CENTRE NATIONAL D'ETUDES SPATIALES

Groupe d'Etude des
Phénomènes Aérospatiaux Non-Identifiés

Toulouse, le 26 Novembre 1982
CT/GEPAN - N° 00148

RECHERCHE STATISTIQUE D'UNE TYPOLOGIE

IDENTIFIEE / NON-IDENTIFIEE

ISSN : 0750-6694

Note Technique N° 13

S O M M A I R E

1. - LIMINAIRES

1.1. - PRÉLIMINAIRES

Cette Note Technique est la troisième abordant la description statistique des témoignages relatifs à des observations de Phénomènes Aérospatiaux Non-identifiés (PAN). Il nous a paru important de préciser les objectifs et limites de cette étude. Ces préliminaires auraient dû trouver leur place antérieurement à tout travail statistique, mais les deux raisons ci-dessous font que des précautions élémentaires doivent être constamment réévaluées au cours de l'analyse.

POPULATION NON-IDENTIFIÉE :

• La première est spécifique à la nature des données : les outils statistiques ont essentiellement pour objectif de décrire, résumer, tester les principaux paramètres d'une *population*, c'est-à-dire, par définition, *d'un ensemble d'individus* présentant une ou des caractéristiques communes.*

On constate alors que la propriété commune à tous les individus définissant la population étudiée est négative** : c'est la *non-identification, par un témoin-observateur et dans un environnement particulier, de la cause de sa perception* ; elle ne définit pas une classe particulière de phénomènes mais seulement un type de comportement ou de relations phénomène/observateur/environnement.

---

*Elément de la population

**Les termes positifs de la définition sont : "phénomènes" (ce qui est perçu par les sens) et "aérospatiaux" mais ils sont d'une telle généralité qu'ils n'apportent guère de précision.

RECHERCHE NON-INDÉPENDANTE :

• La deuxième raison concerne la conceptualisation du problème. D'une part le GEPAN, de par sa position et son rôle, est un facteur actif parmi l'environnement sociologique du "phénomène OVNI" (Cf. BESSE, ESTERLE, JIMENEZ - 81) ; d'autre part, son sujet de travail n'est pas vierge mais, comme toute recherche, il a été précédé par une série de travaux (Cf. Notes d'Information 2, 3 et 4, Note Technique N°3, chapitre 1) qui ont défini une *dynamique de recherche* particulière avec ses propres concepts, discours, systèmes d'hypothèses et préjugés, par rapport auxquels il est nécessaire de se situer.

D'un point de vue général, le GEPAN ne considère pas d'hypothèse globale explicative, il ne développe que des hypothèses ponctuelles et particulières aux situations concrètement rencontrées (Cf. Plaquette GEPAN). Aussi du point de vue spécifique à l'étude statistique ceci nous conduit à reconsidérer l'*héritage* qui est constitué d'une part de la saisie et du codage des informations et d'autre part des priorités accordées aux différentes analyses participant de l'approche statistique.

Le propos de ce travail est donc double :

• Achever la description statistique du fichier des observations recueillies par la Gendarmerie Nationale entre 1974 et 1978, dans sa forme codée actuelle ; c'est la continuation des travaux entâmés dans les notes techniques 2 et 4 ; tous les cas sont maintenant codés.

• Redéfinir une stratégie pour des études statistiques à venir.

1.2. - OBJECTIFS

Un premier travail statistique (BESSE 80) réalisé à partir du fichier de la Gendarmerie Nationale tentait de révéler les biais introduits dans la constitution des fichiers de données. Le travail suivant (BESSE 81) insistait sur les liaisons témoins/phénomènes en faisant remarquer que deux approches et donc deux méthodologies différentes doivent être développées suivant que l'on s'intéresse à la population des témoins ou encore à celle des phénomènes présumés.

Le présent travail vise à décrire les caractéristiques globales de l'ensemble des phénomènes décrits dans les procès-verbaux de la Gendarmerie, tout en sachant bien que celles-ci ont été sévèrement filtrées par les témoins*. Les limites de cette recherche sont donc imposées par notre incapacité actuelle à pouvoir trier, parmi les caractéristiques relevées, celles dues au témoin, celles dues au phénomène-source, celles liées aux conditions d'observation ou encore celles attribuables à d'éventuels stéréotypes sociaux**.

D'autre part, compte tenu de l'imprécision avec laquelle est caractérisée la population étudiée et de l'hétérogénéité de celle-ci (Cf. § 3), l'objectif principal est avant tout de rechercher une *typologie des observations*. Ce n'est que lorsqu'une répartition en classes suffisamment homogènes sera obtenue que l'on pourra envisager la recherche de *modèles explicatifs* pour chacune de celles-ci***.

---

*Il aurait été certainement préférable de commencer par étudier les témoins et leur perception mais l'héritage est tel que...

**Ceux-ci sont actuellement recherchés à l'aide d'un sondage (résultats à paraître).

***Dans la revue INFORESPACE (n° 4 hors-série, décembre 80), on trouve un bon exemple montrant la nécessité d'une telle démarche. Les auteurs y étudiant la corrélation entre la fréquence journalière des observations et la position du soleil trouvent celle-ci significativement élevée. Ils ne donnent que très peu de précisions sur la façon dont a été construit le fichier mais tout semble indiquer qu'aucune sélection n'y a été effectuée. Or (Cf. § 3.1.) les ballons sondes ou Vénus sont des sources de confusion très fréquentes qui ont lieu justement un peu avant ou après le coucher du soleil ! Ceci suffirait peut-être à expliquer les pics importants observés dans les distributions journalières pour le fichier considéré et ainsi les corrélations calculées.

En règle générale, il faut se méfier de ne pas attribuer à l'ensemble des observations les caractéristiques d'un sous-groupe homogène de celles-ci.

De nombreuses classifications ont déjà été proposées dans la littérature spécialisée (HYNECK - SAUNDERS...) mais toutes, extrinsèques, ont été induites sur le corpus des données et reposent essentiellement sur un critère d'éloignement de l'observation (voir de son "étrangeté"). Connaissant l'imprécision de cette estimation et l'importance que peut avoir l'environnement sur celle-ci, les classifications présentées semblent donc concerner beaucoup plus le témoin et les conditions d'observation que le phénomène supposé à l'origine de la description.

En résumé l'objectif reste la recherche d'une *typologie intrinsèque* aux données en tenant compte "au mieux" des imprécisions et écueils signalés ci-dessus. Des éléments permettant d'élaborer une stratégie de recherche plus rigoureuse visant à intégrer les aléas du témoignage humain sont abordés au paragraphe 6.

1.3. - INTRODUCTION

Le fichier des procès-verbaux issus de la Gendarmerie est entièrement codé pour les années 74 à 78 incluses à l'exception de quelques cas qui ont donné lieu à de très nombreux témoignages et qui seront traités à part.

Dans le deuxième paragraphe, sont posés les problèmes concernant le codage des informations afin d'aboutir à une représentation de celles-ci excluant "au mieux" les difficultés rencontrées dans les travaux précédents.

Le but du troisième paragraphe est alors de décrire la famille des cas qui ont été identifiés a posteriori avant d'en esquisser une typologie.

Dans le quatrième paragraphe, il s'agit de préciser la notion de "non-identification" en considérant d'abord chaque variable puis, globalement, en étudiant la répartition des cas restant non-identifiés dans la typologie précédente.

Enfin, après une conclusion provisoire, une nouvelle approche de type probabiliste, fondée sur les résultats de l'annexe 3, est proposée afin d'adapter codage et analyses aux problèmes très particuliers que pose l'étude des phénomènes aérospatiaux non-identifiés.

2. - LES DONNEES

2.1. - LES DONNÉES BRUTES

Les informations brutes contenues dans les procès-verbaux de Gendarmerie ont été résumées à l'aide du codage détaillé par DUVAL - 79 (Cf. Annexe 1). Celui-ci, construit à partir de certaines idées a priori présente des défaillances (Cf. BESSE - 81) lorsqu'il est confronté à la réalité des procès-verbaux ou lors des analyses statistiques :

• Certaines variables ont été abandonnées car jugées à l'usage inutile, trop peu significatives ou trop empreintes de "subjectivité" :
- type d'observation,
- crédibilité,
- intérêt,
- longitude - latitude,
- estimation de l'accélération,
- direction azimutale.

• D'autres apportent ces redondances préjudiciables à la rigueur des analyses (une même information, codée deux fois, prend une importance artificielle) :
- description de la trajectoire et hauteur angulaire (stationnement près du sol, atterrissage...) (hauteur de 0 à 15°)
- description de la trajectoire (nulle puis lente, nulle puis rapide) et estimation de la vitesse (variable).

• Ou encore introduisent des modalités mutuellement exclusives :
- luminosité (principale et secondaire),
- trajectoire (principale et secondaire),

ainsi la linéarité de la trajectoire apparaît dans une variable (principale ou secondaire) mais évidemment jamais dans les deux.

• Des modalités d'une même variable ne sont pas homogènes : estimation de la taille (métrique, angulaire, par comparaison).

• Certaines sont pléthoriques et d'autres vides :
- Trajectoire : ligne droite ou courbe très ample - immobile ou ligne droite avec arrêts (70 %)
  objet pénétrant ou sortant de l'eau (1 cas)
  atterrissage puis décollage immédiat (2 cas)

2.2. - CODAGE RÉVISÉ

Les problèmes évoqués ci-dessus imposent de reconsidérer le codage mais comme il serait trop coûteux de recoder tous les procès-verbaux, on se propose d'en limiter les conséquences en adaptant au mieux la structure des variables aux analyses à réaliser. Ceci nécessite donc le regroupement de certaines modalités et, pour ce faire, la connaissance des répartitions des observations dans les différentes variables.

Les histogrammes représentés en Annexe 2 permettent ainsi d'éliminer les modalités trop peu fréquentes et on obtient finalement le codage ci-dessous qui sera utilisé lors des analyses factorielles décrites aux paragraphes suivants.

VARIABLES	MODALITES	%	MODALITES	%
Mois d'observation	Janvier	7,08	Juillet	7,37
	Février	11,36	Août	9,88
	Mars	11,80	Septembre	9,58
	Avril	5,60	Octobre	11,06
	Mai	3,98	Novembre	7,82
	Juin	7,37	Décembre	7,08
Estimation de l'heure	Inconnue	1,18	Crépuscule	4,57
	Matin	7,23	Début nuit	4,13
	Vers midi	8,85	Fin nuit	15,04
	Après-midi	6,93	Vers minuit	25,66
	Soirée	8,55	Aurore	17,85
Région	Inconnue, Avion, Etranger	2,66	Ouest	14,64
			Nord	19,97
	Sud-Ouest	14,31	Est	22,46
	Sud-Est	16,14	Centre	9,82
Nature du lieu (nombre de témoins potentiels)	Inconnue, autre	0,45	Bourgade, Banlieu	21,39
	Désert, haute montagne, mer	7,37	Ville	14,01
	Habitation isolée	9,44	Très grande ville	6,64
	Hameau petit village	40,71		
Nombre de témoins	Un	26,07	Trois	16,30
	Deux	29,19	Quatre et +	28,44
Catégorie socio-professionnelle du témoin principal	Inconnue	2,37	Employé	8,15
	Agriculteur et salarié agricole	10,82	Personnel de service	3,41
	Patron ind. et comm.	7,42	Ouvrier	17,77
	Prof. Libérale, cadre supérieur	10,82	autre (armée, police)	7,41
	Cadre moyen	11,56	non-actif	20,27
Tranche d'âge	Inconnue	2,22	De 20 à 60 ans	71,64
	Moins de 13 ans	1,62	Plus de 60 ans	13,15
	De 14 à 20 ans	11,37		
Sexe	Masculin	78,91	Féminin	21,09
Conditions météorologiques	Inconnues	46,17	Ciel bas sans pluie	5,45
	Ciel limpide	29,79	Pluie, neige	2,80
	Nuages épars	7,67	Autre	1,18
	Couvert en altitude	6,93		
Durée de l'observation	Inconnue	19,76	de 1 à 19 mn	37,78
	Moins de 10 s	9,30	de 20 à 59 mn	10,03
	de 10 à 59 s	14,16	1 heure et plus	9,00

Estimation de la distance	Inconnue	48,59	En kilomètres	8,45
	en mètres	34,04	Très grande dist.	7,98
	en hectomètres	0,94		
Méthode d'observation	Inconnue	0,59	A bord d'un bateau	0,15
	Oeil nu au sol	66,66	Voiture en mouvement	16,08
	Avec instrument	8,26	Voiture à l'arrêt	5,90
	Trace physique	2,36		
Hauteur angulaire en début	Inconnue	24,93	45/60 degrès	7,12
	0/15 degrès	15,88	60/90 degrès	7,57
	15/30 degrès	10,83	Vu d'avion	0,59
	30/45 degrès	12,46	Vu du sol	20,33
Nombre d'"objet"	Un	85,06	Trois	3,70
	Deux	8,14	Quatre et plus	3,11
Forme principale	Inconnue	5,75	Carré, parallélépipède	2,36
	Disque, soucoupe	4,87	Canotier, couronne	3,98
	Rond, circulaire, sphérique	33,78	Ponctuel	3,24
	Cigare, cylindre	13,42	Croix	0,15
	Oeuf, ovale, ovoïde	11,65	Croissant	0,15
	Conique, toupie	5,90	Autre	14,74
Estimation de la taille	Inconnue	36,87	Non ponctuelle	54,87
	Ponctuelle	8,26		
Luminosité	Inconnue	5,08	Lumineux fluorescent	27,73
	Non lumineux	11,15	Très lumineux	35,0
	Lueur faible	5,48	Eblouissant	15,56
Caractéristiques de la luminosité citée	Variable non périod.	6,50		
	Clignotant	9,17		
	Avec faisceaux	2,67		
Caractéristiques de la trajectoire citée	Linéaire	64,45		
	Virages	3,39		
	Arabesques	2,65		
	Complexe mais analys	11,79		
	Décollage	6,63		
Couleur	Inconnue	13,44	Blanc	14,62
	Rouge	7,24	Gris	1,77
	Orange	22,45	Métallique	3,69
	Jaune	6,50	Plusieurs couleurs	14,77
	Vert	1,62	Couleur changeante	7,53
	Bleu	1,62	Autre	4,74

Estimation de la vitesse	Inconnue Lente, immobile Très rapide Fulgurante	17,58 41,95 16,40 4,58	Comme un avion Variable Autre	4,58 12,70 2,22
Bruit	Inconnu Silence Dans les graves	33,04 58,11 2,80	Dans les aigus Autre	3,25 2,80

REMARQUES :

• Ces résultats concernent 678 cas sur cinq années (74 à 78), aucune distinction n'a encore été faite sur les types (identifié a posteriori, non-identifié...).

• Les variables "luminosité principale et secondaire" ont été reconsidérées en une variable "intensité" et trois autres marquent la présence ou l'absence d'un caractère (variable, clignotant, faisceau).

• De même pour éviter les redondances possibles les variables "trajectoires principale et secondaire" sont remplacées par la présence ou l'absence de 5 caractéristiques.

• Les variables "estimation de la distance" et "estimation de la taille" sont très peu fiables. De plus, la taille est exprimée de façon très diverse (métrique, angulaire, qualitative, par comparaison), elles ont été simplifiées au maximum afin d'éviter une précision factice et donc une répartition ou dispersion non significative des observations dans ces classes de réponses.

• La redondance entre "forme ponctuelle" et "taille ponctuelle" difficilement évitable au niveau du codage sera écartée lors des analyses.

2.3. - CONCLUSION

Les manipulations successives de l'information ne font qu'accroître l'entropie du système étudié. Seul un traitement statistique des textes des procès-verbaux "en clair" serait susceptible de minimiser les pertes et déformations des informations transmises par les témoins mais évidemment pour un coût nettement plus élevé.

3. - LES CAS IDENTIFIES A POSTERIORI

Avant codage, chaque procès-verbal de Gendarmerie reçu au GEPAN a été expertisé par deux personnes qui classent le cas suivant qu'il peut être :

A : explicitement identifié	23 cas	3,4 %
B : probablement identifié	153 cas	22,6 %
C : manque d'information	239 cas	35,3 %
D : non-identifié	263 cas	38,8 %*

compte tenu du caractère limité de moyens mis en oeuvre, ces expertises ne peuvent être considérées comme objectives mais elles apportent une information supplémentaire ou plutôt un commentaire synthétique du procès-verbal, contrairement à la parcellisation systématique introduite par le codage.

3.1. - RÉPARTITION

Les 176 cas qui ont pu être explicitement identifiés (tel avion, telle fusée, telle planète...) ou qui l'ont été probablement (qui se comporte comme un avion, etc.) conduisent à des explications très variées :

- Confusion avec des corps astronomiques :
  Lune, Soleil, Vénus ou autre planète, étoiles, satellites, rentrées de satellites, de météorites ;

- Confusion avec des phénomènes atmosphériques naturels :
  Nuages lenticulaires, réflexion sur des nuages, traînées d'avion, gros grèlons ;

- Ou encore des artéfacts :
  ballons sondes, avions, hélicoptères, fusées (météo, missiles, Tiber, éclairante ou d'artifice) cerf-volant, engin téléguidé, pétard, montgolfière, dirigeable publicitaire, incendie, effet couronne, phares...

---

*Ce ne sont pas les 20 % traditionnement cités car le fichier considéré ici ne prend en compte qu'un témoin jugé "principal" par cas et il ne contient pas les cas de certaines rentrées de satellites ou météorites qui ont donné lieu à de très nombreux procès-berbaux.

- Ou enfin des associations de ces trois types : planète et avion, soleil et fusée...

On remarque que parmi cette longue liste très éclectique certains phénomènes sont rares et peuvent surprendre un observateur non-averti (effet couronne sur une ligne électrique) mais que d'autres beaucoup plus banals, (lune, soleil, avion...), ont dû être observés dans des conditions particulières (lune à l'horizon en partie cachée par des nuages...) (Voir croquis page suivante).

Il est raisonnable de penser que cette liste est loin d'être close et cette hétérogénéité montre bien la nécessité de déterminer une typologie des témoignages pour définir des classes présentant un minimum de cohérence.

