REVUE DE STATISTIQUE APPLIQUÉE

L. TIRET
M. O. LEBEAUX
P. CAZES

Méthode de classification basée sur la règle d'affectation majoritaire. Application à un problème de psychiatrie infantile

Revue de statistique appliquée, tome 28, n° 3 (1980), p. 69-78
<http://www.numdam.org/item?id=RSA_1980__28_3_69_0>

MÉTHODE DE CLASSIFICATION BASÉE SUR LA RÈGLE D'AFFECTATION MAJORITAIRE APPLICATION A UN PROBLÈME DE PSYCHIATRIE INFANTILE

L. TIRET (*), M .O. LEBEAUX (**) et P. CAZES (***)

1. INTRODUCTION

L'objectif de ce travail, portant sur les troubles mentaux de l'enfant, est de déterminer s'il existe une relation plus ou moins forte entre le diagnostic psychiatrique et un ensemble de caractéristiques individuelles recueillies au cours d'un interrogatoire médical et psychologique.

La méthode utilisée consiste à former, à partir d'un échantillon de sujets répartis dans différents diagnostics de trouble mental, de nouvelles classes diagnostiques homogènes quant aux variables descriptives étudiées, c'est à dire regroupant les sujets ayant un profil semblable.

Chaque individu se voit attribuer par cette méthode un nouveau diagnostic, qui devrait être le sien compte tenu de son profil. Le pourcentage de concordance calculé sur l'ensemble des individus, entre le diagnostic théorique et le diagnostic réel porté par le clinicien, indique s'il existe une relation entre l'ensemble des variables composant le profil des sujets et le diagnostic psychiatrique.

2. EXPOSE DE LA METHODE

2.1. Rappels théoriques

La méthode utilisée ici est basée sur la "Valley-Seeking Technique" suggérée par KOONTZ et FUKUNAGA [4] et exposée par BENZECRI dans un problème de Physique Corpusculaire [1].

(*) Statisticien INSERM, Unité 164, 44 Chemin de Ronde – 78110 – Le Vésinet

(**) Ingénieur C.N.R.S., Laboratoire de Statistique (Université de Paris 6)

(***) Maitre-Assistant, Laboratoire de Statistique (Université de Paris 6)

Son but est d'obtenir, sur un ensemble I d'individus décrits par un certain nombre de variables quantitatives ou qualitatives, une partition en classes homogènes.

Voici en quoi consiste l'algorithme :

. à l'étape 0, les individus sont répartis aléatoirement dans des classes dont le nombre est fixé a priori.

. à l'étape 1, on réaffecte chaque individu à une nouvelle classe par la règle d'affectation majoritaire, dont le principe est le suivant :

– chaque individu, décrit par p variables, peut être considéré comme un point de l'espace R^p ; supposant R^p muni d'une métrique, on peut définir une distance entre les individus et calculer pour chacun d'entre eux ses K plus proches voisins dans R^p ; pour attribuer à chaque individu i une nouvelle classe, on considère ses K plus proches voisins et on réaffecte i à la classe qui compte le plus d'individus parmi ces voisins. En cas d'égalité entre s modalités, on affecte i à chacune de ces modalités avec une probabilité 1/s.

. à l'étape 2, on réitère le processus sur l'ensemble des individus affectés aux classes déterminées à l'issue de l'étape 1.

. à l'étape n, de manière générale, on pratique la règle d'affectation majoritaire sur les individus affectés aux classes déterminées à l'issue de l'étape n – 1.

On arrête le processus lorsque plus aucun individu ne change de classe. La convergence n'étant pas certaine, il importe de fixer le nombre maximum d'itérations au-delà duquel on arrête le processus quelque soit le résultat.

Si le processus converge (c'est-à-dire si le nombre d'individus changeant de classe décroît à chaque étape pour atteindre 0), on obtient finalement une partition stable de l'ensemble I, en un nombre de classes inférieur ou égal au nombre initial. En effet, certaines classes peuvent se vider totalement au cours d'une des étapes.

