

INTERVALLES DE CONFIANCE PONCTUELS APPROCHÉS PAR KRIGEAGE DISJONCTIF : UNE ÉTUDE EXPÉRIMENTALE

L.Y. HU†

Table des Matières

Résumé	120	Abstract	120
A - INTRODUCTION.....	121	2 - Construction et statistique des intervalles de confiance	122
1 - Intervalle de confiance déduit de la distribution conditionnelle	121	C - CAS D'ÉTUDE : LUCEDA	125
2 - Approximation de la distribution condi- tionnelle par krigeage disjonctif	121	1 - Description et modélisation des données...	125
3 - Contexte des deux cas d'étude	121	2 - Construction et statistique des intervalles de confiance	125
B - CAS D'ÉTUDE : BERTHOLÈNE	122	D - CONCLUSIONS	128
1 - Description et modélisation des données...	122	Bibliographie	

Illustrations

Figures 1 - Courbes des taux d'appartenance effectif/prévu des teneurs à leurs intervalles de confiance	123
Figure 2 - Longueurs d'intervalles de confiance en fonction des fréquences cumulées des teneurs	123
Figures 3 - Exemples de distributions conditionnelles de teneurs approchées par E.C. trigaussienne et K.D. bigaussien	124
Figures 4 - Courbes des taux d'appartenance effectif/prévu des teneurs à leurs intervalles de confiance	126
Figures 5 - Longueurs d'intervalles de confiance en fonction des fréquences cumulées des teneurs	126
Figures 6 - Exemples de distributions conditionnelles de teneurs approchées par K.D. gamma-beta et K.D. bigaussien	127
Figures 7 - Statistiques des intervalles de confiance approchés par K.D. bigaussien et E.C. trigaussienne	127
Figures 8 - Statistiques des intervalles de confiance approchés par K.D. gamma-beta et K.D. bigamma	128

† Centre de Géostatistique, Ecole Nationale Supérieure des Mines de Paris, 35 rue Saint-Honoré,
77305 FONTAINEBLEAU (France)

RESUME

L'intervalle de confiance pour la valeur en un point x_0 non observé d'une fonction aléatoire $Z(x)$ à partir des échantillons $Z(x_1), Z(x_2), \dots, Z(x_N)$, se déduit de la distribution conditionnelle de $Z(x_0)$ à $Z(x_1), Z(x_2), \dots, Z(x_N)$ fixées. Le krigeage disjonctif (K.D.), basé sur la connaissance des lois bivariées de $Z(x)$, fournit une approximation de cette distribution conditionnelle. En pratique, la mise en œuvre du K.D. s'effectue souvent dans le cadre d'un modèle isofactoriel (e.g. le modèle gaussien ou plus généralement le modèle gamma). Dans cet article, deux cas d'étude sont présentés, tirés de la pratique minière, concernant l'intervalle de confiance au support d'échantillon. On considère, dans les deux cas, les données le long des sondages, ce qui permet de faire une validation croisée des méthodes utilisées.

ABSTRACT

Point Confidence Intervals Approximated by Disjunctive Kriging : an Experimental Study

The confidence interval for the value at an unobserved point x_0 of a random function $Z(x)$ given a set of samples $Z(x_1), Z(x_2), \dots, Z(x_N)$, is deduced from the conditional distribution of $Z(x_0)$ given $Z(x_1), Z(x_2), \dots, Z(x_N)$. The disjunctive kriging (D.K.), based on the knowledge of the bivariate law of $Z(x)$, provides an approximation to this conditional distribution. In practice, D.K. is often used within the framework of an isofactorial model (e.g. the gaussian model or more generally the gamma model). In this paper, two mining case studies, concerning the confidence interval on the sample support, are presented. In both cases, the data along the drillholes are considered, which makes it possible to cross-validate the methods used.

A - INTRODUCTION

1 - INTERVALLE DE CONFIANCE DEDUIT DE LA DISTRIBUTION CONDITIONNELLE

L'intervalle de confiance pour la valeur en un point non observé x_0 d'une fonction aléatoire $Z(x)$ à partir des échantillons $Z^N = (Z(x_1), Z(x_2), \dots, Z(x_N))$, se déduit de la distribution conditionnelle de $Z(x_0)$ à Z^N fixé. Désignons par $F_0(z|Z^N)$ cette distribution conditionnelle :

$$F_0(z|Z^N) = P[Z(x_0) < z | Z^N] = E[1_{Z(x_0) < z} | Z^N]$$

et F_0^{-1} son inverse. Alors, à Z^N fixé, $Z(x_0)$ appartient à l'intervalle (de confiance) :

$$[F_0^{-1}(p|Z^N), F_0^{-1}(1-q|Z^N)]$$

avec la probabilité $1 - p - q$ ($0 \leq p < 1 - q \leq 1$). En pratique, p et q sont choisies selon le degré de risque que l'on veut prendre. Dans ce qui suit, un intervalle de confiance est dit centré si $p = q$, et non centré si $p \neq q$.

