LES CAHIERS DE L'ANALYSE DES DONNÉES

A. HATHOUT

La régression d'après un nombre variable de voisins

Les cahiers de l'analyse des données, tome 8, nº 1 (1983), p. 19-26

<http://www.numdam.org/item?id=CAD_1983__8_1_19_0>

Les Cahiers de l'Analyse des Données
Vol. VIII - 1983 - n° 1 - p. 19-26

LA RÉGRESSION D'APRÈS UN NOMBRE VARIABLE DE VOISINS

[REGR. NVAR. VOIS.]

par A. Hathout (1)

1 *Exposé statistique : critique des méthodes en usage.* Dans les algorithmes de régression basés sur une représentation géométrique de données sans hypothèse de linéarité, on part d'un ensemble S d'individus (appelés données de base) décrits par deux sous-ensembles de variables I et J :

- I est l'ensemble des variables à expliquer (ou plus précisément dans le cas que nous abordons des modalités d'une variable à expliquer).

- J est l'ensemble des variables explicatives.

Nous supposons qu'une analyse des correspondances a porté sur un tableau P croisant I par J ; P(i,j) = le nombre (de S) ayant les modalités $i \in I$ et $j \in J$ (2).

Pour un point p que nous représentons ainsi en supplément dans l'espace parmi les points de l'ensemble I (et, éventuellement, ceux de S en supplément aussi) nous supposons connues les variables explicatives (J) mais non la modalité de la variable à expliquer (ou les modalités des variables à expliquer s'il y en a plusieurs).

Dans ce qui suit, nous raisonnons sur le cas où I est l'ensemble des modalités d'une variable à expliquer unique ; la généralisation à d'autres situations (e.g. au cas de plusieurs variables; ou au cas d'un ensemble de classes) ne posant pas de problème particulier.

Dans l'algorithme présenté par M.O. Lebeaux (cf. C.A.D. Vol. II n° 4, pp 467 sqq ; 1977), on se propose d'estimer la donnée manquante pour p par la moyenne des données correspondantes relatives aux n points de S les plus proches de p. On suggère également qu'on pourrait estimer la valeur manquante de p par la moyenne des valeurs prises par les points de S situés à l'intérieur d'une boule de centre p et de rayon r.

---

(1) *Ingénieur.*

(2) *Si J représente des variables continues (postes de budget par exemple) et I les modalités de la variable à expliquer P(i,j) sera la somme des valeurs que prennent pour la variable j (continue) les individus de S ayant la modalité $i \in I$.*

Ni l'une, ni l'autre de ces deux propositions ne nous satisfont pleinement : nous craignons que n'y persiste l'*a priori* :

Faut-il se fier à notre "pouvoir de prévision" - déjà fragile quand des données inconnues ou inaccessibles peuvent compromettre la "stabilité des axes " ?

Peut-on prétendre entourer p d'une boule de rayon r (solution moins dangereuse que la recherche de voisins car elle laisse la place à l'heureuse possibilité d'une boule vide, i.e. situation pour laquelle il n'est pas licite de "prévoir") alors que reste posée la question fondamentale : comment fixer ce rayon r ?

Les travaux qui nous ont mené à cette position sont d'une part l'étude de l' "agrégabilité" d'un ensemble de points autour d'une "cible de partition" (ou d'un système de k points appelés centres), (cf. A. Hathout, thèse 3e cycle, Paris 1973), inspirée d'un algorithme dû à E. Diday(La méthode des nuées dynamiques , I.R.I.A., 1970) et d'autre part nos travaux sur le Répertoire Français des Emplois qui nous ont amené à concevoir une bibliothèque d'analyse des données (Manuel d'utilisation de la bibliothèque BIB; CEREQ, 1977) axée sur les problèmes d'étude "statique" et "dynamique" d'une partition qui laissait une grande liberté à l'utilisateur dans le choix de sa partition d'un ensemble I à étudier statiquement (sans modification de la partition) ou dynamiquement (avec transformation de la partition $\mathcal{P}$ de départ, proposition d'une partition $\mathcal{P}'$ générée par $\mathcal{P}$ et étude des "migrations" (1) des individus de $\mathcal{P}$ à $\mathcal{P}'$). Nous avons ainsi cru bon de nous contenter de l'étude de la "régressibilité" plutôt que de procéder à la conception d'un algorithme de régression. En effet, l'étude de la " régressibilité" équivaut dans l'optique de la bibliothèque en question (bibliothèque BIB) à une étude dynamique d'une partition :

Supposons que nous ayons à exprimer les k modalités de la variable v en fonction d'un ensemble J de variables explicatives (si la variable v est continue, nous la transformons en variable qualitative en considérant les tranches de variation comme modalités).

