Courbes remplissant l'espace et Clustering dans les Bases de Données Spatiales

Samba Ndiaye

Dépt math-informatique
Fac. Sciences
Université C.A.D de DAKAR-FANN
SENEGAL

MOTS-CLES :

bases de données spatiales, objet, fenêtre, point multidimensionnel, adressage, cluster, accés-disque, courbe remplissant l'espace,

ABSTRACT

Les bases de données spatiales manipulent des objets complexes qui sont des ensembles de points multidimensionnels. On approxime ces objets par des fenêtres. L'accès à un objet spatial dépend donc essentiellement de l'adressage des points multidimensionnels sur le disque magnétique. L'efficacité du clustering) dépend essentiellement de l'adressage choisi. Les adressages les plus performants sont basés sur les courbes remplissant l'espace. Les résultats dans la littérature étaient jusqu'ici uniquement expérimentaux et lse limitaient à la dimension 2. D'abord nous justifions théoriquement les résultats et ensuite nous les généralisons à toute dimension. Et ceux-ci sont confirmés par les expérimentations que nous avons effectuées.

1.INTRODUCTION

Les bases de données actuelles sont incapables de modéliser les données de beaucoup de domaines de la vie courante. C'est le cas, par exemple, en cartographie(White[1981]), en robotique, en C.A.O, en V.L.S.I(Ousterhout[1984]), en vision artificielle(Laurini[1985]), en géologie(Orenstein[1988]), etc...Par exemple, comment représenter un lac sur une carte géographique. On parle alors de données spatiales ou multidimensionnelles. Cette notion de donnée spatiale varie d'une application à l'autre. En effet, les applications géographiques manipulent des données bidimensionnelles alors que celles géologiques travaillent sur des données tridimensionnelles. Par contre, la modélisation des solides en mouvement nécessite quatre dimensions. Pour représenter des données bidimensionnelles par exemple, on fait un grillage du plan.

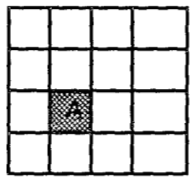

fig.1: donnée spatiale A

Ainsi, une donnée spatiale élémentaire est représentée par un carré ou pixel et la base de données spatiales est l'ensemble des données spatiales. En dimension 3, les données spatiales sont des cubes et plus généralement des hypercubes en dimension n. Chaque pixel est déterminé de façon unique par les coordonnées de son sommet minimal. On parcourt l'ensemble des données spatiales et, à chaque donnée on attribue un numéro qui sera son adresse sur le disque c'est-à-dire le secteur du disque où elle est stockée. On parle d'adressage ou de balayage.

3	7	11	15
2	6	10	14
1	5	9	13
0	4	8	12

fig2: Balayage colonne par colonne

On fait une approximation grossière d'un objet de la base par le plus petit rectangle ou fenêtre le contenant. Bien sûr, des méthodes particulières(utilisation de filtres) permettent d'affiner l'approximation(Orenstein[1988]). Nous travaillerons donc désormais avec ces rectangles que nous appelerons **fenêtres**.

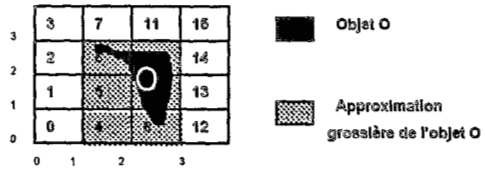

fig3 : Un objet O et son approximation F=[1,3[× [0,3[dans la base.

Un objet est donc approché par une fenêtre F. Cette fenêtre est un sous-ensemble de données spatiales repérées par leur adresse.

Par exemple, sur la fig.3 l'objet O={4,5,6,8,9,10,}={4,5,6} ∪ {8,9,10,}. Ainsi, l'objet O est formé de deux **clusters** ou groupes de données d'adresses consécutives. Celles-ci vont se retrouver dans les mêmes secteurs ou dans des secteurs voisins et pourront être accédées en un nombre minimum d'accès-disque. Il s'agit alors d'avoir le meilleur clustering possible i.e que chaque objet comporte un minimum de clusters. Or le clustering dépend essentiellement de la fonction d'adressage choisie. Beaucoup de balayages furent utilisés(balayage colonne par colonne, balayage ligne par ligne, balayage serpentin, etc...). Orenstein eût l'idée d'utiliser un balayage basé sur une courbe remplissant l'espace : la **z-courbe**. Faloutsos[1987] et Yagadish[1990] introduiront les courbes de Hilbert et Gray. Dans tous ces articles, seul le cas des données bidimensionnelles est abordé. Ensuite, les seuls résultats obtenus le sont à la suite de calculs expérimentaux; enfin aucun résultat théorique n'est établie pour le calcul du nombre de clusters pour les différentes courbes.