2 - APPROXIMATION DE LA DISTRIBUTION CONDITIONNELLE PAR KRIGEAGE DISJONCTIF

En toute rigueur, le calcul de la distribution conditionnelle $F_0(z|Z^N)$ requiert la connaissance de la loi multivariable de $(Z(x_0), Z(x_1), \dots, Z(x_N))$, laquelle est difficilement accessible dans la plupart des cas pratiques, excepté le cas où l'anamorphosée gaussienne de $Z(x)$ peut être considérée comme une fonction aléatoire multigaussienne. Toutefois, par l'estimation de la fonction d'indicatrice $1_{Z(x_0) < z}$, le krigeage disjonctif (K.D.), basé sur la connaissance des lois bivariées de $Z(x)$, fournit une approximation de la distribution conditionnelle $F_0(z|Z^N)$.

En pratique, la mise en œuvre du K.D. s'effectue souvent dans le cadre d'un modèle isofactoriel. A l'heure actuelle, le modèle (isofactoriel) gaussien est le plus utilisé et a donné lieu aux vérifications les plus nombreuses.

Au lieu d'une anamorphose gaussienne, le modèle (isofactoriel) gamma utilise une fonction d'anamorphose gamma qui transforme la fonction aléatoire d'étude $Z(x)$ en une autre $Y(x)$ à loi gamma de paramètre α . Notons que, aux grandes valeurs de α , le modèle gamma tend à s'approcher du modèle gaussien. Ainsi, on peut considérer que le modèle gamma est une généralisation du modèle gaussien. En particulier, le modèle gamma pourrait s'adapter au cas de données très dissymétriques, là où le modèle gaussien s'avère parfois inadapté.

3 - CONTEXTE DES DEUX CAS D'ETUDE

Dans cet article, deux cas d'études sont présentés, tirés de la pratique minière, concernant l'approximation de l'intervalle de confiance de la teneur au support d'échantillon. Dans le premier cas (Bertholène), il s'agit d'un jeu de données peu dissymétrique pour lequel un modèle gaussien a été choisi. Nous comparons le krigeage disjonctif (K.D.) et l'espérance conditionnelle (E.C.). Dans le deuxième cas (Luceda), le jeu de données est beaucoup plus dissymétrique que le premier et un modèle gamma a été retenu. A titre de

comparaison, le modèle gaussien est aussi utilisé pour ce jeu de données. Dans les deux cas, on considère les données le long des sondages, ce qui permet de faire une validation croisée des méthodes utilisées.

La validation croisée est effectuée ici en deux temps. Tout d'abord, on compare le taux d'appartenance des teneurs dans leurs intervalles de confiance prévu par le modèle, et le taux d'appartenance obtenu en réalité. Ensuite, on examine les longueurs des intervalles de confiance. En effet, un modèle doit être préféré à un autre s'il produit un taux plus proche du taux effectif, et s'il fournit de plus les intervalles de confiance plus courts.

B – CAS D'ETUDE : BERTHOLENE

1 - DESCRIPTION ET MODELISATION DES DONNEES

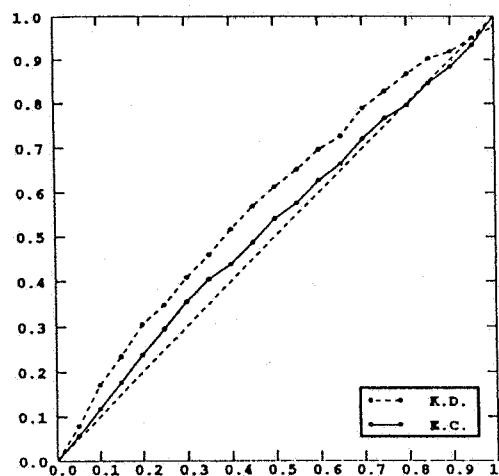
Les données considérées ici ont été prélevées d'un banc de la mine d'uranium de Bertholène. On dispose de 144 sondages verticaux de 6m de longueur, implantés selon une maille régulière de 2m×3m. Les échantillons correspondent à des passes de 1m obtenues par régularisation le long des sondages, et pour lesquels on connaît la position du centre de gravité et la teneur moyenne exprimé en ppm.