3.2. - DESCRIPTION SOMMAIRE

En croisant la variable supplémentaire "identification" avec les autres un certain nombre de celles-ci apparaissent comme plus discriminante :

- mois d'observation
- durée
- estimation de la taille
- luminosité
- vitesse

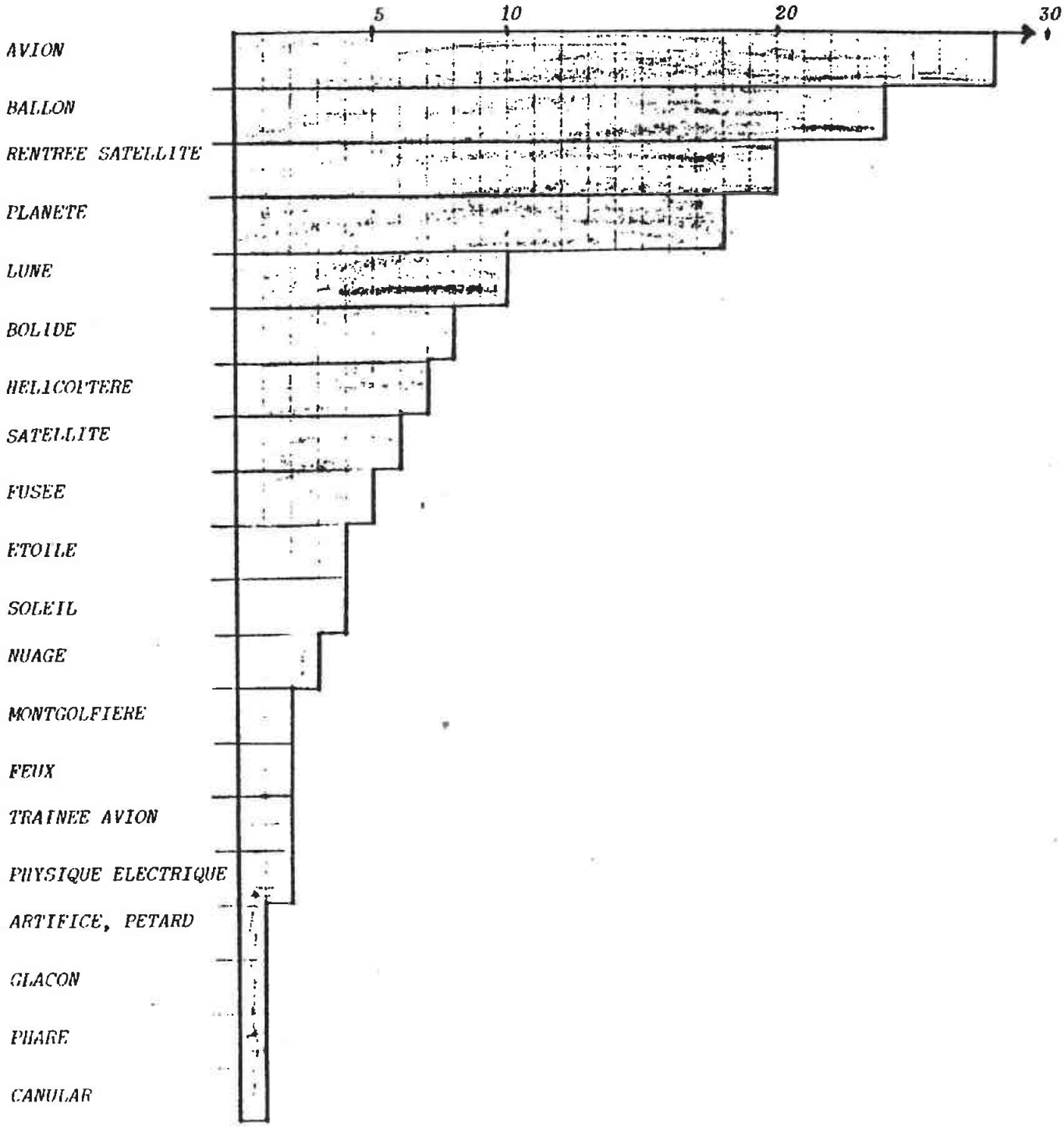
c'est-à-dire que ce sont ces variables qui permettent le mieux de distinguer les phénomènes les uns des autres et donc de les identifier.

Les autres variables sont soit équiréparties suivant les types d'identification, soit trop "creuses" (information non disponible) ou encore trop monolithiques (la majorité des cas prennent la même modalité), par exemple :

trajectoire : linéaire - immobile
estimation de l'heure : la nuit

Le caractère discriminant de ces variables s'explique en fait par un certain nombre de rapprochements banals entre des modalités de ces variables et l'identification la plus souvent associée :

VITESSE : rapide ou fulgurante pour les fusées et météorites.


5
10
20
30
AVION
BALLON
RENTREE SATELLITE
PLANETE
LUNE
BOLIDE
HELICOPTERE
SATELLITE
FUSEE
ETOILE
SOLEIL
NUAGE
MONTGOLFIERE
FEUX
TRAINEE AVION
PHYSIQUE ELECTRIQUE
ARTIFICE, PETARD
GLACON
PHARE
CANULAR

DUREE : très brève, toujours pour les rentrées atmosphériques, très longue pour les planètes et étoiles.

TAILLE : ponctuelle pour les étoiles et satellites.

LUMINOSITE : les artéfacts sont le plus souvent non lumineux tandis que le soleil est plus lumineux que Vénus elle-même plus lumineuse que les autres planètes.

MOIS D'OBSERVATION :

27 % des confusions avec les avions ont lieu en septembre

37 % des confusions avec des ballons sondes ont lieu en juin

24 % des confusions avec les rentrées atmosphériques ont lieu en novembre

38 % des confusions avec Vénus ont lieu en décembre.

3.3. - REPRÉSENTATIONS FACTORIELLES

Afin d'avoir une meilleure idée d'ensemble en tenant compte des liaisons entre les variables prises deux à deux, une analyse des correspondances multiple* a été réalisée et dont le résultat est illustré par les planches 1 et 2 ci-après.

(a) Sur la première planche ne sont représentées que les modalités des variables "actives" qui ont participé au calcul de l'analyse. (recherche des axes factoriels). Ces variables ont été choisies parmi les plus discriminantes observées au paragraphe précédent (le mois d'observation peu interprétable a été écarté) et en y adjoignant deux variables (distance et hauteur angulaire) dont l'importance a été remarquée lors des travaux précédents sur les cas D.

Le premier axe factoriel représente essentiellement la variable "durée de l'observation", c'est-à-dire celle qui permet le mieux de distinguer entre eux les différents types d'observation. L'axe 1 est orienté dans le sens décroissant des durées. La variable "estimation de la vitesse" est également liée à cet axe.

---

*Cette méthode d'analyse est sommairement décrite par BESSE -80. Se reporter également à la bibliographie citée.

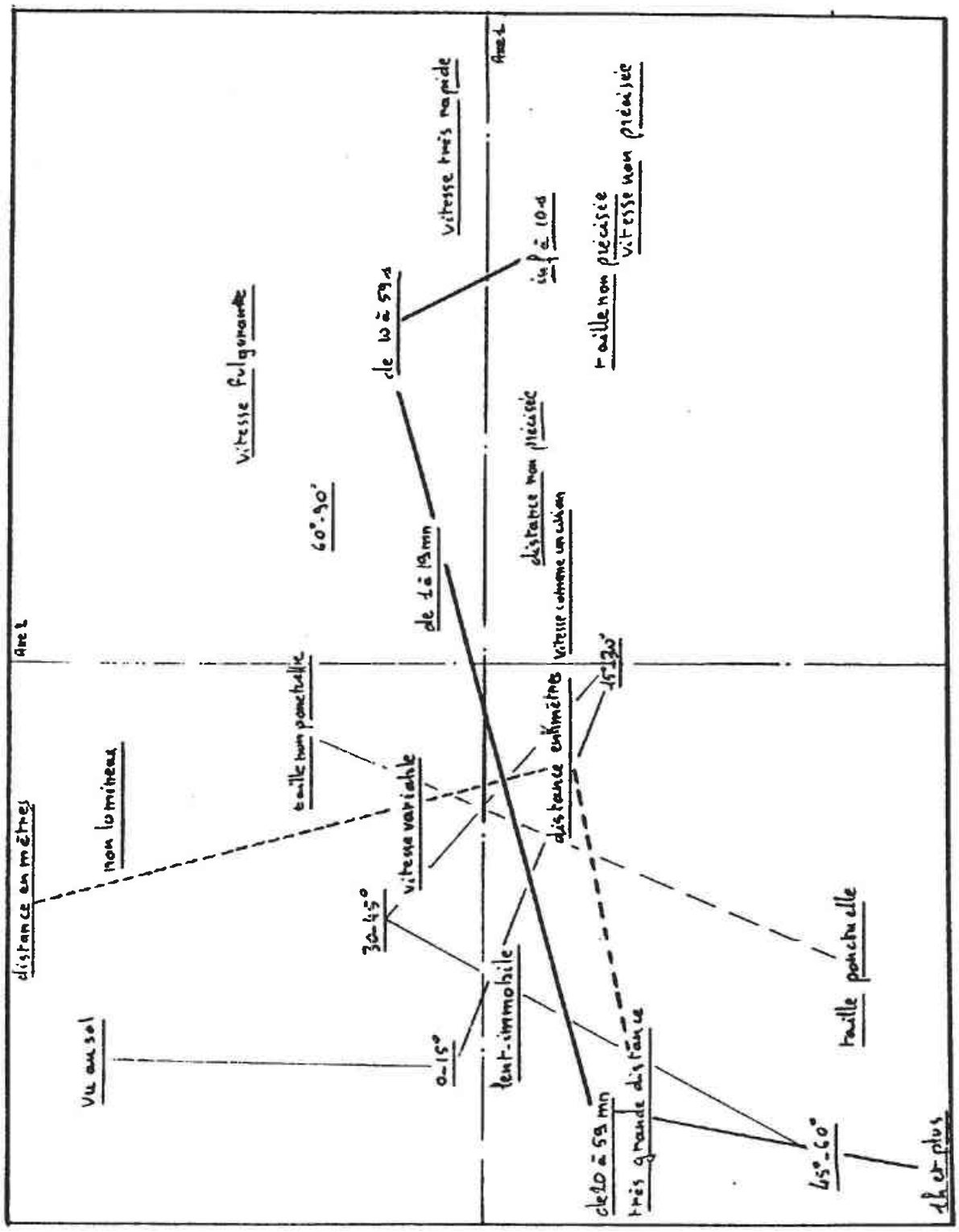

distance en mètres
non lumineux
vu au sol
0-15°
30-45°
taille non ponctuelle
vitesse variable
tout-immobile
Vitesse fulgurante
60°-90°
de 1 à 19 mn
de 10 à 59 mn
très grande distance
45°-60°
taille ponctuelle
de 10 à 59 mn
distance non précisée
Vitesse comme un avion
distance estimée
taille non précisée
Vitesse non précisée
Vitesse très rapide
Axe L
Axe L

Planche 1

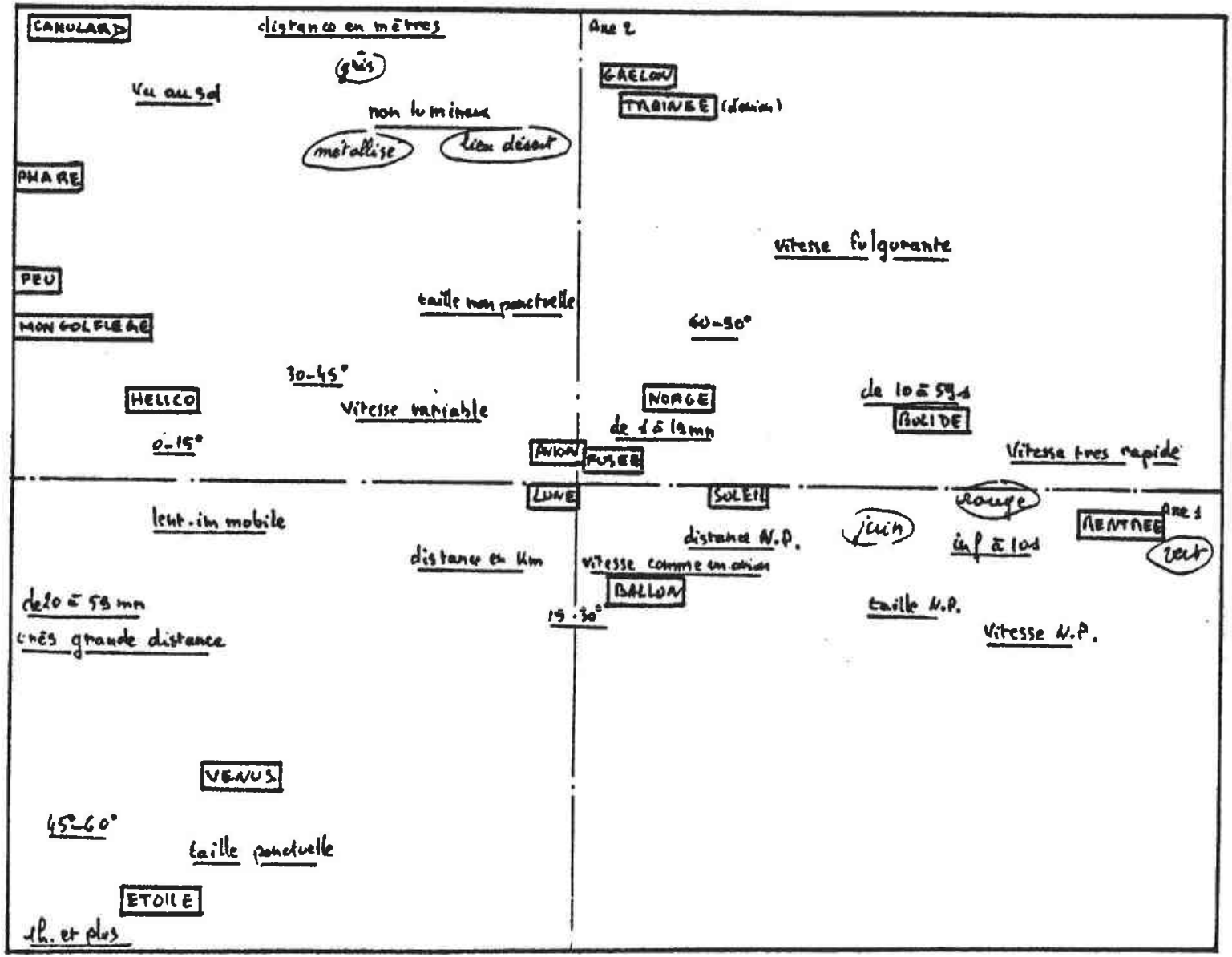

CANULARD
distance en mètres
gris
Vu au sol
non lumineux
métallisé
bien désert
PHARE
GAELOU
TRAINEE
(devant)
Vitesse fulgurante
PEU
taille non ponctuelle
MONGOLFIERE
40-90°
30-45°
HELICO
Vitesse variable
NUAGE
de 10 à 59 s
BOLIDE
de 1 à 19 mn
0-15°
AVION
FUSEE
Vitesse très rapide
LUNE
SOLEIL
rouge
lent. im mobile
distance N.P.
jaune
inf à 10 d
RENTREE
Axe 1
distance en km
vert
vitesse comme en avion
BALLON
de 10 à 59 mn
15-30°
taille N.P.
très grande distance
Vitesse N.P.
VENUS
45-60°
taille ponctuelle
ETOILE
1h. et plus

Planche 2

Aux vitesses estimées très rapides et fulgurantes correspondent alors les durées les plus brèves, tandis qu'à l'autre extrémité de l'axe se trouvent représentées les modalités "lent, immobile" et "vitesse variable". On note de plus que c'est dans le cas d'observations très brèves qu'il manque le plus d'informations.

Le deuxième axe illustre ou résume la liaison entre les variables "estimation de la distance" et "hauteur angulaire", liaison déjà soulignée précédemment (Cf. BESSE - 80) et à laquelle participe également la variable "estimation de la taille". On retrouve donc que, grossièrement, la distance estimée par le témoin croît avec la hauteur angulaire de l'observation à une nuance près : pour les sites les plus élevés (60° à 90°), la distance est le plus souvent non précisée. Un regroupement sommaire des modalités donne alors :

- distance en mètres - vu au sol (h. a. = 0°) - taille non ponctuelle ;
- distance en kilomètres - h. a. de 0° à 45° ;
- très grande distance - h. a. de 45° à 60° - taille ponctuelle

(b) Sur la deuxième planche ont été rajoutées les modalités les plus significatives (i.e. les plus éloignées de l'origine) des variables supplémentaires tandis que la structure de la représentation (variables actives et axes) reste inchangée.

Ceci conduit à trois groupes de modalités qui s'étirent dans trois directions principales de dispersion :

- vitesse très rapide, rouge, vert, Juin, inférieure à 1 mn, pas d'information sur la taille, la vitesse ou la distance, associées aux confusions avec des bolides et autres rentrées atmosphériques ;
- lent, immobile, plus de 20 mn, très grande distance, taille ponctuelle, associées aux confusions avec Vénus ou une étoile ;
- vu au sol, distance en mètres, non lumineux, gris, métallisé, lieu désert, taille non ponctuelle, associées à l'ensemble éclectique des phénomènes les plus rarement cités (montgolfière, phare, canular, traînée d'avion...).

Et enfin, on remarque un "marais statistique" c'est-à-dire l'ensemble du reste des modalités stagnant autour de l'origine sans caractéristique déterminante. Il apparaît donc que des confusions avec certains corps (lune, fusée, avion, nuage) peuvent donner lieu à des descriptions tellement variées qu'il devient impossible de retrouver, à l'état actuel du codage, la typologie antérieure aux observations.

(c) Conclusion :

Ainsi, seules trois classes apparaissent nettement dans la typologie :

- les confusions avec des rentrées atmosphériques ;
- les confusions avec des étoiles ou des planètes ;
- un rassemblement hétéroclite de confusions avec des artéfacts.

L'information qui subsiste après témoignage et codage est insuffisante pour espérer distinguer de façon significative d'autres classes.

4. - COMPARAISON IDENTIFIE / NON-IDENTIFIE

Il est clair que cette distinction identifié / non-identifié concerne le système : phénomène - conditions d'observation - témoin - expert dans toute sa complexité ; aussi, les conclusions se limiteront à deux points particuliers :

• Caractériser les cas non-identifiés (D) par rapport aux cas identifiés a posteriori, c'est-à-dire plus simplement déterminer les spécificités les plus générales de ces cas qui ont, en définitive, conduit l'expert à ne pas les rattacher à des phénomènes connus.

