LES CAHIERS DE L'ANALYSE DES DONNÉES

G. D. MAÏTI
R. FROLOFF

Engagements et partants dans les courses de trot sur les hippodromes français en 1992

Les cahiers de l'analyse des données, tome 18, n° 3 (1993), p. 261-280

<http://www.numdam.org/item?id=CAD_1993__18_3_261_0>

Les Cahiers de l'Analyse des Données
Vol XVIII - 1993 - n° 3 - pp.261-280.

ENGAGEMENTS ET PARTANTS DANS LES COURSES DE TROT SUR LES HIPPODROMES FRANÇAIS EN 1992

[COURSES]

*G. D. MAÏTI**
*R. FROLOFF***

1 L'organisation des courses: engagements et partants

Sur les hippodromes de France, des courses se déroulent tout au long de l'année; avec, pour les principaux hippodromes, plusieurs courses par jour plus de 200 jours par an. Dans la catégorie des trotteurs, les chevaux se comptent par milliers: leur engagement dans les courses ne peut donc être laissé sans règle. Au contraire, un système complexe vise à assurer aux chevaux susceptibles de réaliser de bonnes performances, des occasions de se faire valoir et de rapporter des récompenses à leurs propriétaires.

Ce système est, dans l'ensemble, satisfaisant; mais les experts eux-mêmes qui, suivant une tradition séculaire, en appliquent le programme, ont, depuis plusieurs décennies, désiré s'assurer la collaboration d'une base d'enregistrement informatique de haute performance, dont R. FROLOFF assure la gestion. Et, maintenant, le moment semble venu de confronter l'intuition des experts avec une vue statistique globale des données accumulées.

Il est clair qu'une telle synthèse statistique ne pourra se faire en une seule étape; le but ultime étant d'élaborer un modèle de comportement des entraîneurs susceptible de prévoir l'effet qu'aura un programme de courses national complet sur l'enchaînement des engagements et des départs de tous les chevaux trotteurs. Mais le présent rapport a pour objet de rendre compte d'une première étape qui fait l'objet d'un contrat entre la Société S.E.C.F. et la Société STATMATIC - EUROPE.

Il s'agira, plus précisément, ici, d'un fichier de 453898 engagements dans des courses organisées en 1992. Chaque enregistrement du fichier comporte 18

(*) Société STATMATIC - EUROPE.
(**) Société d'Encouragement à l'Élevage du Cheval Français.

informations, relatives à la course et au cheval. Nous ne présenterons pas immédiatement l'ensemble de ces informations, mais concentrerons d'abord notre attention sur la relation entre la somme d'appel de la course et les gains antérieurs d'un cheval engagé et, éventuellement, partant et gagnant, ou seulement placés parmi les cinq premiers, entre lesquels est communément partagé le prix (suivant la formule usuelle: 50%, 25%, 15%, 6%, 4%).

De ce point de vue, un enregistrement d'engagement se réduit aux 3 informations suivantes:

apl : Somme d'appel de la course (ou plafond des gains antérieurs, que pour être admis à s'engager, un cheval ne doit pas dépasser; sauf dérogations particulières que nous ne prendrons pas en compte ici);

gan : Gains antérieurs du cheval (avec la relation imposée: gan ≤ apl);

plc : Place, variable numérique où l'on peut indiquer, à la fois, si le cheval engagé est, ou non, parti; et, pour un départ effectif, s'il a été classé et dans quel rang;

à quoi on adjoindra le calcul du quotient gan/apl = *g/a ; quotient, en principe, ≤1 ; et dont les experts savent qu'il indique les chances de succès du cheval, et pourrait donc servir à autoriser les refus d'engagements dans des courses encombrées.

Pour découvrir la structure des données, le statisticien crée des graphiques et des tableaux. Ultérieurement, afin de rendre compte des disparités régionales que le modèle définitif devra intégrer, il effectue des classifications et construit des cartes géographiques. Ce premier travail présente, en le commentant, un recueil de graphiques et tableaux; avec un programme d'interpolation qui en facilite la lecture. Des cartes, nous nous bornerons, ici, à donner un aperçu.

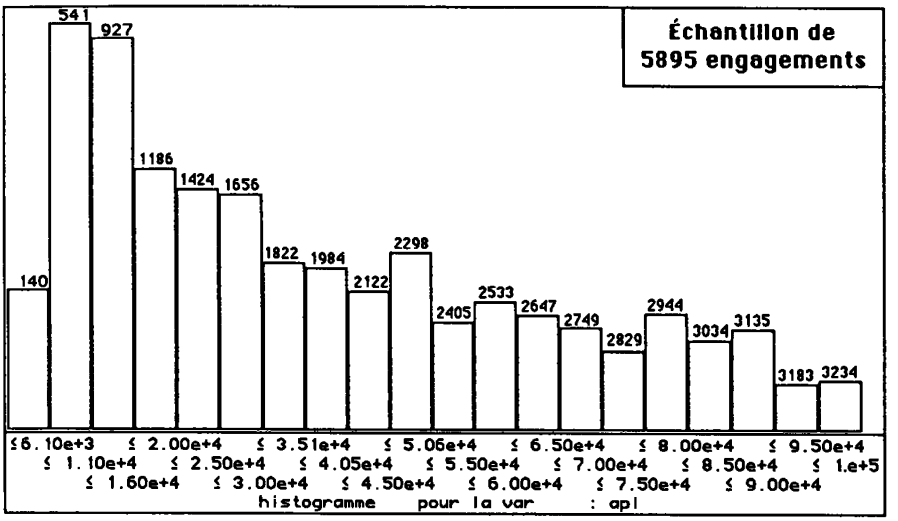
2 Étude des engagements

2.1 Histogrammes des sommes d'appel d'après un échantillon d'engagements

L'admissibilité à une course est déterminée premièrement par la somme d'appel, et deuxièmement par des conditions d'âge et de sexe du cheval; les deux n'étant pas indépendants. Il importe de voir, sur l'ensemble des engagements, le poids relatif des différentes bandes de somme d'appel.

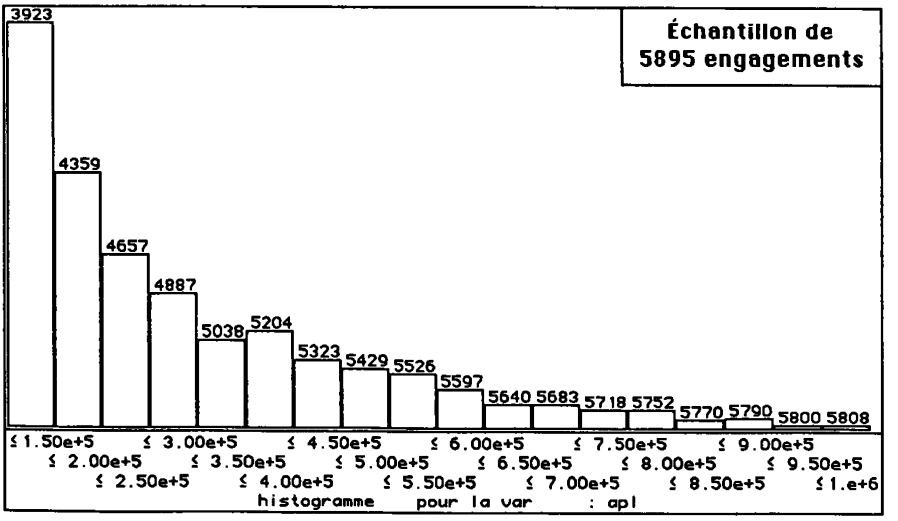
À cet effet, on a extrait du fichier de 453898 engagements plusieurs sous-fichiers représentatifs dont le taux d'échantillonnage varie de 1/100 à 1/70. Il suffira ici de considérer un échantillon de 5895 engagements; lequel s'étend à tous les âges, tous les lieux, tous les niveaux.