Le coefficient de variation des teneurs est de 1.9, ce qui est relativement modeste. Dans un travail précédent (Hu et Lantuéjoul, 1988), il a été montré que le modèle gaussien apparaît satisfaisant pour décrire les lois bivariées de ce jeu de données. Plus précisément, les anamorphosées gaussiennes des teneurs présentent un caractère bigaussien. Ceci nous a conduit à utiliser le K.D. bigaussien pour approcher les intervalles de confiance pour les teneurs au support d'échantillon.

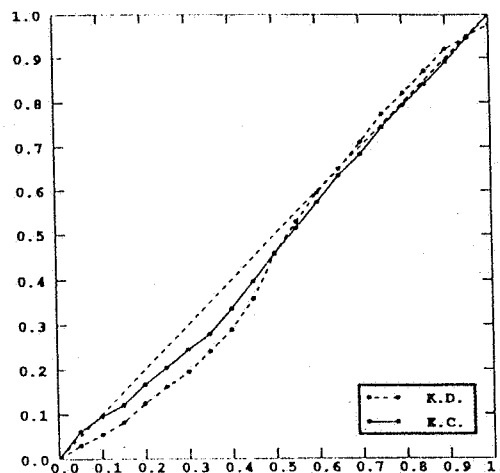
2 - CONSTRUCTION ET STATISTIQUE DES INTERVALLES DE CONFIANCE

Pour pouvoir tester la validité du K.D. bigaussien, on a choisi de travailler le long des sondages. Pour toutes les teneurs des échantillons (excepté les deux extrémités des sondages), on a construit un intervalle de confiance approché par K.D., en utilisant les deux échantillons les plus proches. Par ailleurs, on a aussi essayé de déduire les intervalles de confiance à partir des distributions conditionnelles calculées selon le modèle multigaussien, en considérant dans ce cas que les lois trivariées sont encore gaussiennes.

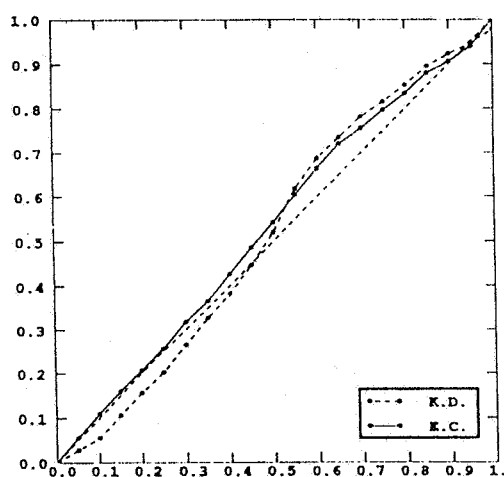
On examine d'abord le taux d'appartenance effectif des teneurs à leurs intervalles de confiance, et le compare au taux prévu. Les figures 1(a), 1(b) et 1(c) montrent les courbes des taux effectif/prévu, fournis par K.D. bigaussien et par E.C. trigaussienne. On voit que les intervalles centrés pas trop grands ($p = q > 0.05$) approchés par K.D. ont recouvert systématiquement plus de teneurs que prévu, et que la différence maximale observée entre les taux effectif et prévu atteint 11.6% ($p = q = 0.3$, intervalles de 40%), ce qui est considérable. Fort heureusement, pour les intervalles supérieurs à 90%, qui sont les plus intéressants en pratique, la différence effectif-prévu est assez faible. Remarquons d'ailleurs que l'E.C. a donné des résultats meilleurs, que les intervalles soient centrés ou non (Figures 1(a), 1(b) et 1(c)).