• Etudier, par des méthodes d'analyses factorielles, la répartition des cas D dans la typologie sommaire et les structures élaborées au paragraphe précédent pour les cas identifiés a posteriori (A ou B). Est-ce que ces cas D se rapprochent de certaines identifications, est-ce qu'une ou plusieurs sous-classes homogènes vont se démarquer ?

4.1. - DESCRIPTION SOMMAIRE

On se propose donc de comparer les distributions des observations sur les différentes variables en fonction du type (A ou B, C et D). Ceci conduit à la représentation par les histogrammes de l'Annexe 2. Comme c'est essentiellement pour l'usage des analyses factorielles que le codage du paragraphe 2.2. a été élaboré, on conserve dans ce cas particulier le codage brut (Cf. Annexe 1) des variables qui est plus précis sans que cela ne nuise à la robustesse de l'analyse.

L'étude des histogrammes de l'Annexe 2 amène les remarques suivantes :

• Pour la plupart des variables les distributions sont relativement homogènes d'un type à l'autre et ce sont les cas D qui sont le plus documentés (i.e. où les informations non disponibles sont les moins nombreuses).

• Des différences significatives apparaissent pour les distributions de certaines variables :

- comparativement moins d'observations de cas D dans les "hameaux, petits villages" et plus dans les zones dépeuplées (habitation isolée, désert, haute montagne et utilisation moindre d'instruments (jumelles, photos...) ;

- moins d'observations de cas D de durées brèves (< 1 mn) souvent interprétées par la suite comme des rentrées atmosphériques et plus d'observations de durée moyenne ;

- la distance est beaucoup plus souvent estimée pour les cas D avec une prépondérance pour l'intervalle 20 m - 1 km ;

- absence de bruit plus marquée pour les cas D ;

- très nettement plus de hauteurs angulaires estimées nulles (vu au sol ou "près du sol") pour ces mêmes cas D ;

- nettement plus d'estimations métriques de la taille entre 2 et 10 m.

Ainsi une observation reste non-identifiée surtout si le phénomène a été *perçu par le témoin dans un cadre très "humain"* ; i.e. jugé à une distance inférieure au km d'une taille "raisonnable" de 2 à 10m, souvent proche de l'horizon et pendant une durée moyenne, suffisante pour une observation détaillée mais insuffisante pour la recherche d'indices matériels (photos avec réseau de diffraction, mesures physiques ...).

4.2. - REPRÉSENTATIONS FACTORIELLES

On reprend donc les mêmes méthodes que celles utilisées au paragraphe 3.3.. La population est cette fois l'ensemble des cas A, B et D ; les cas C ont été volontairement éliminés afin de limiter la confusion déjà importante entre les divers types de cas. Ce sont toujours les mêmes variables (durée, taille, distance, luminosité, vitesse, hauteur angulaire) codées comme au paragraphe 3.3. qui sont considérées comme variables actives.

(a) A une rotation près, on retrouve dans la troisième planche la même représentation que dans la planche 1. Celle-ci amène donc les mêmes commentaires en remarquant que, cette fois, c'est le premier axe qui prend en compte les liaisons entre les variables : estimation de la distance, de la hauteur angulaire, de la taille tandis que le deuxième axe est lié à la durée de l'observation. La présence des cas D dans l'analyse ne modifie donc en rien les "structures" des variables qui peuvent être interprétées en un certain sens (1) comme les caractéristiques des descriptions de phénomènes non-identifiés pour le témoin et que ceux-ci soient identifiés ou non a postériori. La présence des cas D ne fait qu'accroître (axe de plus grande inertie) la distinction entre, d'une part, les phénomènes jugés dans le cadre de référence du témoin (vu au sol, distance en mètres, taille non ponctuelle) et d'autre part la classe de confusions avec des phénomènes astronomiques.

(1) Ces "caractéristiques" sont à approfondire sur le plan expérimental de la psychologie de la perception à l'aide d'outils statistiques décisionnels et non plus seulement descriptifs.

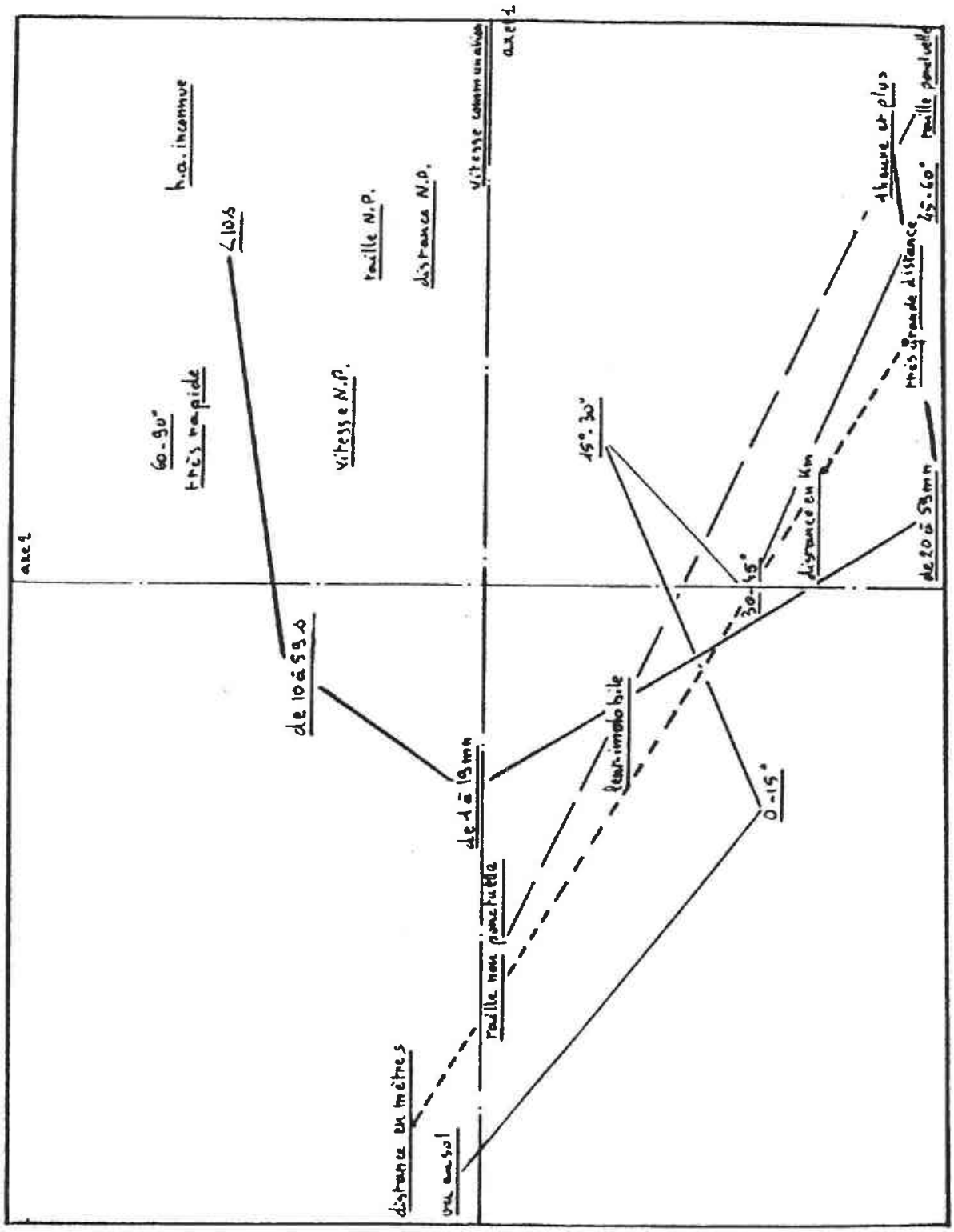

axe 1
axe 2
h.a. inconnue
60-90°
très rapide
<10 b
de 10 à 59 b
de 1 à 19 mn
vitesse N.P.
taille N.P.
distance N.P.
vitesse communication
axe 2
45°-30°
30-15°
15-0°
distance en Km
très grande distance
de 20 à 59 mn
taille nouvelle
taille non pénétrable
pénétrable
distance en mètres
axe 1

Planche 3

(b) Le calcul analogue à celui effectué pour la plance 2 et concernant les autres modalités n'appelle pas de remarques complémentaires.

4.3. - REPRÉSENTATIONS DES OBSERVATIONS (PLANCHE 4)

On cherche maintenant à situer les cas D par rapport aux classes des cas identifiés. Pour ceci, il suffit de reprendre le plan factoriel des planches 1 et 2 et d'y projeter les observations au lieu des modalités des variables. Les cas identifiés qui participent à l'analyse (i.e. au calcul des axes d'inertie du nuage) sont codés de 2 à M tandis que les cas non-identifiés, codés par des "1" sont considérés de poids nul et ainsi n'interviennent pas dans le représentation. Ceci permet donc de comparer les cas non identifiés non pas à la réalité des phénomènes connus mais bien à la réalité de témoignages de ces phénomènes avec toutes les déformations dont il faut bien tenir compte pour rendre les comparaisons possibles.

Parmi les cas identifiés, on retrouve bien les trois axes de dispersions conformément à ceux notés précédemment et sur lesquels est projetée la diaspora des cas non-identifiés. On remarque que les sources de confusions possibles donnent des représentations très enchevêtrées, pas du tout disjointes les unes des autres, montrant ainsi que, témoignage, expertise et codage entraînent de très grosses pertes d'informations (i.e. un accroissement de l'entropie) ; de la vingtaine de classes de la typologie initiale (avant témoignage), il n'en subsiste que trois et peu distinctes les unes des autres.

La projection des cas non-identifiés n'amène en apparence pas de classe nouvelle. Ils se répartissent selon les trois classes ou axes de dispersion déjà existants mais avec une préférence très marquée pour le quart en haut à gauche.. Ceci laisserait donc penser, qu'à l'exception des cas non-identifiés qui sont (ou se comportent comme) des confusions astronomiques, les autres sont à rattacher à l'ensemble * des "confusions hétéroclites". Ceci ne signifiant pas que tous les cas D sont des confusions de même type mais plutôt qu'ils seraient à rattacher à un rassemblement de cas particuliers en marge des confusions les plus fréquentes. Mais il faut relativiser cette dernière remarque en rappelant que la population étudiée n'est pas celle des phénomènes mais celles de comportements de témoins variés dans des conditions diverses. De plus, l'une des variables prépondérantes pour caractériser la troisième classe est l'estimation de la distance (distance en mètres) et comme le note JIMENEZ - 82 lors de l'étude de cas à témoignages multiples (rentrée de satellites avec plus de 40 observations par exemple) "ces estimations sont toujours fausses par rapport à la réalité du phénomène : il s'agit toujours de sous estimation traduisant un rapprochement subjectif très prononcé du phénomène".

---

* Il s'agit là d'un ensemble et non pas d'une classe bien définie.

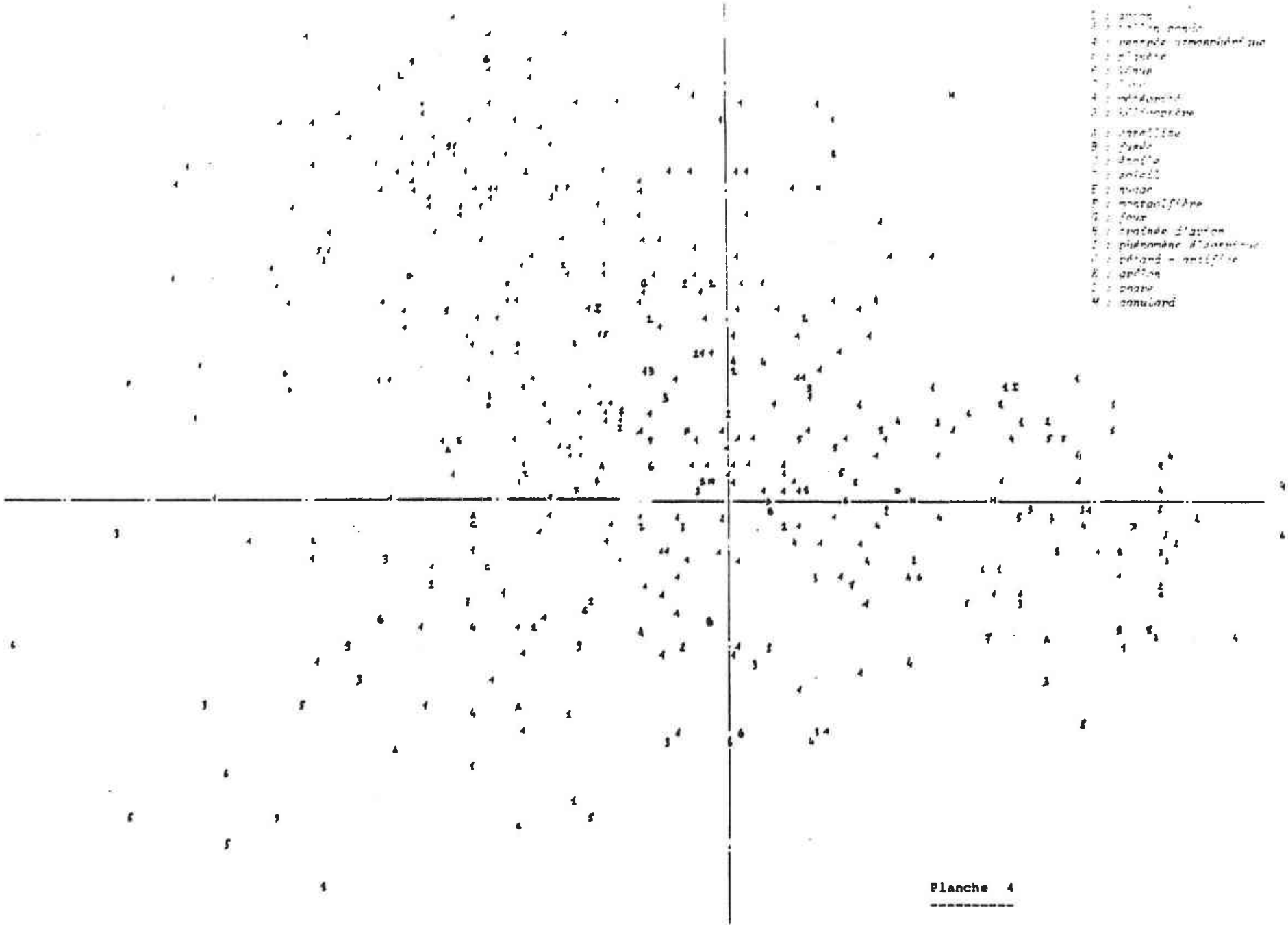


Planche 4

5. - CONCLUSION

La complexité de l'objet de l'étude alliée au manque de fiabilité des informations rendent les études statistiques très délicates. La classification obtenue au paragraphe 3 faisant apparaître principalement 3 groupes :

- confusion avec des rentrées atmosphériques
- confusion avec des étoiles ou planètes
- confusion avec des artéfacts très divers

reflète bien les diffucultés rencontrées. De la typologie initiale des phénomènes supposés à l'origine des observations, il ne subsiste que quelques classes mal distinctes concernant le comportement du système "phénomène/témoin/situation/expert" et non plus seulement les phénomènes.

En ajoutant les cas non-identifiés (§ 4), ces difficultés ne font que s'accroître car la variable "estimation de la distance", paramètre le plus subjectif, prend une place prépondérante.

En l'état actuel des choses, c'est-à-dire tant qu'il reste impossible de produire une typologie fiable, le plus prudent, à l'exception des cas relevant de confusions banales (planète, rentrée atmosphérique) est de considérer chaque cas comme un cas particulier à traiter individuellement.

6. - PERSPECTIVES

6.1. - PROBLÈMES

Les problèmes rencontrés lors de cette étude sont de deux ordres :

- Manque d'informations ;
- Subjectivité ou déformation de celles-ci.

La réponse au premier nécessiterait d'une part une action auprès de la Gendarmerie Nationale en vue de donner aux procès-verbaux une forme plus adaptée, et peut-être même une action auprès du public (i.e. auprès des témoins potentiels) et, d'autre part, un codage exhaustif de ces procès-verbaux.
Il ne s'agit donc plus de faire rentrer un procès-verbal dans un moule (le codage) mais bien de prendre en compte le maximum d'informations en respectant la forme employée par le témoin.
On est donc conduit à relever, pour chaque procès-verbal, tous les mot-clés en clair tels que les a cités le témoin* sans faire de regroupement a priori. Ce n'est qu'a posteriori et selon les besoins d'une analyse qu'il sera possible d'établir des tableaux de synonimes afin de permettre la comparaison des procès-verbaux.
Cette démarche nécessite des outils informatiques plus sophistiqués que ceux utilisés jusqu'alors et donc évidemment plus coûteux.

Une réponse théorique est proposée ci-dessous pour aborder le deuxième problème.

---

*Cette gestion des procès-verbaux est en cours de mise en place.

6.2. - MODÈLE THÉORIQUE

La situation est la suivante :

Un phénomène, dont les caractéristiques sont représentées par une variable aléatoire multidimensionnelle Y (forme, couleur, luminosité...), est observé par un témoin qui en fait une description à l'aide d'une variable de même type X. Que peut-on dire des valeurs prises par Y connaissant X ?

Concernant un phénomène isolé, il n'y a pas de solution mais si on se place au niveau d'un ensemble (une population) d'observations, ce qui importe ce ne sont plus les caractéristiques d'un phénomène mais leurs fréquences d'apparition ou encore la loi de probabilité de Y (par exemple : nombre de phénomènes rouges ou probabilité qu'un phénomène soit rouge).

Dans ce cas, connaissant (ou sachant estimer) la loi de probabilité de X (par exemple : probabilité pour que le témoignage relate un phénomène rouge) et la loi conditionnelle de Y à X (par exemple : probabilité que Y soit rouge sachant que X le décrit vert ou bleu ou rouge...) alors la loi de Y peut être estimée.

Ainsi, dans ce cadre théorique idéal, les analyses factorielles calculées comme aux paragraphes 3 et 4 et les classifications ne concerneront plus les témoignages mais décriront bien la population des phénomènes étudiés (pour plus de détails, cf. BESSE-VIDAL - 82 ou encore l'annexe 3).