Nous posons comme partition de départ, le résultat de l'action suivante : la classe de $i \in I$ est c avec $1 \leq c \leq k$ si la valeur de la variable v pour l'individu i est dans la tranche c.(2)

Dans quelle mesure cette partition est-elle "optimale" ?

A titre d'exemple, on se reportera à l'étude de la "*régressivité*" des salaires, pp 283 à 294 dans "Contributions à l'analyse des relations entre système éducatif et système productif" par J.J. PAUL, IREDU 1978 (3).

Nous formons ainsi une matrice carrée que nous appelons "matrice des migrations" où à la rencontre de la ligne j_1 et de la colonne j, nous inscrivons le nombre d'individus qui , ayant la

---

(1) ou "transferts", en bref, disons que la partition $\mathcal{P}'$ ne diffère de la partition $\mathcal{P}$ qu'en ce que, quelques individus sont passés d'une classe numéro q à une classe numéro q' : on édite donc en matrice des nombres k(q,q') recensant ces transferts.

(2) k = nombre de tranches de variation de v.

(3) où sous le titre " régressivité *", Monsieur J.J. PAUL parle de* "régressibilité".

modalité j_1 pour la variable v à expliquer ($1 \leq j_1 \leq k$, k modalités pour v) se trouvent plus proches du centre de gravité des individus dont la modalité pour la variable v est j que du centre de gravité de toute autre classe j' (i.e. du centre de gravité de l'ensemble des individus dont la modalité pour la variable v est $j' \neq j$). Si la diagonale de cette matrice est bien "fournie", la pratique de la régression est justifiable. Si au contraire cette diagonale est vide, procéder à une régression serait illicite.

Remarque complémentaire. En réalité, le problème est bien plus complexe car parfois la pratique de la régression est licite pour certaines modalités et injustifiable pour d'autres (les exemples de tels cas ne manquent pas).

Par ailleurs, il convient de distinguer entre deux cas :

1er cas : Les modalités de la variable à expliquer sont sémantiquement différentes, (variable nominale) exemple : la variable à expliquer est l'emploi et ses modalités sont les noms de l'emploi (agriculteur, dessinateur...).

2ème cas : Les modalités de la variable à expliquer relèvent de la même signification, (variable ordinale) exemple : la variable à expliquer est le salaire et ses modalités sont les tranches de salaire.

Dans chacun de ces deux cas, il convient d'examiner la matrice des migrations (1) mais, cependant que dans le premier cas, les associations qu'elle révèle sont *a priori* "ignorées" (par définition une nomenclature des emplois détermine des postes à contenu différent), dans le deuxième cas, ces associations sont connues (la 6ème tranche de salaire est plus proche de la 7ème tranche que de la 20ème). C'est ici particulièrement que nous avons toujours vu s'imposer une analyse des correspondances de la matrice des migrations (cf. *supra*) car même si sa diagonale n'est pas "aussi bien fournie qu'on le voudrait", la pratique de la régression peut être licite : c'est en particulier lorsque sur les axes d'inertie la suite ordonnée des modalités ne subit pas de perturbations importantes.

2 Description d'un nouvel algorithme

2.1 Avant l'utilisation de l'algorithme : Rappelons les conclusions du § 1 : en résumé, nous dirons qu'avant de procéder à une *régression d'après un nombre variable de voisins* que nous présentons ci-après, il convient :

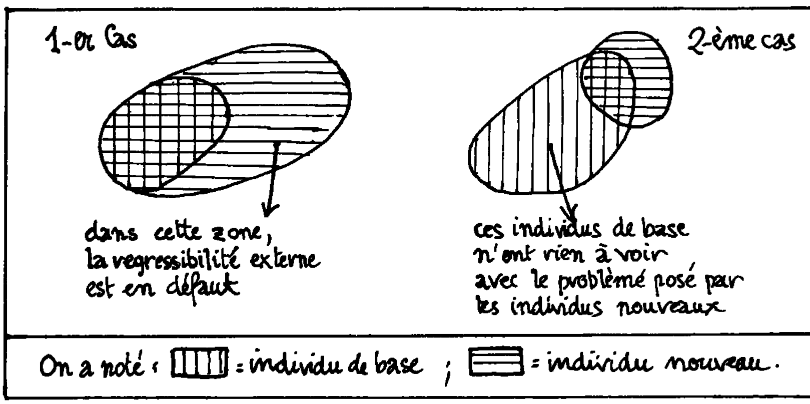
1°) - D'avoir déjà étudié "la régressibilité interne" ou la capacité des variables dites "explicatives", d'expliquer la variable dite "à expliquer" au sein de l'échantillon des individus de base ; donc indépendamment d'une situation nouvelle où nous "devons" compléter une donnée manquante (situation extérieure).