Dans ce papier, nous nous restreignons au cas où les fenêtres sont carrées(voir pour le cas général Ndiaye[1993]). Nous introduisons la courbe de Peano, et comparons les performances des différentes courbes pour toutes les dimensions. Nous obtenons des résultats nouveaux confortés par les calculs effectués. On peut calculer le nombre de clusters réalisés par chaque courbe. Et nous montrons que la courbe de Peano réalise de meilleures performances que les autres courbes.

Dans le paragraphe 2, nous faisons un bref survol des courbes remplissant l'espace. Dans le paragraphe 3 nous définissons les critères de performance demandés aux fonctions et donnons nos propositions. Dans le paragraphe 4, nous donnons les résultats des calculs expérimentaux. Enfin nous concluons et donnons quelques directions de recherche dans le paragraphe 5.

2. COURBES REMPLISSANT L'ESPACE

Ce sont des applications de **R** dans $\mathbf{R^n}$, bijectives, continues(courbe t—>x(t)) passant par tout point de l'espace $\mathbf{R^n}$ une et seule fois). Leurs applications réciproques ont des ensembles de discontinuité de mesure nulle. Cela veut dire concrètement que si deux points sont voisins dans $\mathbf{R^n}$, leurs images dans **R** par l'application réciproque d'une courbe remplissant l'espace seront voisines dans **R**. Les balayages sont basés sur ces applications réciproques.

La première courbe fut introduite par G.Peano[1890], la deuxième par Hilbert[1891] et il y en eut beaucoup d'autres(Sierpinski[1912], Gray[1953], etc...) Toutes ces courbes peuvent être construites à partir d'algorithmes aussi bien analytiques que géométriques.

2.1 La z-courbe

Elle fut proposée par Orenstein[1986]]. On se restreint sans perte de généralité à l'intervalle [0,1]. Chaque élément t de [0,1] est exprimé en base 2 sous forme binaire $t=t_1t_2t_3t_4.....$L'image(x,y) de t est obtenue par

$$x=t_1t_3....$$
$$y=t_2t_4....$$

En fait, on construit cette courbe géométriquement, de manière récursive comme le montre la figure 4.

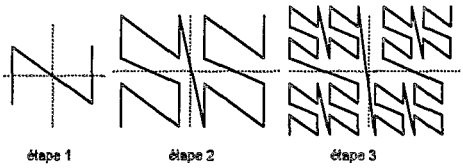

fig.4 Construction récursive de la z-courbe.

On construit géométriquement la z-courbe selon l'algorithme suivant. A l' étape 1, on a une figure de base. A l'étape 2, on subdivise le carré en 4 quadrants, dans chaque quadrant on reproduit à une échelle de 1/4 la courbe de base et on relie les quatre figures obtenues. Et ainsi de suite...La courbe se construit donc par approximations successives. On peut construire la courbe avec le degré d'approximation voulu. A l'étape m, on a 2^{2m} points multidimensionnels i.e la taille de la base de données spatiales est égale à 2^{2m}.

2.2 La courbe de Hilbert

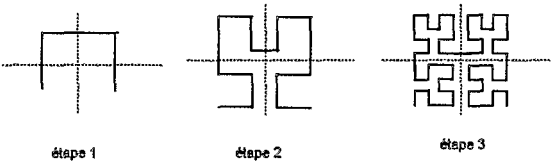

fig.5 Construction récursive de la courbe de Hilbert.

Contrairement à la z-courbe, la courbe de Hilbert n'a pas de "saut" i.e on passe d'un point au suivant en changeant une seule coordonnée et seulement d'une unité Ceci permet de mieux préserver la distance. La construction géométrique se fait de la même façon que pour la z-courbe. Butz[1969][1971] exhibe un algorithme analytique de construction de cette courbe. Quand les points sont exprimés en base 2, on passe d'un point à l'autre en changeant un seul bit.

2.3 La courbe de Peano

Elle fût la première courbe remplissant l'espace. Peano[1890] donne un algorithme analytique de construction de la courbe. Les points sont exprimés dans une base b impaire. Dans la figure 6, on a b=3. Dans ce cas, à chaque étape, on subdivise la région considérée en 9 sous-régions. Comme pour la courbe de Hilbert, on passe d'un point au suivant en changeant une seule coordonnée et d'une seule unité.