(a) $p = q$



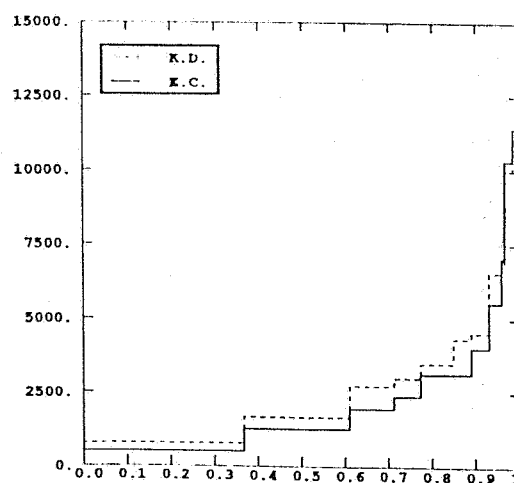
(b) $p = 0$



(c) $q = 0$

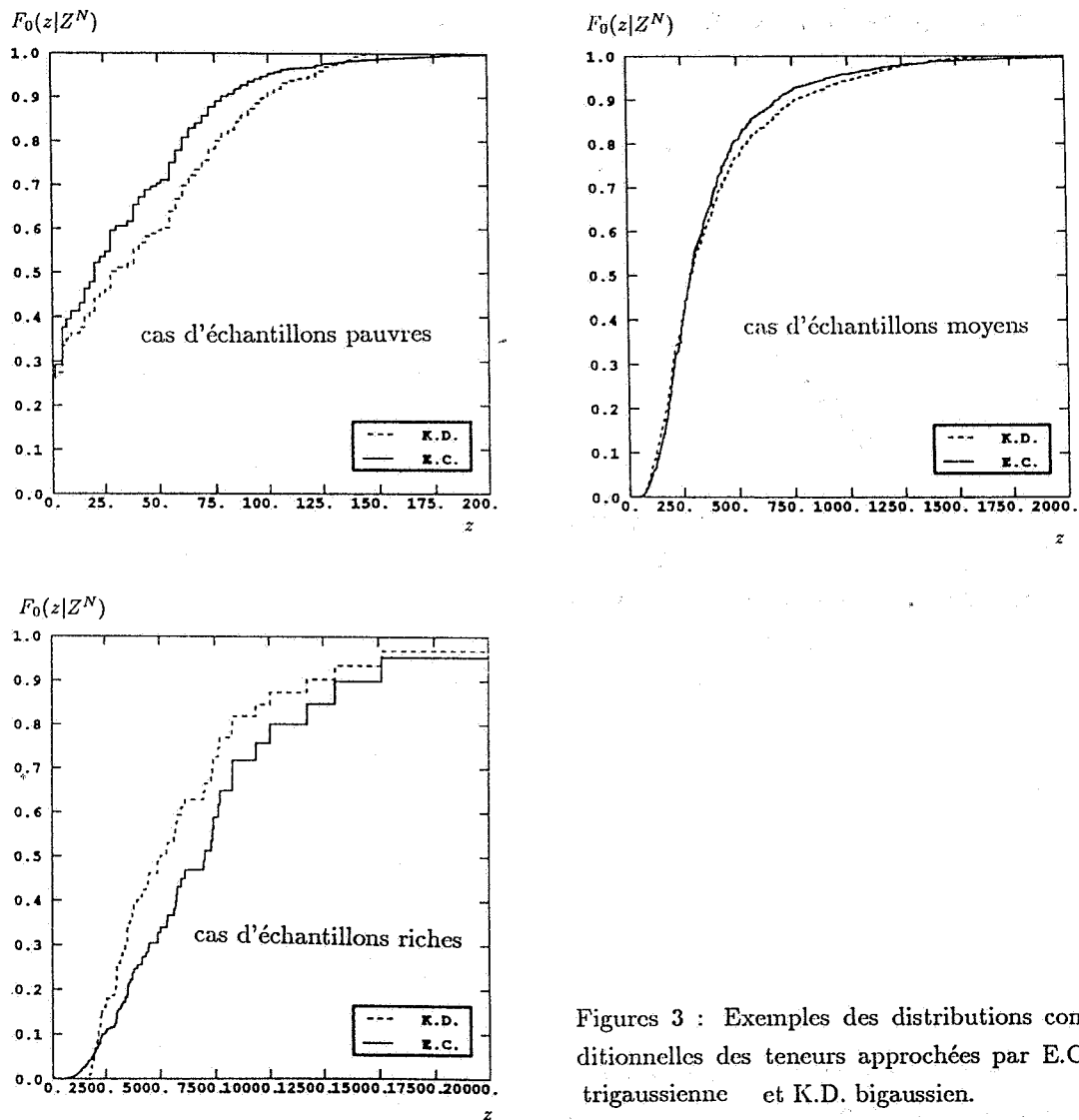
Figures 1 : Courbes des taux d'appartenance effectif/prévu des teneurs à leurs intervalles de confiance.

Figure 2 : Longueurs d'intervalles de confiance en fonction de fréquences cumulées des teneurs.



Pour deux méthodes qui donnent les taux d'appartenance de teneurs proches, on préfère évidemment celle qui fournit les intervalles de confiance les plus courts. La figure 2 montre les longueurs des intervalles fournis par K.D. et E.C. en fonction de fréquences cumulées des teneurs pour un taux d'appartenance prévu de 90% ($p = q = 0.05$). Pour 96% des teneurs, les intervalles fournis par K.D. sont les plus longs. Pour les teneurs plus fortes, la comparaison est difficile, mais elles ne représentent que 4%.

Enfin, il n'est pas inintéressant de visualiser quelques exemples de distribution conditionnelle et de pseudo-distribution du K.D. (Figures 3). Ils montrent directement la différence existant entre les intervalles de confiance fournis par ces deux méthodes (K.D. bigaussien et E.C. trigaussienne).