En pratique, l'estimation de la loi conditionnelle de Y à X pose de gros problèmes et nécessite de nombreuses expériences avant d'être opérationnelle surtout lorsque, comme c'est le cas pour les variables hauteur angulaire, distance, taille, les erreurs d'estimations ne sont pas indépendantes. (Cf. Annexe 3). Mais, au minimum, cette démarche permet de pondérer les variables en fonction de leur fiabilité. Dans le cas de la distance, par exemple, la probabilité qu'un phénomène soit en réalité très éloigné (distance astronomique) alors qu'il a été estimé proche (quelques centaines de mètres) est loin d'être négligeable (cf. les cas de rentrée de satellite in JIMENEZ - 82). La dispersion artificielle introduite par cette variable sera alors très sensiblement atténuée comparativement aux variables qui, à l'expérience, se montreront plus fiables.

6.3. - STRATÉGIE

L'approche proposée nécessité alors trois étapes :

- estimation de la loi conditionnelle par des expériences en laboratoire et à l'aide des cas d'observations multiples où, d'une part, les caractéristiques réelles du phénomène sont connues, et, d'autre part, de nombreux témoignages permettent de faire des estimations ;

- test ou qualification de ces estimations en les appliquant à l'étude des cas identifiés a posteriori. Si les résultats obtenus sont insuffisants (mauvaise classification par exemple), il faut affiner l'étape précédente sinon :

- application à l'étude des cas non-identifiés.

6.4. - CONCLUSION

Il est clair que les outils statistiques classiques ne sont guère adaptés à l'étude de phénomènes rares et non reproductibles pour laquelle ils n'ont pas été conçus.

L'approche proposée dans ce paragraphe devrait permettre de remédier à certains des problèmes rencontrés mais celle-ci sera évidemment longue et coûteuse ; c'est le prix à payer si l'on veut espérer limiter l'accroissement de l'entropie observé tout au long du cheminement de l'information.

REFERENCES

BESSE Ph. - 1980
Etude comparative de résultats statistiques élémentaires
CNES/GEPAN
Note Technique n° 2
Avril 1980

BESSE Ph. - 1981
Recherche statistique d'une typologie des descriptions de phénomènes aérospatiaux non-identifiés
CNES/GEPAN
Note Technique n° 4
Mars 1981

BESSE Ph., ESTERLE A., JIMENEZ M. - 1981
Eléments d'une méthodologie de recherche
CNES/GEPAN
Note Technique n° 3
Avril 1981

DUVAL P. - 1979
Règles de codage (4ème version)
CNES/GEPAN
Note Technique n° 1
Octobre 1979

ESTERLE A. - 1981
Le problème des phénomènes aérospatiaux non-identifiés
CNES/GEPAN
Note Technique n° 3
Avril 1981

JIMENEZ M. - 1982
Quelques expériences en psychologie de la perception
CNES/GEPAN
(A paraître)

BESSE Ph. - VIDAL cl. - 1982
Analyse des correspondances et codage par une probabilité de transition.
Statistique et analyse des données - décembre 1982 -

ANNEXE 1

RÈGLES DE CODAGE (VERSION 4)

(cf. N. T. n° 1)

2.9. DATE - 8 cases

Date non disponible : ********

Si la date est disponible, les deux premières cases contiennent le quantième du jour, les deux suivantes le numéro du mois et les quatre dernières, l'année (ex. 1er février 1978 = 01021978). Si la date est partiellement connue, on codera les parties connues et on mettra des * pour les parties inconnues.

2.10. HEURE - 5 cases

Utilisation des 5 caractères de la façon suivante :

a) 4 premiers : Heure locale suivant le format : hhmm
Si cette heure est inconnue : ****

b) 5ème : estimation grossière de l'heure suivant le code :
M = matin
V = vers midi
P = après midi
S = soirée
C = crépuscule
D = début de la nuit
Z = vers minuit
F = fin de la nuit
L = aurore

Exemple : 20 heures 40 minutes = 2040D
Donc, en général, l'heure sera codée selon les deux modes.

2.11. DEPARTEMENT - 2 cases

On mettra le numéro du département dans lequel se trouve le lieu de l'observation. Exemple : Haute-garonne = 31. Si l'information n'est pas disponible, on mettra **

Si l'observation a été faite hors de la métropole, dans les territoires d'outre mer, on mettra deux zéros (00). Pour l'étranger, on mettra deux point (..).

2.12. LONGITUDE - 4 cases

Il s'agit de la longitude mesurée en degrés et fractions décimales (au dixième de degré près), à partir de Greenwich , positivement vers l'Est.

Exemple :
+ 2 degrés et 15 mn = +022
- 2 degrés et 25 centièmes de degré = -022 ou 3578

S'il n'y a pas d'information de longitude disponible, on mettra quatre * (****). NOTA : en France métropolitaine, les longitudes sont comprises entre - 9 et + 9 degrés. La longitude de Paris est de 2,33 degrés.

.../...

2.13. LATITUDE - 4 cases

Il s'agit de la latitude mesurée en degrés et centièmes de degré (au dixième de degré près) positivement vers le Nord. Exemple : 49 degrés et 20 minutes = + 493. S'il n'y a pas d'information de latitude, on mettra : **** .

2.14. NATURE DU LIEU - 1 case

Témoins potentiels :

* ..... : inconnue
0 .....D ..... : désert, haute montagne, mer
0 à 10 .....S ..... : habitation isolée
10 à 100 .....H ..... : hameau, petit village
100 à 1000.....B ..... : bourgade, banlieue
1000 à 10 000 ....V ..... : ville
10000..... A ..... : vue d'avion
. : connue mais n'entre pas dans les rubriques précédentes

2.15. NOMBRE DE TEMOINS - 1 case

Code	*	1	2	3	4	5	6	7	8	9
Nbre témoins	inconnu	1	2	3	4	5	dizaine	centaine	millier	+ millier

2.16. PROFESSION DU TEMOIN PRINCIPAL - 2 cases

Voir annexe 2 - Si la profession est inconnue, on mettra : **

2.17. AGE DU TEMOIN PRINCIPAL - 3 cases

L'âge du témoin étant en général connu, il semble intéressant de le coder sur 3 caractères afin de limiter l'effet de l'arbitraire de la classification.

2 premiers caractères : âge en clair
3ème caractère : classe (suivant le code actuel suivant) :

* : âge inconnu
E : enfant de 0 à 13 ans
J : adolescent de 14 à 20 ans
A : adulte de 21 à 59 ans
C : vieillard de 60 ans et plus.

2.18. SEXE DU TEMOIN - 1 case

On utilisera les numéros correspondants suivants :
1 : masculin
2 : féminin

2.19. CONDITIONS METEOROLOGIQUES - 1 case

* : pas d'indication
1 : très beau temps, ciel pur
2 : nuages épars
3 : ciel couvert mais à haute altitude
4 : ciel bas, mauvais temps sans pluie
5 : pluie, grêle, neige, orage, faible visibilité
6 : 1 avec vent
7 : 2 avec vent
8 : 3 avec vent
9 : 4 avec vent
A : 5 avec vent
. : autres conditions

2.20. DUREE DE L'OBSERVATION - 5 cases

Ce critère sera codé en clair et de la façon suivante :
- les 4 premières cases pour la durée en chiffre
- la 5ème case pour la classe (S : secondes, M : minutes, H : heures).

Toutes les cases sont remplies :
* : pas d'indication de durée
S : de 0 à 59 secondes
M : de 1 à 1440 minutes (soit de 1 mn à 24 h)
H : supérieure à 24 heures
. : autres cas

Exemple : 15 secondes..... 0015S
2 heures 45 minutes..... 0165M

2.21. DISTANCE MINIMALE D'OBSERVATION - 4 cases

Ce critère sera codé en clair et de la façon suivante :
- les 3 premières cases pour la distance chiffrée
- La 4ème pour la classe (M-mètres, K-kilomètres, A > 3 km).

* : pas d'indication
M : de 0 à 999 mètres
K : de 1 à 3 kilomètres
A : supérieure à 3 kilomètres

Exemple : 55 mètres 055M
1,5 km 1.5K

2.22. METHODE D'OBSERVATION - 1 case

* : pas d'indication
A : oeil nu au sol
B : jumelles, longue vue, théodolite
C : lunette astronomique
D : télescope
E : photographie ou film
F : radar
G : jumelles + photo
H : visuel + radar
J : oeil nu à partir d'un avion

.../...

K : oeil nu à partir d'un bateau
L : jumelles à partir d'un bateau
M : télescope + photo
N : à bord d'un véhicule automobile en marche
P : à bord d'un véhicule automobile à l'arrêt
. : autres méthodes

2.23. NOMBRE D'OBJETS - 2 cases

** : pas d'indication
OO : aucun objet
de Oi à 98 : nombre d'objets si inférieur à 99
99 : si 99 objets ou plus

NOTA : au-delà de 10 objets, on arrondit au nombre de dizaines si le nombre n'est spécifié explicitement nulle part.

2.24. FORME DE L'OBJET PRINCIPAL - 1 case

* : pas d'indication
A : disque, soucoupe lenticulaire
B : ronde, circulaire, boule
C : cigare, cylindre, fusée
D : oeuf, ovale, ovoïde, ballon de rugby
E : conique, triangulaire, chapeau asiatique, trapézoïdale
F : toupie
G : carrée, rectangulaire, parallélépipédique
H : soucoupe à coupole, chapeau de canotier
J : couronne, pneumatique
K : ponctuelle, étoile, grosse planète
L : dôme, tasse, parachute, parapluie, meule de foin
M : méduse, champignon
N : croix
P : croissant
Q : cigare accompagné de disques
R : nuée, nuage, halo
S : nid d'abeilles
. : autres formes

2.25. TAILLE - 2 cases

1er cas : évaluation non métrique

Pour la 1ère case : C - par comparaison
A - angulaire
* - pas d'indication
. - autres types non métriques

Pour la 2ème case : A - immense, très gros
B - comme une pièce de 5F
C - comme une orange
D - comme une assiette, un melon
E - comme une citrouille
F - comme un avion
G - comme la lune
H - comme une voiture
J - comme une grosse étoile
K - petit - tout petit
. - autres comparaisons

} Par comparaison ex. comme la lune = CG

0 à 9 dizaines de minutes d'arc
. non codable (sup. à 1°30')] Angulaire ex. A6 (1°)

2ème cas : évaluation métrique
On utilise les 2 cases pour coder la plus grande dimension évaluée
01 à 98 : taille en mètres
99 : 99mètres ou plus
00 : inférieur à 1 mètre

3ème cas : pas d'indication, on utilisera : ★★

NOTA : Dans le cas où plusieurs types d'indications sont fournis, on gardera l'indication la plus crédible, cette appréciation étant laissée au codeur. De toutes façons, les deux cases doivent être remplies.

2.26. LUMINOSITE - 2 cases

Combinaison de 2 paramètres si nécessaire :
- 1ère case : paramètre paraissant le plus important,
- 2ème case : paramètre apportant un enrichissement.

Si un seul paramètre, laisser la 2ème case vierge.

★ : pas d'indication
1 : lueur, faiblement lumineux
2 : lumineux, fluorescent
3 : brillant, très lumineux
4 : intense, éblouissant, éclatant
5 : non lumineux
6 : réfléchit la lumière du soleil ou autre lumière
7 : halo seulement
8 : variable en intensité (de 0 à ∞ mais non périodique)
9 : clignotant
A : non lumineux mais avec faisceaux
. : autres types

2.27. COULEUR - 1 case

★ : pas d'indication
A : rouge sombre
B : rouge
C : orangé, feu
D : jaune, ambre
E : vert
F : bleu
G : bleu sombre, métallique, indigo
H : violet
J : blanc
K : noir
L : gris
M : métallique (argent, aluminium poli)
N : plusieurs couleurs
P : couleur(s) changeante(s)
Q : marron
R : or
. : autres couleurs

2.28. TRAJECTOIRE - 2 cases

Combinaison de 2 paramètres si nécessaire. Si un seul paramètre, laisser la 2ème case vierge.

* : pas d'indication
A : ligne droite, ou courbe très ample - immobile ou ligne droite avec arrêts
B : virages brusques
C : arabesques compliquées
D : trajectoire complexe mais analysable (périodicité, suivi de route, de fleuve, etc...)
E : stationnement près du sol
F : atterrissage et arrêt prolongé avant décollage
G : atterrissage puis décollage immédiat
H : objet vu au sol qui décolle
J : objet pénétrant ou sortant de l'eau
K : objet qui monte et se perd dans les étoiles
L : nulle puis lente
M : nulle puis rapide
. ; autres types de trajectoire

2.29. VITESSE - 4 cases

Ce critère sera codé en clair de la façon suivante :
- les 3 premières casespour la vitesse chiffrée (en centaines Km/h)
- la 4ème case, pour la classe.

* pas d'indication
A lente ou très lente ou immobile
B très rapide
C variable
D fulgurante
E vitesse d'un avion
. autres types de vitesse

Si l'information est variable, on code la plus grande suivie de la lettre C. Exemple : très rapide 1000 km/h 010B
immobile 000A
variable sans précision ..... ***C

2.30. ACCELERATION - 1 case

* : pas d'indication
1 : faible
2 : variable
3 : élevée
. : autres types

2.31. BRUIT - 1 case

* : pas d'indication
A : aucun bruit, silence total, objet silencieux
B : bourdonnement, vrombrissement, bruit d'abeilles, grondement

C : sifflement aigu
D : bruit d'air comprimé s'échappant
E : bruit de moteur électrique démarrant ou de machine centrifuge
F : explosions violentes
G : bruit de vent violent sous l'objet
H : aucun bruit perceptible
. : autres bruits

2.32. HAUTEUR ANGULAIRE - 2 cases

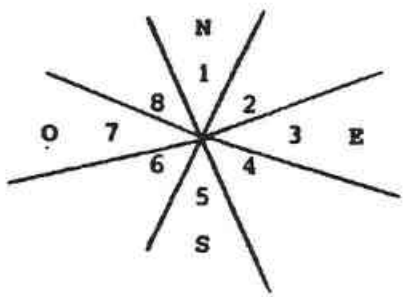
On notera la hauteur en début et en fin d'observation :
-1ère case : début
-2ème case : fin
Si un seul paramètre, laisser la 2ème case vierge.

* : pas d'indication
1 : de 0 à 15° (ou bas sur l'horizon)
2 : de 15 à 30°
3 : de 30 à 45°
4 : de 45 à 60°
5 : de 60 à 90° (proche du zénith)
6 : au-dessous de l'horizon, sous un avion
7 : observé d'un avion, à la même hauteur ou au-dessus
8 : objet vu au sol ou près du sol
. : autres types

2.33. DIRECTIONS AZIMUTALES - 2 cases

Les directions azimutales seront estimées dans les secteurs angulaires suivants : (azimut en degrés à partir du Nord, positivement vers l'Est) :

* : pas d'indication
1 : -22,5 à + 22,5 vers le nord
2 : 22,5 à 67,5 nord-est
3 : 67,5 à 112,5 est
4 : 112,5 à 157,5 sud-est
5 : 157,5 à 202,5 sud
6 : 202,5 à 247,5 sud-ouest
7 : 247,5 à 292,5 ouest
8 : 292,5 à 337,5 nord-ouest
O : à la verticale (vers le zénith)
. : pas codable



On donne ainsi la direction azimutale au début de l'observation puis en fin d'observation.

3. CONCLUSION

En régle générale, quand l'information n'est pas disponible, on code * et quand l'information est disponible, mais non compatible, avec les régles de codage prévues, on code . (point). Par conséquent, il ne peut pas en principe, y avoir de blancs dans les 79 cases exceptées les cases 66, 69 et 77.

Il s'agit donc, en général, d'une grille de signification. Les indications contenues dans les témoignages peuvent n'être identiques à aucun des cas proposés, mais se rapprocher fortement de certains. Ceci reste à l'appréciation du codeur.