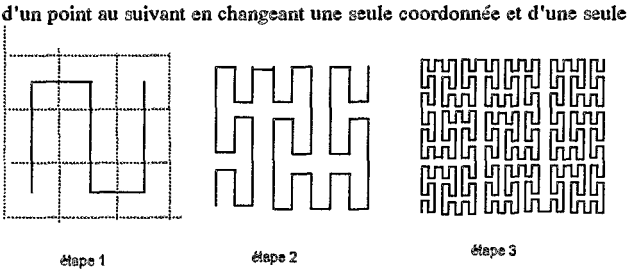

fig.6 Construction récursive de la courbe de Peano.

2.4 Construction des courbes en dimension 3

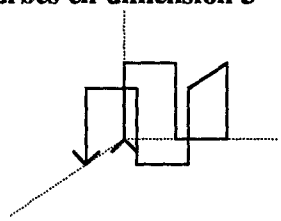

fig.7 Figure de base de la courbe de Peano en dimension 3.

On peut faire de même pour la courbe de Hilbert et les autres courbes.

3. PROPOSITIONS

3.1 Critères de performance.

Les données spatiales sont confondues avec leur sommet minimal. Chaque fenêtre est donc un ensemble de sommets ou points. Un cluster sera ainsi un ensemble de points d'adresses consécutives. On choisit comme mesure de coût le nombre de clusters dans une fenêtre F. En effet, le nombre d'accès-disques nécessaires pour accéder à cette fenêtre i.e l'amener en mémoire vive dépend essentiellement de ce nombre. On s'intéresse à deux situations. D'abord on veut mesurer le nombre moyen de clusters pour toutes les fenêtres F d'une taille donnée, quand varie la taille de la base. Ensuite on veut mesurer la sensibilité de ce nombre de clusters quand varie la taille des fenêtres, celle de la base restant constante.

3.2 Résultats.

Proposition1 : Pour une fenêtre F donnée, pour chacune des trois courbes considérées, pour une taille de la base fixée, le nombre de clusters ne dépend que de la nature des points qui sont à sa frontière (au sens topologique).

Preuve: Elle est évidente. Un point qui est sur la fenêtre est un point d'entrée e de la courbe dans la fenêtre , ou un point de sortie s , ou les deux à la fois, ou enfin ni l'un, ni l'autre. Un couple formé par un point d'entrée et un point de sortie détermine un cluster.

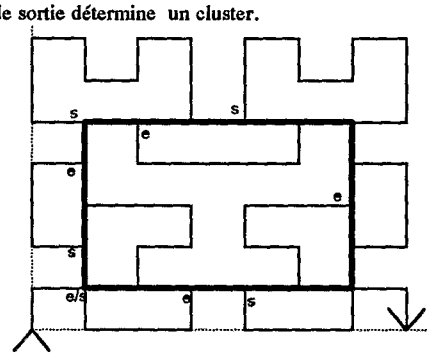

fig.8: Nombre de clusters pour F=[1,6[× [1,5[pour la courbe de Hilbert en dimension 2. On dénombre 5 clusters dans F.

Proposition 2: Le nombre de clusters pour une fenêtre donnée F ne dépend pas de la taille de la base de données.

Preuve: C'est une conséquence de la construction récursive des courbes. Quand on passe de l'ordre d'approximation m à m+1, la fenêtre F se retrouve toute entière dans un seul des quatre quadrants. Elle n'a pas d'intersection commune avec les trois autres. Donc l'augmentation de la taille de la base n'a aucune influence sur le nombre de clusters contenus dans la fenêtre F.

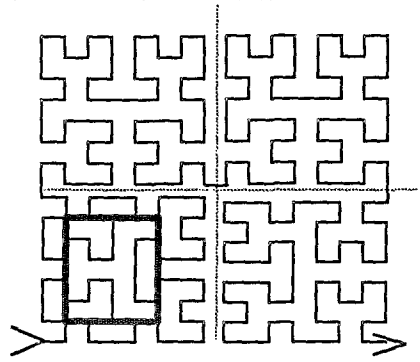

fig.9 Le nombre de clusters pour F=[1,6[× [1,5[pour la courbe de Hilbert en dimension 2 ne varie pas quand la taille de la base augmente.

Proposition 3: Le nombre de clusters pour les fenêtres F de taille donnée dépend aussi de la position de celles-ci dans la base. De plus il varie entre le cas le plus favorable qui vaut 1 et le cas le plus défavorable qui vaut 2c si c=côté de la fenêtre.

