

J. P. BENZÉCRI

**Algorithme accéléré de recherche des plus proches voisins après agrégation en boules de rayon borne**

*Les cahiers de l'analyse des données*, tome 9, n° 1 (1984), p. 119-122

[http://www.numdam.org/item?id=CAD\\_1984\\_\\_9\\_1\\_119\\_0](http://www.numdam.org/item?id=CAD_1984__9_1_119_0)

© Les cahiers de l'analyse des données, Dunod, 1984, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

## ALGORITHME ACCÉLÉRÉ DE RECHERCHE DES PLUS PROCHES VOISINS APRÈS AGRÉGATION EN BOULES DE RAYON BORNÉ

[VOIS. BOUL.]

par J.P. Benzécri

Le principe de l'algorithme ci-dessous (qui présuppose l'exécution de l'algorithme d'agrégation [ALG. AGR. RAY.] ; cf. *Cahiers* Vol IV n° 3 pp 365 sqq) est de borner la recherche des plus proches voisins d'un point PJ, à l'ensemble des points POSIJ inclus dans des boules dont le centre est assez proche de PJ.

entier KI, KJ, KN, KVMIN, KVMAX, KBO, KB, KV; KVO, I, N, M, KZ ;

*Commentaire* : KI est le nombre des points POSIJ parmi lesquels on recherche des voisins au point PJ ; KJ est la dimension de l'espace ambiant (ici on ne détaillera pas les calculs de coordonnées, comprises dans la procédure DISTANCE ; cf. *infra*) ; KN est le nombre des boules en lesquelles ont été agrégés les KI points ; KVMIN et KVMAX sont deux bornes entre lesquelles on désire que soit compris le nombre des plus proches voisins effectivement sortis (éventuellement ce nombre est imposé : KVMIN = KVMAX ; en général, pour plus de souplesse KVMIN < KVMAX) ; KB est le nombre des boules au sein desquelles s'effectue la recherche des voisins ; KBO est ce que valait ce nombre avant l'appel en cours de la procédure de recherche (éti-quetée BOULIER) ; KV est le nombre des voisins envisagés, KVO est le nombre des voisins éventuellement trouvés à un passage précédent, KZ le numéro d'un voisin considéré sur la liste de ceux-ci ; I est l'indice d'un point ; N, l'indice d'une boule ; M, une position sur la liste des boules retenues. On notera que KVMIN et KVMAX sont des données (les bornes du nombre de v. demandés) ; KV un résultat (le nombre des voisins sortis) ; KB un résultat auxiliaire (le nombre de boules où il a fallu choisir ces voisins).

Réel RAY, DIS, DISO, DSU, CD, AD ;

*Commentaire* : comme dans [ALG. AGR. RAY.], RAY est le rayon des boules, ou distance maxima d'un point au centre auquel il est agrégé ; DIS est la distance en deçà de laquelle on cherche présentement les voisins ; DISO est la valeur de DIS avant l'étape en cours ; CD (que nous proposons de fixer à 0,1) est le coefficient définissant en fonction de RAY la valeur initiale de DIS ; AD (que nous proposons de fixer à 1,5) est le coefficient par lequel DIS est modifié quand on élargit la recherche des voisins ; DSU est la distance seuil en deçà de laquelle doivent être pris de nouveaux voisins, compte tenu de ceux présentement retenus :

entier tableau IDN, IFN, NB[1:KN], IV[1:KVMAX] ;

---

(1) J.P. Benzécri, professeur de statistique. Université Pierre et Marie Curie.

*Commentaire* : à la différence de ce qui est dans [ALG. AGR. RAY], on suppose que les points ont été numérotés suivant des segments successifs correspondant chacun à l'une des boules : en sorte que les points rattachés au centre N ont des indices I variant de IDN[N] (début) à IFN[N] (fin) ; NB est le tableau de numérotage des KB boules retenues pour la recherche des voisins, les indices N de ces boules étant NB[1], NB[2], ..., NB[KB] ; IV est le tableau de numérotage des KV voisins retenus, les indices de ces points étant IV[1], IV[2], ..., IV[KV].