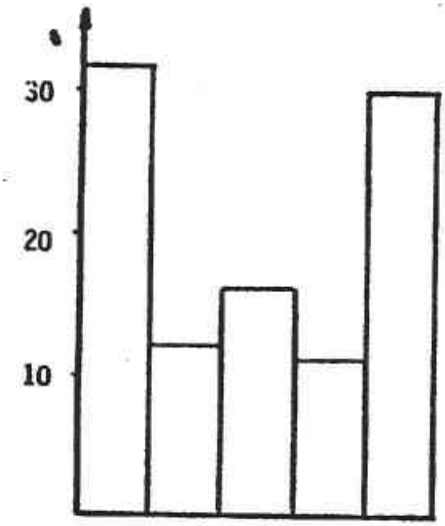
ANNEXE 2

RÉPARTITION DES OBSERVATIONS SELON LE TYPE

CAS A OU B

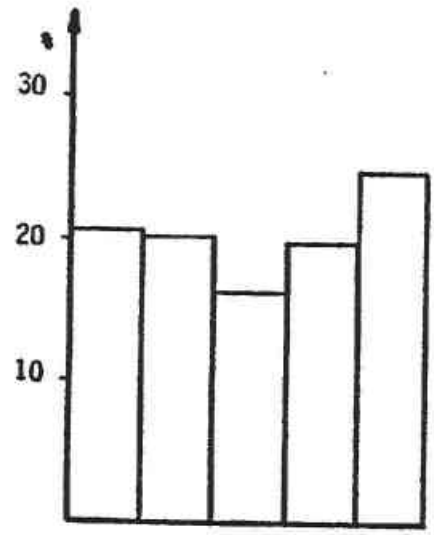
CAS D

CAS C

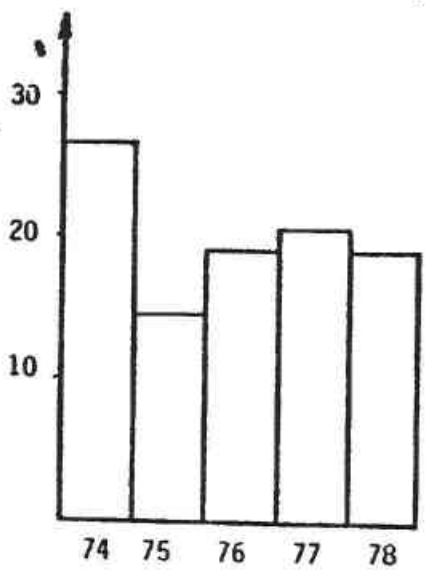


ANNEE DE L'OBSERVATION

A ou B

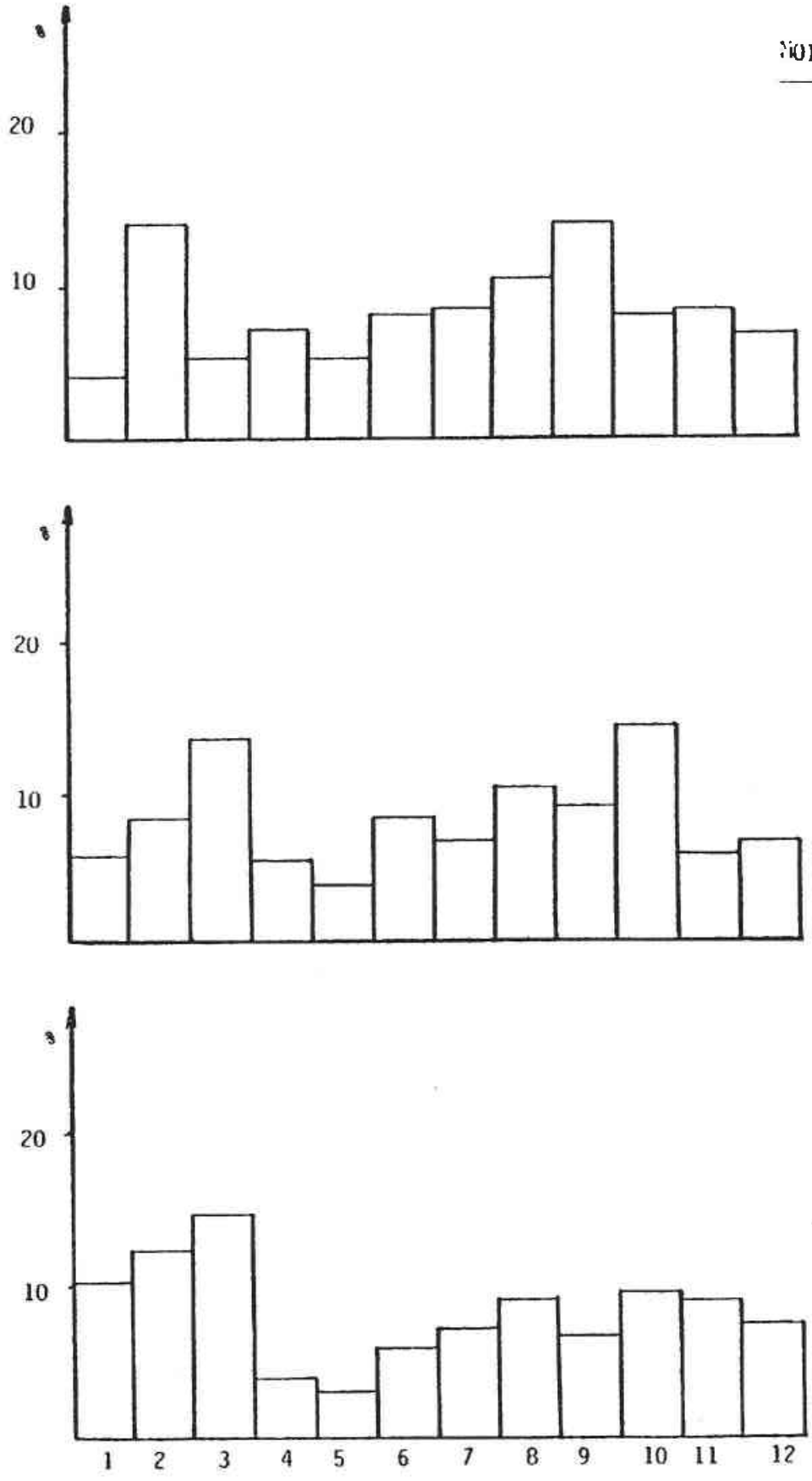


D



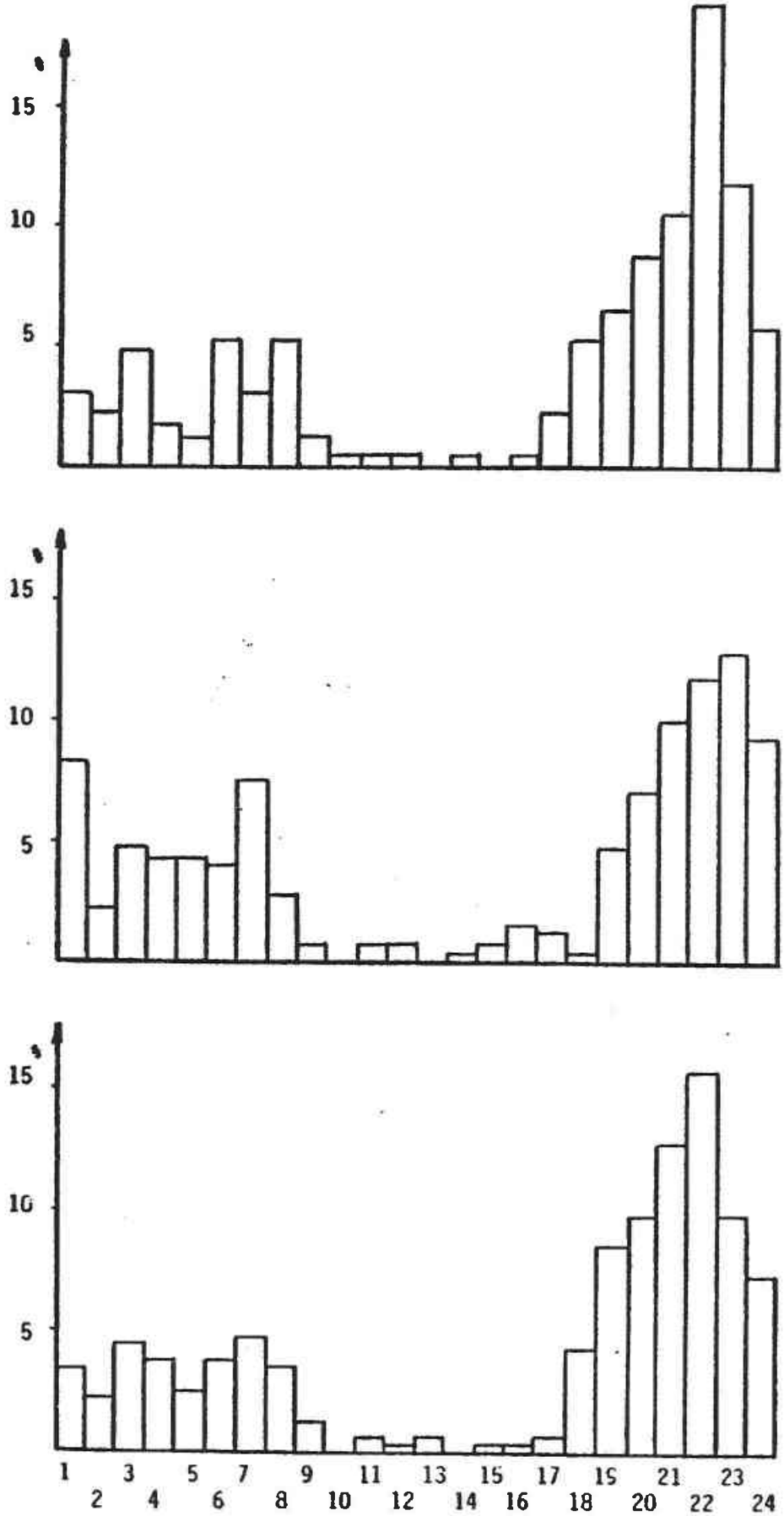
C

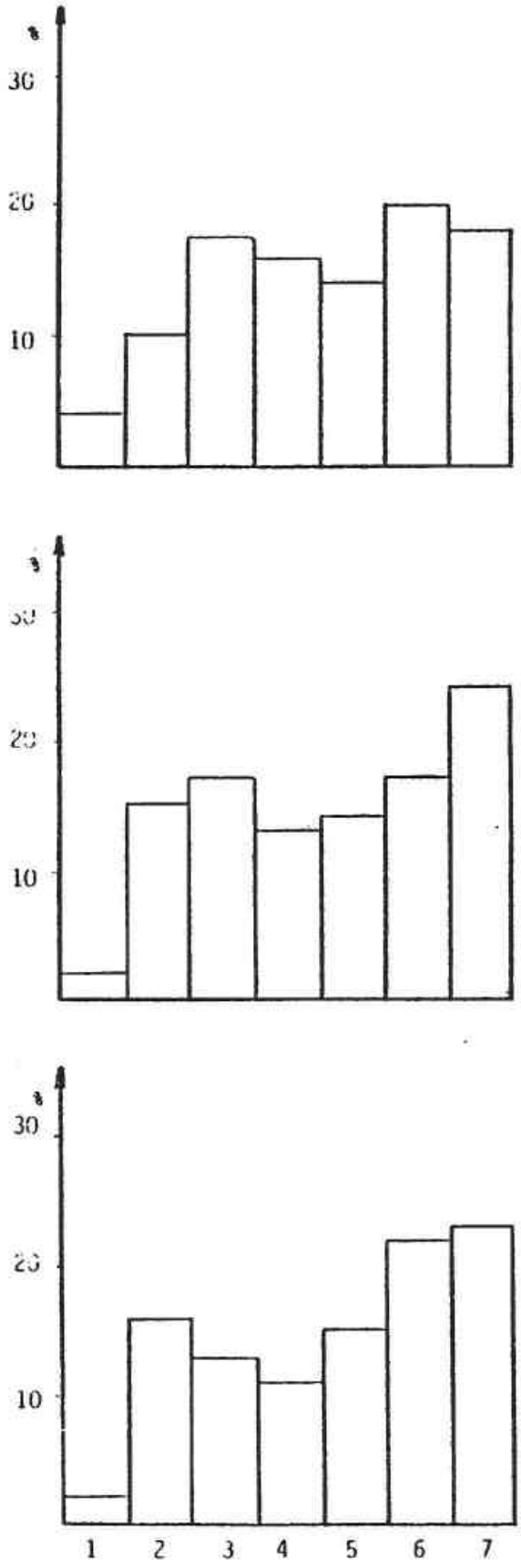
MOIS DE L'OBSERVATION



HEURE LOCALE

D'OBSERVATION

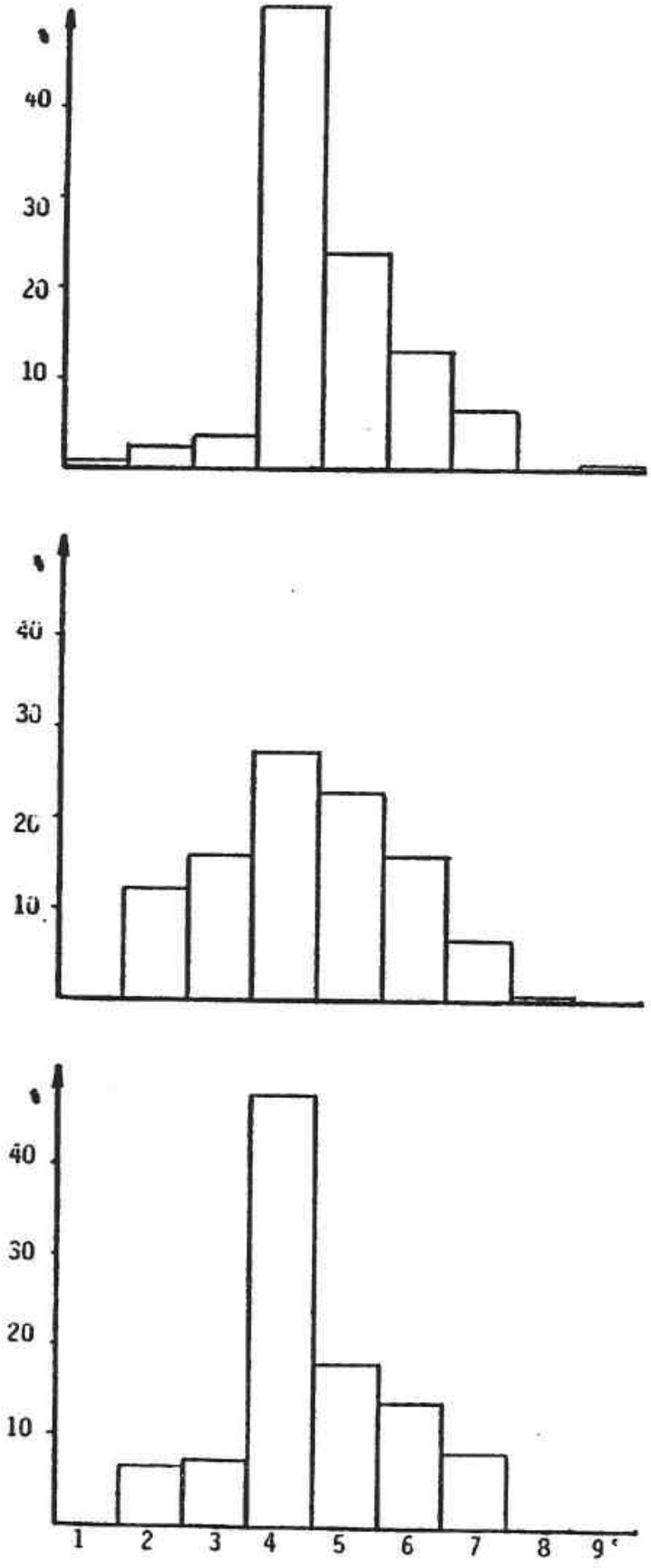




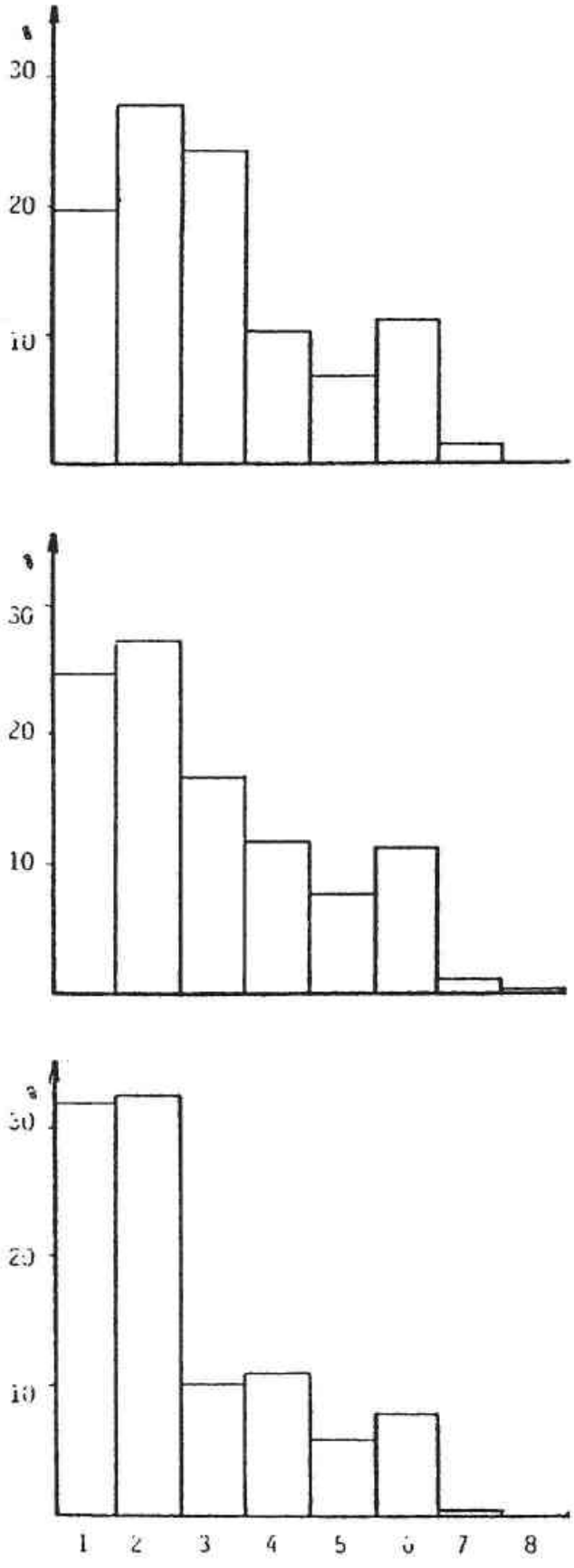
REGION D'OBSERVATION

1. DOM-TOM
2. sud-ouest
3. sud-est
4. centre
5. ouest
6. nord
7. est

NATURE DU LIEU

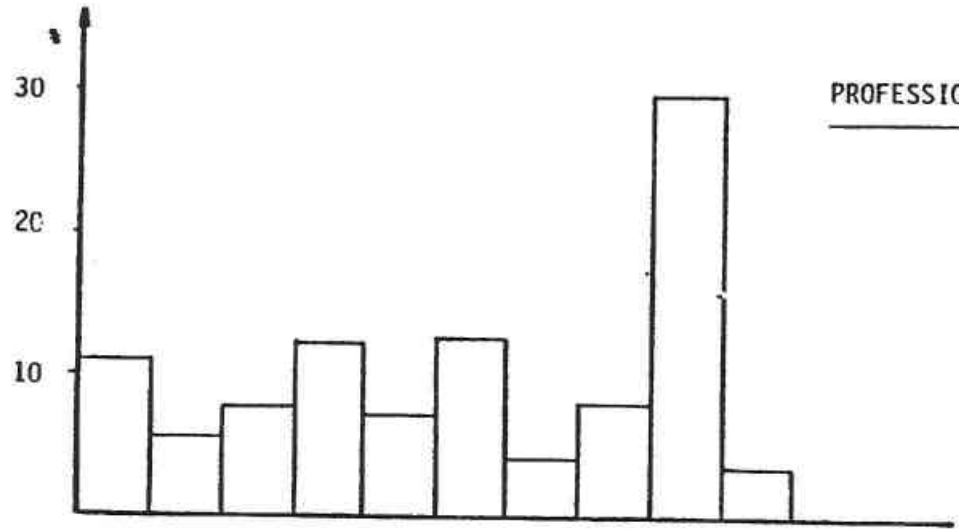


1. information non disponible
2. désert, haute montagne, mer
3. habitation isolée
4. hameau, petit village
5. bourgade, banlieue
6. ville
7. grande ville, métropole
8. vue d'avion
9. information non codable

