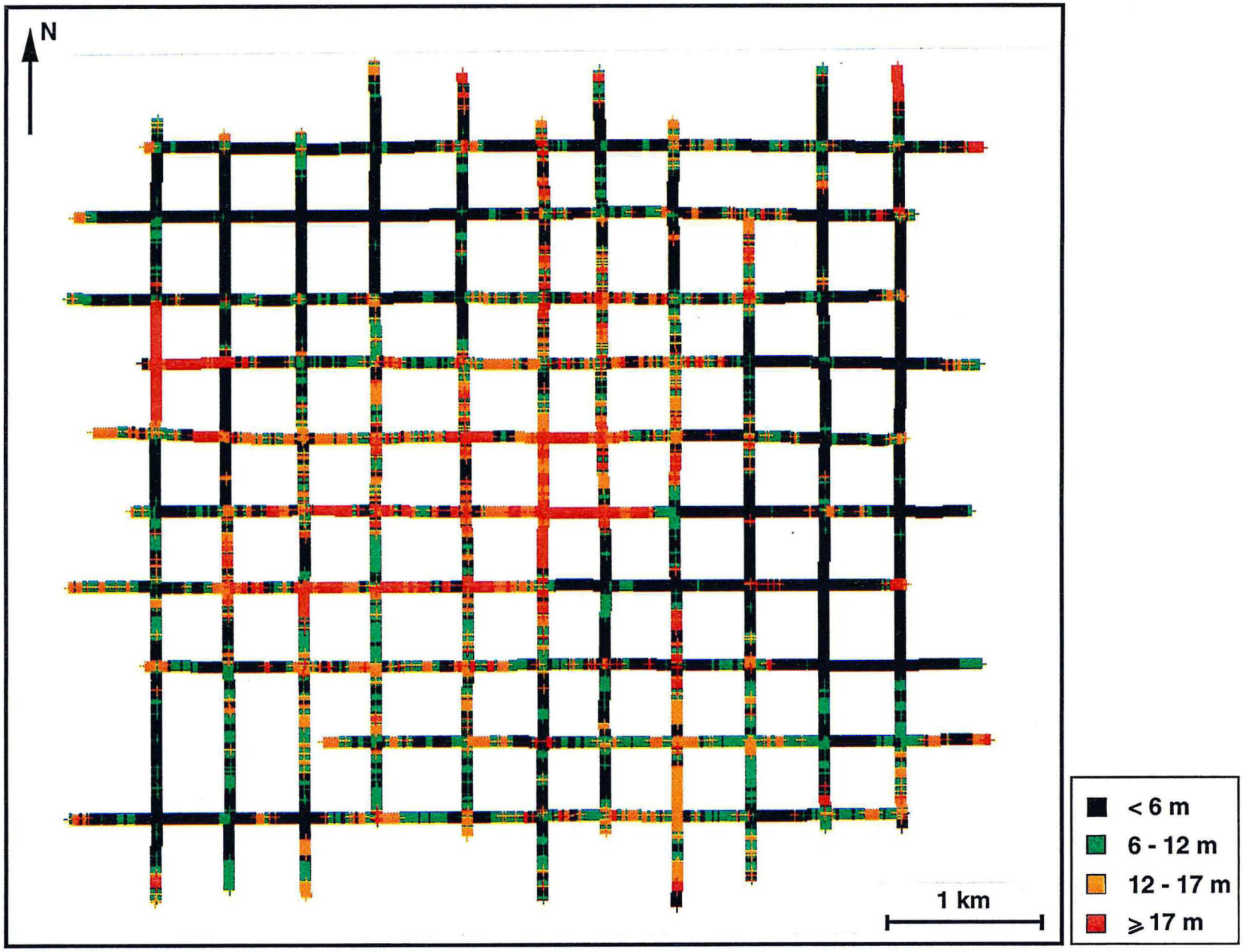


Fig. 54 Distribution spatiale de l'épaisseur de dolomies vacuolaires
Prédiction conjointe des épaisseurs de grès et de dolomies vacuolaires

4 Conclusion

Nous avons appliqué la méthodologie de régression non paramétrique pour caractériser lithologiquement les données sismiques au niveau du réservoir d'un champ pétrolier. Nous avons obtenu des prédictions d'épaisseurs de lithofaciès sur l'ensemble du champ. Après comparaison aux puits des prédictions obtenues avec celles fournies antérieurement par d'autres méthodes de calibrage statistique, il apparaît une amélioration notable de la qualité des prédictions ainsi qu'une meilleure quantification des incertitudes. Par contre, nous avons constaté la forte influence du choix de la décomposition en classes gaussiennes sur la qualité des prédictions.

Par ailleurs, cette méthodologie présente d'autres limitations qui sont en fait communes à toutes les méthodes de calibrage statistique. Tout d'abord, ces méthodes de calibrage ne font pas intervenir de modèle physique ce qui est gênant. Il se pose aussi le problème de la constitution de la population de calibrage, nécessitant l'association de puits et de traces adjacentes : dans le cas d'une sismique 2D, les traces sismiques adjacentes à un puits sont-elles réellement représentative de la géologie à ce puits ? Enfin, ces méthodes ne peuvent être utilisées que sur un champ développé (et non en cours d'acquisition), car elles nécessitent un grand nombre de puits.

Ces problèmes étant posés, la méthodologie de régression non paramétrique nous semble tout de même très intéressante pour améliorer la représentation du réservoir entre les puits, et donc pour permettre l'implantation de nouveaux forages ou pour servir de contraintes supplémentaires dans des simulations stochastiques de réservoir ou de modèle d'écoulement.

CHAPITRE 6

CONCLUSION

Les travaux de ce mémoire s'inscrivent dans le cadre de l'interprétation lithologique des données sismiques. Du fait de la forte densité spatiale des données sismiques, ce type d'interprétation peut être très intéressant pour améliorer la caractérisation des réservoirs pétroliers entre les puits. Généralement, les méthodes utilisées sont des méthodes de calibrage, qui font en fait intervenir un calage de la réponse sismique sur l'information géologique connue aux puits. Dans ce cadre, nous nous sommes intéressés aux méthodes de **calibrage statistique** permettant d'obtenir de l'information quantitative. Entre autres, des méthodes statistiques telles que la régression linéaire ou l'analyse canonique ont déjà été utilisées pour calibrer géologiquement les données sismiques. Mais ces méthodes ne permettent que la recherche de relations linéaires entre données sismiques et données géologiques. Nous avons donc cherché à mettre en place des méthodologies statistiques de calibrage qui ne présentent pas cette limitation.

La base de nos travaux a consisté à développer l'application d'une méthode de **décomposition en classes gaussiennes** sur les données étudiées. De là, deux méthodologies de calibrage ont été développées ; la première fait intervenir l'**analyse canonique sur données codées** et la seconde la **régression non paramétrique**.

En ce qui concerne la **méthode de décomposition en classes gaussiennes**, nous avons retenu une technique de décomposition itérative appelée méthode du maximum de vraisemblance. Nos travaux ont en fait porté sur les différentes façons d'obtenir une décomposition initiale, nécessaire pour cette technique itérative. Nous avons aussi mis en place des critères de qualité portant sur l'adéquation entre la fonction de répartition empirique de la population étudiée et la fonction de répartition issue du mélange gaussien. Nous avons testé la méthode de décomposition sur différentes populations synthétiques ; et en considérant les critères de qualité, nous avons cherché à utiliser cette méthodologie dans le cadre du calibrage par des méthodes statistiques.

Dans un premier temps, nous avons utilisé cette méthode de décomposition dans le cadre de l'**analyse canonique**. En effet, toute décomposition nous permet de coder les données étudiées par codage disjonctif ou probabiliste (par exemple). Nous avons donc étudié le calibrage par analyse canonique sur données codées. En fait, nous avons constaté que la formule de reconstitution des données, permettant la prédiction d'un groupe de variables en fonction d'un second groupe de variables n'est pas applicable en pratique puisqu'elle fait intervenir une inverse généralisée. L'analyse canonique ne peut donc pas être utilisée directement pour faire du calibrage. Cependant, il nous semble que l'analyse canonique sur données codées pourrait servir pour **quantifier les correspondances** entre données géologiques et données sismiques, et

notamment entre faciès géologiques et faciès sismiques dans le cadre du calibrage qualitatif. Ceci n'a pas encore été testé en pratique.

Dans un second temps, nous avons considéré la méthode de **régression non paramétrique** avec utilisation des résultats de la décomposition en classes gaussiennes, dans le cadre d'un calibrage quantitatif. En fait, la méthode de décomposition en classes gaussiennes, appliquée sur la population générée aux puits par les propriétés géologiques à prédire et les attributs sismiques, nous fournit une approximation de la fonction de densité de cette population. En utilisant cette approximation dans le cadre de la régression non paramétrique, il est possible de calculer la fonction de densité conditionnelle des propriétés géologiques connaissant les attributs sismiques. On peut alors prédire les propriétés géologiques en prenant par exemple le mode, l'espérance mathématique... de cette fonction de densité. De plus, cette méthodologie permet de prendre en compte des relations non linéaires existant entre les propriétés géologiques et les attributs sismiques. Enfin, elle permet d'obtenir la quantification complète des incertitudes associées aux prédictions.