On voit sur l'histogramme supérieur que plus de la moitié des engagements concernent des courses dont la somme d'appel est ≤100000 F. (De façon



précise, le nombre entier, écrit en haut de chaque créneau, donne le nombre total des individus, i.e. des engagements, compris dans ce créneau ou à sa gauche). Les créneaux successifs, larges de 5kF, montrent une décroissance régulière; seul le premier créneau est plus léger que celui qui le suit.

Le deuxième histogramme, par tranches de 50kF, s'étend de 100kF à un million: la décroissance y est encore plus régulière (les rares remontées locales s'expliquant par la surcharge des valeurs rondes: multiples exacts de 100kF).



De ce premier examen du fichier, on conclura que de très nombreux engagements concernent des courses à basse somme d'appel: il s'agit essentiellement de chevaux débutants, dont la plupart ne poursuivront pas une très longue carrière; mais, précisément, c'est de la bonne gestion des courses ouvertes aux débutants que dépend la sélection des chevaux de hautes performances, et, plus généralement la prospérité de l'élevage: il faut donc étudier toute la largeur de la distribution.

2.2 Croisement entre somme d'appel et gains antérieurs

Avec le graphique supérieur, croisement entre apl (somme d'appel) et gan (gains antérieurs), on aborde la question posée: relation entre apl et gan. À chacun des engagements de notre échantillon, il correspond un point; l'ensemble de l'échantillon est représenté par un nuage.

De par le réglement, gan doit être inférieur à apl, ce qui laisse aux points représentatifs des engagements la possibilité de remplir tout le triangle situé au-dessous de la diagonale. En fait, il y a une zone assez dense, immédiatement en dessous de la diagonale; mais la densité décroît rapidement quand on va vers l'axe horizontal; lequel correspond à des gains antérieurs nuls.

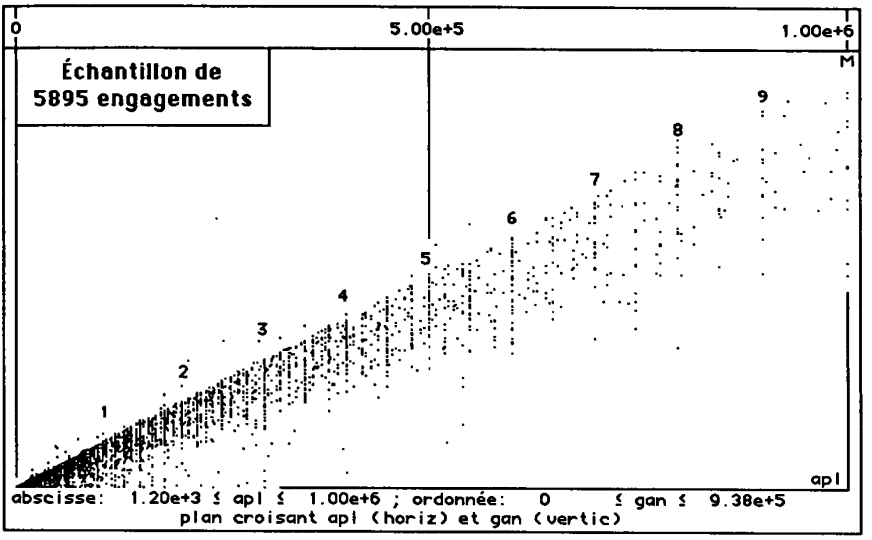
Examinons le graphique plus en détail. Au coin inférieur gauche, sont les engagements dans des courses à faible somme d'appel. Pour ces courses, les gains antérieurs descendent vonlontiers jusqu'à zéro: le coin limité par la bissectrice et l'axe horizontal est d'un noir intense. Il est en effet normal que ces courses attirent des débutants n'ayant strictement aucun gain antérieur.

Mais au-delà d'une somme d'appel de 100kF (chiffre 1 sur la bissectrice), les points avec gan=0 sont très rares; il n'y en a plus (dans l'échantillon) au-delà de 300kF; et c'est pourquoi on a pu tracer l'axe horizontal, sans craindre de recouvrir le nuage des points.

Quand on se déplace vers la droite, le nuage devient clairsemé; ce qui correspond à la distribution de apl montrée sur les histogrammes de la première planche: les courses à forte somme d'appel sont rares. Dans cette partie du nuage, on distingue nettement des stries verticales: celles-ci correspondent à des sommes d'appel rondes: multiples entiers de 100kF, voire de 50kF. On notera que le graphique ne va pas au-delà de M, apl=1000kF.

Les points situés au-dessus de la diagonale sont rares; mais on en voit quelques-uns: ils correspondent à de rares engagements au-dessus de la somme d'appel. Il s'agit généralement d'engagements annulés, finalement, du fait d'un dépassement survenu avant le départ.

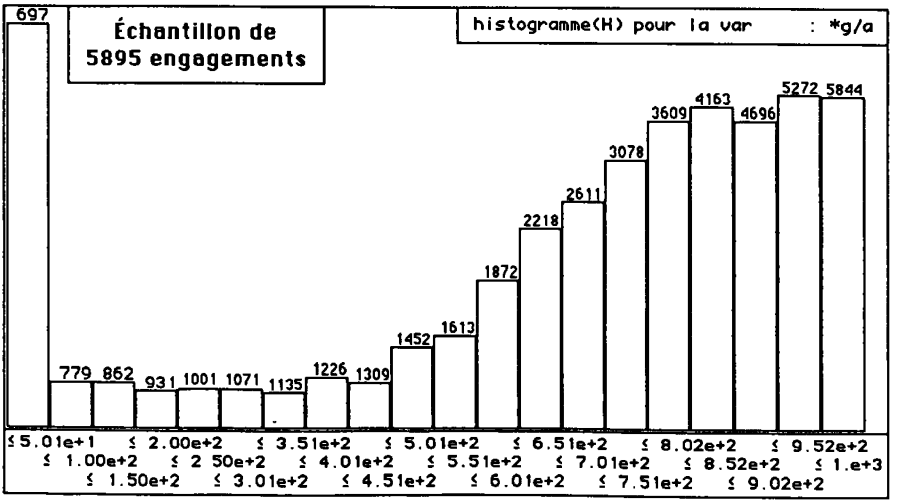
On cherche à caractériser l'estompage du nuage entre la diagonale et l'axe horizontal: tout ce que l'œil peut apprécier est que cet estompage est beaucoup



plus net au-delà de 200kF qu'il ne l'est pour les faibles sommes d'appel. Pour plus de précision, il faut considérer explicitement le rapport gan/apl =*g/a

L'histogramme qui occupe le bas de la page offre une première vue de la variable *g/a: on s'est restreint aux 5844 valeurs inférieures ou égales à 1.