2.2. Application au cas de la classification diagnostique

Le problème étudié ici est le suivant : partant d'un ensemble de sujets répartis au départ dans différents diagnostics, on cherche à obtenir une nouvelle partition de cet ensemble en classes diagnostiques homogènes par rapport à un certain nombre de critères descriptifs.

Ce problème se rapproche donc de celui qui vient d'être exposé, avec la différence que les individus ne sont pas classés au départ aléatoirement, mais affectés a priori à un diagnostic, celui du clinicien. Il n'était pas possible d'affecter aléatoirement les individus à des classes quelconques, car il fallait que leur attribution finale ait une signification diagnostique afin de pouvoir la comparer au diagnostic clinique.

Introduire une classification a priori reviendrait à sauter l'étape 0 de l'algorithme si les classes avaient des effectifs égaux, comme c'est le cas dans la méthode de K. et F. où elles sont attribuées aléatoirement avec la même probabilité.

Mais lorsque les classes ont des effectifs très dissemblables, elles n'ont pas la même probabilité d'être représentées dans un voisinage donné ; les classes faibles ont donc moins de chance d'être attribuées que les classes fortes, puisque l'on se base sur des effectifs absolus et non des effectifs relatifs dans la règle d'affectation majoritaire.

Certains diagnostics peuvent ainsi disparaître au cours de l'analyse car ils sont peu fréquents dans la population, bien qu'ils aient une signification propre sur le plan clinique.

Pour pallier cet inconvénient, on peut pondérer les individus par l'inverse de l'effectif de la classe à laquelle ils appartiennent. Dans ce cas, l'affectation majoritaire pour chaque individu se fait en multipliant le nombre de ses voisins dans chaque classe par le poids correspondant à cette classe, puis en retenant la classe pour laquelle ce produit est maximum ; en cas d'égalité entre S classes, on retient chacune d'elles avec une probabilité de 1/S.

2.3. Choix de la distance

Plutôt que l'espace R^P des variables descriptives initiales, on préfère souvent, pour la recherche des voisins, se placer dans l'espace des premiers facteurs obtenus après analyse factorielle de ces variables descriptives. Les coordonnées sur ces axes sont en effet préférables aux variables primaires chargées de bruit.

L'analyse factorielle se fait sur le tableau croisant les individus avec les variables descriptives, ou – ce qui est équivalent dans le cas où les variables sont codées de manière disjonctive – sur le tableau de BURT croisant les variables avec elles-mêmes, les individus étant placés en éléments supplémentaires ; on sait en effet que dans ce cas les facteurs obtenus sur les individus sont identiques [2].

On prendra alors comme distance entre deux individus l'approximation sur les q premiers facteurs de la distance du χ^2 :

$$d^2(i, i') = \Sigma \{ [F_\alpha(i) - F_\alpha(i')]^2 / \alpha = 1, q \}$$

formule où F_α (i) (resp. F_α (i')) désigne la valeur du $\alpha^{\text{ième}}$ facteur sur l'individu i (resp. i').

On peut également utiliser la métrique d'inertie (ou métrique de MAHALANOBIS) :

$$d^2(i, i') = \Sigma \{ (1/\lambda_\alpha)[F_\alpha(i) - F_\alpha(i')]^2 / \alpha = 1, q \}$$

λ_α désignant la variance de F_α (i.e. la racine carré de la valeur propre associée au $\alpha^{\text{ième}}$ facteur du tableau de BURT)

2.4. Choix du nombre de voisins

Il n'existe pas de règle théorique pour déterminer le nombre de voisins à prendre en compte. Celui-ci est déterminé de manière empirique : il doit être suffisant pour s'affranchir des fluctuations d'échantillonnage (en général supérieur à 10) et d'autant plus grand qu'il y a davantage de classes et que certaines d'entre-elles ont des effectifs plus faibles (en effet, un voisinage trop petit ne permettrait pas à ces classes d'être représentées).

Mais il ne faut pas non plus élargir le voisinage plus qu'il n'est indispensable pour décider de l'affectation d'un individu (en général, on ne dépasse pas 10 % de la taille de l'échantillon). Des contraintes de coût peuvent également intervenir, la recherche des voisins nécessitant un temps d'ordinateur important.