Cette méthodologie a été **testée** avec succès **sur un champ pétrolier**, afin de prédire au niveau du réservoir les épaisseurs de grès et de dolomies vacuolaires (les deux lithofaciès présentant les meilleures caractéristiques pétrophysiques) en fonction de quatre attributs sismiques. Ces épaisseurs ont été prédites conjointement ou séparément ; si la prédiction conjointe permet théoriquement de prendre en compte les corrélations pouvant exister entre ces deux propriétés géologiques, il est apparu plus intéressant, en pratique, de prédire séparément ces propriétés. Par ailleurs, nous avons pu comparer aux puits les prédictions obtenues et celles fournies par d'autres méthodes de calibrage statistique, et nous avons constaté une notable amélioration de la qualité de ces prédictions. Nous avons donc confirmé sur un cas réel l'intérêt de cette méthodologie de régression non paramétrique utilisant une décomposition en classes gaussiennes, dans le cadre de la caractérisation lithologique des réservoirs pétroliers.

Cependant, des **travaux complémentaires** pourraient être effectués. Tout d'abord, puisque c'est l'un des objectifs du calibrage géologique des données sismiques, il nous paraît nécessaire d'intégrer les résultats de calibrage, obtenus par régression non paramétrique, dans des simulations probabilistes de réservoir, afin de démontrer l'importance de ces données complémentaires pour améliorer la caractérisation du réservoir.

Il pourrait être aussi intéressant de prendre en compte les corrélations spatiales dans la méthodologie de régression non paramétrique. En théorie, cela est tout à fait possible. Mais dans la pratique, cela suppose (comme pour les méthodes géostatistiques) un grand nombre de puits disponibles, ainsi qu'une bonne répartition spatiale de ces puits, ce qui est rarement le cas.

Enfin, nous avons constaté la forte influence de la décomposition en classes gaussiennes sur la qualité de la prédiction par régression non paramétrique. Nous pourrions donc appliquer des méthodes de statistiques robustes pour estimer les paramètres des classes gaussiennes de façon plus fiable. Et parallèlement, il serait important de pouvoir mettre en place un critère permettant de choisir de façon plus discriminante la décomposition en classes gaussiennes optimale.

De nombreux travaux restent donc possibles dans le cadre de la caractérisation lithologique des réservoirs à partir des données sismiques, la richesse de l'information sismique présentant un grand intérêt en géophysique de gisement au stade de développement des champs pétroliers.

RÉFÉRENCES BIBLIOGRAPHIQUES

ABRAMOWITZ M. and STEGUN I.A., 1972, *Handbook of Mathematical functions with formulas, graphs, and mathematical tables*, Dover Publications, New York, p. 933.

ANGELERI G.P. and CARPI R., 1982, Porosity prediction from seismic data, *Geophysical Prospecting*, v. **30**, p. 580-607.

BOIS P., 1980, Autoregressive pattern recognition applied to the delimitation of oil and gas reservoirs, *Geophysical Prospecting*, v. **28**, p. 572-591.

BOIS P., 1981, Determination of the nature of reservoirs by use of pattern recognition algorithm with prior learning, *Geophysical Prospecting*, v. **29**, p. 687-701.

BRIDGES N.J. and MCCAMMON R.B., 1980, Discrim: a computer program using an interactive approach to dissect a mixture of normal or lognormal distributions, *Computers & Geosciences*, v. **6**, n°1, p. 361-396.

CELEUX G. et DIÉBOLT J., 1986, L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités, *Revue de Statist. Appliquées*, v. **34**, n°2, p. 35-52.

CHARLIER C.V.L. and WICKSELL S.D., 1924, *On the dissection of frequency functions*, Arkiv för Matematik, Astronomi och Fysik, **Bd. 18**, n°6.

COLLOMB G., 1977, Quelques propriétés de la méthode du noyau pour l'estimation non paramétrique de la régression en un point fixé, *C. R. Acad. Sc. Paris, Series A*, t. 285, p. 289-292.

DAY N.E., 1969, Estimating the components of a mixture of normal distributions, *Biometrika*, v. **3**, n°56, p. 463-474.

DEMPSTER A.P., LAIRD N.M. and RUBIN D.B., 1977, Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statist. Soc., Series B*, v. **39**, p. 1-38.

DIDAY E., 1979, *Optimisation en classification automatique*, INRIA, 2 tomes.

DOETSCH G., 1928, Die elimination des dopplereffekts bei spektroskopischen feinstrukturen und exakte bestimming der komponenten, *Zeitschr. f. Phys.*, v. **49**, p. 705-730.

DOYEN P.M., 1988, Porosity from seismic data: a geostatistical approach, *Geophysics*, v. **53**, n°10, p. 1263-1275.

DOYEN P.M., GUIDISH T.M. and DE BUYL M.H., 1988, Lithology prediction from seismic data, a Monte-Carlo approach, 57th Ann. Mtg. and Intern. Exp., Soc. Expl. Geophys., Expanded Abstracts, p. 873-876.

DUMAY J. et FOURNIER F., 1988, Multivariate statistical analyses applied to seismic facies recognition, *Geophysics,* v. **53**, n°9, p. 1151-1159.

EPITALON J.M., 1985, *Indicatrices floues appliquées à la reconnaissance automatique des diagraphies*, Thèse de 3ème cycle, INPL.

EVERITT B.S. and HAND D.J., 1981, *Finite mixture distributions*, Monographs on Statistics and Applied Probability n°15, Chapman and Hall, London New York.

FORGY E.W., 1965, Cluster analysis of multivariate data: efficiency versus interpretability of classifications, *Biometric Society Meetings* (Riverside, California), abstract in *Biometrics*, v. **21**, n°3, p. 768.

FOURNIER F., 1989, *Extraction of quantitative geologic information from seismic data with multidimensional statistical analyses, Part 1: Methodology, Part 2: A case history*, 59th Ann. Mtg. and Intern. Exp., Soc. Expl. Geophys., Expanded Abstracts, p. 726-733.

FOURNIER F., 1992, *Arco/Elf/IFP soft inversion project*, Rapport interne Institut français du Pétrole, n°40131.

FOURNIER F. et DERAIN J.F., 1992-a, *A statistical methodology for the geological calibration of a 2D seismic data set*, 54th Mtg. and Techn. Exhib., European Assoc. Expl. Geophys., Expanded Abstracts, p. 114-115.

FOURNIER F. et DERAIN J.F., 1992-b, *Seismic data integration in reservoir simulations through a multivariate statistical calibration approach*, 62nd Ann. Mtg. and Intern. Exp., Soc. Expl. Geophys., Expanded Abstracts, p. 95-98.

FOURNIER F. et DERAIN J.F., 1994, *A statistical methodology for deriving reservoir properties from seismic data*, Rapport interne Institut français du Pétrole, n°41133.

HAGEN D.C., 1982, The application of principal components analysis to seismic data sets, *Geoexploration*, v. **20**, p. 93-111.

HARDING J.P., 1949, The use of probability paper for the graphical analysis of polymodal frequency distributions, *J. of the Marine Biol. Ass. of the UK*, v. **28**, p. 141-153.

HÄRDLE W. and MARRON J.S., 1985, Optimal bandwidth selection in nonparametric regression function estimation, *The Annals of Statistics* , v. **13**, n°4, p. 1465-1481.

HASSELBLAD V., 1966, Estimation of parameters for a mixture of normal distributions, *Technometrics*, v. **8**, p. 431-444.

HOLGERSSON M. and JORNER U., 1978, Decomposition of a mixture into normal components: a review, *Int. J. Bio-Medical Computing*, v. **9**, p. 367-392.

HOTELLING H., 1936, Relations between two sets of variables, *Biometrika*, v. **28**, p. 321377.

JOSEPH C., FOURNIER F. et ROYER J.J., 1993, *Seismic data calibration in terms of reservoir properties with a multivariate gaussian segmentation technique*, 63rd Ann. Mtg. and Intern. Exp., Soc. Expl. Geophys., Expanded Abstracts, p. 285-288.

JUSTICE J.H., HAWKINS D.J. and WONG G., 1985, Multidimensional attribute analysis and pattern recognition for seismic interpretation, *Pattern Recognition*, v. **18**, n°6, p. 391-407.

KENDALL M. and STUART A., 1979, *The advanced theory of statistics*, v. **2**, Griffin, London New York.

KUBICHEK R.F. and QUINCY E.A., 1985, Statistical modeling and feature selection for seismic pattern recognition, *Pattern Recognition*, v. **18**, n°6, p. 441-448.

LENDZIONOWSKI V., WALDEN A.T. and WHITE R.E., 1990, Seismic character mapping over reservoir intervals, *Geophysical Prospecting*, v. **38**, p. 951-969.

MCQUEEN J.B., 1967, Some Methods for Classification Analysis of Multivariate Observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, v. **1**, p. 281-297.

MALLET J.L., EPITALON J.M. et DE BEAUCOUR , 1985, *Discrimination non linéaire par indicatrices floues. Application à la reconnaissance des formes*, Congrès INRIA, Versailles, Proceedings p. 267-274.

MALLET J.L., 1988, *Analyse des données*, Cours INPL, tome 2.

MEZGHACHE H., 1989, *Cartographie automatique et interprétation géostatistique d'une campagne de prospection géochimique sur sol. Application à la zone mercurielle nord numidique (Algérie)*, Thèse de 3ème cycle, INPL.

MITCHUM R.M. *et al.*, 1977, A.A.P.G. Memoir n°26.

PEARSON K., 1894, Contribution to the mathematical theory of evolution, *Phil. Trans. A*, v. **185**, p. 71-110.