2°) - D'avoir déjà éliminé parmi les situations extérieures celles qui ne sont pas "apparentables" à ce que nous connaissons dans les

(1) Dans "Trois versions pour l'utilisation de la méthode des nuées dynamiques avec application à l'analyse des échanges" CEREQ 1975, on trouve que l'agrégation de la nomenclature des emplois en 62 postes autour de la nomenclature des activités économiques en 32 postes (P(i,j) = nombre d'individus exerçant un emploi i dans un établissement dont le secteur d'activité économique est j) donne une matrice des migrations où les postes d'emploi de "dessinateurs d'études" et de "dessinateurs d'exécution" sont échangés par les mêmes classes, ce qui suggère de les regrouper en un seul poste.

données de base. C'est ce que nous pouvons appeler l'étude de la "régressibilité externe" (cf. *NOTA, in fine*).

3°) - D'avoir déjà décidé de l'ensemble des points - réels ou fictifs - parmi lesquels, lors de la pratique de la régression, nous recherchons les voisins de l'individu externe. Cet ensemble peut être celui des individus de base ; il peut également être celui (I) des centres de gravité des classes des individus de base définies chacune par une modalité de la variable à expliquer : c'est cette dernière option que nous considérons seule dans la suite. D'autres variantes sont possibles : il convient de les aborder à propos d'applications concrètes.



La représentation spatiale (analyse factorielle) des individus nouveaux doit se faire sur des axes ajustés à ceux des individus de base qui tombent dans le domaine d'application de la régression. M. Roux quant à lui n'a pas hésité à utiliser dès cette étape le programme (POUBEL) pour ne retenir des individus de base que ceux apparus au moins une fois parmi les plus proches voisins de l'un ou de l'autre des individus nouveaux. Ceci fait, le problème de la régressibilité interne est lui-même mieux circonscrit : il est limité à la zone où la régression sera ultérieurement appliquée. Et l'on répond aussi dans la mesure du possible, à la critique formulée ci-dessus que la régressibilité peut n'être pas uniforme sur tout le champ d'une étude (voir aussi ; A. Abdel Shahid : [*REGR. CLIMATS*] ; C.A.D. Vol VII n° 1 ; pp 93-111 ; 1982).

2.2 <u>L'algorithme de régression d'après un nombre variable de voisins</u> : Soit I l'ensemble des modalités de la variable à expliquer (I = ensemble des lignes ; ensemble où l'on recherchera les voisins : cf. §2.1) et J l'ensemble des variables explicatives (J = ensemble de colonnes). Nous supposons connaître pour un point p les valeurs des variables du groupe J, ce qui permet d'associer à p une ligne supplémentaire du tableau k_{IJ} et de calculer ses coordonnées factorielles $G_x(p)$ dans l'espace issu d'une analyse des correspondances.

NOTA : Comme le montre un travail de M. Roux (cf. C.A.D. Vol IV, n°1, pp 61-81, 1979), il peut être également utile d'éliminer de l'échantillon de base les individus trop éloignés du domaine auquel on doit appliquer la régression. C'est ce qu'explique le schéma ci-dessus).

Action de l'algorithme.

Nous la définissons en fixant deux notations et en énonçant deux conditions c_1 et c_2.

- *Définition de I(n,p)* *

On désignera par I(n,p) l'ensemble des n points de I les plus proches de p et par d(n,p) la distance de p au centre de gravité de I(n,p).