À gauche, le premier créneau renferme les engagements sans gains antérieurs (ou presque): ce créneau est le plus haut de notre histogramme, il



comprend plus du dixième des engagements. Les créneaux suivants restent très bas jusqu'à *g/a=0,5 ; puis ils montent régulièrement jusqu'à *g/a=0,8 ; et se stabilisent ensuite. (De façon précise, les bornes des créneaux sont notées en marge inférieure des graphiques).

Il reste à examiner le lien entre la distribution du rapport *g/a et la valeur absolue de la somme d'appel, apl.

2.3 Croisement entre somme d'appel et le rapport gain/appel

Le graphique en nuage présente le croisement entre apl (somme d'appel) et le rapport *g/a = gan/apl; ce graphique est construit comme celui de la deuxième planche: dans l'un et l'autre, l'axe horizontal est celui de la somme d'appel, apl; mais en introduisant le quotient *g/a, au lieu de la valeur brute des gains, gan, on obtient un meilleur étalement des points; donc une appréciation plus précise de la distribution globale.

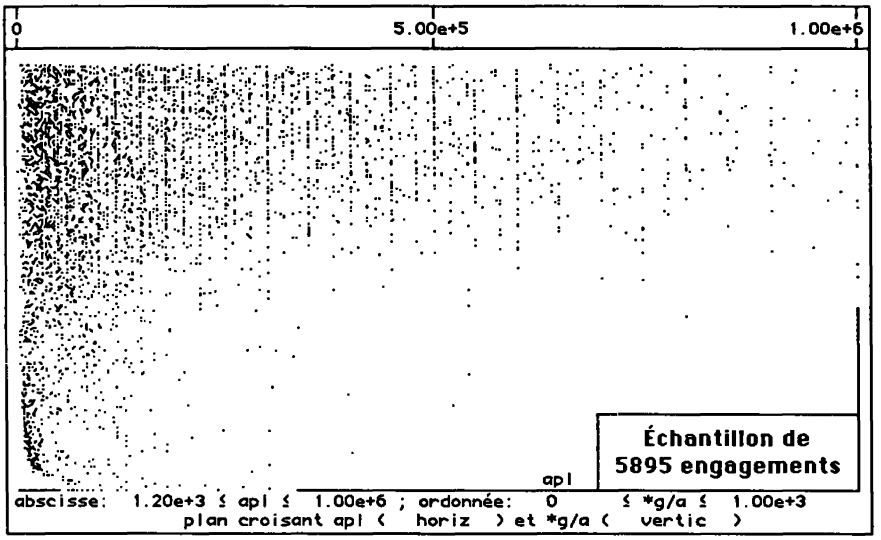
Comme précédemment, des stries verticales marquent les valeurs rondes de apl, de 100kF en 100kF, ou de 50kF en 50kF. En dessous de 200kF, malgré l'étalement vertical des points entre 0 et 1, la densité du nuage est telle que les stries se perdent; d'autant plus que les valeurs intermédiaires de apl sont nombreuses, au-dessous de la maille de 50kF.

Il est clair que l'étalement vers le bas (à partir du maximum de principe : gan = apl, *g/a=1), est plus net à gauche qu'à droite: on a déjà remarqué que les valeurs minima de *g/a sont le fait de débutants, confinés dans les courses à faible somme d'appel. Un graphique chargé de près de 6000 points ne permet toutefois pas de mesurer exactement cet étalement: pour plus de précision, on a produit un tableau du tri croisé.

Dans ce tableau, il y a 11 colonnes. Les dix premières correspondent aux tranches de sommes d'appel par 100kF, jusqu'à 1MF=1000kF; la dernière concerne les courses avec somme d'appel supérieure à 1MF. De même, il y a 21 lignes: les 20 premières partagent l'intervalle de variation du rapport *g/a de 0 jusqu'à 1000 millièmes, par tranches de 50 millièmes; la dernière tranche recense les cas où gan est supérieur à apl: plus de la moitié de ces cas se présentent dans des courses dont la somme d'appel est ≤100kF.

Chaque colonne de nombres du tableau correspond à une bande verticale du nuage des points: il faut seulement noter que, sur le nuage, les fortes valeurs de *g/a sont en haut; tandis que sur le tableau, elles sont en bas. À cela près, les colonnes de nombres offrent un bilan précis des faits ponctuels recensés dans le nuage.

Le total d'une colonne n'est autre que le nombre des points dans la bande; et la décroissance des nombres dans la colonne (décroissance à partir du bas...) exprime l'estompage de la densité dans la bande.



ValSup	1.01e+5	3.00e+5	5.00e+5	7.00e+5	9.00e+5	4.00e+6
5.0 e+1	679 13 5					
1.0 e+2	74 6 1	1				
1.5 e+2	68 12 3					
2.0 e+2	62 7					
2.5 e+2	60 6 1	2	1			
3.0 e+2	59 10		1			
3.5 e+2	49 11 2	2				
4.0 e+2	81 9	1				
4.5 e+2	60 17 4		1	1		
5.0 e+2	100 26 12	1	1	1	1	1
5.5 e+2	114 29 12		1	1	2	1
6.0 e+2	164 46 17	9	6	7	4 1	1 4
6.5 e+2	168 83 35	19	12	10	4 5	2 8
7.0 e+2	196 89 38	16	17	11	7 5	1 3 10
7.5 e+2	200 106 51	32	31	16	10 8	3 3 7
8.0 e+2	227 118 58	36	25	16	13 6	10 5 17
8.5 e+2	241 108 71	37	28	34	12 8	5 3 7
9.0 e+2	208 110 69	49	36	19	11 6	9 5 11
9.5 e+2	235 119 69	55	34	25	16 9	4 4 6
1.0 e+3	223 129 76	55	31	22	12 14	3 2 5
2.1 e+3	26 11 4	3	2	3	2	

tri croisant apl (colonnes) et *g/a (lignes)

Le poids des colonnes diminue de la première à la dixième; la onzième est un peu plus lourde que celle-ci; mais elle correspond à une largeur plus grande sur l'axe apl. Quant à l'étalement: il est manifestement plus prononcé pour les faibles sommes d'appel (les seules où comptent les valeurs de *g/a dans la tranche de 0 à 50 millièmes); mais, au-delà, les fluctuations d'échantillonnage et les différences d'ordre de grandeur rendent difficile la comparaison.

C'est pourquoi la planche suivante est fondée sur des calculs de pourcentages.

2.4 Profils cumulés, par somme d'appel, du rapport gain/appel

un échantillon de 5895 engagements
12011

	G≤e	G≤1	G≤2	G≤3	G≤4	G≤5	G≤6	G≤7	G≤8	G≤9	G≤X
A<%	37	41	47	52	56	59	66	74	80	89	100
A1%	2	5	6	8	12	18	29	43	62	80	100
A2%	1	2	4	5	7	11	18	34	54	75	100
A3%	1	1	2	2	2	5	11	24	45	70	100
A4%	0	0	0	1	2	2	5	16	37	64	100
A5%	0	0	0	0	0	1	4	16	42	70	100
A6%	0	0	0	1	1	2	7	20	38	70	100
A7%	0	0	0	0	0	0	1	14	38	66	100
A8%	0	0	0	0	0	1	9	23	42	62	100
A9%	0	0	0	0	0	0	3	11	45	76	100
AX%	0	0	0	0	0	0	11	21	50	79	100
A>%	0	0	0	0	0	1	8	31	60	84	100

Le Tableau proposé compte 12 lignes et 11 colonnes.