Figures 3 : Exemples des distributions conditionnelles des teneurs approchées par E.C. trigaussienne et K.D. bigaussien.

C - CAS D'ETUDE : LUCEDA

1 - DESCRIPTION ET MODELISATION DES DONNEES

Les données proviennent d'un autre gisement d'uranium reconnu par une cinquantaine de sondages verticaux. Chaque sondage contient en moyenne 40 échantillons. Le coefficient de variation des teneurs en uranium ($\sigma/m = 3.7$) est beaucoup plus élevé que celui de Bertholène étudié précédemment ($\sigma/m = 1.9$). Selon un travail déjà cité (Hu et Lantuéjoul, 1988), les données de Luceda sont compatibles avec un modèle gamma, mais pas avec un modèle gaussien. Rappelons que la distribution bivariable des anamorphosées gamma $Y(x_i) = \phi_\alpha^{-1}[Z(x_i)]$ ($i = 1, 2, \dots$) de paramètre $\alpha = 0.0541$ peut être approchée par une loi de Laguerre de type beta (modèle gamma-beta), à savoir que les polynômes de Laguerre $l_n^\alpha[Y(x_i)]$ et $l_n^\alpha[Y(x_i + h)]$ admettent le coefficient de corrélation de forme :

$$\rho_n(h) = \frac{\Gamma(\beta) \Gamma[\beta \rho(h) + n]}{\Gamma(\beta + n) \Gamma[\beta \rho(h)]} \quad \text{avec } \beta = 1.$$

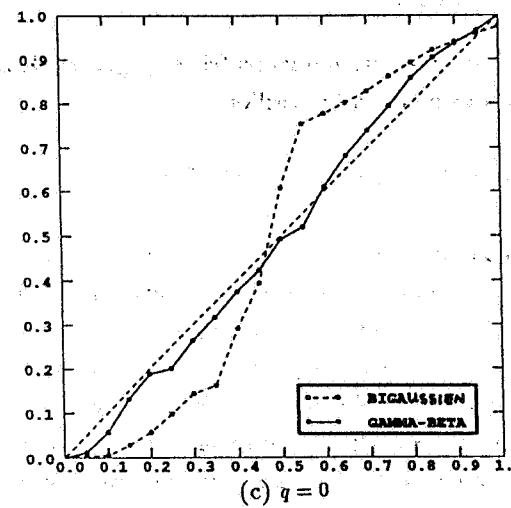
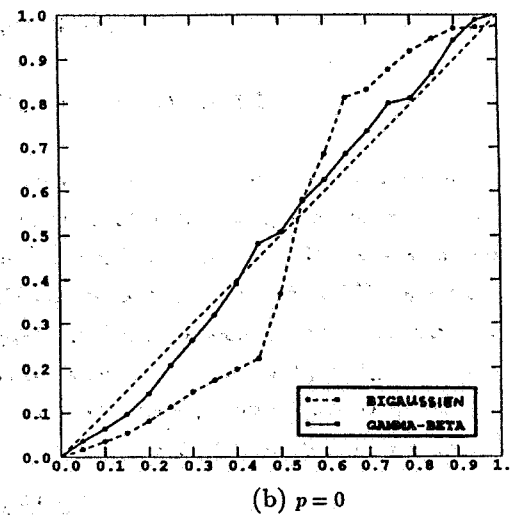
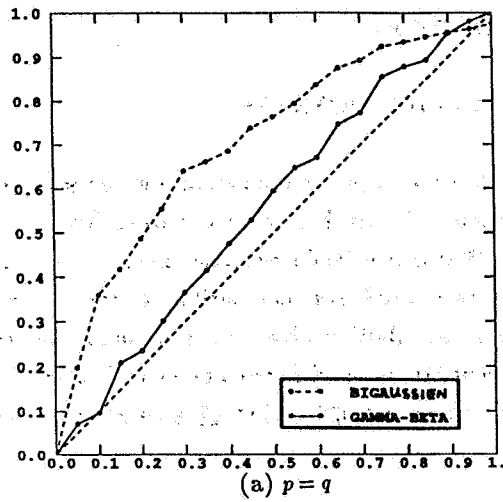
A partir de cette modélisation de loi bivariable, on peut estimer la fonction d'indicatrice $1_{Z(x_0) < z}$ par K.D., ainsi obtenir une approximation des intervalles de confiance au support d'échantillon.