PREISENDORFER R.W., 1988, *Principal Component Analysis in Meteorology and Oceanography*, Developments in Atmospheric Sciences n°17, Elsevier, Amsterdam New York.

RAO C.R., 1948, The utilization of multiple measurements in problems of biological classification, *J. Royal Statist. Soc., Series B*, v. **10**, p. 159-203.

RAO C.R., 1965, *Linear Statistical Inference and its Applications*, John Wiley & Sons Inc., London New York Sidney.

SAPORTA G., 1990, *Probabilités, analyse des données et statistique*, Technip, Paris.

SHEATHER S.J. and JONES M.C., 1991, A reliable Data-based Bandwidth Selection Method for Kernel Density Estimation, *J. Royal Statist. Soc., Series B*, v. **53**, n°3, p. 683-690.

SILVERMAN B.W., 1986, *Density estimation for statistics and data analysis*, Monographs on Statistics and Applied Probability n°26, Chapman and Hall, London New York.

SINCLAIR A.J., 1976, *The Application of probability plots to mineral exploration*, Assoc. of Exploration Geoch., v. **4**.

SINVHAL A. and KHATTRI K., 1983, Application of seismic reflection data to discriminate subsurface lithostratigraphy, *Geophysics*, v. **48**, n°11, p. 1498-1513.

SINVHAL A., KHATTRI K., SINVHAL H. and AWASTHI A.K., 1984, Seismic indicators of stratigraphy, *Geophysics*, v. **49**, n°8, p. 1196-1212.

STANULONIS S.F. and TRAN H.V., 1992, Method to determine porosity-thickness directly from 3D seismic amplitude within the Lisburne carbonate pool, Prudoe Bay, *Geophysics: The Leading Edge of Exploration*, v. **11**, n°1, p. 14-20.

THADANI S.G., ALABERT F. et JOURNEL A.G., 1987, *An integrated geostatistical/pattern recognition technique for characterization of reservoir spatial variability*, 57th Ann. Mtg. and Intern. Exp., Soc. Expl. Geophys., Expanded Abstracts, p. 372-375.

WOLFE J.H., 1970, Pattern clustering by multivariate mixture analysis, *Multivariate Behavioral Research*, v. **5**, p. 329-350.

ANNEXE A

Résultats des tests de la méthode de décomposition en classes gaussiennes sur des données synthétiques

NOTATIONS

- **PMIN** : poids minimal d'une classe pour qu'elle soit conservée.

- **Init Mono** : initialisation monovariable

  FR

  Application de la méthode de Harding automatisée sur la fonction de répartition empirique de chaque variable.

  FRE *h1=... h2 =...*

  Application de la méthode de Harding automatisée sur la fonction de répartition intégrée à partir de l'estimation de la densité par la méthode des noyaux pour chaque variable.
  h1 et *h2* sont respectivement les tailles des noyaux d'Épanechnikov utilisées pour les variables X^1 et X^2.

  FRL *t1=... t2=... p=...*

  Application de la méthode de Harding automatisée sur la fonction de répartition empirique lissée par moyennes mobiles.
  t1 et *t2* sont les tailles respectives des fenêtres utilisées pour les variables X^1 et X^2.
  p est le pourcentage de recouvrement des fenêtres quelle que soit la variable.

  MEff *j*

  Recherche de *j* classes de même effectif sur chaque variable.

- **Init Multi** : initialisation multivariable

  5
  10
  15
  20

  } classes : partition de la population en 5, 10, 15 ou 20 classes *a priori*.

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol0	1	0.5	toutes	1.
sol1	2	0.15	**Init Mono** FRE h1=0.14 h2=0.22	0.76 0.24
		0.1	**Init Mono** MEff 10 FRE h1=0.14 h2=0.22 FRE h1=0.5 h2=0.8 FRE h1=2. h2=2. FRE h1=2. h2=3.2 FRL t1=0.2 t2=0.3 p=0 FRL t1=0.2 t2=0.3 p=95 **Init Multi** 5 classes	
		0.05	**Init Mono** FRE h1=0.5 h2=0.8 FRE h1=1. h2=1.6	
sol2	2	0.1	**Init Multi** 10 classes	0.80 0.20
		0.05	**Init Multi** 5 classes	

Tableau A-1-1 : résultats du fichier 1

Tableau A-1-2 : résultats du fichier 1 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol3	3	0.1	**Init Mono** MEff 5 FR FRE h1=0.05 h2=0.05 FRE h1=0.05 h2=0.08 FRE h1=0.5 h2=0.5 FRE h1=1. h2=1. FRE h1=1. h2=1.6 **Init Multi** 15 / 20 > classes	0.72 0.16 0.12
		0.05	**Init Mono** MEff 5 MEff 10 FR FRE h1=0.05 h2=0.05 FRE h1=0.05 h2=0.08 FRE h1=0.14 h2=0.22 FRE h1=0.5 h2=0.5 FRE h1=2. h2=3.2 **Init Multi** 15 / 20 > classes	

Tableau A-1-3 : résultats du fichier 1 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol4	3	0.05	**Init Mono** FRE h1=2. h2=2. FRL t1=0.2 t2=0.3 p=0 FRL t1=0.2 t2=0.3 p=95 **Init Multi** 10 classes	0.59 0.35 0.06
sol5	5	0.05	**Init Mono** FRE h1=1. h2=1.	0.27 0.23 0.22 0.18 0.10

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol0	1	0.5	toutes	1.
sol1	3	0.15	**Init Mono** FRE h1=0.92 h2=1.08	
		0.1	**Init Mono** MEff 10 FRE h1=1.25 h2=1.5 FRE h1=1.5 h2=1.5 FRE h1=1.75 h2=2. FRE h1=2. h2=2. FRL t1=0.25 t2=0.25 p=0 FRL t1=0.25 t2=0.25 p=96 **Init Multi** 5, 10, 15, 20 > classes	0.33 0.33 0.33
sol2	4	0.1	**Init Mono** MEff 5 FR FRE h1=0.1 h2=0.1 FRE h1=0.5 h2=0.5 FRE h1=0.5 h2=0.7 FRE h1=0.92 h2=1.08	0.33 0.33 0.21 0.12

Tableau A-2-2 : résultats du fichier 2 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol3	4	0.05	**Init Multi** 5 classes	0.33 0.33 0.27 0.06
sol4	5	0.05	**Init Mono** MEff 5 FRE h1=0.5 h2=0.7 FRE h1=0.92 h2=1.08 FRE h1=1.75 h2=2. **Init Multi** 15 classes	0.33 0.27 0.21 0.12 0.06
sol5	5	0.05	**Init Mono** MEff 10 FRE h1=0.1 h2=0.1 FRE h1=1.25 h2=1.5 FRE h1=1.5 h2=1.5 FRE h1=2. h2=2. **Init Multi** 10 > classes 20 >	0.33 0.27 0.25 0.08 0.06

Tableau A-2-3 : résultats du fichier 2 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol6	6	0.05	**Init Mono** FR FRE h1=0.5 h2=0.5	0.33 0.27 0.19 0.08 0.07 0.06
sol7	7	0.05	**Init Mono** FRL t1=0.25 t2=0.25 p=0 FRL t1=0.25 t2=0.25 p=96	0.33 0.25 0.11 0.10 0.08 0.07 0.06

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol0	1	0.5	toutes	1.
sol1	3	0.15	**Init Mono** FRE h1=0.98 h2=1.35	
		0.1	**Init Mono** MEff 5 MEff 10 FR FRE h1=0.1 h2=0.1 FRE h1=0.5 h2=0.5 FRE h1=0.5 h2=0.7 FRE h1=0.98 h2=1.35 FRE h1=1.5 h2=1.5 FRE h1=1.5 h2=2. FRE h1=2.5 h2=2.5 FRL t1=0.25 t2=0.25 p=0 FRL t1=0.25 t2=0.25 p=96 **Init Multi** 5, 10, 15, 20 classes	0.35 0.33 0.32
sol2	4	0.05	**Init Mono** FRL t1=0.25 t2=0.25 p=0 FRL t1=0.25 t2=0.25 p=96	0.35 0.33 0.26 0.06

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol3	4	0.05	**Init Multi** 5 classes	0.33 0.31 0.28 0.08
sol4	5	0.05	**Init Mono** MEff 5 FRE h1=0.5 h2=0.5 FRE h1=0.5 h2=0.7 FRE h1=0.98 h2=1.35 FRE h1=1.5 h2=2. FRE h1=2.5 h2=2.5 **Init Multi** 10 / 20 > classes	0.33 0.29 0.25 0.07 0.06
sol5	5	0.05	**Init Mono** FRE h1=1.5 h2=1.5	0.33 0.23 0.23 0.13 0.08
sol6	6	0.05	**Init Mono** FR **Init Multi** 15 classes	0.33 0.22 0.21 0.09 0.09 0.06

Tableau A-3-3 : résultats du fichier 3 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol7	6	0.05	**Init Mono** MEff 10	0.33 0.22 0.17 0.11 0.09 0.08
sol8	6	0.05	**Init Mono** FRE h1=0.1 h2=0.1	0.33 0.23 0.16 0.12 0.09 0.07