La première colonne, G≤e, concerne les cas où *g/a est quasi nul (inférieur à un seuil e=0,1 proche de zéro): il s'agit des cas, assez nombreux d'après le §3.3, où les gains antérieurs sont pratiquement nuls. Dans les colonnes suivantes, de G≤1 à G≤X, il s'agit de bornes du rapport *g/a de 0,1 à 1; (plus précisément, compte tenu de l'objectif de l'étude, cf infra, les rares cas d'engagements au-dessus de la somme d'appel ont été ajoutés à la dernière colonne G≤X).

La première ligne, A<% concerne les courses dont la somme d'appel est ≤30kF; la deuxième ligne, A1%, est pour les courses où apl varie de 30kF à 100kF; A2%, s'étend de 100kF (exclu) à 200kF (inclus); ... ; AX% s'étend de 900kF (exclu) à 1MF (inclus); le reste (au-delà de 1MF) est dans A>%. Donc, chaque ligne se rapporte à un sous-ensemble de notre échantillon d'engagements, délimité par une condition sur la somme d'appel.

Le tableau du §2.4 diffère à plusieurs titres de celui du §2.3.

Il faut d'abord noter que, dans le tableau du §2.4, la somme d'appel définit des lignes, et non des colonnes, comme au §2.3: la raison est qu'au §2.3, on voulait, sur le nuage, étaler au maximum l'axe apl en prenant la dimension horizontale; tandis qu'au §2.4, la ligne individuelle est associée, en perspective, à une courbe; laquelle, par comparaison avec celle afférente aux sujets gagnants (cf. §3) et non aux engagés en général (comme au §2), peut suggérer des règles de gestions des engagements pour une classe de course de niveau d'appel donné.

Ensuite, relativement au tableau du §2.3, celui du §2.4 supprime, par des calculs de pourcentages, les inégalités de poids entre les diverses tranches de

sommes d'appel (inégalités apparues dès le §2.1). Pour passer d'une colonne du tableau du §2.3 à une ligne du tableau du §2.4, il faut d'abord diviser chaque nombre de la première par le total; et puis effectuer des cumuls dont nous préciserons l'utilité.

En bref, supposons que l'on fixe pour le rapport *g/a un seuil = 0,8: c'est-à-dire qu'on n'accepte d'engagements que pour des chevaux dont les gains antérieurs sont compris entre 80% et 100% de la somme d'appel: il importe de savoir dans quelle proportion on aura ainsi réduit le nombre total des engagements acceptés.

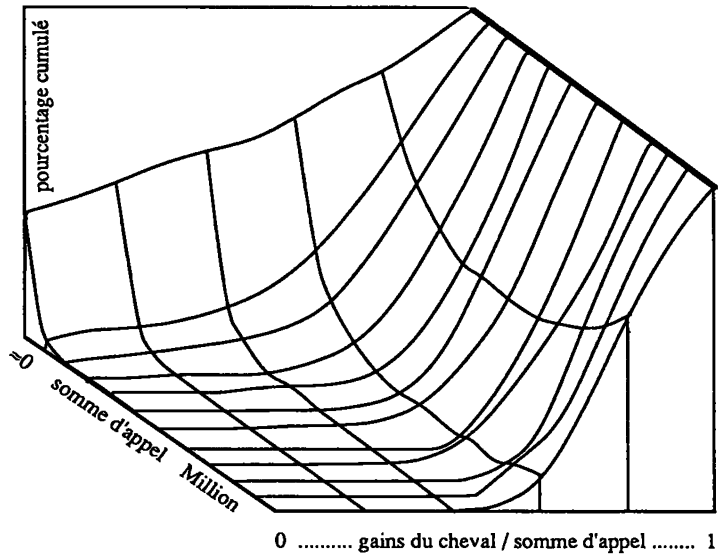
Le tableau du présent § permet de répondre directement à cette question, sous l'hypothèse simplificatrice que le comportement d'engagement est régi par le niveau d'appel; (hypothèse que l'on peut aisément modifier, pour prendre en compte d'autres variables: e.g. la zone géographique, délimitée par une classification des départements où se situent les hippodromes).

De façon précise, pour une course dont le niveau d'appel est entre 600kF et 700kF, *g/a est ≤0,8 dans 38% des cas: k(A7%, G≤8) = 38; c'est l'exemple choisi dans le titre. Donc fixer un seuil à 0,8 éliminerait plus du tiers des engagements: la piste serait déblayée, mais la question est évidemment de savoir combien de gagnants potentiels seraient ainsi éliminés!

On voit sur le tableau que c'est précisément dans la bande A7% qu'est maxima la concentration du rapport *g/a vers les valeurs proches de 1: au-delà et en deçà, la concentration est moindre. Cela se comprend: pour les faibles sommes d'appels, il y a beaucoup de chevaux non confirmés, voire sans gain antérieur; dans les courses à forte somme d'appel, qui sont aussi des courses offrant de grands gains (nominal élevé), peuvent s'engager des chevaux dont les gains antérieurs, tout en étant notables en valeur absolue, sont bien en dessous du plafond fixé.

Il nous a paru bon de mettre en perspective cavalière l'ensemble des 12 courbes caractéristiques afférentes aux zones d'appel, de A<% à A>%: le dessin est moins précis qu'un tableau, mais il suggère une synthèse globale des nombres. Au fond, est la courbe correspondant aux très faibles sommes d'appel: A<%: cette courbe est la seule qui parte nettement au-dessus de zéro, parce que les engagements de trotteurs sans gain antérieur y sont très nombreux: plus du tiers, k(A<%, G≤e) = 37 dans le tableau. Les autres courbes partent, du niveau zéro, à peu près à plat. Au premier plan est la courbe afférente aux très fortes sommes d'appel (au-delà de 1MF). Cette courbe débute certes, à gauche, horizontalement, à zéro, mais certaines courbes intermédiaires ont un palier plus long, et, corrélativement, une montée finale plus accentuée encore.

Les rares cas d'engagements au-dessus de la somme d'appel ayant été ajoutés à la colonne G≤X, *g/a=1, les courbes atteignent, pour cette abscisse, le niveau 100%.



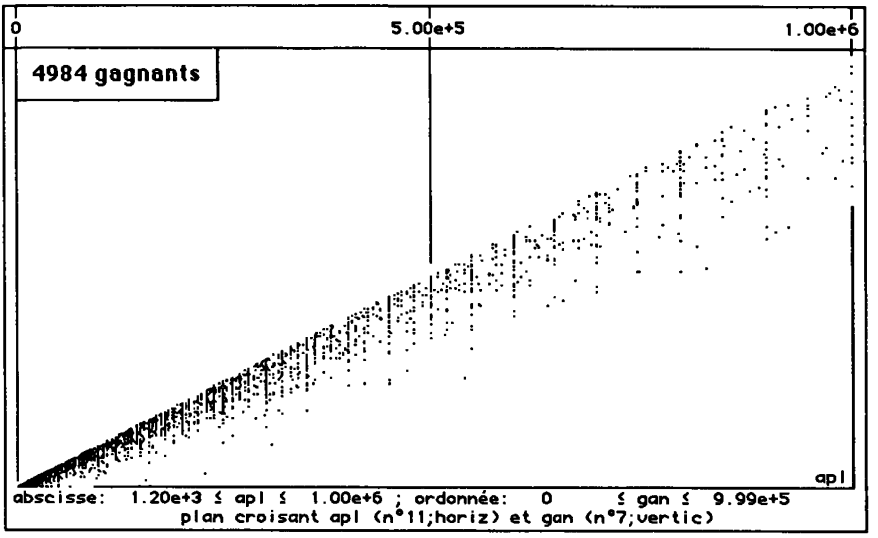
Relation entre somme d'appel et gains du cheval pour 5895 engagements

Reste à comparer l'ensemble des engagés au sous-ensemble des gagnants. C'est ce qu'on a fait en reprenant l'étude des §§2.2 à 2.4, non pour un échantillon d'engagements, mais pour deux fichiers exhaustifs: le fichier des partants placés 5-ème; et le fichier des partants placés premier, i.e., des gagnants. Les résultats obtenus dans les deux cas diffèrent peu: dans le présent article, on se bornera à publier ceux afférents aux gagnants.