réel tableau POSIJ[1:KI,1:KJ],CENJ[1:KN;1:KJ],PJ[1:KJ],DI[1:KI],DN[1:KN];

*Commentaire* : conformément aux notations d'[ALG. AGR. RAY.] les tableaux POSIJ et CENJ donnent respectivement les coordonnées des points du nuage (points numérotés de 1 à KI) et des centres des boules (numérotées de 1 à KN) ; de même PJ est le vecteur de coordonnées du point dont on recherche les voisins. Les tableaux DI et DN contiennent respectivement les distances à PJ, des points I et des centres N : les distances des centres sont toutes calculées dès le début de l'exécution du programme ; les distances des points sont calculées boule par boule au fur et à mesure que celles-ci sont explorées.

étiquette BOULIER,ETI1,ETI2,ETI3 :

*Commentaire* : l'étiquette BOULIER est un repère essentiel : c'est par elle qu'on passe chaque fois qu'on entreprend d'étendre à de nouvelles boules la recherche des voisins ; les étiquettes ETIx distinguent diverses éventualités qui se rencontrent dans le tri des voisins.

réel procédure DISTANCE(PJ1[1:KJ],PJ2[1:KJ]) ;

*Commentaire* : Cette procédure que nous ne cherchons pas à décrire ici, donne la distance entre deux points PJ1, PJ2 de l'espace de dimension KJ.

lire entier KI,KJ,KN,KVMIN,KVMAX ;  
lire réel RAY,CD,AD ;  
lire entier tableau IDN,IFN[1:KN] ;  
lire réel tableau POSIJ[1:KI,1:KJ],CENJ[1:KN,1:KJ], PJ[1:KJ] ;

*Commentaire* : répétons que  $KVMIN \leq KVMAX$  ; et que les valeurs proposées pour CD et AD sont  $CD=0,1$  ;  $AD=1,5$  ; on commentera ces valeurs ultérieurement.

pour N:=1 pas 1 jusqu'à KN faire  
DN[N]:=DISTANCE(PJ[.],CENJ[N,.]) ;

*Commentaire* : les distances des centres à PJ pouvant chacune servir plusieurs fois on les range dès le départ en un tableau.

DIS:=CD\*RAY;KBO:=0;DISO:=-RAY;KVO:=0

*Commentaire* : en choisissant initialement  $DIS:=CD*RAY=0,1RAY$ , on est conduit (cf. *infra*) à explorer les boules dont les centres sont à une distance de PJ inférieure à  $DIS+RAY$  : on a donc fixé DIS petit afin de n'avoir pas à trier les points d'un trop grand nombre de boules ; d'ailleurs avec le coefficient d'accroissement AD (1,5) choisi pour DIS, on explorera rapidement un nombre suffisant de boules. Initialement, KBO est nul parce qu'on n'a encore considéré aucune boule ; le choix de DISO (distance à l'étape précédente, qui en fait n'existe pas...) est fait pour que  $DISO+RAY$  (cf. *infra*) ait la valeur zéro ; KVO est nul puisqu'aucun voisin n'a encore été trouvé.

BOULIER ;  
KB:=KBO;KV:=KVO;DSU:=DIS;

*Commentaire* : à chaque passage par l'étiquette BOULIER, on envisage de nouvelles boules à explorer, et celles-ci sont numérotées à partir du rang maximum KBO atteint antérieurement ; en revanche le numérotage des voisins éventuels est chaque fois repris à zéro, avec le nouveau seuil DIS en deçà duquel les points I sont seuls considérés.

pour N:=1 pas 1 jusqu'à KN faire  
si DISO+RAY≤DN[N]<DIS+RAY alors début  
KB:=KB+1;NB[KB]:=N fin ;

*Commentaire* : Les points I situés à une distance de PJ inférieure strictement à DIS, sont certainement affectés à un centre N dont la distance à DN[N]PJ est inférieure strictement à DIS+RAY : ce sont les boules afférentes à ces centres, qu'il faut explorer ; d'ailleurs si DISO+RAY≤DN[N], la boule N a déjà été retenue donc il ne convient pas de l'adjindre à la liste.