Tableau A-4-1 : résultats du fichier 4

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol0	1	0.5	toutes	1.
sol1	4	0.1	**Init Mono** FRE h1=2. h2=2. FRE h1=3. h2=2. **Init Multi** 10 classes	0.26 0.25 0.25 0.24
sol2	5	0.15	**Init Mono** FRE h1=0.52 h2=0.34	0.22 0.22 0.20 0.19 0.17
sol3	5	0.1	**Init Mono** FR	0.26 0.25 0.18 0.16 0.15
sol4	5	0.1	**Init Multi** 5 / 15 classes	0.26 0.24 0.18 0.16 0.16
		0.05	**Init Multi** 5 classes	

Tableau A-4-2 : résultats du fichier 4 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol5	6	0.1	**Init Mono** MEff 5 MEff 10 FRE h1=0.1 h2=0.1 FRE h1=0.25 h2=0.1 FRE h1=0.25 h2=0.25 FRE h1=0.52 h2=0.34 FRE h1=1. h2=0.7 FRE h1=1. h2=1. FRE h1=1.5 h2=1.5 FRL t1=0.1 t2=0.1 p=0 FRL t1=0.1 t2=0.1 p=96 **Init Multi** 20 classes	0.18 0.18 0.17 0.16 0.155 0.155
sol6	7	0.05	**Init Mono** MEff 10	0.18 0.17 0.15 0.14 0.13 0.12 0.11
sol7	9	0.05	**Init Multi** 10 classes	0.14 0.13 0.12 0.11 0.11 0.10 0.10 0.10 0.09

Tableau A-4-3 : résultats du fichier 4 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol8	9	0.05	**Init Mono** FRE h1=0.25 h2=0.25 FRE h1=0.52 h2=0.34 FRE h1=3. h2=2.	0.13 0.13 0.12 0.12 0.12 0.10 0.10 0.09 0.09
sol9	9	0.05	**Init Mono** MEff 5 **Init Multi** 20 classes	0.14 0.13 0.13 0.12 0.11 0.10 0.10 0.09 0.08
sol10	10	0.05	**Init Mono** FRE h1=0.1 h2=0.1 FRE h1=0.25 h2=0.1 FRE h1=1. h2=0.7 FRE h1=1.5 h2=1.5 FRE h1=2. h2=2. FRL t1=0.1 t2=0.1 p=0 FRL t1=0.1 t2=0.1 p=96	0.12 0.11 0.10 0.10 0.10 0.10 0.10 0.09 0.09 0.09

Tableau A-4-4 : résultats du fichier 4 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
soll1	10	0.05	**Init Mono** FR FRE h1=1. h2=1.	0.13 0.12 0.10 0.10 0.10 0.10 0.09 0.09 0.09 0.08
soll2	11	0.05	**Init Multi** 15 classes	0.12 0.11 0.10 0.10 0.10 0.10 0.09 0.09 0.085 0.055 0.05

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol0	1	0.5	toutes	1.
sol1	2	0.15	**Init Mono** FRE h1=0.26 h2=0.40	0.71 0.29
		0.1	**Init Mono** MEff 5 MEff 10 FR FRE h1=0.05 h2=0.05 FRE h1=0.1 h2=0.1 FRE h1=0.1 h2=0.2 FRE h1=0.26 h2=0.4 FRE h1=0.5 h2=0.5 FRE h1=0.5 h2=0.75 FRE h1=1. h2=1. FRE h1=1. h2=1.25 FRE h1=2. h2=2. FRL t1=0.01 t2=0.02 p=0 FRL t1=0.01 t2=0.02 p=75 **Init Multi** 5 10 15 20 > classes	

Tableau A-5-2 : résultats du fichier 5 (suite)

Solution	Nombre de classes	PMIN	Initialisation	Poids des classes
sol2	3	0.05	**Init Mono** MEff 10 **Init Multi** 20 classes	0.52 0.39 0.09
sol3	4	0.05	**Init Mono** MEff 5 FR FRE h1=0.05 h2=0.05 FRE h1=0.1 h2=0.1 FRE h1=0.1 h2=0.2 FRE h1=0.26 h2=0.4 FRE h1=0.5 h2=0.5 FRE h1=0.5 h2=0.75 FRE h1=1. h2=1. FRE h1=1. h2=1.25 FRE h1=2. h2=2. FRL t1=0.01 t2=0.02 p=0 FRL t1=0.01 t2=0.02 p=75 **Init Multi** 5, 10, 15 classes	0.42 0.30 0.21 0.07

ANNEXE B

Régression dans le cas d'un couple de variables suivant un loi normale bidimensionnelle

Démontrons que, dans le cas où X et Y sont deux variables suivant une loi bidimensionnelle, la régression de Y en X est une régression linéaire (cf. FIG. B-1).

Supposons que le couple de variables (X,Y) suive une loi normale de moyenne M et de matrice de variance-covariance Σ, où :

- $M = \begin{bmatrix} m_x \\ m_y \end{bmatrix}$
- $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$ et $\begin{cases} \rho = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \\ \sigma_x^2 = \mathrm{Var}(X) \\ \sigma_y^2 = \mathrm{Var}(Y) \end{cases}$

Alors,

$$\det\Sigma = \left(1-\rho^2\right)\sigma_x^2\,\sigma_y^2$$

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} \dfrac{1}{\sigma_x^2} & -\dfrac{\rho}{\sigma_x\sigma_y} \\ -\dfrac{\rho}{\sigma_x\sigma_y} & \dfrac{1}{\sigma_y^2} \end{bmatrix} \qquad \text{définie si } \begin{cases} \rho \neq \pm 1 \\ \sigma_X \neq 0 \\ \sigma_Y \neq 0 \end{cases}$$

Dans le cas général, la régression de Y en X s'écrit :

$$E(Y/X) = \int_R y\, f(y/X)\,dy = \int_R y\,\frac{f(X,y)}{f(X,.)}\,dy$$

Il nous faut donc calculer $f(y/X)$, soit chacun des termes $f(X,y)$ et $f(X,.)$. Le couple (X,Y) suivant une loi normale, on a :

$$f(X,y) = \frac{1}{2\pi\,\sigma_X\sigma_Y\left(1-\rho^2\right)} \cdot \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\cdot\left[\left(\frac{X-m_X}{\sigma_X}\right)^2 + \left(\frac{Y-m_Y}{\sigma_Y}\right)^2 - 2\rho\frac{\left(X-m_X\right)\left(Y-m_Y\right)}{\sigma_X\,\sigma_Y}\right]\right]$$

et

$$f(X,.) = \frac{1}{\sqrt{2\pi}\,\sigma_X}\cdot\exp\left[-\frac{1}{2}\left(\frac{X-m_X}{\sigma_X}\right)^2\right]$$

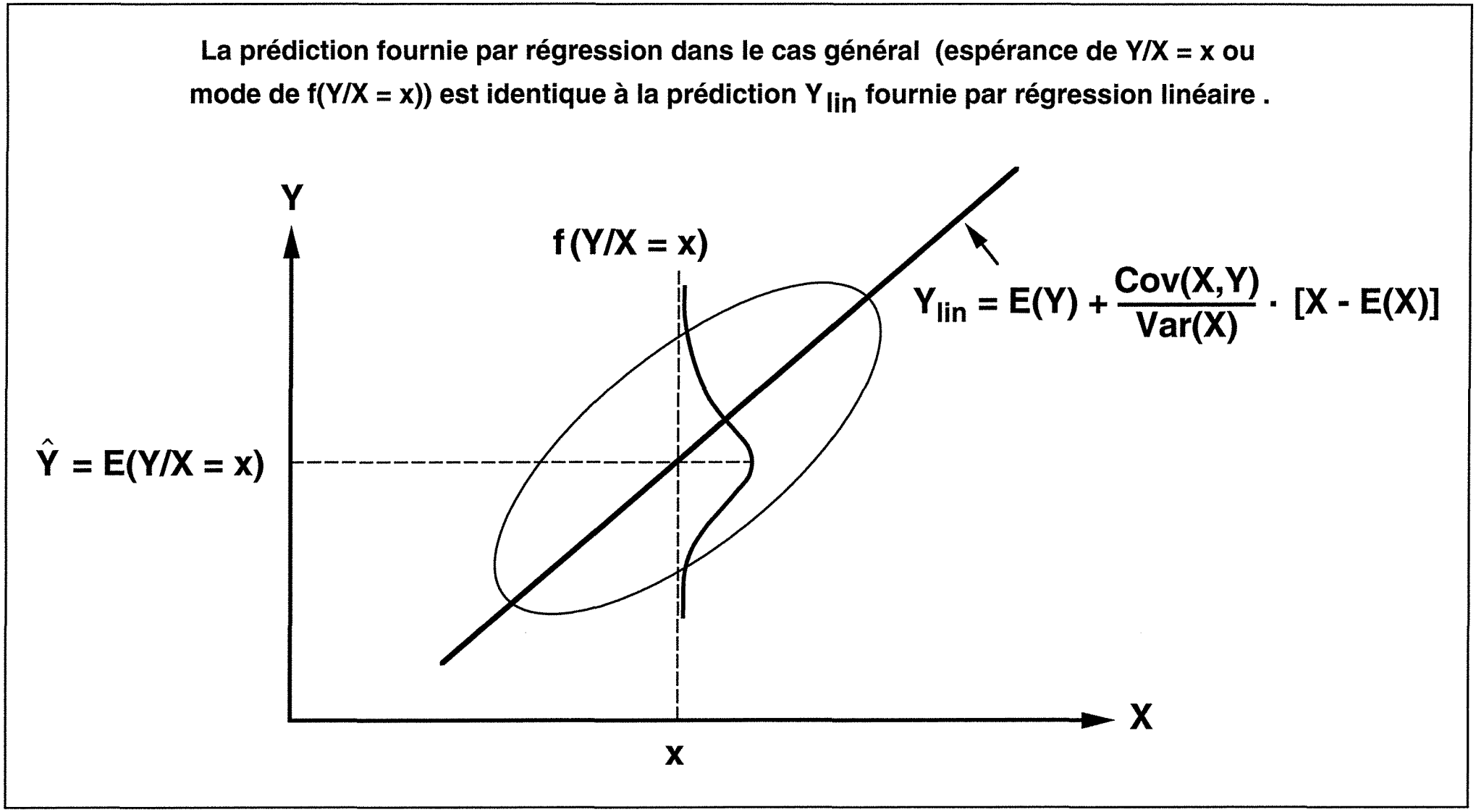


Fig. B-1 Cas de la régression sur deux variables suivant une loi gaussienne bidimensionnelle