Dans la pratique, on peut choisir un nombre de voisins maximum tenant compte de toutes les contraintes précédentes, puis faire plusieurs essais successifs en réduisant ce nombre jusqu'à atteindre 10 voisins. On garde finalement le nombre de voisins minimum permettant la convergence.

3. RESULTATS

3.1. Matériel d'étude

Les données utilisées dans ce travail ont été recueillies dans les consultations du Secteur d'Hygiène Mentale Infantile du 14ème arrondissement de Paris, entre Janvier 1970 et Juillet 1974. La population étudiée comprend 1159 enfants et adolescents venus consulter pour un trouble d'ordre mental.

Le diagnostic psychiatrique était codé initialement selon 12 rubriques. Après suppression des rubriques ayant des effectifs insuffisants pour l'analyse, 8 diagnostics ont finalement été gardés, concernant 1120 enfants. Ils se répartissent de la manière suivante dans l'échantillon (Tableau 1).

TABLEAU 1

Répartition des diagnostics

	Effectif	Pourcentage
Variations de la normale	177	16 %
Troubles réactionnels	174	15 %
Troubles spécifiques du développement	179	16 %
Troubles névrotiques	282	25 %
Troubles de la personnalité et du caractère	83	7 %
Troubles psychotiques	53	5 %
Troubles de l'évolution libidinale	130	12 %
Retard mental	42	4 %
Total	1120	100 %

Les variables descriptives retenues sont :

- la nature et le nombre de facteurs étiologiques de la maladie
- la nature et le nombre de troubles notables relevés dans les antécédents de l'enfant
- la nature et le nombre de troubles associés au diagnostic principal
- le degré de niveau intellectuel, découpé en cinq classes
- la présence d'antécédents psychiatriques familiaux
- l'âge des sujets, découpé en cinq classes.

Toutes ces variables ont été mises sous forme disjonctive ; on peut voir leur répartition dans l'échantillon, en annexe.

3.2. Résultats

L'espace dans lequel s'est effectué la recherche des voisins est celui des 6 premiers axes après analyse factorielle du tableau de Burt (variables x variables).

La distance utilisée est la distance du χ^2 calculée sur les 6 premiers facteurs, qui expliquent 45 % de l'inertie totale du nuage.

Le nombre de voisins initialement choisi était de 30, en raison du nombre élevé de diagnostics. Une première analyse a été faite en donnant le même poids à tous les individus. Mais les diagnostics ayant des fréquences très inégales (tableau 1), on a recommencé l'analyse en pondérant les individus par l'inverse de l'effectif de leur classe, comme cela a été expliqué précédemment. Dans ce cas, deux solutions sont possibles :

– soit les individus gardent le même poids tout au long de l'analyse
– soit on recalcule le système de pondération à chaque étape en fonction des nouveaux effectifs des classes.

Les résultats des 3 méthodes ont été comparés.

3.2.1. Comparaison des résultats obtenus selon les 3 systèmes de pondération (30 voisins)

Le critère d'arrêt du processus était un peu moins strict que celui de la méthode initiale : en effet, il suffisait qu'il n'y ait pas plus de 11 individus (1 %) qui changent de classe pour que l'on arrête. Le nombre maximum d'itérations était fixé à 15.

On constate que, quelque soit le système de pondération appliqué la convergence est atteinte au bout d'une dizaine d'itérations. (Tableau 2).