Preuve: évidente. Quand la base toute entiére est contenue dans une fenêtre, alors le nombre de clusters vaut 1. Si, par contre, chaque point de la frontière de la fenêtre est un point d'entrée/sortie, alors le nombre de clusters vaut $\frac{4c}{2} = 2c$ .C.Q.F.D.

Il s'agit donc d'observer une moyenne. Celle-ci dépend de la distribution des fenêtres dans la base. Plus la base est grande, plus la dispersion est forte. Nous étudions le cas de la distribution uniforme.

Proposition 4: En moyenne, en dimension 2, pour les courbes de Peano et Hilbert, pour les fenêtres de forme carrée, de coté c, le nombre de clusters varie de c à 2c quand varie la taille de la base.

Preuve: A chaque point d'entrée correspond un point de sortie d'une part, d'autre part à tout point n'étant pas un point d'entrée ou sortie correspond nécessairement un deuxième point de même nature. En moyenne, un point sur 2 de la frontière de la fenêtre est soit un point d'entrée, soit un point de sortie, ou les deux à la fois. Or un cluster correspond à un couple de points d'entrée/sortie. On a donc $\frac{4c}{4} = c$ clusters. Quand la taille de base augmente, on tend vers le cas le plus défavorable i.e tout point de la frontière de la fenêtre est un point d'entrée/sortie d'où $\frac{4c}{2} = 2c$ clusters. D'où le résultat.

Proposition 5: En moyenne, en dimension 2, pour la z-courbe , pour les fenêtres de forme carrée et de coté c, le nombre de clusters est voisin de 2c quand varie la taille de la base.

Preuve: La z-courbe a une configuration géométrique (voir fig.9) qui fait que tout point de la frontière de la fenêtre est un point d'entrée/sortie. D'où le résultat.

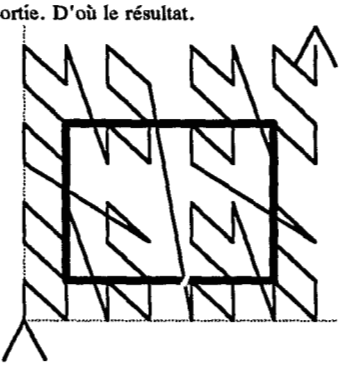

fig.10: Le nombre de clusters pour F=[1,6[× [1,5[pour la z-courbe en dimension 2. On dénombre 11 clusters dans F.

Proposition 6 :En moyenne, en dimension 2, pour les fenêtres de forme carrée et de coté c, les courbes de Peano et de Hilbert ont des performances égales. Celles-ci sont meilleures que celles de la z-courbe.

Preuve: C'est une conséquence des propositions 4 et 5.

Proposition 7: En moyenne, pour toute dimension $n>2$, pour la courbe de Peano, pour les hypercubes de taille c^n, le nombre de clusters est proche de c^{n-1} quand varie la taille de la base.

Preuve: En dimension 3, par exemple, la courbe de Peano parcourt la fenêtre plan par plan. Comme il y a c plans à parcourir et que pour chaque plan il y a c clusters, on trouve c^2 clusters. Le cas général se traite par récurrence sur n. Supposons qu'en dimension n-1, le nombre de clusters soit égal à c^{n-2} clusters. En dimension n, la courbe doit parcourir c hyperplans l'un après l'autre. D'où le résultat.

Proposition 8: En moyenne, pour toute dimension $n>2$, pour la courbe de Hilbert, pour les hypercubes de taille c^n, le nombre de clusters est compris entre $\frac{2nc^{n-1}}{4}=\frac{2nc^{n-1}}{2}$ et $\frac{2nc^{n-1}}{2}=nc^{n-1}$ quand varie la taille de la base.

Preuve: Contrairement à la courbe de Peano, celle de Hilbert parcourt la fenêtre dans tous les sens possibles. On se retrouve dans le cas où il faut considérer la nature de tous les points de la frontière de la fenêtre. Donc le nombre de clusters sera compris entre $\frac{x}{4}$ et $\frac{x}{2}$ si x est la taille de la frontière de l'hypercube. Or une hypercube $[0,c]^n$ de taille c^n a une frontière de taille $2nc^{n-1}$. D'où le résultat.

Proposition 9: En moyenne, pour toute dimension $n>2$, pour la z-courbe, pour les hypercubes de taille c^n, le nombre de clusters est voisin de $\frac{2nc^{n-1}}{2}=nc^{n-1}$ quand varie la taille de la base.

Preuve: Du fait de la nature géométrique de la z-courbe, chaque point de la frontière de l'hypercube est un point d'entrée/sortie. D'où le résultat.