Alors,

$$f(y/X) = \frac{f(X,y)}{f(X,.)}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\cdot\left[\rho^2\left(\frac{X-m_X}{\sigma_X}\right)^2+\left(\frac{Y-m_Y}{\sigma_Y}\right)^2-2\rho\frac{(X-m_X)(Y-m_Y)}{\sigma_X\,\sigma_Y}\right]\right]$$

A $(X=x)$ fixé, on constate que, quel que soit y,

$$f(y/X=x)=\frac{1}{\sqrt{2\pi}\,\hat{\sigma}}\cdot\exp\left[-\frac{1}{2}\left(\frac{y-\hat{m}}{\hat{\sigma}}\right)^2\right]$$

avec :

- $\hat{m}=m_y-\rho\frac{\sigma_Y}{\sigma_X}(x-m_X)$
- $\hat{\sigma}^2=(1-\rho^2)\sigma_Y^2$

Donc la variable $(Y/X=x)$ suit une loi normale unidimensionnelle de moyenne $\hat{m}$ et de variance $\hat{\sigma}^2$. De ce fait, son espérance $E(Y/X=x)$ est confondue avec le mode de sa fonction de densité $f(Y/X=x)$, que nous allons évaluer car il est plus simple à calculer.

$$\frac{\partial f(Y/X=x)}{\partial Y}=0=\frac{1}{\sqrt{2\pi}\,\hat{\sigma}}\cdot\exp\left[-\frac{1}{2}\left(\frac{Y-\hat{m}}{\hat{\sigma}}\right)^2\right]\cdot\left(\frac{\hat{m}-Y}{\hat{\sigma}^2}\right)$$

Alors, le mode de $f(Y/X=x)$ (ou l'espérance mathématique $E(Y/X=x)$) est défini par :

$$Y=\hat{m}=m_y-\rho\frac{\sigma_Y}{\sigma_X}(x-m_X)$$

D'où :

$$\boxed{E(Y/X=x)=E(Y)-\frac{Cov(X,Y)}{Var(X)}\cdot[x-E(X)]}$$

si on reprend les notations en fonction de $Cov(X,Y)$ et $Var(X)$. On retrouve finalement l'**équation de la régression linéaire** de Y en $(X=x)$.

ANNEXE C

Seismic data calibration in terms of reservoir properties with a multivariate gaussian segmentation technique. C. JOSEPH, F. FOURNIER et J.J. ROYER, 1993.

Seismic data calibration in terms of reservoir properties with a multivariate gaussian segmentation technique

C. Joseph and F. Fournier, Institut Français du Pétrole ;
J.J. Royer, CRPG/CNRS, France

SUMMARY

This paper presents a statistical calibration technique to derive geological properties from seismic traces at the reservoir level.The calibration method is based on a gaussian segmentation of the calibration points in the multivariate space generated by the geological properties and seismic attributes measured at the wells and their neighboring traces. The multivariate probability density function is viewed as a finite mixture of gaussian components. This approximation of the probability density function can be used to compute the conditional distribution of the geological properties given the measurement of a vector S of seismic attributes. Various parameters characterizing the conditional distribution can be derived to quantify the geological prediction and its uncertainty at the trace under consideration. The proposed calibration technique also allows to account for non linear relationships between the geological and the seismic attributes. The method is illustrated on the geological calibration of a 2D seismic data set.

INTRODUCTION

The use of seismic data to improve the knowledge of the reservoir properties between wells is a key point for a reliable geological model and therefore for a good assessment of the reservoir behavior during production.

Seismic data can be directly integrated in the reservoir model (Bortoli *et al.*, 1992). However, because of the complexity of the relationship between the reservoir properties and its seismic response, it is often more reliable to firstly extract meaningful geological information from seismic data, secondly integrate it in the reservoir model (Fournier and Derain, 1992).

Geological information extraction from seismic data can be achieved with statistical (Fournier and Derain, 1992) or geostatistical (Doyen, 1988) calibration techniques between well properties at the reservoir level and seismic attributes of the neighboring traces. Calibration techniques should be able to easily handle multivariate problems, to account for possibly non linear relationships between the seismic and geological parameters and to provide a reliable quantification of the uncertainties attached to the predicted values. Also, computation efficiency is very important with the increasing volumes of the seismic data sets to be studied.

This paper presents a calibration technique based on an approximation of the joint geological/seismic multivariate probability density function with a mixture of gaussian populations. The technique will be discussed then illustrated on the geological calibration of a 2D marine seismic data set.

CALIBRATION WITH A MULTIVARIATE GAUSSIAN SEGMENTATION

The calibration technique is based on the segmentation of the population into sub-sets of gaussian clusters in the multivariate space generated by the geological and seismic properties measured at the wells and their associated traces. This segmentation provides an approximation of the joint seismic/geological probability density function (p.d.f.) with a finite mixture of multivariate gaussian densities.

The technique we have developed is based on an extension of the method proposed by Harding (1949) and used by Sinclair (1976) to the multivariate case. Initial thresholds on each variable are determined through the computation of the inflexion points of the empirical cumulative density function (c.d.f.). These 1D thresholds allow to form initial clusters in the multivariate space. Then a gaussian probability density function is fitted on each cluster. An optimization technique (Everitt and Hand, 1981, Mezghache, 1989) is carried out in the multivariate space to modify the thresholds and possibly suppress classes.

The approximation $\hat{F}$ of the empirical cumulative density function is given by :

$$\hat{F}(x) = \sum_{j=1}^{K} p_j \hat{F}_j(x)$$

where K is the number of gaussian components, $\hat{F}_j$ is the gaussian c.d.f. of the jth component, and p_j is the jth estimated component weight.

In this method, special attention has to be paid to a possible influence of the initial solution on the final optimum which is found and also to a quantification of the approximation quality with the analysis of various criteria such as :

$$C_1 = \sum_{i=1}^{N} \left[F(x_i) - \hat{F}(x_i)\right]^2 \text{ or } C_2 = \sum_{i=1}^{N} \left|F(x_i) - \hat{F}(x_i)\right|$$

where $F(x_i)$ is the empirical c.d.f. value at the calibration point x_i and N is the number of calibration points.

In case where the approximation $\hat{F}$ is enough satisfying, it can be used for an inference of the geological information G from the seismic information S measured at a particular trace location (Figure 1) with the computation of the geological conditional distribution probability density function :

$$f(G \text{ given } S = s) = \frac{f(G, S = s)}{f(S = s)}$$

$$f(G \text{ given } S = s) = \frac{\sum_{j=1}^{K} p_j f_j(G, S = s)}{\sum_{j=1}^{K} p_j f_j(S = s)}$$

Notice that the denominator is the seismic approximate marginal distribution p.d.f. at point s.

Various parameters computed on the conditional distribution can be used to characterize the geological prediction at the trace s such as the expected value, the mode or quantiles.

The proposed technique applied to calibration purposes allows to easily handle non linear relationships betwen the sets of seismic and geological attributes and provides a quantification of the uncertainties since the full p.d.f. is available. The method is also efficient from the computing time point of view even with a highly multivariate problem.

Difficulties are linked to a possible lack of fit of the approximated c.d.f. to the empirical one. From a pratical point of view, enough calibration points (wells and associated traces) should be available for a reliable decomposition. In case of a multivariate geological calibration (p geological properties), it should also be kept in mind that the geological conditional distribution p.d.f. is a p dimension surface, the analysis of which can be heavy to carry out.

APPLICATION TO THE LITHOLOGICAL CALIBRATION OF A 2D MARINE SEISMIC DATA SET

The 70 m thick reservoir under study consists of several levels of tidal to supratidal sandstones, dolomitic sandstones and vuggy dolomites, interbedded with dolomicrites and anhydritic shales. The 44 wells available on the field are interpreted in terms of six major lithofacies. The lithofacies distribution rules the petrophysical properties. In particular, the sandstones are associated with high porosities. Therefore their thicknesses cumulated on the whole reservoir interval is the parameter chosen to illustrate the seismic data calibration process with the multivariate gaussian segmentation.

Four seismic attributes are extracted from the seismic traces on a time window at the reservoir level. Each well is associated with three neighboring traces. The relationship between the sandstone thickness and the four seismic attributes appears non linear as it is illustrated on figure 2

The gaussian segmentation of the probability density function was carried out in a 5 dimension space generated by the sandstone thickness and the four seismic attributes. The optimal decomposition with regard to the previously discussed criteria corresponds to a mixture of six multivariate gaussian populations. The marginal distribution p.d.f.s of the approximation and its six components for the sandstone thickness are represented on figure 3 as also the histogram of this variable.