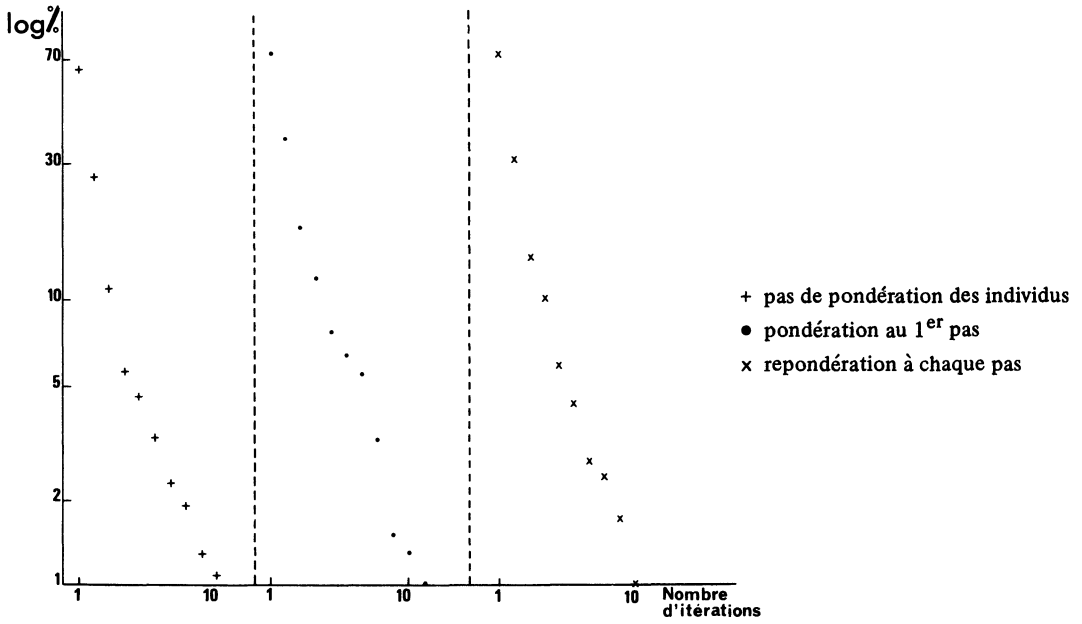
TABLEAU 2

Comparaison des résultats obtenus selon les trois systèmes de pondération (30 voisins)

	Pas de pondération des individus	Pondération en début d'analyse	Repondération à chaque pas
Nombre d'itérations effectuées (maximum = 15)	11	11	10
Nombre d'individus changeant de classe après la dernière itération	10	7	10
Nombre et % d'individus bien classés en fin d'analyse	418 *37 %*	210 *19 %*	316 *28 %*

Le pourcentage d'individus changeant de classe à chaque étape décroît de manière quasi-exponentielle dans les 3 cas (graphique).

Par contre, le pourcentage d'individus bien classés – c'est-à-dire dont le diagnostic final correspond au diagnostic initial du clinicien – varie selon les 3 mé-



Pourcentage d'individus changeant de classe en fonction du nombre d'itérations (échelle semi-logarithmique).

thodes, le meilleur résultat étant obtenu lorsqu'on ne pondère pas les individus (Tableau 2).

On serait donc tenté de préférer cette méthode, mais en examinant la répartition finale des diagnostics (Tableau 3), on observe qu'avec les 2 premiers systèmes de pondération, certains diagnostics disparaissent au cours de l'analyse : dans un cas, il s'agit des diagnostics peu fréquents, donc défavorisés par l'absence

TABLEAU 3

Comparaison de la répartition finale par diagnostic selon les trois systèmes de pondération (30 voisins)

	Répartition initiale	Répartition finale		
		Pas de pondération des individus	Pondération en début d'analyse	Repondération à chaque pas
Variations de la normale	16	24	16	18
Tr. Réactionnels	15	19	4	14
Tr. du Développement	16	13	3	10
Névroses	25	42	0	9
Tr. Personnalité	7	0	2	11
Psychoses	5	0	33	11
Tr. Evol. Libidinale	12	2	2	18
Retard mental	4	0	40	9
Total	100 %	100 %	100 %	100 %

de pondération, et dans l'autre, au contraire, il s'agit du diagnostic le plus fréquent, celui de névrose, pénalisé par un poids constant car son effectif diminue à chaque étape.

C'est finalement en repondérant les individus à chaque étape que l'on obtient les résultats les plus satisfaisants, bien que ce système de pondération ait l'inconvénient de fabriquer des classes d'effectifs à peu près égaux, ce qui n'est pas non plus très réaliste.

3.2.2. Comparaison des résultats obtenus en modifiant le nombre des voisins

En gardant donc le dernier système de pondération, plusieurs essais ont été faits en modifiant le nombre de voisins.