Proposition 10: Pour toute dimension $n>2$, en moyenne, la courbe de Peano est meilleure que la courbe de Hilbert qui est elle-même supérieure à la z-courbe.

Preuve: C'est une conséquence immédiate des trois précédentes propositions.

Variation du nombre de clusters par rapport à la taille des fenêtres

On maintient constante la taille de la base et on fait varier la taille des fenêtres. Etudions alors le comportement du nombre de clusters. Nous avons la proposition suivante.

Proposition 11: Soit une fenêtre de coté c et NC le nombre de clusters qu'elle contient, alors la variation relative de NC est proportionnelle à celle de $\Delta c/c$ quand varie la taille de la fenêtre et que celle de la base reste constante. Le rapport de proportionnalité depend de la dimension n de l'espace. Plus précisément, $\frac{\Delta NC}{NC} = (n-1)\frac{\Delta C}{C}$.

Preuve:On a $NC=pc^{n-1}$ où $p=1$ pour Peano, $p=\frac{n}{2}$pour Hilbert et $p=\frac{n}{4}$ pour la z-courbe d'où $d(Nc)=p(n-1)c^{n-2}d(c)$ d'où $cd(NC)=p(n-1)c^{n-1}d(c)=(n-1)NCd(c)$ d'où le résultat.

4. EXPERIMENTATIONS

4.1 Conditions des expérimentations:

Pour les calculs en dimension 2, 200 expériences sont effectuées systématiquement, pour chaque courbe, pour chaque taille de fenêtre et pour taille de la base. On a fait le choix d'une distribution uniforme des fenêtres dans la base. En dimension >2, nous avons effectué 50 à 100 expériences selon les cas. Nous présentons ici quelques cas.

4.2 Expériences:

Seuls quelques cas sont présentés ici. Voir Ndiaye[1993] pour des résultats plus complets. Les expériences faites et l'état de l'art confirment nos propositions. Dans toutes les figures ci-après, on a en abscisse le logarithme décimal de la taille de la base.

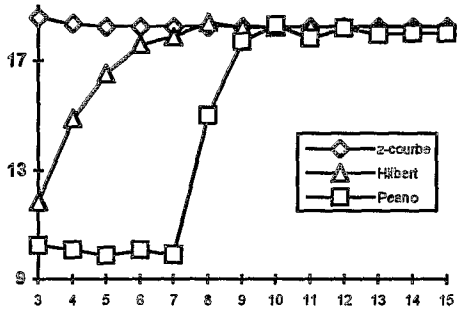

fig.11 Le nombre de clusters pour les fenêtres de coté c=10 en dimension 2 quand varie la taille de la base. On a en abscisse le logarithme décimal de cette taille.

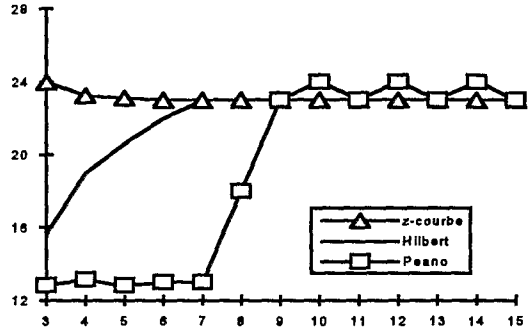

fig.12 Le nombre de clusters pour les fenêtres de coté c=13 en dimension 2 quand varie la taille de la base.

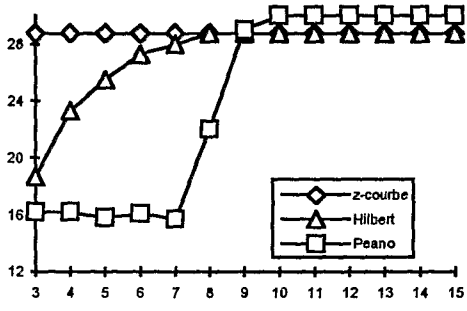

fig.13 Le nombre de clusters pour les fenêtres de coté c=16 en dimension 2 quand varie la taille de la base.

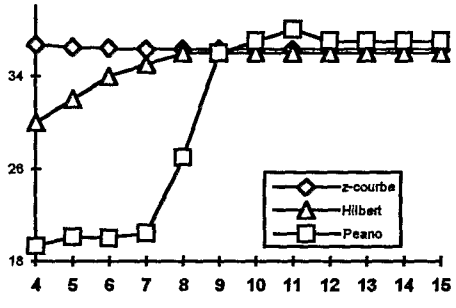

fig.14 Le nombre de clusters pour les fenêtres de coté c=20 en dimension 2 quand varie la taille de la base.