2 - CONSTRUCTION ET STATISTIQUE DES INTERVALLES DE CONFIANCE

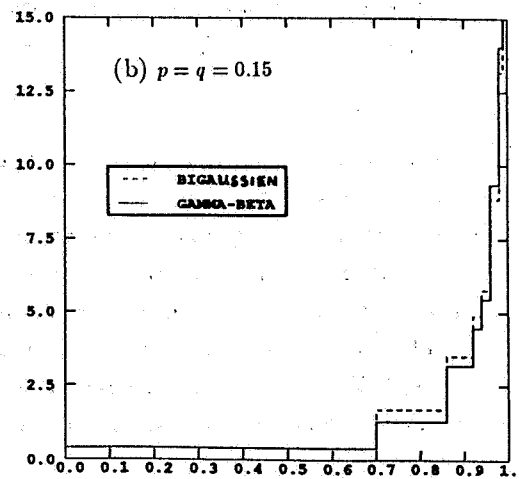
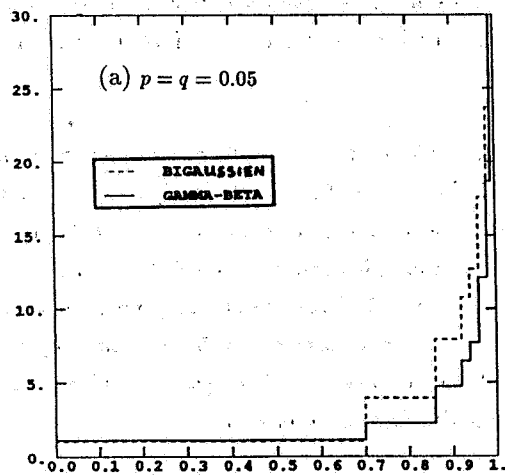
On s'intéresse, comme précédemment, aux données le long des sondages, et on essaie de construire les intervalles de confiance des teneurs (excepté celles aux deux extrémités des sondages), en utilisant les deux échantillons les plus proches. Les figures 4(a), 4(b) et 4(c) montrent que les taux d'appartenance effectifs des teneurs à leurs intervalles approchés par K.D. gamma-beta sont assez proches de ceux qui ont été prévus, et cette approximation est surtout satisfaisante pour les intervalles non centrés. Par contre, si l'on utilise le K.D. bigaussien, les courbes des taux d'appartenance effectifs/prévus sont bien plus éloignées de la diagonale, sauf en quelques points. En particulier, les intervalles centrés, hormis ceux qui sont proches de 0% et de 100%, contiennent systématiquement plus de teneurs que prévu. La différence maximale observée atteint 34% (intervalles de 30%), ce qui est catastrophique. Pour les intervalles non centrés, on observe un taux d'appartenance effectif plus faible que prévu aux intervalles courts, et plus fort que prévu aux intervalles longs.

Examinons maintenant les longueurs d'intervalles. Par exemple, pour $p = q = 0.05$ (intervalles centrés de 90%), les taux d'appartenance effectifs ne varient guère du K.D. gamma-beta au K.D. bigaussien, mais les intervalles fournis par K.D. gamma-beta sont nettement plus courts que ceux fournis par K.D. bigaussien (Figure 5(a)). Prenons un autre exemple avec $p = q = 0.15$ (intervalles centrés de 70%). Cette fois, les intervalles du K.D. gamma-beta sont légèrement plus courts que ceux du K.D. bigaussien pour les classes de fréquences inférieures à 96%, contrairement pour le 4% de teneurs restantes (Figure 5(b)). Mais rappelons surtout que le K.D. bigaussien a entraîné un taux d'appartenance effectif (89%) beaucoup plus fort que prévu (70%), ainsi qu'une estimation plus pessimiste que celle du K.D. gamma-beta.

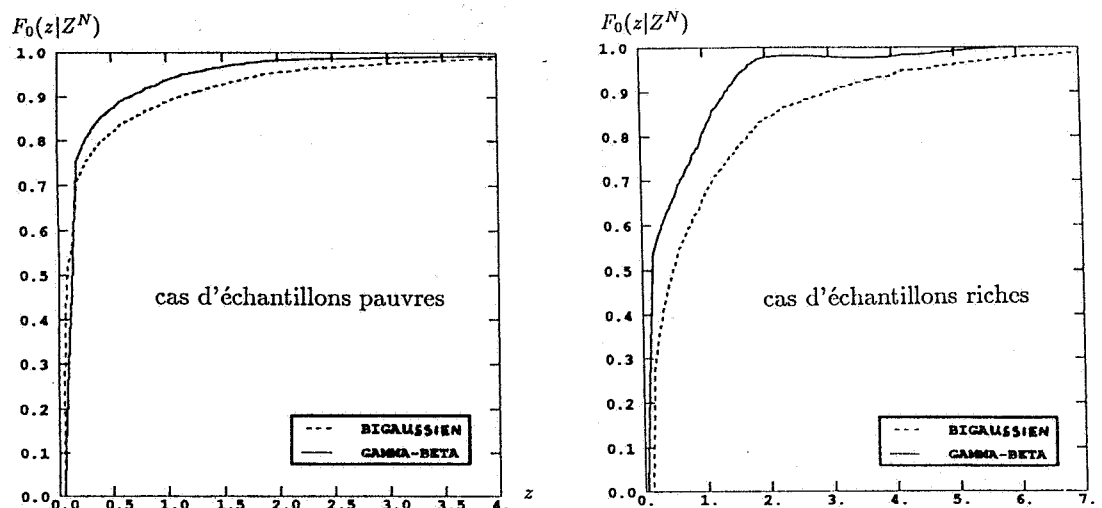
Quelques pseudo-distributions présentées aux figures 6(a) et 6(b) montrent bien la différence entre le K.D. gamma-beta et le K.D. bigaussien.