3 Étude des gagnants

3.1 Croisement entre somme d'appel et gains antérieurs

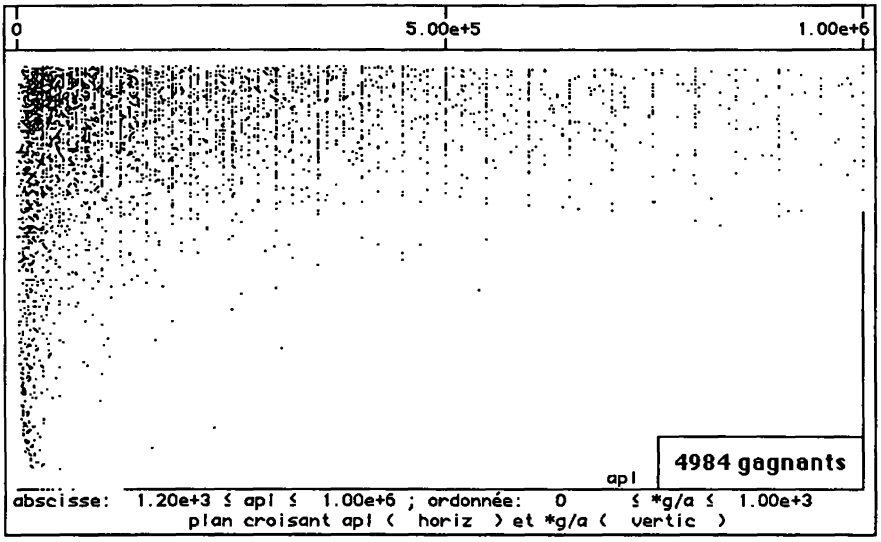
Le graphique en nuage du §3.1 est construit comme celui du §2.2: dans l'un et l'autre, l'axe horizontal représente la somme d'appel, apl; et l'axe vertical, les gains antérieurs, gan. Les deux graphiques se ressemblent et appellent les mêmes remarques: notamment quant aux stries parallèles afférentes aux sommes rondes de 50kF en 50kF, à droite; et quant à l'accumulation des points vers la pointe, à gauche, entre l'axe horizontal et la première diagonale. Mais les deux graphiques diffèrent assez nettement quant à l'étalement du nuage: pour les gagnants, §3.1, l'accumulation vers la diagonale est plus nette, particulièrement dans la pointe, à gauche; l'étalement est, au contraire, plus prononcé pour l'échantillon des engagements §2.2. De plus, ici, il n'y a pas de point au-dessus de la diagonale (apl<gan).



ValSup	1.00e+5	2.00e+5	3.00e+5	4.00e+5	5.00e+5	6.00e+5	7.00e+5	8.00e+5	9.00e+5	1.00e+6	3.50e+6
5.00e+4	1935	4	1								
1.00e+5	623	142									
1.50e+5		489	13	1							
2.00e+5		228	116	3							
2.50e+5			254	22	2						
2.99e+5			147	83	3	1					
3.50e+5				149	17	1					
4.00e+5				80	51	6					
4.50e+5					85	13	2				
4.98e+5					42	43	10				
5.48e+5						49	15	3	1		
5.97e+5						26	31	5	2		
6.50e+5							24	12	2	2	
7.00e+5							13	21	5		
7.50e+5								16	12	3	
7.98e+5								12	13	4	3
8.46e+5									10	3	2
8.98e+5									8	5	4
9.43e+5										8	9
1.00e+6										1	12
2.90e+6											80

tri croisant apl (col) et gan (lignes)

Le tableau de croisement comprend 11 colonnes, correspondant à 11 classes de valeurs de apl: les dix premières ont chacune une largeur de 100kF; la dernière comprend tout ce qu'il y a au-delà de 1MF. De même, on a 21 lignes; dont les 20 premières correspondent à des intervalles larges de 50kF pour gan; et la dernière s'étend jusqu'au maximum, au-delà de 1MF.



ValSup	1.00e+5	2.00e+5	3.00e+5	4.00e+5	5.00e+5	6.00e+5	7.00e+5	8.00e+5	9.00e+5	1.00e+6	4.00e+6
4.8 e+1	272										
1.0 e+2	18	1									
1.5 e+2	34		1								
2.0 e+2	26										
2.5 e+2	29										
3.0 e+2	36										
3.5 e+2	30	2		1							
4.0 e+2	26	2									
4.5 e+2	37	1	1								
5.0 e+2	43	6	2			1					
5.5 e+2	45	7		2	1						
6.0 e+2	76	10	6	2	3	1					
6.5 e+2	93	17	6					1	2		1
7.0 e+2	88	34	14	13	2	5	5	2	2	2	8
7.5 e+2	166	61	28	15	6	5	3	3		2	4
8.0 e+2	197	69	37	24	14	10	5	4	7	4	13
8.5 e+2	230	109	84	45	20	14	16	5	3	6	16
9.0 e+2	301	148	97	56	39	30	18	16	12	4	15
9.5 e+2	399	188	110	82	43	33	28	16	10	7	17
1.0 e+3	412	208	145	98	72	40	20	22	17	10	27

tri croisant apl (colonnes) et *g/a (lignes)

3.2 Croisement entre apl (somme d'appel) et le rapport *g/a = gan/apl pour 4984 gagnants

Le graphique en nuage et le tableau du §3.2 sont construits comme ceux du §2.3. Dans les deux cas, en introduisant le quotient *g/a, au lieu de la valeur brute des gains, gan, on obtient, sur les graphiques, un meilleur étalement des

points; donc une appréciation plus précise de la distribution globale; cette appréciation peut même être chiffrée en consultant les tableaux de tris croisés.

Le tableau atteste que, pour les gagnants, les très faibles valeurs de *g/a sont rares: 272 cas, au total, pour un rapport <5 millièmes; 19 autres cas, pour un rapport ≤100millièmes=0,1. Et les valeurs jusqu'à *g/a=0,45 ne se trouvent que dans 9 cas pour des sommes d'appel supérieures à 100kF, et ces exceptions elles-mêmes sont relatives à des sommes d'appel ≤400kF.

Quant à l'étalement, la différence entre les graphiques des §§2.3 et 3.2 est manifeste. D'une part, le long du bord supérieur, *g/a=1, l'ensemble des gagnants est plus reserré que l'échantillon des engagements; d'autre part, le long du bord vertical droit, pour les valeurs de *g/a<0,5, les gagnants sont rares, et confinés dans les valeurs très faibles de apl; tandis que l'échantillon des engagements est assez dense pour apl≤50kF et d'un poids significatif jusqu'à apl=200kF.