The process of calibration by computing the probability density function of the sandstone thickness conditioned by the seismic attributes is illustrated at four wells (Figure 4). The three curves correspond to the p.d.f. at the three traces associated to the well. The p.d.f.s are often multimodal but in most cases, a mode is clearly prevailing. On this data set, the modes are very close to the sandstone thickness actual values except for a few traces (Figure 5). On the conditional distribution p.d.f. of the sandstone thickness which was estimated at each seismic trace, the mode and interquartile range were computed. The sandstone thickness mode values derived from the whole survey traces are grey coded on the map of figure 6.

CONCLUSION

The calibration technique we have developped is based on an approximation of the joint geological/seismic p.d.f. with a mixture of a finite number of multivariate gaussian components. This segmentation allows to take into account non linear trends between the sets of properties under study. The technique is also efficient for multivariate calibration problems provided a sufficient number of wells is available for the reliability of the statistical inferences which are requested. Finally, various parameters can be computed to quantify the uncertainties attached to the predicted values, because the approximated conditional distribution p.d.f. is fully available.

These advantages were corroborated by the good results we obtained for the geological calibration of the traces of a 2D seismic survey in terms of the lithofacies thicknesses, at the reservoir level.

REFERENCES

Bortoli, L.J., Alabert, F., Haas, A., Journel, A.G., 1992, Constraining stochastic images to seismic data, to be published in Proceedings of the 4th International Geostatistics Congress, Troia, Portugal, ed. A.Soares, Kluwer Publishers.

Fournier, F., and Derain, J.F., 1992, Seismic data integration in reservoir simulations through a multivariate statistical calibration approach, 62nd Ann. Intern. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 95-98.

Doyen, P.M., 1988, Porosity from seismic data : a geostatistical approach, Geophysics, 53, n° 10, 1263-1275.

Harding, J.P., 1949, The use of probability paper for the graphical analysis of polymodal frequency distributions, J. of the Marine Biol. Ass. of the UK, 28, 141-153.

Sinclair, A.J., 1976, Application of probability graphs in mineral exploration, Assoc. of Expl. Geoch., vol. 4.

Everitt, B.S., Hand, D.J., 1981, Finite mixture distributions, Monographs on Applied Probability and Statistics, Chapman and Hall, London New York.

Mezghache, H., 1989, Cartographie automatique et interprétation géostatistique d'une campagne de prospection géochimique sur sol. Application à la zone mercurielle nord numidique (Algérie), Thèse de 3ème cycle, INPL.

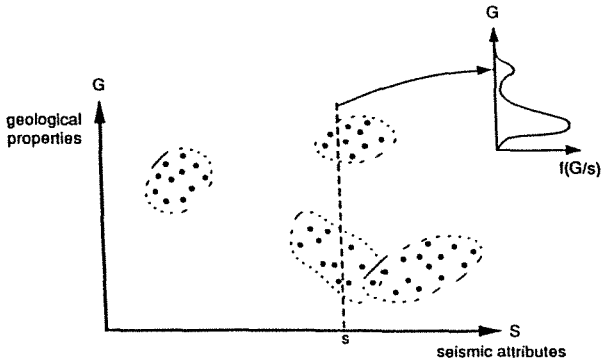


Fig. 1 Calibration methodology with the multivariate gaussian segmentation

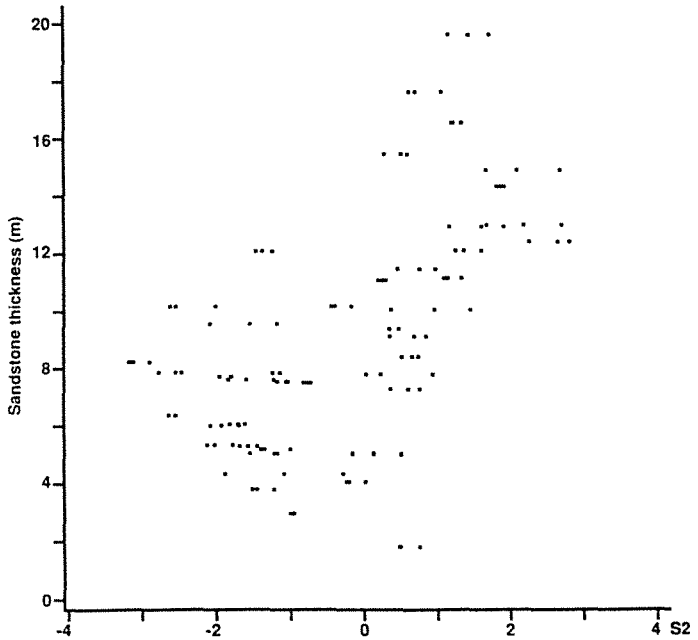


Fig. 2 Sandstone thickness versus seismic attribute number 2 cross-plot
The three closest traces are associated to each well

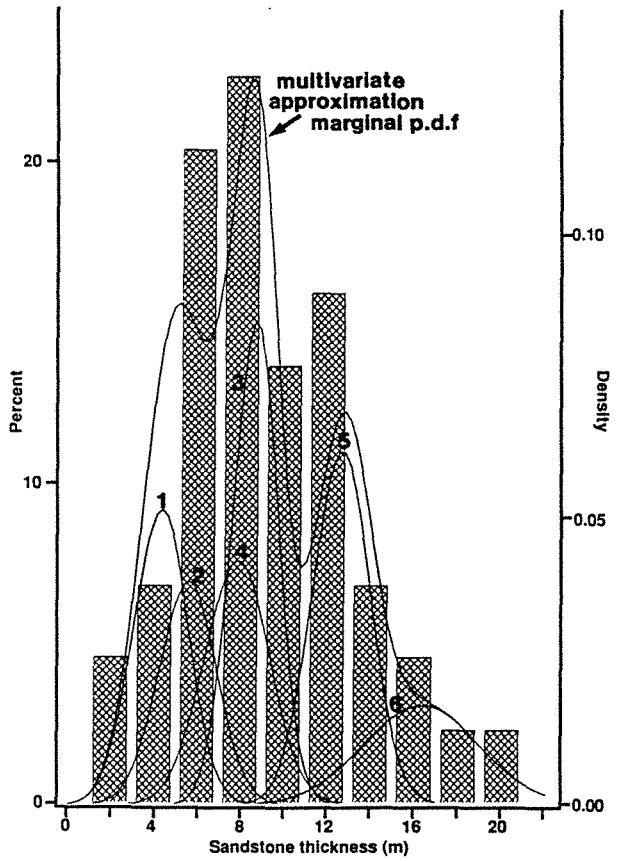


Fig. 3 Histogram of the sandstone thickness and associated marginal distribution probability density functions of the multivariate approximation and its components

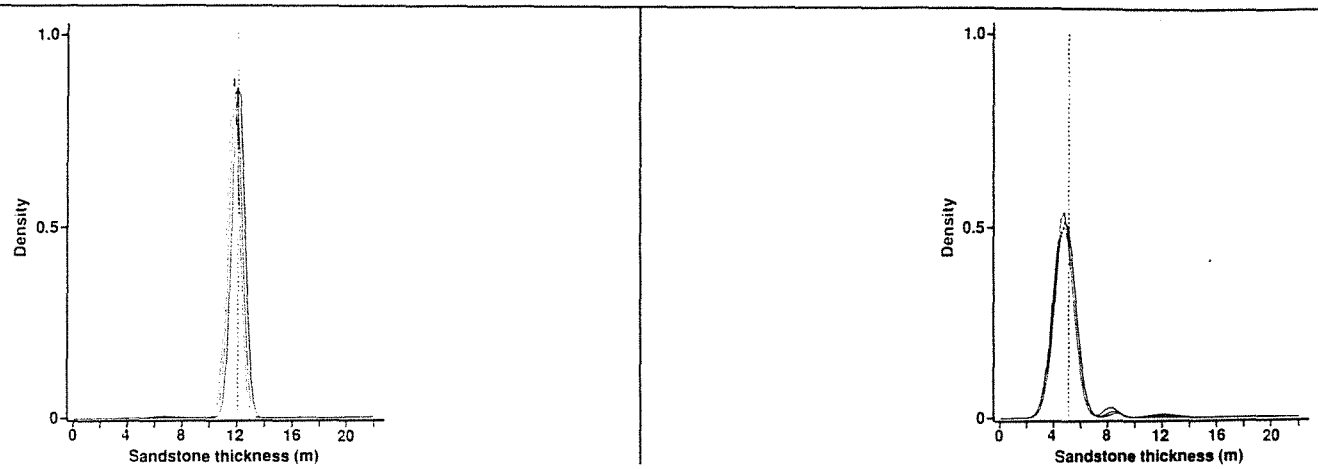
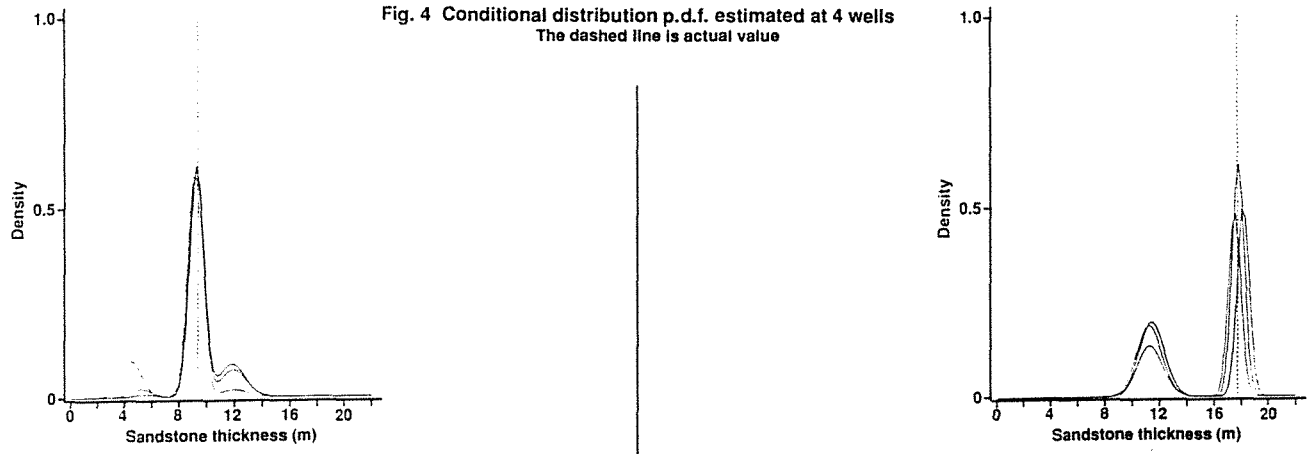