Les figures 11 à 14 montrent, en dimension 2, le comportement du nombre de clusters quand la taille de la base augmente. Pour de petites valeurs de la taille de la base, la courbe de Peano est meilleure que la courbe de Hilbert qui, elle-même, est meilleure que la z-courbe. Mais, dès que la taille de la base

augmente, les performances s'égalisent et restent à peu constantes. Ceci était prévu par l'analyse et déjà observé dans l'état de l'art(Yagadish[1990]).

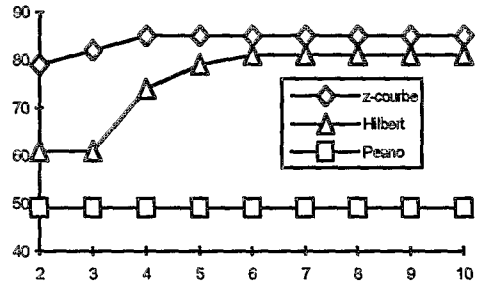

fig.15 Le nombre de clusters pour les fenêtres de coté c=7 en dimension 3 quand varie la taille de la base.

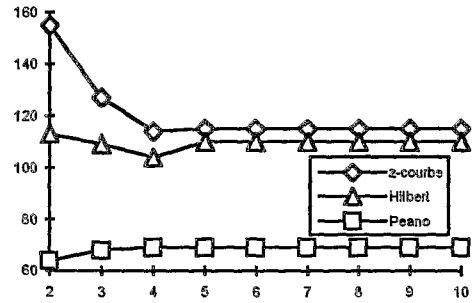

fig.16 Le nombre de clusters pour les fenêtres de coté c=8 en dimension 3 quand varie la taille de la base.

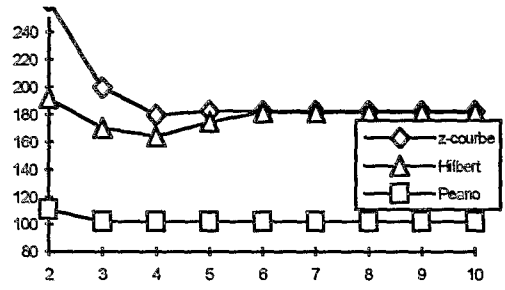

fig.17 Le nombre de clusters pour les fenêtres de coté c=10 en dimension 3 quand varie la taille de la base.

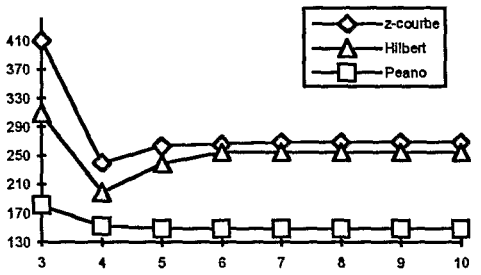

fig.18 Le nombre de clusters pour les fenêtres de coté c=12 en dimension 3 quand varie la taille de la base.

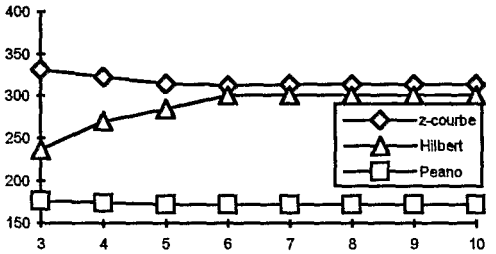

fig.19 Le nombre de clusters pour les fenêtres de coté c=13 en dimension 3 quand varie la taille de la base.

Les figures 15 à 19 montrent le comportement en dimension 3, pour des fenêtres de taille c^3 , des différentes courbes. On constate immédiatement que la courbe de Peano réalise de meilleures performances que les 2 autres. De plus, les calculs coïncident avec l'analyse pour dire le nombre de clusters pour la courbe de Peano est voisin de c^2.

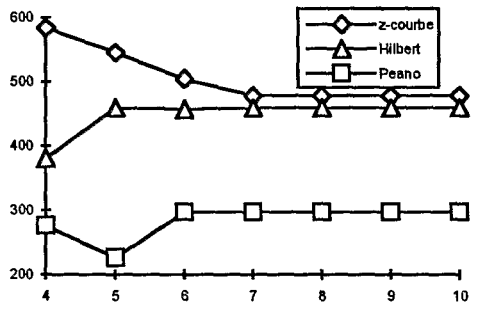

fig.20 Le nombre de clusters pour les fenêtres de coté c=6 en dimension 4 quand varie la taille de la base.