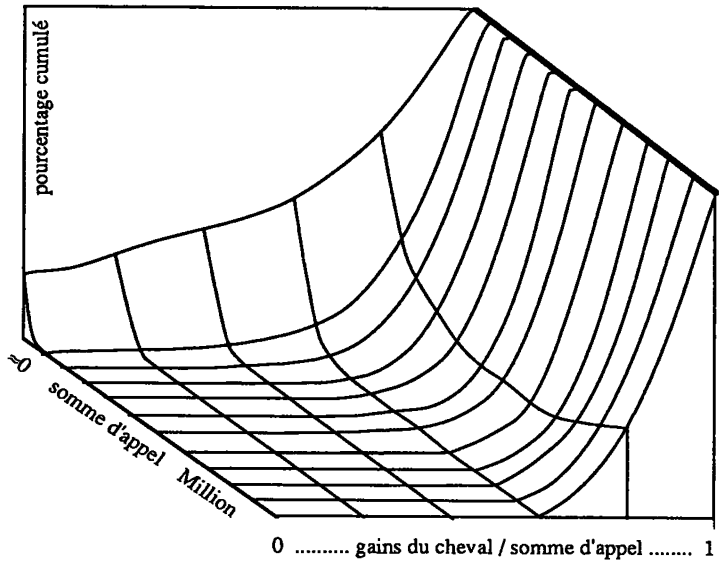
3.3 Tableau en pourcentages cumulés du croisement entre sommes d'appels (A) et rapport gain/appel (G)

chevaux gagnants
12011

	G≤e	G≤1	G≤2	G≤3	G≤4	G≤5	G≤6	G≤7	G≤8	G≤9	G≤X
A<%	20	21	26	30	33	37	42	51	62	78	100
A1%	1	1	1	1	2	5	8	14	30	56	100
A2%	0	0	0	0	1	1	3	9	24	54	100
A3%	0	0	0	0	0	1	2	5	18	51	100
A4%	0	0	0	0	0	0	1	5	16	46	100
A5%	0	0	0	0	0	0	2	3	13	41	100
A6%	0	0	0	0	0	1	1	5	15	47	100
A7%	0	0	0	0	0	0	0	5	14	49	100
A8%	0	0	0	0	0	0	0	4	14	45	100
A9%	0	0	0	0	0	0	0	8	19	49	100
AX%	0	0	0	0	0	0	0	6	23	51	100
A>%	0	0	0	0	0	0	0	9	26	55	100

Le tableau, dont l'étude fait l'objet du présent §, est construit comme celui du §2.4 et se lit suivant les mêmes règles.

Reprenons le même exemple. On a dit que fixer un seuil à 0,8 éliminerait plus du tiers (38%) des engagements; ce qui posait la question de savoir combien de gagnants potentiels seraient ainsi éliminés! Réponse à cette question se trouve sur le tableau ci-dessus: de façon précise, pour les gagnants dans une course dont le niveau d'appel est entre 600kF et 700kF, *g/a est ≤0,8 dans 14% des cas: k(A7%, G≤8) = 14. En moyenne, le pourcentage de gagnants éliminés serait donc 14%. Le taux des partants placés cinquièmes, donc susceptibles de jouer un rôle dans la course, et qu'on éliminerait, par la même règle, est, à peu près le même: 15%. Il revient aux organisateurs des courses d'orienter leur décision d'après de telles indications.



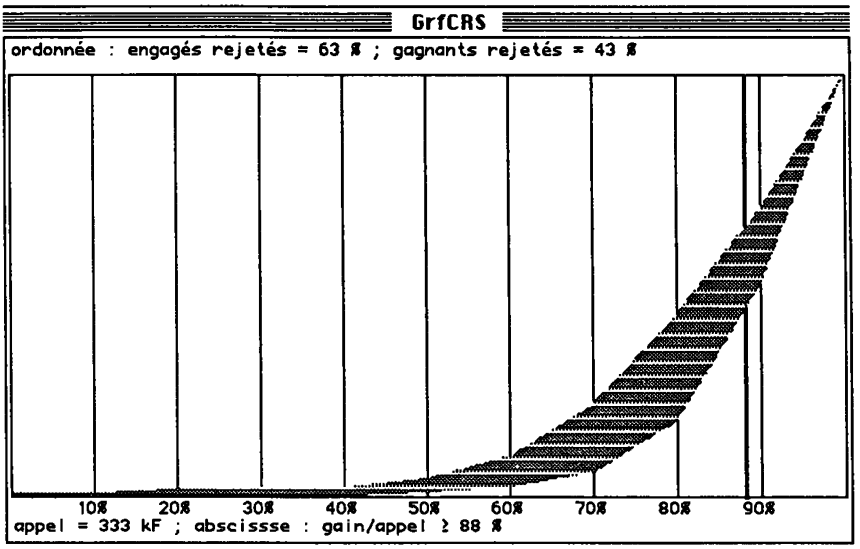
Relation entre somme d'appel et gains antérieurs pour les gagnants

Même si elle est nettement décalée vers les fortes valeurs de *g/a = gan/apl, la surface, qui, dans ce §3.3, présente en vue cavalière l'ensemble du contenu du tableau de pourcentages, reproduit celle du §2.4.

Au fond, est la courbe correspondant aux très faibles sommes d'appel: A<%: cette courbe est la seule qui parte nettement au-dessus de zéro, parce que les engagements de trotteurs sans gain antérieur y sont assez nombreux: exactement un cinquième, k(A<%, G≤e) = 20 dans le tableau. Les autres courbes partent à plat du niveau zéro.

Au premier plan, est la courbe afférente aux très fortes sommes d'appel (au delà de 1MF). Cette courbe débute, à gauche, horizontalement, à zéro, mais certaines courbes intermédiaires ont un palier plus long, et, corrélativement, une montée finale plus accentuée encore.

Les graphiques en perspective cavalière qui illustrent les §§2.4 et 3.3 offrent une image suggestive des taux d'engagés ou de gagnants susceptibles d'être éliminés, si, outre la majoration usuelle des gains antérieurs par la somme d'appel, on introduisait une borne inférieure pour les gains antérieurs. Mais il est difficile de tirer d'une semblable image des pourcentages précis: afin de calculer de tels pourcentages, on a écrit un programme conversationnel d'interpolation.



3.4 Estimation, par interpolation, des taux d'engagés et de gagnants éliminés en exigeant un minimum pour le rapport gain/appel

Les données sur lesquelles se fonde le calcul d'interpolation sont celles des tableaux des §§2.4 et 3.3; à ceci près que, pour les faibles sommes d'appels, on a fait des bilans plus détaillés: au lieu de deux classes, (apl≤30kF) et (30kF<gan≤100kF), on a distingué dix classes, définies par paliers de 10kF: (apl≤10kF), (10kF<apl≤20kF),... Toutes ces données sont réunies en un tableau unique.

Le programme d'interpolation 'GrfCRS' s'ouvre sur une légende explicative:

```
 NB ce programme montre, en fonction de
la somme d'appel, quel serait, en moyenne,
le % de gagnants et le % d'engagés éliminés
en exigeant un minimum fixé pour gain/appel
```

et, après lecture du tableau de données, commence un dialogue qui se répète un aussi grand nombre de fois que le demande l'utilisateur, avec l'affichage de résultats numériques illustrés par un graphique, susceptible de guider les demandes ultérieures.