- *Condition $c_1(n,p)$* *

Pour tous entiers n' et n" tels que : $1 < n' < n'' < n$, nous avons : $d(n'',p) < d(n',p)$; autrement dit : quand le nombre de voisins varie de 1 à n, le centre de gravité des sujets les plus proches se rapproche de p. Ce qui s'écrit :

$$\forall\, n',n'' : 1 < n' < n'' < n \Rightarrow d(n'',p) \leq d(n',p)$$

- *Condition $c_2(n,p)$* *

En adjoignant un voisin supplémentaire, la condition $c_1(n,p)$ n'est plus satisfaite : le centre de gravité s'éloigne :

$$c_2(n,p) : d(n,p) \underset{s}{\leq} d(n+1,p)$$

- *Définition de n(p)*

C'est la suite des n points les plus proches de p tels que :

$$c_1(n,p) \wedge c_2(n,p)$$

dont nous proposons de prendre les valeurs pour estimer la donnée non connue pour p.

Posons : NV = n(p)
DNV = d(p,n)

La procédure de calcul s'écrit alors :

```
    PROC : CALCUL VOISINS ;

       NV:=1;
      DIS:=DNV(p,1);

DEB ; DES:= DNV(p,NV+1);

    Si DES>DIS ou si NV>NMAX aller à FIN;
    sinon DIS:=DES, NV:=NV+1,aller à DEB;

FIN;FIN PROC;
```

---

* *Dans le texte nous disons plus brièvement PE ou point d'estimation pour p et SEBE ou sous-ensemble de base à l'estimation pour I(n,p).*

2.3 Cas d'une variable continue. Quand les modalités de la variable à expliquer sont des tranches de variation d'une variable continue : l'estimation se fait comme suit :

Soit μ_j la valeur centrale de la tranche j, i.e. la valeur à laquelle nous assimilons toutes celles des individus de la tranche de variation j. Soit n le nombre de voisins de p et J(n,p) son sous ensemble voisin :

(Z_{21}) $\mu_p = \Sigma\{\mu_j/n \mid j \in J(n,p)\}$ est une première estimation.

Soit $d^2(p,j)$ le carré de la distance de p à l'un de ses voisins $j \in J(n,p)$, la deuxième estimation sera :

(Z_{22}) $\mu'_p = 1/s \ \Sigma\{c_j \mu_j \mid j \in J(n,p)\}$

où $c_j = 1/d^2(p,j)$; et $s = \Sigma\{c_j \mid j \in J(n,p)\}$.

L'intérêt de l'estimation μ'_p par rapport à μ_p est que dans (Z_{22}) nous prenons en considération l'inégalité des proximités de p à ses voisins : ce qui est logique puisque p, tout en étant plus proche du c. de g. de J(n,p) que de tout autre sous-ensemble de J(n,p) n'est pas situé au centre de gravité de J(n,p) d'où la nécessité d'une correction. (Une autre estimation de c_j est proposée au § 2.4.4).

L'expérimentation nous permet de conclure que cette 2-ème estimation μ' est meilleure que la première (μ).

2.4 *Objections et suggestions :*

2.4.1 "Un piège" théoriquement possible nous a été tendu par J. P. Benzécri : une suite de voisins s'éloignant indéfiniment de p tandis que leur centre de gravité s'en rapproche pas à pas de ε ! Tel est le cas en effet si nous supposons p à l'origine de la droite et prenons comme suit les voisins, leur somme de coordonnées et la distance du centre de gravité à p :

Rang	Voisins	Somme des coordonnées	Centre de gravité = distance à p.
1	$1-\varepsilon$	$1-\varepsilon$	$(1-\varepsilon)$
2	$-3+5\varepsilon$	$-2(1-2\varepsilon)=-2+4\varepsilon$	$-(1-2\varepsilon)$
3	$5-13\varepsilon$	$3(1-3\varepsilon)=3-9\varepsilon$	$(1-3\varepsilon)$
4	$7+25\varepsilon$	$-4(1-4\varepsilon)=-4+16\varepsilon$	$-(1+4\varepsilon)$
5	$+9-49\varepsilon$	$5(1-5\varepsilon)=5+25\varepsilon$	$(1+5\varepsilon)$
6	$-11+61\varepsilon$	$-6(1-6\varepsilon)=-6-36\varepsilon$	$-(1-6\varepsilon)$
n	$(2n-1)(-1)^{n+1}+(-1)^n((n-1)^2+n^2)\varepsilon$	$(-1)^{n+1}n(1-n\varepsilon)$	$(-1)^{n+1}(1-n\varepsilon(-1)^{n+1})$

pour se prémunir contre une telle situation on pourra par exemple, fixer un seuil de distance au-delà duquel un voisin ne sera pas accepté, même si son introduction améliore la position du centre de gravité.