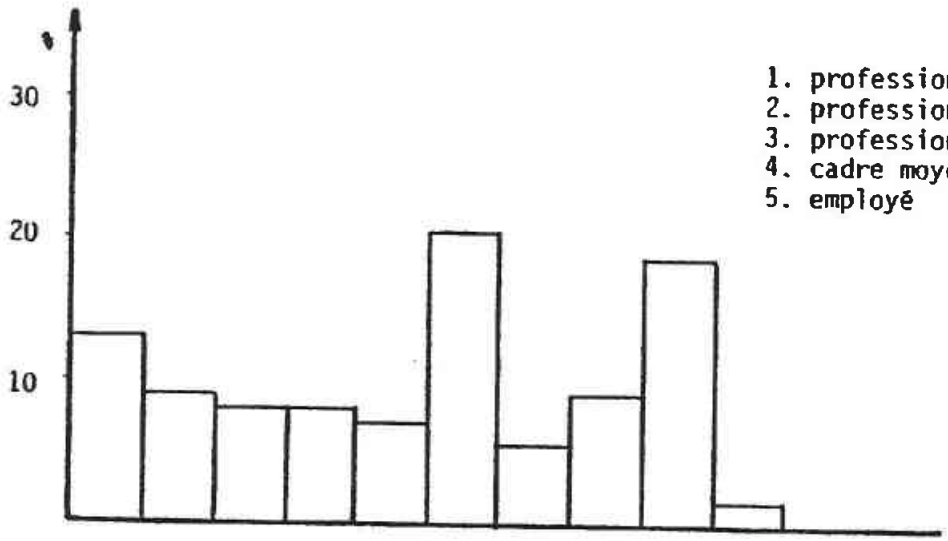


NOMBRE RELATE DE TEMOINS

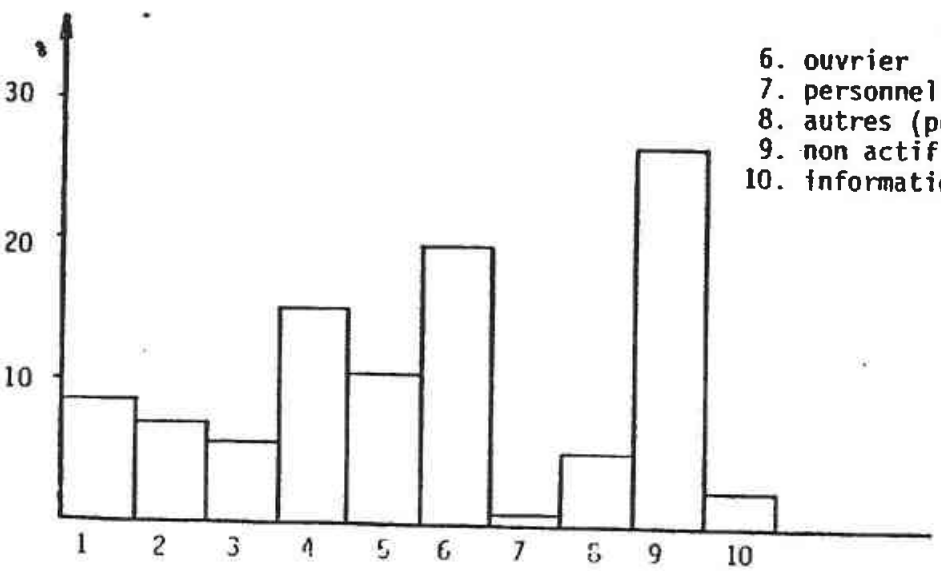
1. un témoin
2. deux témoins
3. trois témoins
4. quatre témoins
5. cinq témoins
6. dizaine
7. centaine
8. millier



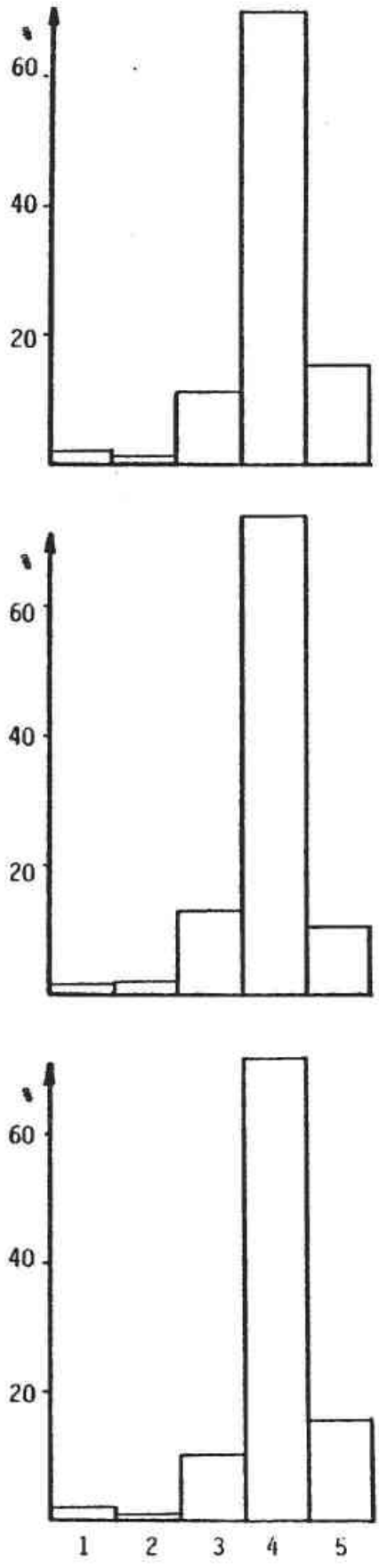
PROFESSION DU TEMOIN



1. profession agricole
2. profession de l'industrie et du commerce
3. profession libérale et cadre supérieur
4. cadre moyen
5. employé

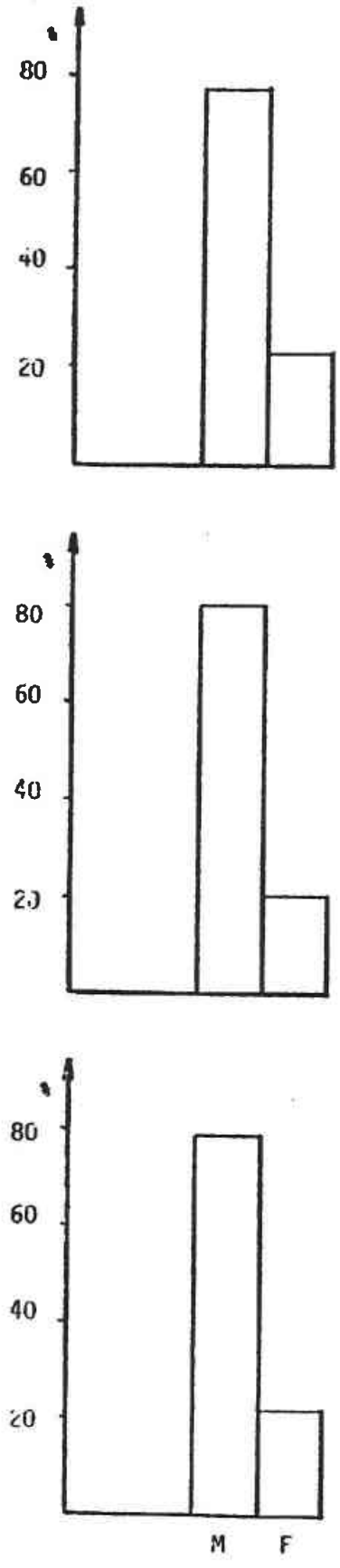


6. ouvrier
7. personnel de service
8. autres (police, armées)
9. non actif
10. information non disponible



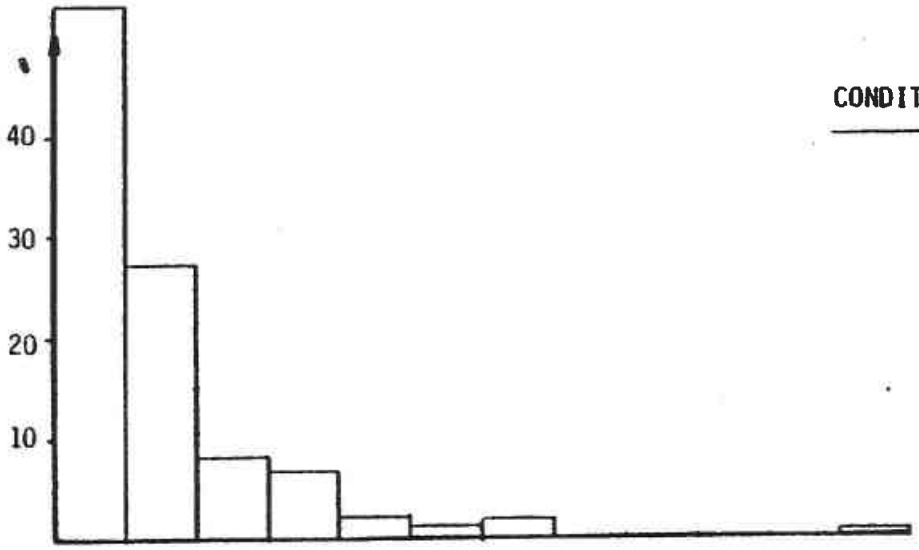
AGE DU TEMOIN

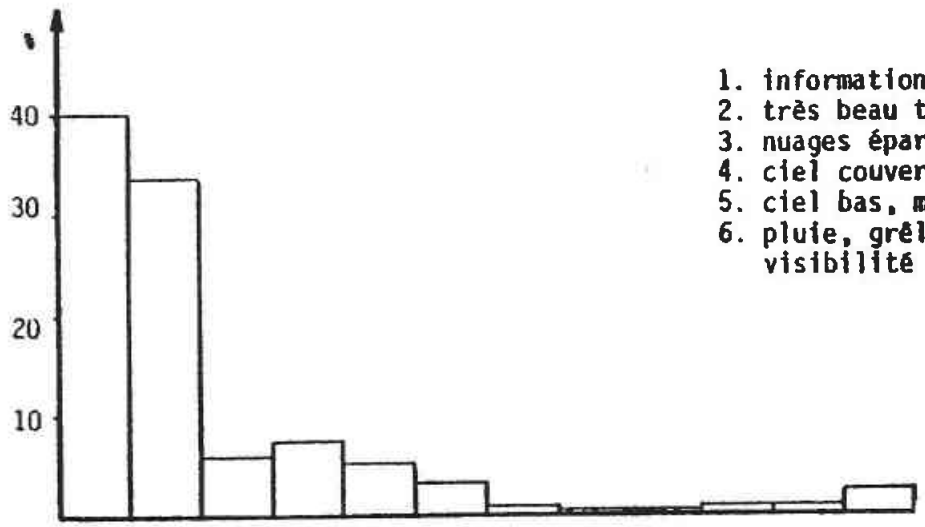
1. information non disponible
2. enfant de 0 à 13 ans
3. adolescent de 14 à 20 ans
4. adulte de 21 à 59 ans
5. 60 ans et plus



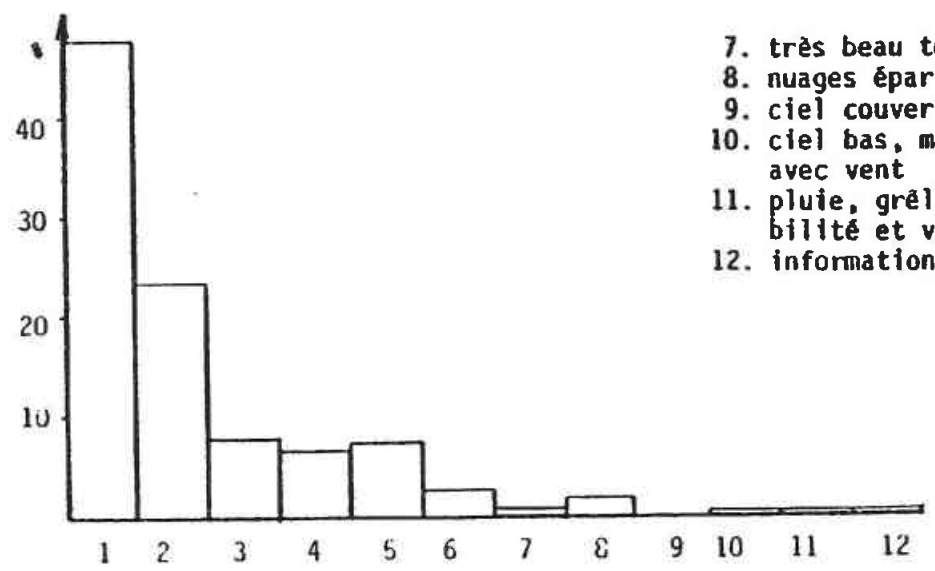
SEXE DU TEMOIN

CONDITIONS METEOROLOGIQUES



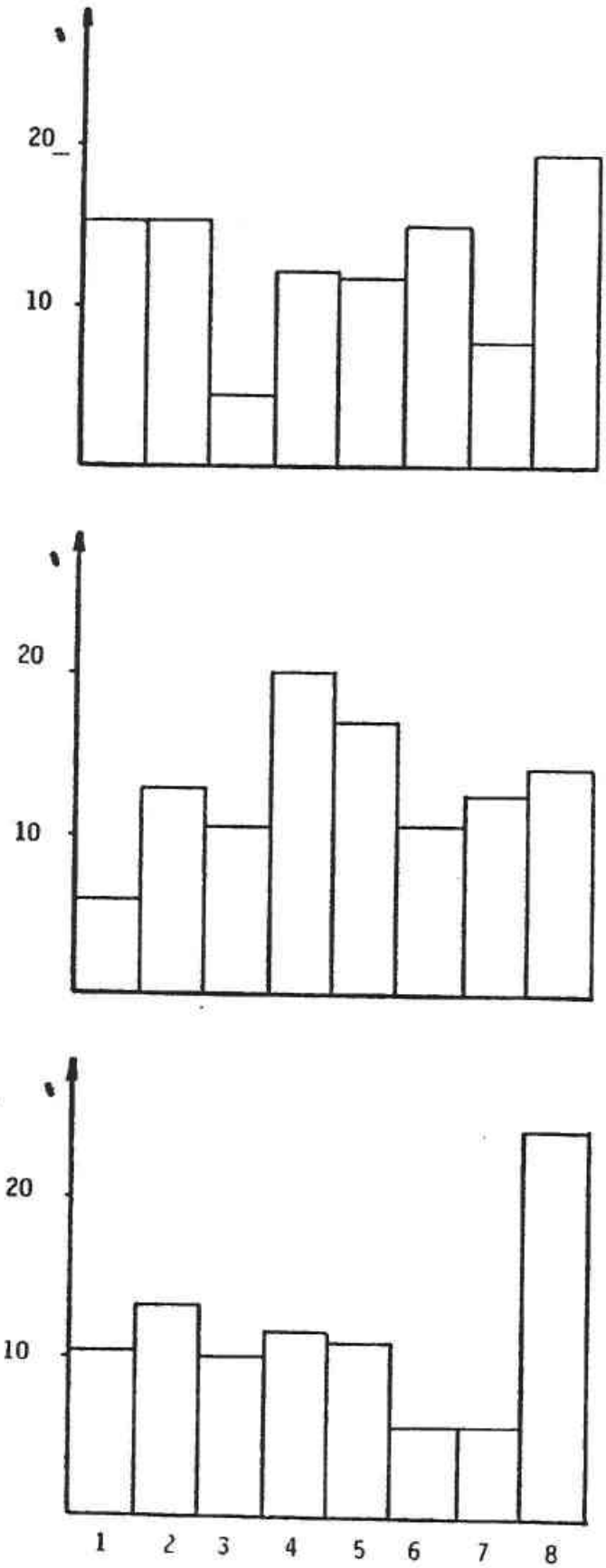


1. information non disponible
2. très beau temps, ciel pur
3. nuages épars
4. ciel couvert mais à haute altitude
5. ciel bas, mauvais temps, sans pluie
6. pluie, grêle, neige, orage, faible visibilité



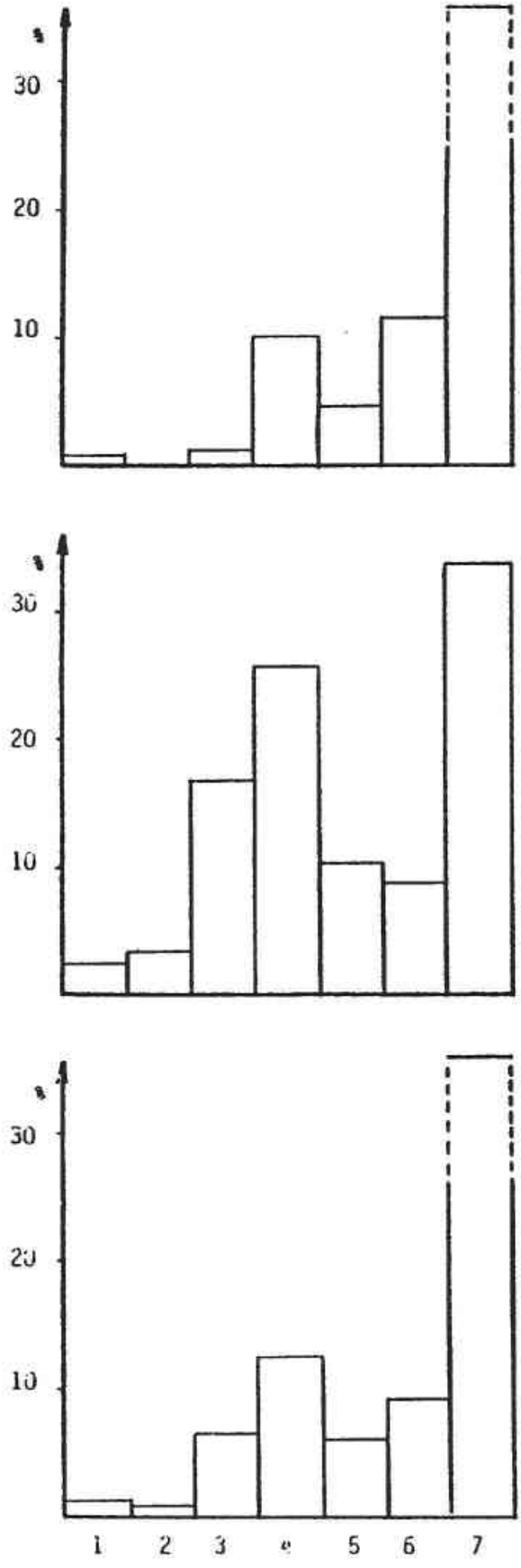
7. très beau temps, ciel pur, avec vent
8. nuages épars, avec vent
9. ciel couvert à haute altitude avec vent
10. ciel bas, mauvais temps, sans pluie, avec vent
11. pluie, grêle, neige, orage, faible visibilité et vent
12. information non codable