De façon précise, à la demande:

```
la somme d'appel de la course est (en kF)
```

l'utilisateur répond par un nombre auquel le programme impose de rentrer dans l'intervalle [10..1000] (i.e. de 10kF à un million); puis vient une seconde demande:

```
le pourcentage min gain/appel exigé (de 10 à 90) est
```

la réponse de l'utilisateur étant, ici encore ramenée dans les limites fixées.

S'affiche alors un graphique, qui comme celui publié en tête du présent §, montre, pour la valeur choisie de la somme d'appel, une zone grise ayant pour frontières les deux courbes donnant, en fonction du pourcentage (gain/appel), porté en abscisse, les taux respectifs de refus pour les engagés - courbe supérieure - et les gagnants - courbe inférieure. Une barre verticale, en trait gras, est placée à la valeur choisie pour le minimum de gain/appel; et les valeurs numériques des taux de rejet s'inscrivent en haut du cadre du graphique.

D'après le contour de la zone grise, l'utilisateur peut apprécier les taux de rejet qu'entraînerait le choix d'un autre minimum. Il efface le graphique en entrant un caractère quelconque; et s'il répond par 'O', Oui, à la question:

```
faut-il considérer d'autres cas O ou N
```

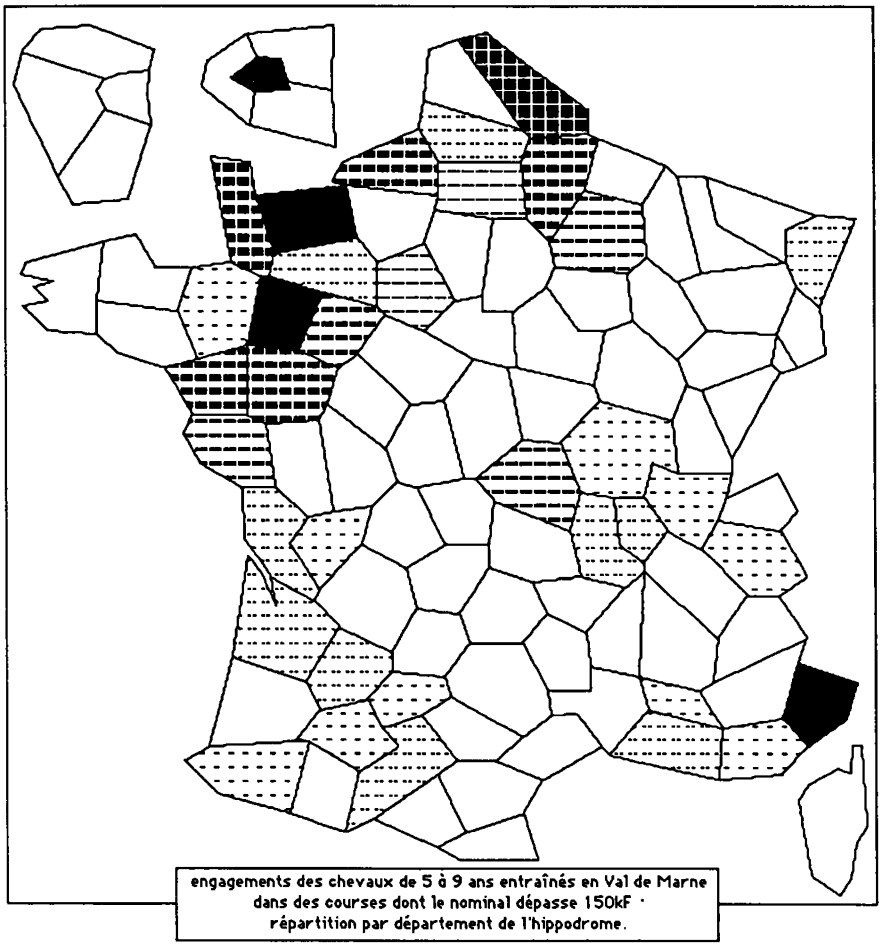
le dialogue reprend avec d'autres valeurs qui commandent l'affichage d'un nouveau graphique.

4 Cartographie

Classifications et cartes géographiques servent à rendre compte des disparités régionales que le modèle définitif devra intégrer. Dans ce premier article, nous nous bornerons à donner un aperçu de telles cartes. D'une part, au §4.1, une carte qui traduit simplement par l'intensité du noir des trames une colonne d'un tableau dont l'ensemble des lignes n'est autre que l'ensemble des 95 départements français (plus exactement, les 94 départements du continent et la Corse). D'autre part, au §4.2, une carte qui représente une partition en classes de ces mêmes 95 départements, obtenue en soumettant un tableau à l'analyse des correspondances et à la classification ascendante hiérarchique.

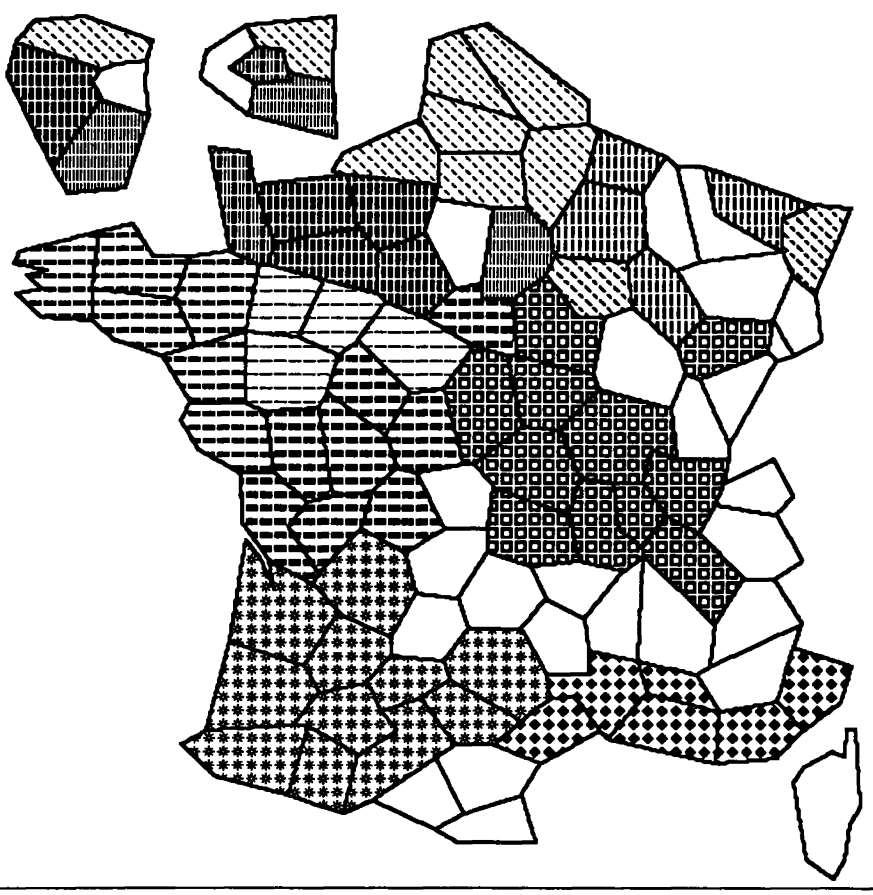
4.1 Exemple de représentation cartographique d'une colonne d'un tableau de contingence

L'ensemble du fichier traité contient 453898 engagements. On a filtré le sous-ensemble de ce fichier constitué des engagements de chevaux âgés de 5 à 9 ans dans des courses dont le nominal (ou somme à partager entre les gagnants) dépasse 150kF. Ce sous-fichier a lui-même été mis sous la forme d'un tableau de contingence, croisant avec lui-même l'ensemble des 95 départements. De façon précise, k(i, j), nombre inscrit à l'intersection de la ligne i et de la colonne j, est le nombre d'engagements relevé, pour des chevaux entraînés dans le département j, dans une course ayant lieu dans le département i : e.g.



k(35,94)=1, parce qu'il y a eu un engagement de cheval entraîné dans le Val-de-Marne (94) dans une course ayant lieu en Ile-et-Vilaine (35).