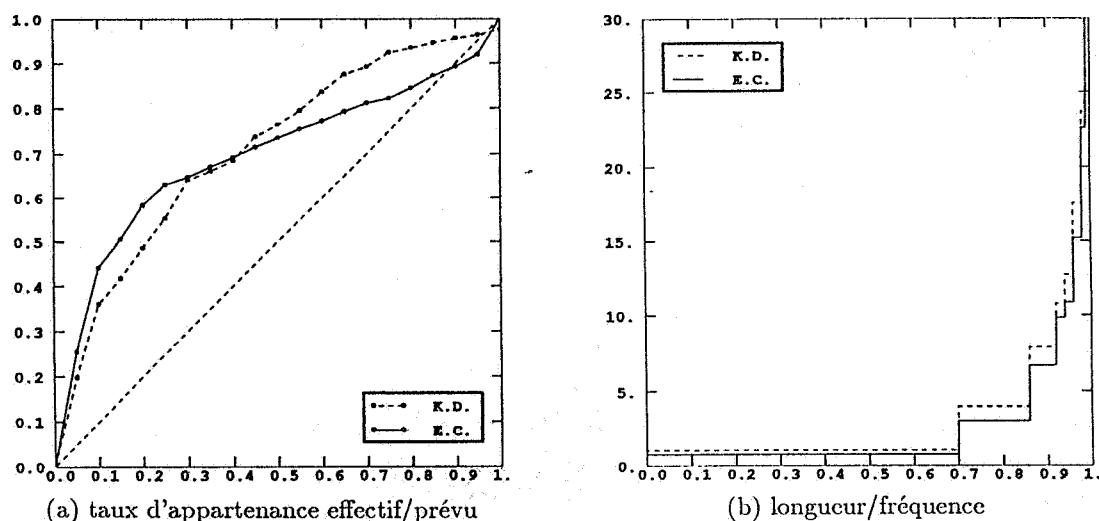
Figures 4 : Courbes des taux d'appartenance effectif/prévu des teneurs à leurs intervalles de confiance.



Figures 5 : Longueurs d'intervalles de confiance en fonction de fréquences cumulées des teneurs.



Figures 6 : Exemples de distributions conditionnelles de teneurs approchées par K.D. gamma-beta et K.D. bigaussien.

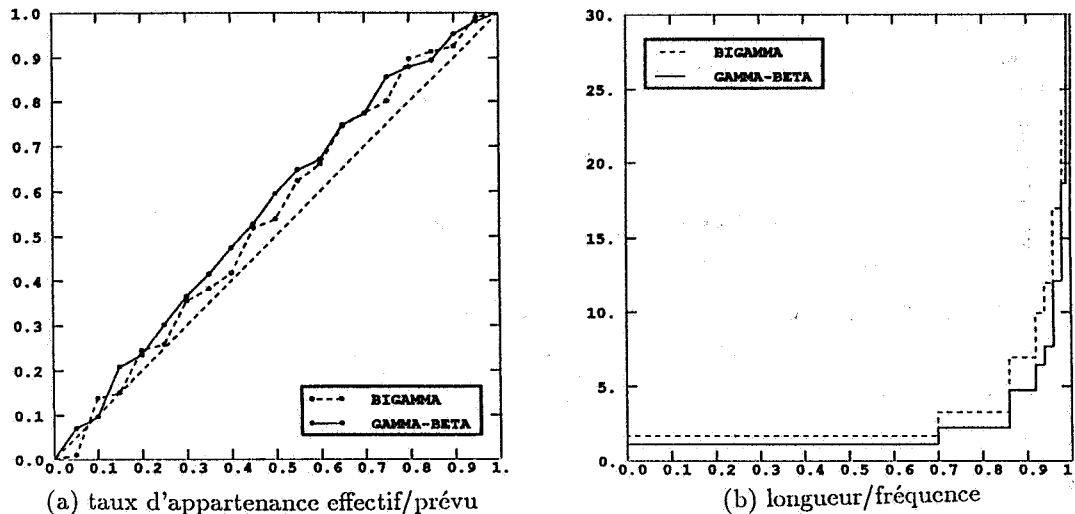


Figures 7 : Statistiques des intervalles de confiance approchés par K.D. bigaussien et E.C. trigaussienne.