DUREE DE L'OBSERVATION

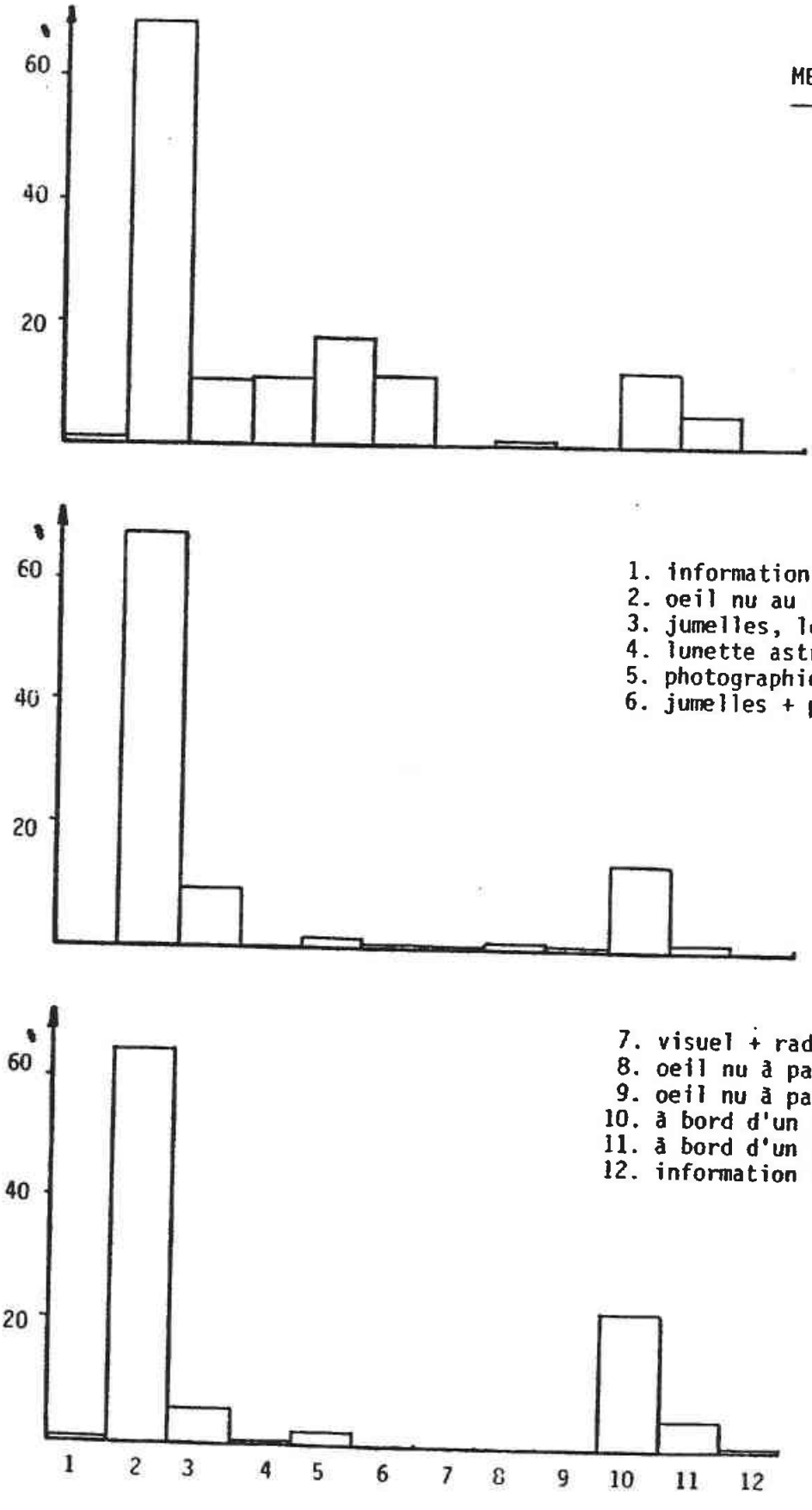


1. $<$10 s
2. $\geqslant$10 s et $\leqslant$ 59 s
3. $\geqslant$1 mn et $<$ 3 mn
4. $\geqslant$3 mn et $<$ 10 mn
5. $\geqslant$10 mn et $<$ 20 mn
6. $\geqslant$20 mn et $<$ 1 h
7. $\geqslant$ 1 h
8. non disponible

DISTANCE D'OBSERVATION

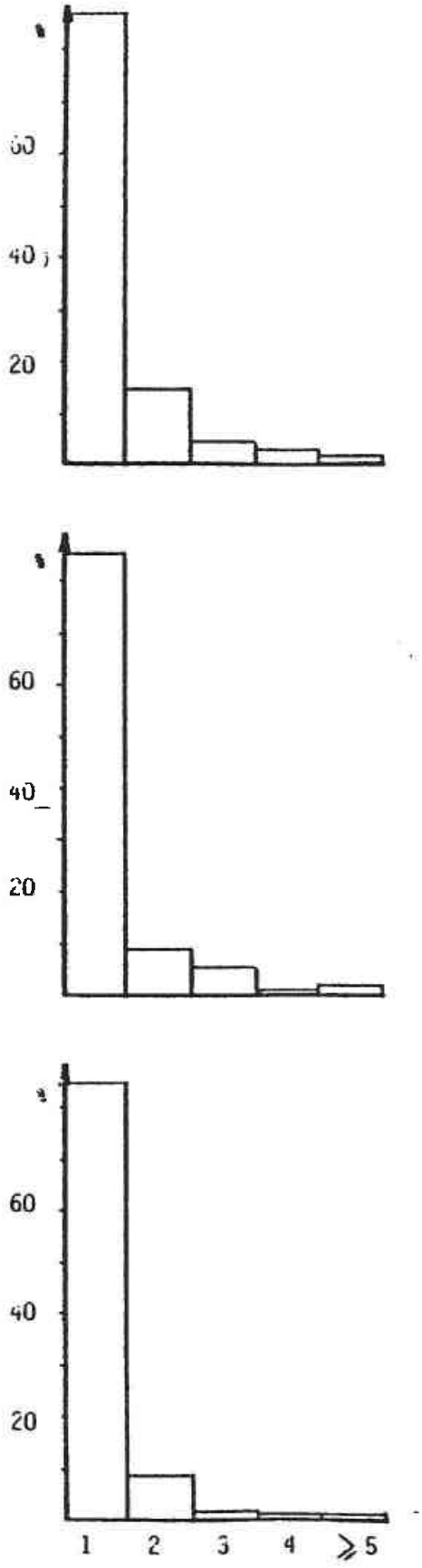


1. < 10 m
2. ⩾10 m et < 20 m
3. ⩾20 m et <150 m
4. ⩾150 m et <1 km
5. ⩾ 1 km et <3 km
6. >3 km
7. non disponible



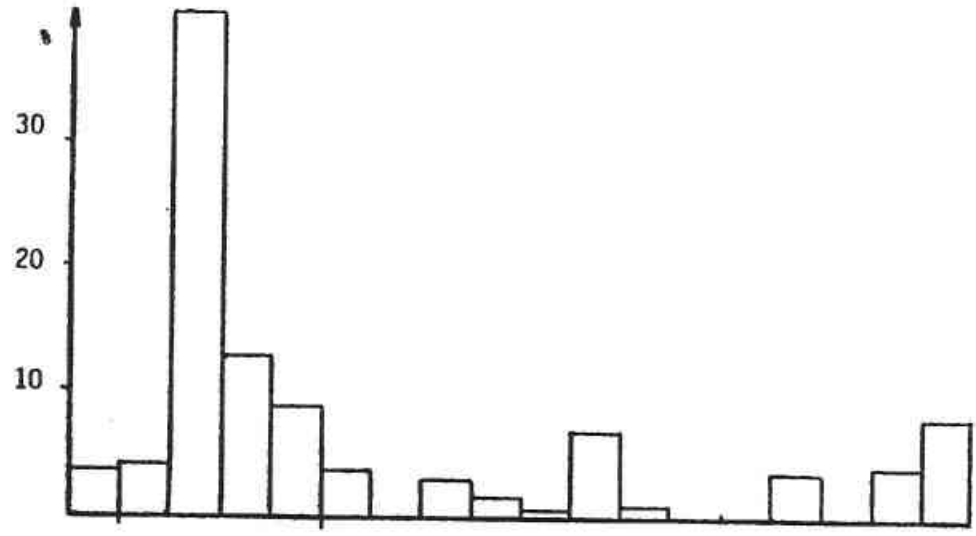
METHODE D'OBSERVATION

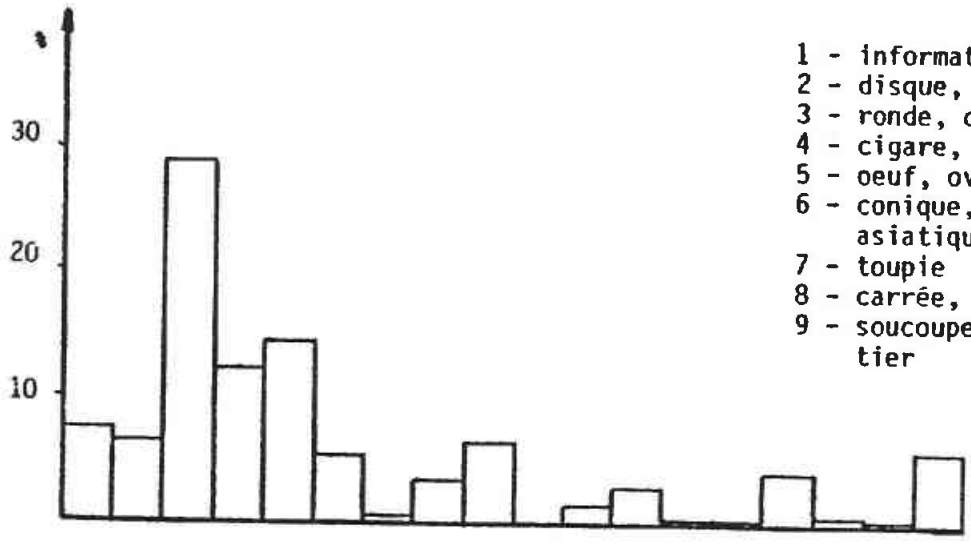
1. information non disponible
2. oeil nu au sol
3. jumelles, longue vue, théodolite
4. lunette astronomique
5. photographie ou film
6. jumelles + photo
7. visuel + radar
8. oeil nu à partir d'un avion
9. oeil nu à partir d'un bateau
10. à bord d'un véhicule automobile en marche
11. à bord d'un véhicule automobile à l'arrêt
12. information non codable



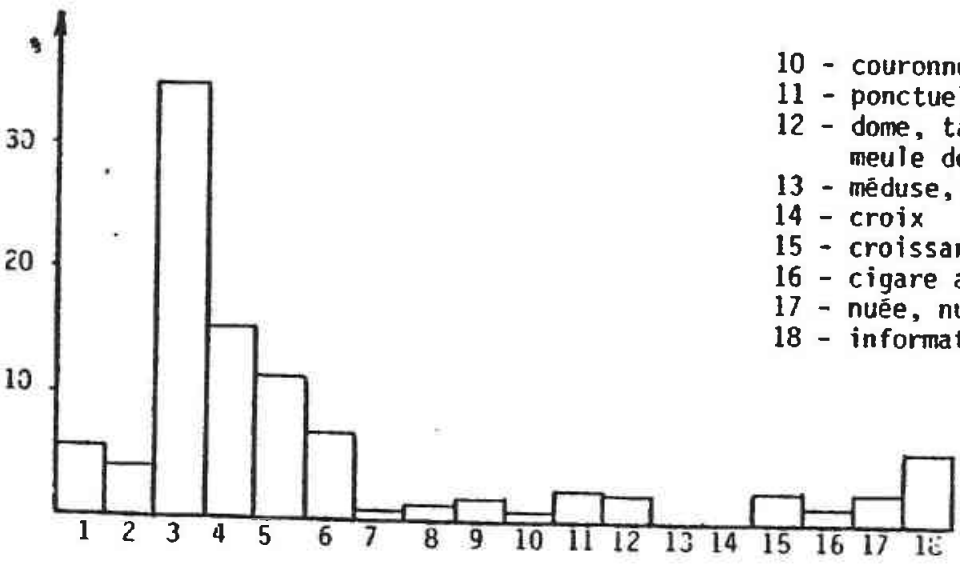
NOMBRE "D'OBJETS"

FORME PRINCIPALE



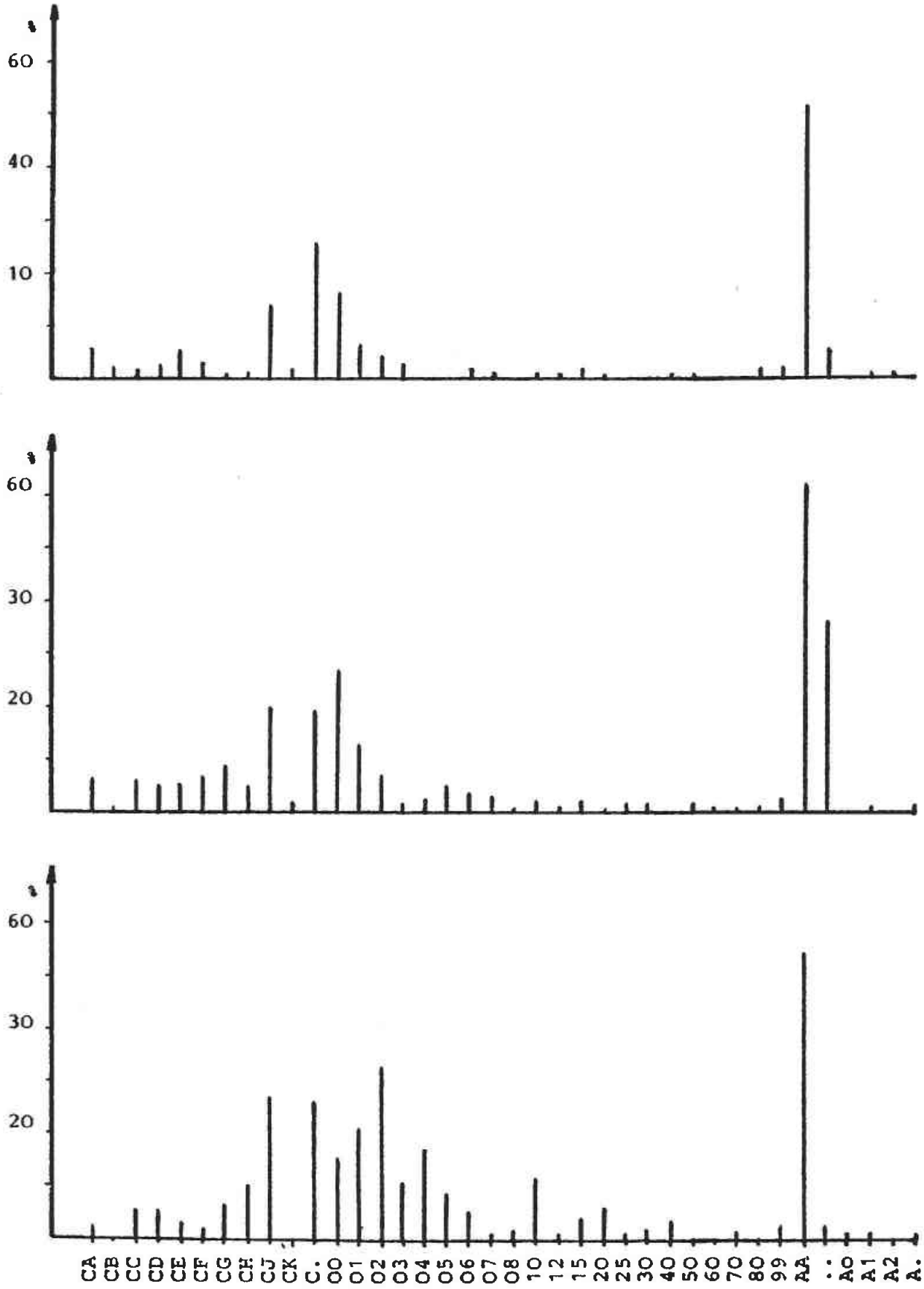


1 - information non disponible
2 - disque, soucoupe lenticulaire
3 - ronde, circulaire, boule
4 - cigare, cylindre, fusée
5 - oeuf, ovale, ovoide, ballon de rugby
6 - conique, triangulaire, chapeau asiatique
7 - toupie
8 - carrée, rectang., parallélépipédique
9 - soucoupe à coupole, chapeau de canotier