Chacune des colonnes, j, de ce tableau (comme aussi chacune des lignes) se prête à une représentation cartographique: il suffit de traduire la valeur k(i, j) par une intensité de noir, sur le département i. C'est ce qu'on a fait ici pour le département du Val-de-Marne, qui contient un important centre d'entraînement de chevaux. On voit que les engagements sont particulièrement nombreux dans les 4 départements, couverts en noir, des Alpes-Maritimes, du Calvados, de la Mayenne et de la Seine. Mais ils sont absents dans de nombreux départements laissés en blanc: beaucoup de ces départements n'ont, d'ailleurs, aucune course.

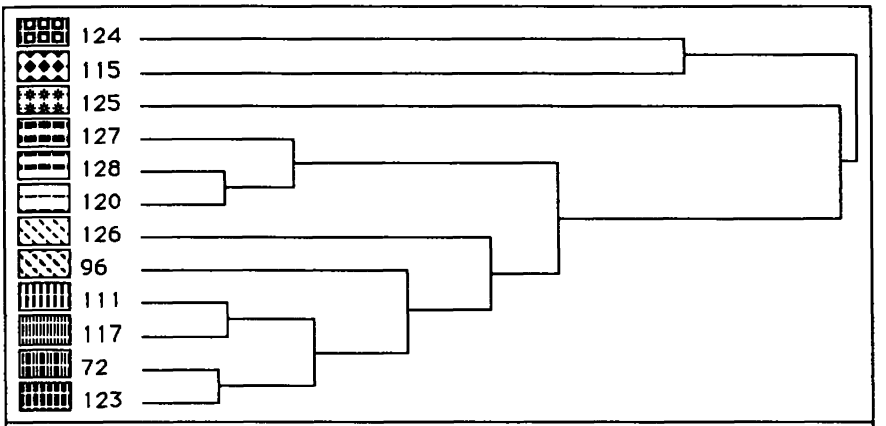


Classification des départements d'après les hippodromes où sont partants les chevaux d'âge 4 qui y sont entraînés

les départements sans entraînement et la Corse sont en blanc

4.2 Exemple de représentation cartographique d'une classification fondée sur une matrice de flux

La carte qui illustre le présent § est fondée sur un tableau de contingence en tout analogue à celui dont une colonne est rendue sur la carte du §4.1. On a filtré le sous-ensemble constitué des engagements de chevaux âgés de 4 ans partis effectivement dans la course considérée. Ce sous-fichier a été mis sous la forme



Classification des départements d'après les hippodromes où sont partants les chevaux d'âge 4 qui y sont entraînés

les départements sans entraînement et la Corse sont en blanc

d'un tableau de contingence, croisant avec lui-même l'ensemble des 95 départements: k(i, j), nombre inscrit à l'intersection de la ligne i et de la colonne j, est le nombre d'engagements relevé, pour des chevaux entraînés dans le département i, et partants dans une course ayant lieu dans le département j. On peut dire que k(i, j) représente un flux de chevaux de i (entraînement) vers j (hippodrome).

L'analyse des correspondances fournit une synthèse de cette matrice de flux; i.e. une vue globale des affinités que met entre des départements le fait que les chevaux qui y sont entraînés courrent, dans des proportions similaires, sur les mêmes hippodromes de France; que sur leurs hippodromes courent des chevaux entraînés dans les mêmes lieux; ou, enfin, que les chevaux entraînés dans l'un courent volontiers sur les hippodromes de l'autre. La classification automatique exprime ces proximités en constituant un système de régions.

De façon précise, la carte du §4.2, est fondée sur une partition de l'ensemble des départements considérés comme lieux d'entraînement (ensemble I des lignes); deux départements étant agrégés si leurs chevaux courent dans les mêmes lieux, dans des proportions similaires. Ainsi, se trouvent constituées des régions qui, bien que déterminées par l'analyse des flux et non par des calculs de distance sur le terrain, se trouvent, pour la plupart, recouvrir, sur la carte, des zones d'un seul tenant assimilables à de grandes provinces.

Le graphique arborescent complète la carte du §4.2. On voit que, sur la carte, des trames similaires ont été choisies pour les classes qui s'agrègent entre

elles à un bas niveau: e.g. lignes horizontales pour 128, 120, Bretagne et Vallée de la Loire, proches entre elles et proches de 127, Poitou...

5 Conclusion: traitement informatique et recherches ultérieures

Les fichiers de base de la présente étude ont été créés sur un matériel de grande capacité, adapté à l'interrogation des bases de données relationnelles. Ils ont été ensuite traités sur microordinateur par le logiciel MacSAIF d'analyse multidimensionnelle; en recourant, le cas échéant, à des programmes spéciaux.

Dans leur quasi-totalité, les programmes du logiciel MacSAIF d'analyse des données sont conçus pour traiter des tableaux susceptibles d'être logés dans une mémoire centrale limitée à 5 ou à 8 mégaoctets. À partir d'un tel tableau sont créés, en mode conversationnel, des diagrammes, des courbes, des cartes, des classifications, des tableaux de cumuls; éventuellement on adjoint des colonnes ou lignes nouvelles: e.g., dans la présente étude, la colonne afférente à la variable quotient *g/a.

Mais ici, nous devions traiter un fichier de base de quelque 50 mégaoctets; ce qui dépasse la capacité des micro-ordinateurs présentement répandus. On a donc écrit des programmes paramétrés particuliers afin de filtrer des sous-fichiers du fichier de base et de créer des tableaux de contingence, tels que les matrices de flux qui ont servi à tracer les cartes proposées en exemple. Ainsi que l'explique l'article [ÉTAT MacSAIF], certaines fonctions des programmes spéciaux ont pu ensuite être introduits, à titre d'option, dans les programmes généraux du logiciel MacSAIF.

Pour annoncer la suite de nos travaux, c'est encore le terme de programme que nous utiliserons, mais dans un un autre sens que celui reçu en informatique!

Il s'agit, en effet, de prévoir, avec sûreté, l'effet d'un programme de courses national complet sur l'enchaînement des engagements et des départs de tous les chevaux trotteurs. À cette fin, il faudra prendre en compte explicitement la diversité des chevaux et non seulement la variable gan, gain antérieur; construire une typologie des chevaux pour chaque âge; et suivre, au fil des années, les performances des chevaux résumées selon cette typologie.

C'est en poursuivant de telles expériences que statisticiens et experts en programmation des courses pourront élaborer conjointement des méthodes nouvelles.