pour M:=1 pas 1 jusqu'à KB faire  
pour I:=IDN[NB[M]] pas 1 jusqu'à [FN[NB[M]] faire début

*Commentaire* : on passe successivement en revue toutes les boules à explorer, (depuis la boule NB[1], jusqu'à la boule NB[KB]) ; les points de ces boules étant comme on l'a dit préalablement rangés en séquence.

si KBO<M alors DI[I]:=DISTANCE(PJ[.],POSIJ[I,.]) ;

*Commentaire* : s'ils s'agit d'une nouvelle boule non encore explorée, il faut calculer les distances à PJ de ses points I : sinon ce-là est déjà fait. Ici apparaît l'intérêt qu'il y a à distinguer entre "nouvelle boule" et "boule déjà vue".

si DI[I]≥DSU ou DI[I]<DSO aller à ETI3 ;

*Commentaire* : au-delà du seuil DSU le point I n'est pas à retenir même temporairement ; de même si DI[I] DISO, il s'agit d'un voisin déjà reconnu et bien placé.

si KV<KVMAX alors KV:=KV+1 ;

*Commentaire* : si le nombre KV des points déjà retenus pour pouvoir être des voisins de PJ, est strictement inférieur au maximum admissible KVMAX, on peut retenir un point de plus sans éliminer aucun de ceux en attente ; sinon c'est le dernier point (numéroté IV[KV]) qui étant le plus éloigné de PJ (cf. *infra*) est éliminé ; de toute façon, le numéro initial du nouveau point admis sera KV, d'où ce qui suit :

IV[KV]:=I;KZ:=KV;  
ETI1;  
si KZ<2 aller à ETI2;

*Commentaire* : il faut que les points retenus soient rangés de 1 à KV par ordre de distance croissante au centre PJ : i. e. que l'on ait  $DI[IV[1]] \leq DI[IV[2]] \leq \dots \leq DI[IV[KV]]$  ; or rien ne prouve que le dernier point admis (d'indice IV[KV]), soit à sa place : on le fait donc remonter par échange avec les points plus éloignés que lui de PJ. Cette partie de saute-mouton s'arrête quand I parvient à un rang KZ où il rencontre un point IV[KZ-1] plus proche que lui de PJ ; à moins que KZ ne soit inférieur à 2 (i.e. ne vale 1) auquel cas I s'arrête ayant atteint le premier rang !

```

si DI[I]<DI[IV[KZ-1]] alors début
  IV[KZ ]:=IV[KZ-1];IV[KZ-1]:=I;KZ:=KZ-1;
  aller à ETI1 fin;

```

*Commentaire* : ôte-toi de là que je m'y mette ; et l'on remonte à ETI1 pour tenter de faire avancer encore le point I dans la liste des IV ;

```

ETI2;
si KV:=KVMAX alors DSU:=DI[IV[KV]];

```

*Commentaire* : au départ, DSU=DIS : c'est-à-dire que tout point I dont la distance à PJ est inférieure à DIS, est candidat à être un voisin ; mais dès lors qu'on a déjà retenu un ensemble de points I en nombre égal au maximum KVMAX, un nouveau point I n'est admissible que s'il est plus proche de PJ que ne l'est le plus écarté des points présentement retenus, (i.e. compte tenu du rangement effectué, le point IV[KV]) ; et en ce cas (cf. *supra*) son admission se fait en chassant IV[KV].

```

ETS fin ;

```

*Commentaire* : ici s'achève l'exploration des boules qu'on a retenues.

```

si KV<KVMIN alors début
  DISO:=DIS;DIS:=AD*DIS;KBO:=KB;KVO:=KV;
  aller à BOULIER fin

```

*Commentaire* : si le nombre KV des voisins qu'on a trouvé est trop faible, il faut explorer de nouvelles boules, en se souvenant du nombre KBO de celles déjà considérées ainsi que de celui de KVO des voisins déjà trouvés ; les points I qu'on cherche pouvant être à une nouvelle distance DIS du point PJ. On notera que le coefficient AD (e.g. 1,5) doit être suffisant pour qu'on accède rapidement à un nombre suffisant de voisins ; sans toutefois être retardé par l'examen de trop nombreux candidats.