2.4.2 Cet exemple nous amène à formuler une 2ème critique à l'algorithme : il n'est pas impossible qu'on trouve un sous ensemble

$I(m+n, p)$ de I avec $m > 1$ de façon à ce que $d(m+n,p) <_s d(n,p)$; mais ce sous ensemble ne donnera pas satisfaction à la condition $c_1(n,p)$ et ses points risquent d'être trop dispersés *par rapport à p*.

2.4.3 Une autre critique à la fois technique et méthodologique nous a également été formulée par J.P. Benzécri qui propose de procéder comme suit :

a) - Rechercher les voisins de p parmi les individus de base (l'ensemble S) et à l'intérieur d'une classe S(k;t) déterminée ainsi : si KM est le nombre de modalités de la variable à expliquer et si g(k;1) est le profil du centre de gravité des individus de S dont la modalité pour la variable à expliquer est k (dont la classe au temps 1 est k) alors :

S(k;1) est l'ensemble des individus de S qui appartiennent à la classe k (ou qui ont la modalité k pour variable à expliquer).

g(k,1) = centre de gravité de S(k,1).

. . . .

S(k,t) est l'ensemble des individus de S qui sont plus proches de g(k,t-1) que de tout autre centre g(k',t-1) (pour k' différent de k, et compris entre 1 et KM).

b) - Ne rechercher ces voisins que dans la classe S(k,t) dont le centre de gravité g(k,t) est plus proche de p que tout autre g(k',t) avec $k' \neq k$.

Quel temps t prendre ? Nous discuterons ultérieurement de ce choix, mais rappelons déjà :

1°) qu'au temps t + 1 la partition fournie possède un moment d'ordre 2 *total* inférieur à celui de la partition du temps t (cf. E. Diday, *op. cit.*).

2°) que S(k,t +1) *n'a pas forcément* un moment d'ordre 2 inférieur à celui de S(k,t).

3°) que l'optimisation de la partition de S (au sens de 1°)) se fait au détriment de la variable "à expliquer" (transferts ou migrations - en particulier au temps 2 - vers une classe constituée d'individus ayant une modalité différente de la variable à expliquer).

4°) qu'une solution discutable serait de prendre comme "cible de régression" :

- soit les KM centres de gravité de $S(k,1) \cap S(k,2)$ pour k = 1 à KM.

- soit les individus constituant ces centres i.e. le sous ensemble de S constitué par les individus qui, tout en ayant la modalité k pour la variable à expliquer ($\forall$ k = 1 à KM) sont plus proches du centre de gravité g(k,1) que de tout autre centre g(k',1) $\forall k' \neq k$ (individus "non rebelles", pour lesquels IREBEL(I) = 0 dans nos programmes BIB déjà cités).

Dans ce cas, il s'impose pratiquement d'écrire les coordonnées des points de S sur un fichier séquentiel indexé qui nous évite la lecture de la totalité du fichier pour chaque calcul de voisins. Ceci représente une économie très appréciable tant en calculs qu'en "entrées-sorties".

Notons que cette technique s'impose lorsque le fichier de base atteint plusieurs dizaines de milliers d'individus.

N. B. Un fichier partitionné résoudrait parfaitement notre problème mais tant en FORTRAN qu'en PL/I, ces fichiers sont réservés aux bibliothèques.

Notons toutefois que le FORTRAN UNIVAC offre la possibilité d'utiliser plusieurs fichiers de travail dans le programme sans en donner la définition physique au niveau des cartes de contrôle. Avec ce FORTRAN notre problème se résoudrait en créant autant de fichiers de travail que de classes (de modalités de la variable à expliquer par exemple). On peut objecter à cette remarque que FORTRAN IV d'I.B.M. offre aussi la possibilité d'utiliser autant de disques qu'on le voudrait mais il nous paraît inconcevable d'imposer à l'utilisateur la perforation de 80 cartes de définition de fichier pour 40 modalités de la variable à expliquer.

2.4.4 Projection et régression barycentrique : On a vu au § 2.3 qu'il était naturel de ne pas affecter tous les voisins j retenus d'une même masse. J.P. Benzécri nous a suggéré de déterminer les coefficients positifs c_j de telle sorte que le barycentre des points j affectés des masses c_j soit aussi proche que possible de p. C'est là un problème classique en programmation linéaire : chercher le point m d'un simplexe (ici le simplexe ayant pour sommets les points j) le plus proche d'un point donné : m est appelé projection de p sur le simplexe.

Nous pensons que cette méthode pourrait améliorer la précision ; mais elle allongerait le temps de calcul.