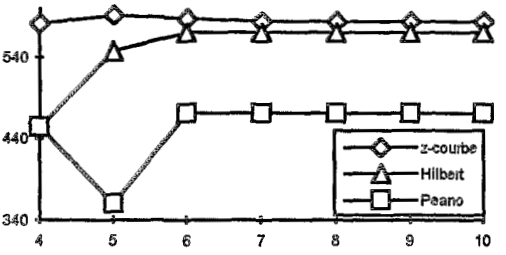

fig.21 Le nombre de clusters pour les fenêtres de coté c=7 en dimension 4 quand varie la taille de la base.

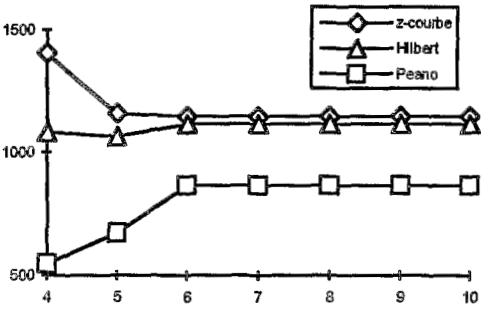

fig.22 Le nombre de clusters pour les fenêtres de coté c=8 en dimension 4 quand varie la taille de la base.

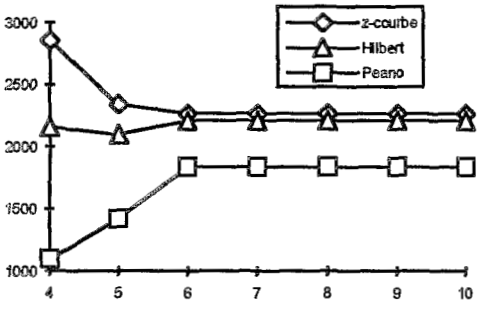

fig.23 Le nombre de clusters pour les fenêtres de coté c=10 en dimension 4 quand varie la taille de la base.

Les figures 20 à 23 montrent le comportement en dimension 4, pour des fenêtres de taille c^4 , des différentes courbes. On constate immédiatement que la courbe de Peano est meilleure que les 2 autres. De plus, les calculs coïncident avec l'analyse pour dire que le nombre de clusters pour la courbe de Peano est voisin de c^3.

Variation du nombre de clusters par rapport à la taille des fenêtres quand la taille de la base reste constante

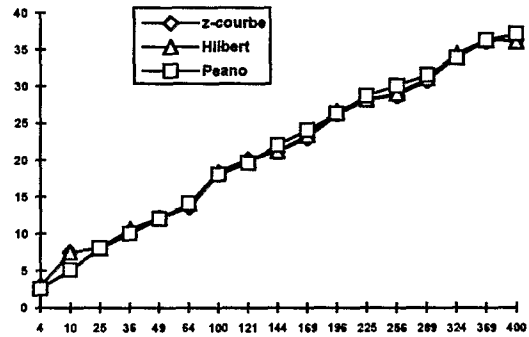

fig.24 Le nombre de clusters en dimension 2 quand varie la taille des fenêtres et que la taille de la base reste à 3^{11}.

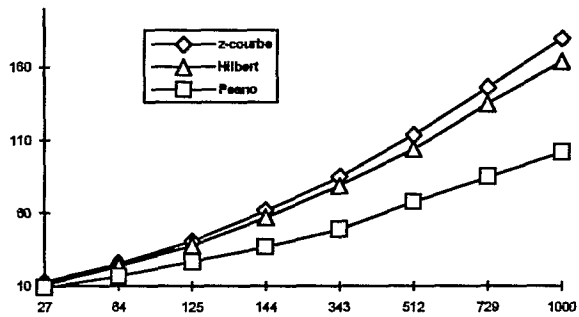

fig.25 Le nombre de clusters en dimension 3 quand varie la taille des fenêtres et que la taille de la base reste à 3^{12}.

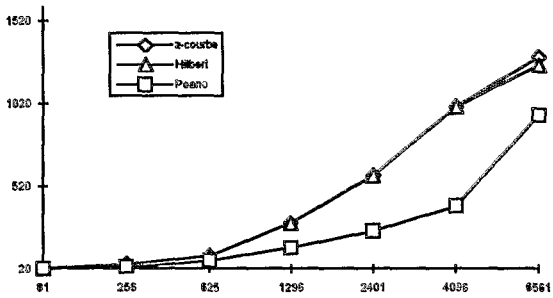

fig.26 Le nombre de clusters en dimension 4 quand varie la taille des fenêtres et que la taille de la base reste égale à 3^{14}.