Fig. 4 Conditional distribution p.d.f. estimated at 4 wells
The dashed line is actual value



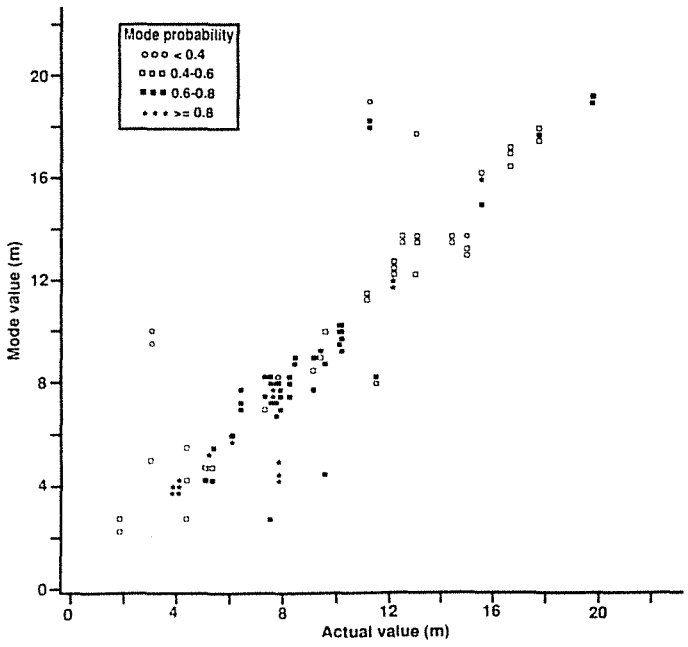


Fig. 5 Sandstone thickness prediction at the wells

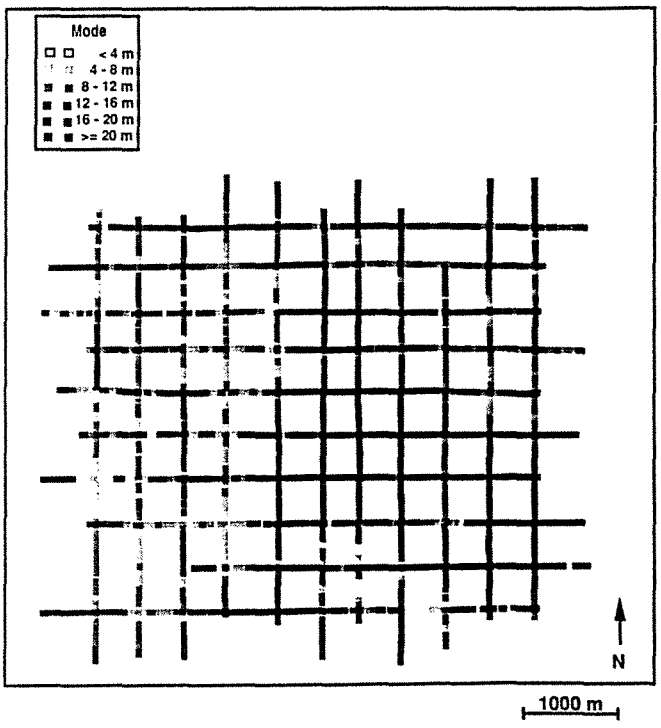


Fig. 6 Sandstone thicknesses predicted from all traces (conditional distribution p.d.f. mode values)

ANNEXE D

Résultats de la méthode de décomposition en classes gaussiennes dans le cadre de la prédiction de l'épaisseur de grès :

- *paramètres de la décomposition* ***Sol6*** *à 5 classes gaussiennes retenue.*
- *caractéristiques des décompositions en classes gaussiennes obtenues.*

Paramètres de la décomposition *Sol6* à 5 classes gaussiennes

Notons p_k le poids de la classe k. Et pour une variable donnée, notons m_k et σ_k respectivement la moyenne et l'écart-type de la classe k.

$p_1 = 0.18$
$p_2 = 0.27$
$p_3 = 0.18$
$p_4 = 0.16$
$p_5 = 0.21$

Variable S_1

$m_1 = -1.13$	$\sigma_1 = 1.16$
$m_2 = 1.51$	$\sigma_2 = 1.05$
$m_3 = -2.18$	$\sigma_3 = 1.12$
$m_4 = 1.41$	$\sigma_4 = 1.51$
$m_5 = 0.9$	$\sigma_5 = 1.72$

Variable S_2

$m_1 = -0.99$	$\sigma_1 = 1.13$
$m_2 = 1.09$	$\sigma_2 = 0.68$
$m_3 = 0.53$	$\sigma_3 = 1.39$
$m_4 = -1.54$	$\sigma_4 = 1.38$
$m_5 = -1.15$	$\sigma_5 = 0.50$

Variable S_3

$m_1 = -0.38$	$\sigma_1 = 0.86$
$m_2 = 0.14$	$\sigma_2 = 1.21$
$m_3 = 0.00$	$\sigma_3 = 1.45$
$m_4 = 0.24$	$\sigma_4 = 1.21$
$m_5 = 0.35$	$\sigma_5 = 1.09$

Variable S_4

$m_1 = -0.12$ $\sigma_1 = 0.97$

$m_2 = 0.04$ $\sigma_2 = 0.36$

$m_3 = -1.07$ $\sigma_3 = 0.59$

$m_4 = -0.20$ $\sigma_4 = 1.22$

$m_5 = 0.88$ $\sigma_5 = 0.92$

Variable G_1

$m_1 = 4.94$ $\sigma_1 = 1.38$

$m_2 = 11.06$ $\sigma_2 = 2.75$

$m_3 = 13.94$ $\sigma_3 = 3.66$

$m_4 = 9.14$ $\sigma_4 = 1.08$

$m_5 = 5.58$ $\sigma_5 = 1.71$

Caractéristiques des décompositions en classes gaussiennes obtenues
- Notations -

- **PMIN** : poids minimal d'une classe pour qu'elle soit conservée.

- **Initialisations monovariables :**

  FR

  Application de la méthode de Harding automatisée sur la fonction de répartition empirique de chaque variable.

  FRE

  Application de la méthode de Harding automatisée sur la fonction de répartition intégrée à partir de l'estimation de la densité par la méthode des noyaux pour chaque variable.

  MEff *j*

  Recherche de *j* classes de même effectif sur chaque variable.

- **Initialisation multivariable :**

  Multi *j* classes

  Partition de la population en *j* classes *a priori.*

Solution	Nombre de classes	Poids des classes - minimal - maximal	Critères C_1	 C_2	PMIN	Initialisation
Sol1	7	0.11 0.17	4.24	0.52	0.1	Multi 15 classes
Sol2	6	0.13 0.23	4.18	0.60	0.1	MEff 5
Sol3	6	0.11 0.22	4.12	0.86	0.1	MEff 10
Sol4	6	0.14 0.24	4.35	0.69	0.1	MEff 3
Sol5	4	0.13 0.39	4.21	0.90	0.1	Multi 5 classes
Sol6	5	0.16 0.27	4.39	0.76	0.15	Multi 10 classes
Sol7	5	0.16 0.25	4.32	0.96	0.15	MEff 5
Sol8	6	0.13 0.23	4.76	0.65	0.1	FR
Sol9	5	0.15 0.24	4.66	0.90	0.1	FRE
Sol10	4	0.18 0.32	4.91	0.89	0.15	Multi 15 classes
Sol11	4	0.19 0.33	4.40	1.11	0.15	MEff 3

Tableau D-1

Solution	Nombre de classes	Poids des classes - minimal - maximal	Critères C_1	 C_2	PMIN	Initialisation
Sol12	3	0.26 0.41	4.63	1.22	0.2 0.25	MEff 3
Sol13	5	0.16 0.24	4.90	0.96	0.1	Multi 20 classes
Sol14	3	0.23 0.41	4.83	1.12	0.15 0.2	Multi 5 classes
Sol15	1	1.00	4.82	1.27	0.1 0.15 0.2 0.25	FRE
Sol16	4	0.18 0.43	4.95	1.11	0.15	FRE
Sol17	3	0.27 0.41	5.14	1.32	0.2	Multi 20 classes
Sol18	3	0.28 0.39	5.13	1.35	0.2	MEff 10
Sol19	4	0.23 0.26	5.19	1.30	0.15	FR
Sol20	4	0.22 0.28	5.29	1.32	0.2	Multi 15 classes