On constate qu'un nombre de voisins trop faible ne permet pas d'obtenir la convergence du processus, sans doute à cause du nombre élevé de diagnostics (Tableau 4).

C'est donc la partition obtenue en prenant en compte 30 voisins, qui est finalement gardée.

Rappelons que dans ce cas, le pourcentage total d'individus bien classés était de 28 %.

TABLEAU 4

Comparaison des résultats obtenus selon le nombre de voisins pris en compte (repondération à chaque pas).

	30 Voisins	20 Voisins	10 Voisins
Nombre d'itérations effectuées (maximum = 15)	10	15 Pas de convergence	15 Pas de convergence
Nombre d'individus changeant de classe après la dernière itération	10	16	19
Nombre et % d'individus bien classés en fin d'analyse	316 *28 %*	320 *29 %*	359 *32 %*

TABLEAU 5

Répartition du diagnostic final dans les différentes classes de diagnostic initial (30 voisins ; repondération à chaque pas).

Diag. final / Diag. initial	Variat. Normale	Troubles Réaction	Troubles Dévelop.	Névroses	Troubles personnalité	Psychoses	Troubles Libido	Retard Mental	Total
Var. Normale	56	7	8	4	2	7	7	9	100 %
Tr. Réaction.	10	36	10	8	9	6	14	7	100 %
Tr. Dévelop.	30	10	16	4	11	7	14	8	100 %
Névroses	7	15	10	12	11	10	23	12	100 %
Tr. Personnal.	2	11	5	16	29	13	13	11	100 %
Psychoses	7	7	4	19	25	21	15	2	100 %
Tr. Libido	6	12	10	9	5	16	39	3	100 %
Retard Mental	0	2	0	2	19	41	19	17	100 %

Ce pourcentage varie beaucoup en fonction du diagnostic initial (Tableau 5, chiffres lus sur la diagonale).

Seul le diagnostic de "Variations de la normale" a plus d'une moitié de sujets bien classés.

Lorsque les individus sont mal classés, on observe toutefois une attirance vers certains diagnostics dont les proximités avaient été mises en évidence au cours d'analyses précédentes [7] :

- troubles du développement et variations de la normale
- névroses et troubles de l'évolution libidinale
- psychoses et troubles de la personnalité
- retard mental et psychoses

4. CONCLUSION

Les résultats obtenus sont assez décevants, puisque moins d'un tiers des individus se voient attribuer par l'analyse un diagnostic concordant avec celui du clinicien.

Ces résultats témoingnent néanmoins de la relation existant entre le diagnostic et les variables étudiées. En effet, s'il n'existait aucune relation, un individu aurait une chance sur huit de se voir attribuer son diagnostic initial par l'analyse, ce qui conduirait à un pourcentage global de bien classés de 12,5%, soit 2,2 fois moins que le pourcentage obtenu.

Deux raisons peuvent être avancées pour expliquer ces résultats médiocres :

. les variables retenues pour l'analyse ne constituaient pas la totalité du tableau clinique, mais simplement les éléments d'un premier interrogatoire. Il serait intéressant de voir si l'on obtiendrait de meilleurs résultats en adjoignant à ces variables les données de l'examen psychiatrique approfondi.

. mais une autre raison réside, semble-t-il, dans le concept même de diagnostic utilisé ici. En effet, les controverses existant actuellement sur la classification diagnostique en Psychiatrie Infantile laissent présumer que le diagnostic n'est pas une donnée objective, sur laquelle on peut se baser de manière fiable, comme c'est davantage le cas en médecine somatique.

Aux différences d'orientations théoriques viennent s'ajouter l'extrême difficulté de porter un diagnostic définitif chez un enfant en évolution continue.

REMERCIEMENTS

Nous adressons nos plus vifs remerciements à Monsieur LELLOUCH, directeur de l'unité de recherche "Méthodes Statistiques et Epidémiologiques et application à l'étude des maladies chroniques" (I.N.S.E.R.M. unité 169), pour les conseils qu'il a bien voulu nous donner et le temps qu'il nous a consacré pour ce travail.