Les figures 24 à 26 montrent le comportement du nombre de clusters quand on maintient constante la taille de la base et qu'on fait varier la taille des fenêtres. Elles confirment que, pour toute dimension, la variation relative de ce nombre est proportionnelle à la variation relative du coté c des fenêtres.

5.CONCLUSIONS ET PERSPECTIVES.

Dans cet article, nous avons introduit une courbe, celle de Peano dans le domaine du clustering dans les bases de données spatiales. Nous nous sommes limités à des approximations de forme "carrée" des objets de la base. Cette limitation est levée dans Ndiaye[1993]. Nous obtenons d'abord des résultats théoriques nouveaux qui montrent que le nombre de clusters, pour une fenêtre donnée, dépend de la nature des points qui sont sur sa frontière. Ceci nous permet des donner des expressions analytiques du nombre de clusters, pour toute dimension n. Nous montrons ensuite pourquoi ce nombre est indépendant de la taille de la base. Enfin nous prouvons que, quand la taille de la base est constante, le nombre de clusters dans une fenêtre carrée de coté c a une variation relative égale à celle de c.

Dans un travail en cours, nous appliquons ces résultats au domaine du hachage multiclé(Faloutsos[1987]).

REMERCIEMENTS:

Tous mes remerciements au professeur G.Lèvy qui m'a initié à la recherche en informatique et à Tidiane Seck et Boualem Bounar qui ont bien voulu relire le manuscrit.

REFERENCES:

Butz[1968]: Space-filling curves and mathematical programming.
Information and Control 12, 314-330 (1968)

Butz[1969]:Convergence with Hilbert's space filling curve.
Journal of computer and System Sciences(3) 128-146(1969)

Butz[1971]: Alternative algorithm for Hilbert's space-filling curve.
I.E.E.E Trans.Computer vol C-20 April 1971

Faloutsos[1986]: Multiattribute Hashing using Gray codes
SYGMOD 1986 ACM 0-89791-191-1/86/0500/0227

Faloutsos[1987]: Gray codes for partial match and range queries
IEEE trans. on Software Engineering 1987

Faloutsos[1989]:Fractals for secondary key retrieval.
Proc.of the A.C.M on the Principles of Database syst. Mars 1989, 247-252

Faloutsos[1990]:Spatial access methods using Fractals:algorithm and perfomance evaluation.
Rapport technique Univ.Maryland

Gray[1957]:Pulse code ommunications,
U.S Patent 2632058, March 17, 1953

Hilbert[1891] : Ueber Stetege Abbildumg einer linie auf ein, Flashentruck 1981 pp 403
Math.Annalen(38) 457-460 (1891)

Laurini[1985]:Graphical databases built on Peano space-filling curves.
Eurographics 1985 Elsevier Sciences Publishers(North-Holland)

Ndiaye[1993]:Thèse:"Utilisation des courbes de Peano-Hilbert dans la gestion des objets dans les bases de données spatiales" sous la direction de G.Lèvy, Universite Paris-IX Dauphine, Sept. 1993

Orenstein & Merett[1984]:A class of data structures for associative searching.
1984 ACM 0-89791-128-8/84/004/0181 3

Orenstein[1986]: Spatial query processing in object-oriented database system.
1986 ACM 0-89791-191-1/86/0500/0326

Orenstein & Manola[1988]: PROBE spatial data modeling and query processing in an image database application.
I.E.E.E Trans.on Softw. Eng. vol 14,n°5 May 1988

Ousterhout, Hamachi, Mayo,Scott and Taylor[1984]:"Magic:A VLSI Layout System" 21st Design Automation Conference pp. 152-159, Alburqueque, NM, June 1984

Peano[1890] : Sur une courbe qui remplit toute une aire plaine.
Math.Annalen(36) 157-160 (1890)

Sierpinski[1912]: Sur une nouvelle courbe continue qui remplit tout une aire plane.
Bull.Acad.Sci.Cracovie.Série A 462-478 (1912)

White[1981]:N-trees:Large ordered Indexesfor Multi-dimensional Space, Application Mathematics Research Staff, Statiscal Research Division, U.S. Bureau of the Census, Dec. 1981

Yagadish[1990]: Linear clustering of objects with multiple attributes.
1990 ACM 089791-365--5/90/0005/6332-51-50