Par ailleurs, on peut se demander ce qui se passe lorsque l'on utilise l'E.C. multigaussienne (trigaussienne dans le cas présent) ou le K.D. bigamma.

La figure 7(a) montre les deux courbes des taux d'appartenance effectif/prévu des teneurs à leurs intervalles de confiance centrés, fournis par K.D. bigaussien et E.C. trigaussienne. On constate que la courbe de l'E.C., comme celle du K.D., est très éloignée de la diagonale (ceci est également vrai pour les intervalles non centrés). Ce biais vient évidemment de ce que l'hypothèse trigaussienne du modèle est inadéquate (on sait déjà que les données de Luceda ne manifestent même pas de caractère bigaussien). Quant aux longueurs des intervalles, on présente ici l'exemple des intervalles centrés de 90% pour lesquels l'E.C. améliore le K.D. (Figure 7(b)).

En ce qui concerne le modèle bigamma, la courbe des taux d'appartenance effectif/prévu du K.D. bigamma est tout à fait proche de celle du K.D. gamma-beta (Figure 8(a)). Mais nous pouvons gagner en longueur d'intervalle de confiance en utilisant le K.D. gamma-beta, comme montre l'exemple des intervalles centrés de 90% (Figure 8(b)).



Figures 8 : Statistiques des intervalles de confiance approchés par K.D. bigamma et K.D. gamma-beta.

D - CONCLUSIONS

Quelques observations peuvent être tirées de cette étude expérimentale :

- Pour l'exemple de Bertholène, l'espérance conditionnelle en modèle multigaussien a donné des résultats mieux conformes à la réalité par rapport au krigeage disjonctif. Ceci suggère que, d'une part, le modèle multigaussien (plus précisément trigaussien dans ce cas) s'adapte aux données considérées, d'autre part, l'espérance conditionnelle améliore le krigeage disjonctif.
- Pour l'exemple de Luceda, le krigeage disjonctif bigaussien et l'espérance conditionnelle trigaussienne ont donné tous de mauvais résultats à cause de l'inadaptation du modèle gaussien. Or, le krigeage disjonctif en modèle gamma a fourni des résultats meilleurs.
- Pour les deux cas traités, le krigeage disjonctif, dans le cadre d'un modèle correcte de loi bivariable, a fourni une assez bonne approximation d'intervalle de confiance. Mais cette méthode peut être très biaisée lorsque le modèle est mal choisi. D'où l'importance du choix de modèle lors de l'utilisation du krigeage disjonctif.

REMERCIEMENTS

L'auteur tient à exprimer sa gratitude à Dr Ch. Lantuéjoul, Dr Ch. Lajaunie et Dr J. Rivoirard pour leurs suggestions et commentaires pendant la préparation de cet article.

BIBLIOGRAPHIE

Dowd, P.A., 1989 : Some observations on confidence intervals and kriging errors. Proceedings of the Third International Geostatistics Congress, Avignon, Sept. 5-9 1988.

Hu, L.Y. et Lantuéjoul, Ch., 1988 : Recherche d'une fonction d'anamorphose pour la mise en œuvre du krigeage disjonctif isofactoriel gamma. "Etudes Géostatistiques", Séminaire CFSG, 15-16 Juin 1987, dans Sc. de la Terre, Sér. Inf. N.28.

Hu, L.Y., 1988 : Mise en l'œuvre du modèle gamma pour l'estimation de distributions spatiales. Thèse de Doctorat en Géostatistique, ENSMP, Fontainebleau.

Journel, A., 1983 : Nonparametric estimation of spatiale distributions. Mathematical Geology, Vol.15, N.3.

Kleingeld, W., 1987 : La géostatistique pour des variables discretes. Thèse de Docteur-Ingénieur, ENSMP, Fontainebleau.

Lajaunie, Ch., 1989 : Comparing some approximate methods to build local confidence interval. to be appeared in "Mathematical Geology".

Matheron, G., 1973 : Le krigeage disjonctif. Note interne, N-360, Centre de Géostatistique, ENSMP, Fontainebleau.

Switzer, P. et Xiao, H., 1988 : Approximate prediction intervals for spatial interpolation using bivariate information: an experiment. Technical report No.118, Stanford University.