ANNEXE

Répartition des variables descriptives dans l'échantillon

	Nombre	Pourcentage
Nombre de facteurs étiologiques :		
Aucun facteur étiologique	513	*46*
Un facteur étiologique	458	*41*
Deux facteurs étiologiques	105	*9*
Trois facteurs étiologiques	26	*2*
Quatre facteurs étiologiques ou plus	18	*2*
	1120	*100*
Nature des facteurs étiologiques		
Etiologie organique démontrable	30	*2*
Epilepsie	15	*1*
Prématurité ou poids < 2,5 Kg	60	*5*
Affection somatique	33	*3*
Carence affective	35	*3*
Relation affective pathogène	302	*27*
Traumatismes affectifs	93	*8*
Placements hors famille	98	*9*
Carence éducative	58	*5*
Facteurs d'environnement	94	*8*
Antécédents psychiatriques familiaux		
Antécédents psychiatriques constatés chez un ou plusieurs membres de la famille	228	*20*
Aucun antécédent psychiatrique familial précisé	892	*80*
Total	1120	*100*
Nombre de troubles antécédents		
Aucun trouble antécédent	426	*38*
Un trouble antécédent	417	*37*
Deux troubles antécédents	196	*18*
Trois troubles antécédents	70	*6*
Quatre troubles antécédents ou plus	11	*1*
Total	1120	*100*
Nature des troubles antécédents		
Troubles néonataux	47	*4*
Troubles du sommeil	199	*18*
Troubles de l'alimentation	183	*16*
Retards du développement	274	*24*
Spasmes du sanglot	13	*1*
Troubles psychosomatiques	42	*4*
Enurésie	145	*13*
Encoprésie	24	*2*
Episodes aigus de troubles du comportement	136	*12*
Niveau intellectuel		
Supérieur (QI supérieur à 110)	102	*9*
Moyen (entre 90 et 110)	306	*27*
Limite (entre 70 et 90)	142	*13*
Débilité mentale (inférieur à 70)	31	*3*
Indéterminable ou non fait	539	*48*
Total	1120	*100*

Nombre de troubles associés			
Aucun trouble associé		573	*51*
Un trouble associé		471	*42*
Deux troubles associés		69	*6*
Trois troubles associés ou plus		7	*1*
	Total	1120	*100*
Nature des troubles associés			
Troubles de la parole et du langage		283	*25*
Troubles psychosomatiques		36	*3*
Troubles psychomoteurs		109	*10*
Troubles réactionnels		143	*13*
Retard mental		59	*5*
Age			
0-3 ans		79	*7*
4-6 ans		552	*49*
7-10 ans		327	*29*
11-14 ans		110	*10*
15-19 ans		52	*5*
	Total	1100	*100*

BIBLIOGRAPHIE

[1] BENZECRI J.P. – Analyse des données en Physique Corpusculaire. III. Les méthodes multidimensionnelles. *Les cahiers de l'Analyse des Données.* Vol. III – 1978 – N° 1 – PP 79-94.

[2] BENZECRI J.P. – Sur l'analyse des tableaux binaires associés à une correspondance multiple. *Les Cahiers de l'Analyse des Données.* Vol. II – 1977 – N° 1 – pp. 55-71

[3] CAZES P. – Méthodes de régression. III. L'Analyse des données. *Les cahiers de l'Analyse des Données.* Vol. III-1978-N° 4 pp. 385-391.

[4] KOONTZ W., FUKUNAGA K. – A non parametric Valley-seeking Technique for Cluster Analysis. *IEEE Transactions on Computers.* Vol. C-21, 1972, N°2 – pp. 171-178.

[5] LEBEAUX M.O. – Notice sur l'utilisation du programme POUBEL. *Les Cahiers de l'Analyse des Données.*. Vol. II-1977-N°4 pp. 467-481.

[6] LEBOVICI S., SADOUN R. – L'enregistrement du diagnostic au centre de santé mentale A. BINET (Paris 13ème). *La Psychiatrie de l'enfant.* Vol. XI-Fascicule 2- 1968. pp. 533-550.

[7] TIRET L. – *Apport de l'Analyse de Données à l'orientation diagnostique en Psychiatrie Infantile.* Thèse de 3ème cycle-Paris VI. 1978.