Tableau D-2

Solution	Nombre de classes	Poids des classes - minimal - maximal	Critères C_1	 C_2	PMIN	Initialisation
Sol21	3	0.25 0.41	5.29	1.33	0.2	FRE
Sol22	2	0.48 0.52	5.30	1.35	0.25	FR
					0.25	MEff 10
					0.25	Multi 5 classes
					0.2 0.25	Multi 10 classes
					0.25	Multi 20 classes
Sol23	3	0.24 0.38	5.34	1.31	0.2	MEff 5
Sol24	4	0.17 0.43	5.47	1.24	0.1	Multi 10 classes
Sol25	3	0.21 0.41	5.53	1.51	0.15	Multi 20 classes
Sol26	2	0.48 0.52	5.66	1.47	0.25	Multi 15 classes

Tableau D-3

Solution	Nombre de classes	Poids des classes - minimal - maximal	Critères C_1	 C_2	PMIN	Initialisation
Sol27	2	0.43 0.57	5.71	1.57	0.25	MEff 5 FRE
Sol28	3	0.20 0.41	6.06	1.55	0.2	FR
Sol29	4	0.16 0.36	6.21	1.65	0.15	MEff 5

Tableau D-4

ANNEXE E

Résultats de la méthode de décomposition en classes gaussiennes dans le cadre de la prédiction de l'épaisseur de dolomies vacuolaires :

- *paramètres de la décomposition* ***Sol2*** *à 6 classes gaussiennes retenue.*
- *caractéristiques des décompositions en classes gaussiennes obtenues.*

Paramètres de la décomposition *Sol2* à 6 classes gaussiennes

Notons p_k le poids de la classe k. Et pour une variable donnée, notons m_k et σ_k respectivement la moyenne et l'écart-type de la classe k.

$p_1 = 0.15$
$p_2 = 0.18$
$p_3 = 0.16$
$p_4 = 0.17$
$p_5 = 0.14$
$p_6 = 0.20$

Variable S_1

$m_1 = 0.42$	$\sigma_1 = 0.95$
$m_2 = -1.93$	$\sigma_2 = 1.76$
$m_3 = 2.85$	$\sigma_3 = 0.75$
$m_4 = 1.34$	$\sigma_4 = 0.87$
$m_5 = -0.61$	$\sigma_5 = 1.60$
$m_6 = -0.50$	$\sigma_6 = 1.19$

Variable S_2

$m_1 = 0.26$	$\sigma_1 = 1.18$
$m_2 = 0.03$	$\sigma_2 = 1.75$
$m_3 = 1.13$	$\sigma_3 = 0.87$
$m_4 = -1.35$	$\sigma_4 = 0.74$
$m_5 = 0.25$	$\sigma_5 = 0.99$
$m_6 = -1.50$	$\sigma_6 = 0.94$

Variable S_3

$m_1 = -0.77$	$\sigma_1 = 0.40$
$m_2 = -0.39$	$\sigma_2 = 1.45$
$m_3 = 0.61$	$\sigma_3 = 1.40$
$m_4 = -0.55$	$\sigma_4 = 0.84$
$m_5 = -0.34$	$\sigma_5 = 0.91$
$m_6 = 0.88$	$\sigma_6 = 0.73$

Variable S_4

$m_1 = 0.33$	$\sigma_1 = 0.51$
$m_2 = -1.19$	$\sigma_2 = 0.64$
$m_3 = -0.41$	$\sigma_3 = 0.77$
$m_4 = 0.52$	$\sigma_4 = 1.35$
$m_5 = 0.01$	$\sigma_5 = 0.60$
$m_6 = 0.44$	$\sigma_6 = 0.73$

Variable G_2

$m_1 = 7.20$	$\sigma_1 = 5.41$
$m_2 = 17.41$	$\sigma_2 = 3.39$
$m_3 = 4.32$	$\sigma_3 = 2.14$
$m_4 = 4.36$	$\sigma_4 = 2.34$
$m_5 = 9.03$	$\sigma_5 = 7.67$
$m_6 = 4.17$	$\sigma_6 = 4.36$

Caractéristiques des décompositions en classes gaussiennes obtenues - Notations -

- **PMIN** : poids minimal d'une classe pour qu'elle soit conservée.

- **Initialisations monovariables :**

  FR

  Application de la méthode de Harding automatisée sur la fonction de répartition empirique de chaque variable.

  FRE

  Application de la méthode de Harding automatisée sur la fonction de répartition intégrée à partir de l'estimation de la densité par la méthode des noyaux pour chaque variable.

  MEff *j*

  Recherche de *j* classes de même effectif sur chaque variable.

- **Initialisation multivariable :**

  Multi *j* classes

  Partition de la population en *j* classes *a priori.*

Solution	Nombre de classes	Poids des classes - minimal - maximal	Critères C_1	 C_2	PMIN	Initialisation
Sol1	6	0.11 0.25	4.61	0.49	0.1	Multi 20 classes
Sol2	6	0.14 0.19	5.26	0.51	0.1	Multi 15 classes
Sol3	4	0.18 0.40	5.48	0.58	0.15	FRE
Sol4	5	0.15 0.30	5.57	0.61	0.1	MEff 10
Sol5	5	0.15 0.28	5.68	0.58	0.1 0.15	Multi 5 classes
Sol6	6	0.14 0.20	5.80	0.72	0.1	Multi 10 classes
Sol7	3	0.24 0.51	5.61	1.01	0.15	MEff 5
Sol8	5	0.13 0.26	5.84	0.97	0.1	MEff 5
Sol9	6	0.12 0.23	6.39	0.81	0.1	FRE
Sol10	5	0.13 0.30	6.38	0.95	0.1	FR

Tableau E-1

Solution	Nombre de classes	Poids des classes - minimal - maximal	Critères C_1	 C_2	PMIN	Initialisation
Sol11	4	0.20 0.33	6.56	0.84	0.15	MEff 10
Sol12	3	0.24 0.51	6.53	1.03	0.2	MEff 3
Sol13	4	0.21 0.33	6.81	1.10	0.15	Multi 10 classes
Sol14	2	0.49 0.51	6.81	1.42	0.2 0.25	MEff 5
					0.25	MEff 10
Sol15	3	0.29 0.37	6.98	1.32	0.25	Multi 15 classes
Sol16	5	0.16 0.25	7.39	0.99	0.15	FR
Sol17	3	0.23 0.44	7.13	1.36	0.2	MEff 10
Sol18	4	0.17 0.31	7.25	1.25	0.15	Multi 15 classes
Sol19	3	0.25 0.42	7.61	1.17	0.2	FRE
Sol20	4	0.21 0.31	7.69	1.22	0.15	MEff 3

Tableau E-2

Solution	Nombre de classes	Poids des classes - minimal - maximal	Critères C_1	 C_2	PMIN	Initialisation
Sol21	4	0.15 0.33	7.71	1.24	0.1	MEff 3
Sol22	3	0.28 0.40	7.49	1.66	0.2	Multi 10 classes
Sol23	3	0.27 0.39	8.28	1.34	0.2 0.25	Multi 5 classes
Sol24	3	0.30 0.36	8.57	1.58	0.15	Multi 20 classes
Sol25	3	0.28 0.38	8.69	1.73	0.2	FR
Sol26	1	1.00	8.95	2.45	0.1 0.15 0.2 0.25	FRE
Sol27	2	0.30 0.70	10.28	2.62	0.25	Multi 20 classes
Sol28	2	0.37 0.63	9.86	3.10	0.2	Multi 15 classes
Sol29	2	0.40 0.60	10.10	2.96	0.25	FRE

Tableau E-3

Solution	Nombre de classes	Poids des classes - minimal - maximal	Critères C_1	 C_2	PMIN	Initialisation
Sol30	2	0.40 0.60	10.14	2.89	0.25	FR
Sol31	2	0.47 0.53	12.23	3.84	0.2	Multi 20 classes
					0.25	MEff 3
					0.25	Multi 20 classes

Tableau E-4

INSTITUT NATIONAL
POLYTECHNIQUE
DE LORRAINE

AUTORISATION DE SOUTENANCE DE THESE
DU DOCTORAT DE L'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE

o0o

VU LES RAPPORTS ETABLIS PAR :

Monsieur DEPAIX, Professeur, Université NANCY I,

Monsieur BAYER, Maitre de Conférence, Université Montpellier 2.

Le Président de l'Institut National Polytechnique de Lorraine, autorise :

Mademoiselle JOSEPH Caroline

à soutenir devant l'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE,

une thèse intitulée :

"Application de l'analyse des mélanges gaussiens au calibrage géologique des données sismiques".

en vue de l'obtention du titre de :

DOCTEUR DE L'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE

Spécialité : **"GÉOSCIENCES & MATIERES PREMIERES"**

Fait à Vandoeuvre le, **09 Septembre 1994**

Le Président de l'I.N.P.L.,

M. LUCIUS

INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE NANCY

NANCY BRABOIS
2, AVENUE DE LA
FORET-DE-HAYE
BOITE POSTALE 3
F - 5 4 5 0 1
VANDŒUVRE CEDEX

TEL. 33/83.59.59.59
FAX. 33/83.59.59.55