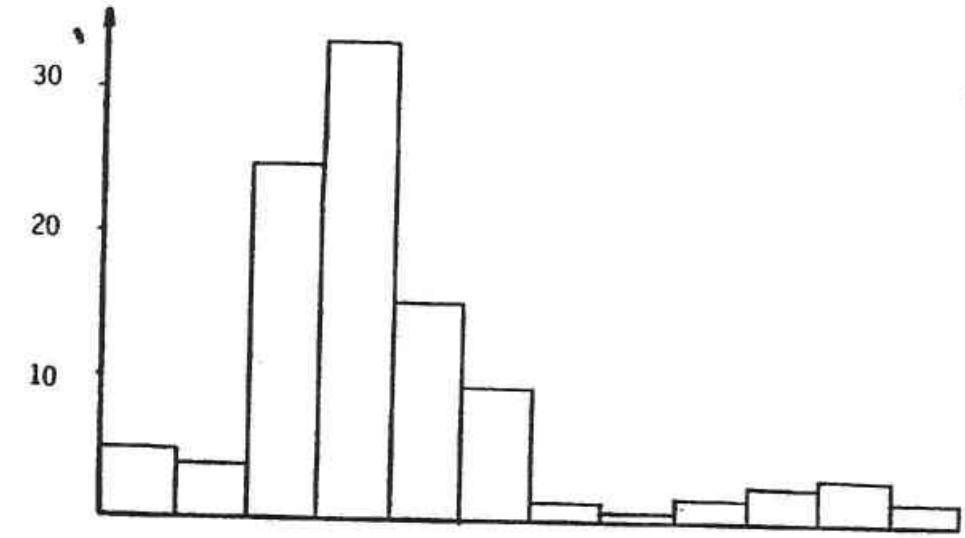


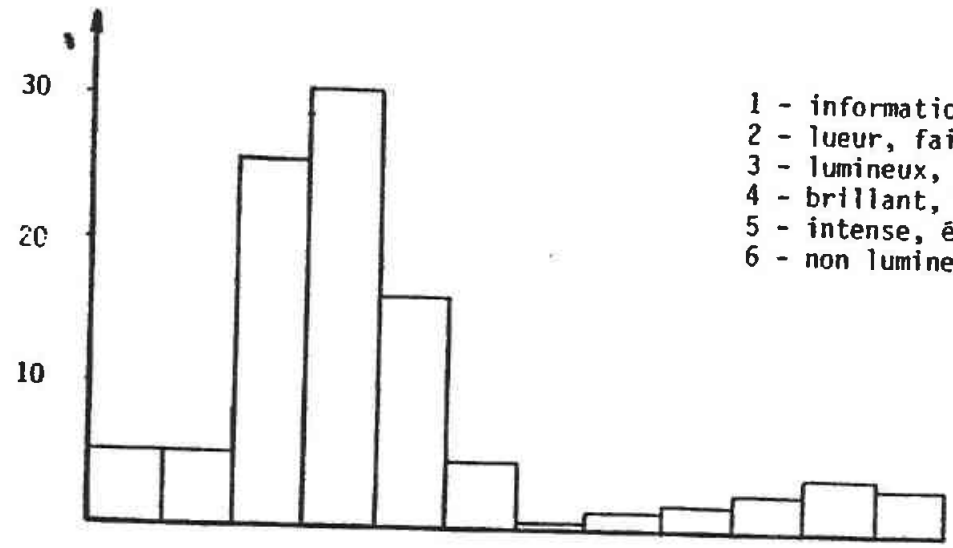
10 - couronne, pneumatique
11 - ponctuelle, étoile, grosse planète
12 - dome, tasse, parachute, parapluie, meule de foin
13 - méduse, champignon
14 - croix
15 - croissant
16 - cigare accompagné de disques
17 - nuée, nuage, halo
18 - information non codable

Estimation de la taille
(signification des codes en annexe 1)

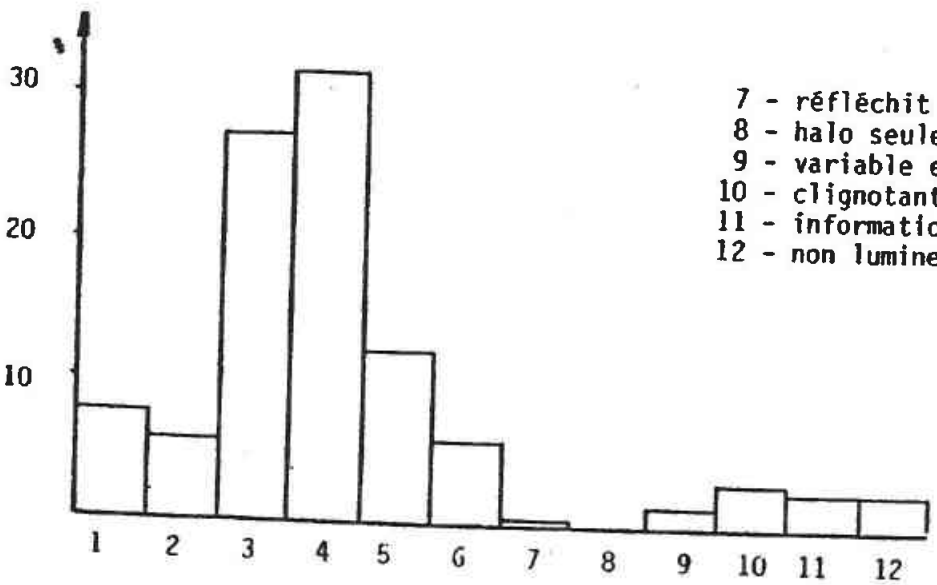


LUMINOSITE PRINCIPALE





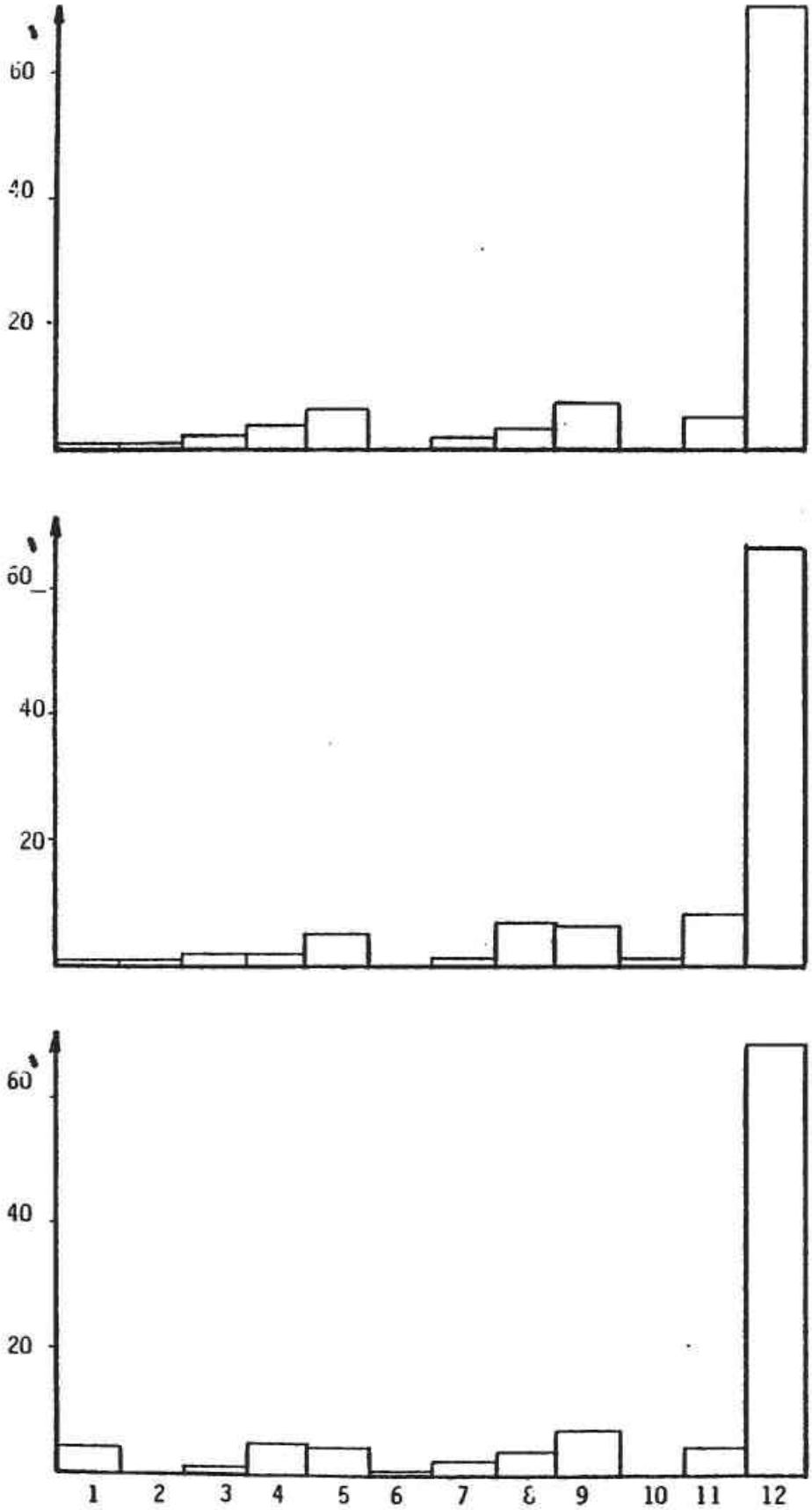
1 - information non disponible
2 - lueur, faiblement lumineux
3 - lumineux, fluorescent
4 - brillant, très lumineux
5 - intense, éblouissant, éclatant
6 - non lumineux



7 - réfléchit la lumière du soleil
8 - halo seulement
9 - variable en intensité
10 - clignotant
11 - information non codable
12 - non lumineux, mais avec faisceaux

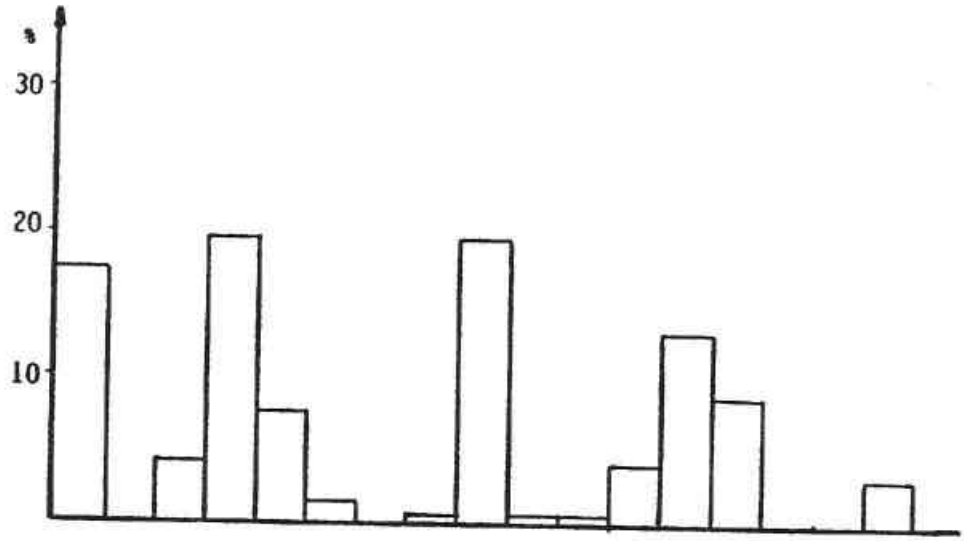
LUMINOSITE

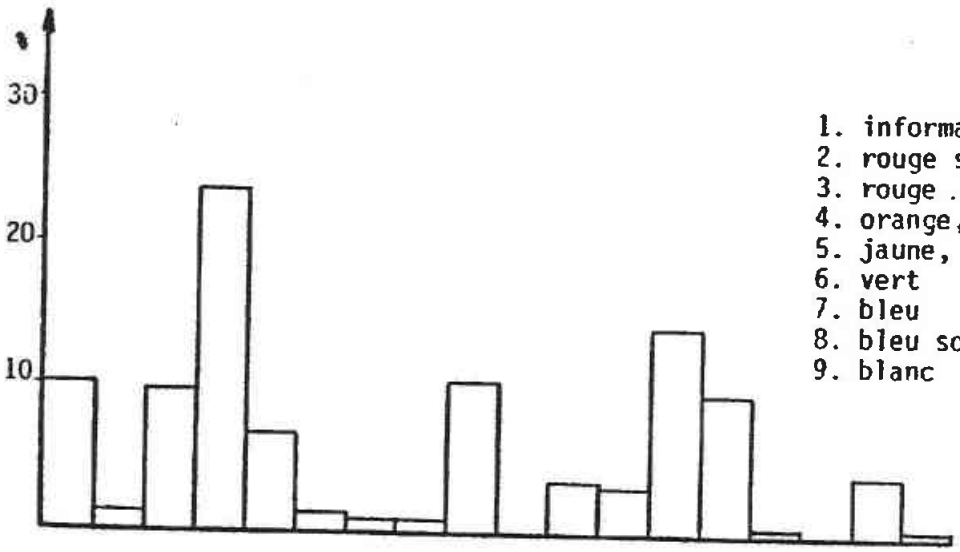
SECONDAIRE



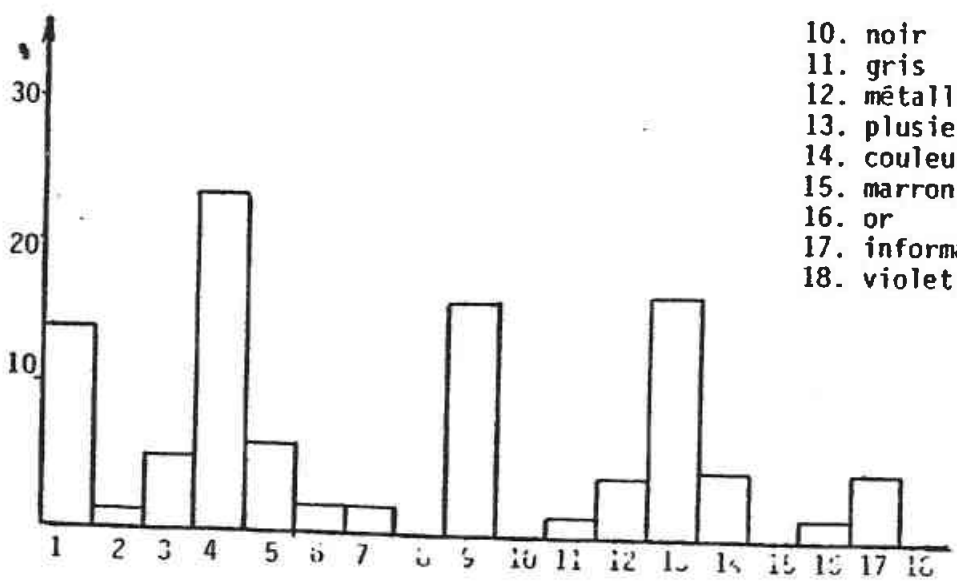
1. Information non disponible
2. Lueur, faiblement lumineux
3. lumineux, fluorescent
4. brillant, très lumineux
5. intense, éblouissant, éclatant
6. non lumineux
7. réfléchit la lumière du soleil
8. variable en intensité
9. clignotant
10. non lumineux, mais avec faisceaux
11. information non codable
12. pas d'information secondaire

COULEUR



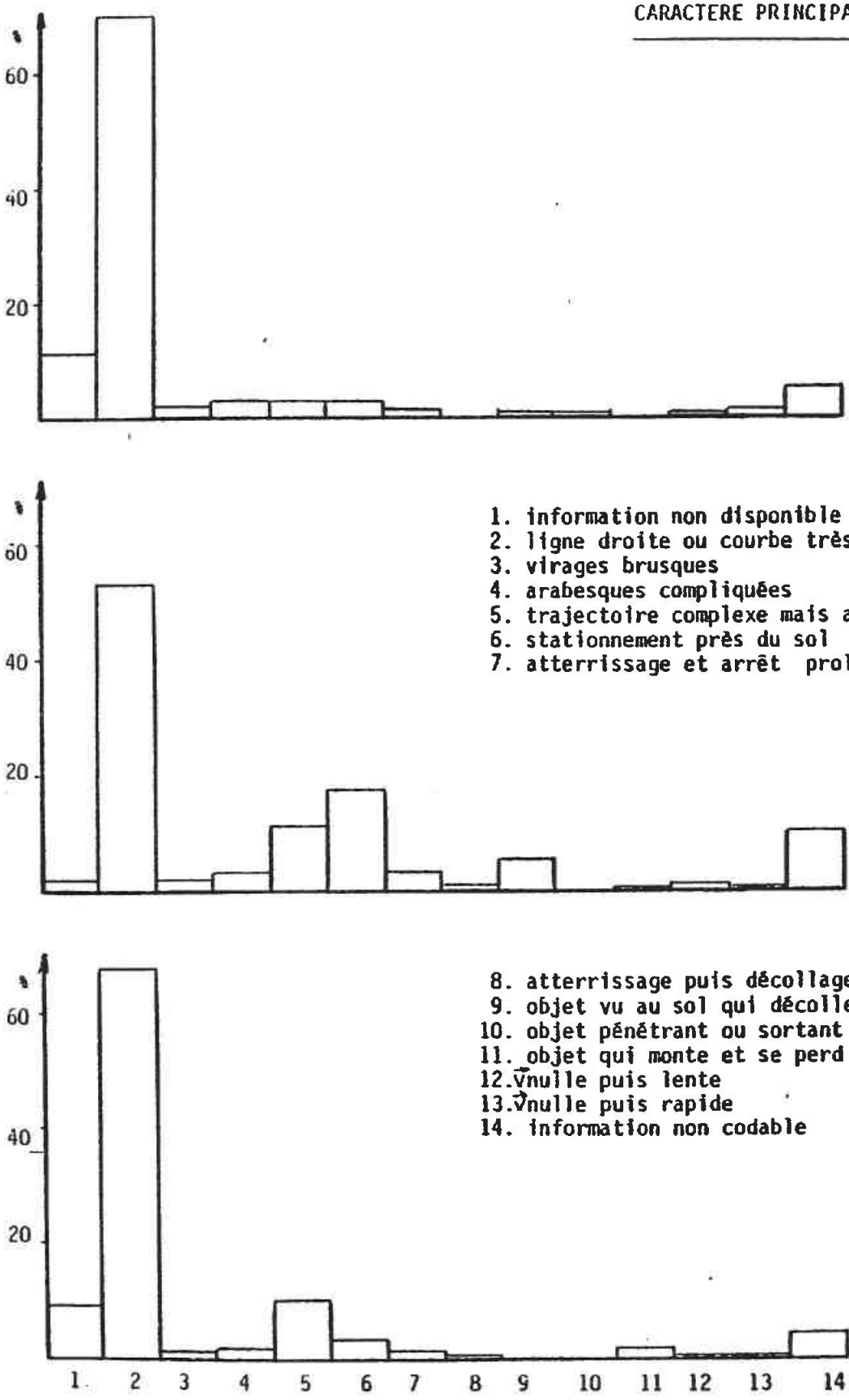


1. information non disponible
2. rouge sombre
3. rouge .
4. orange, feu
5. jaune, ambre
6. vert
7. bleu
8. bleu sombre, métallique, indigo
9. blanc

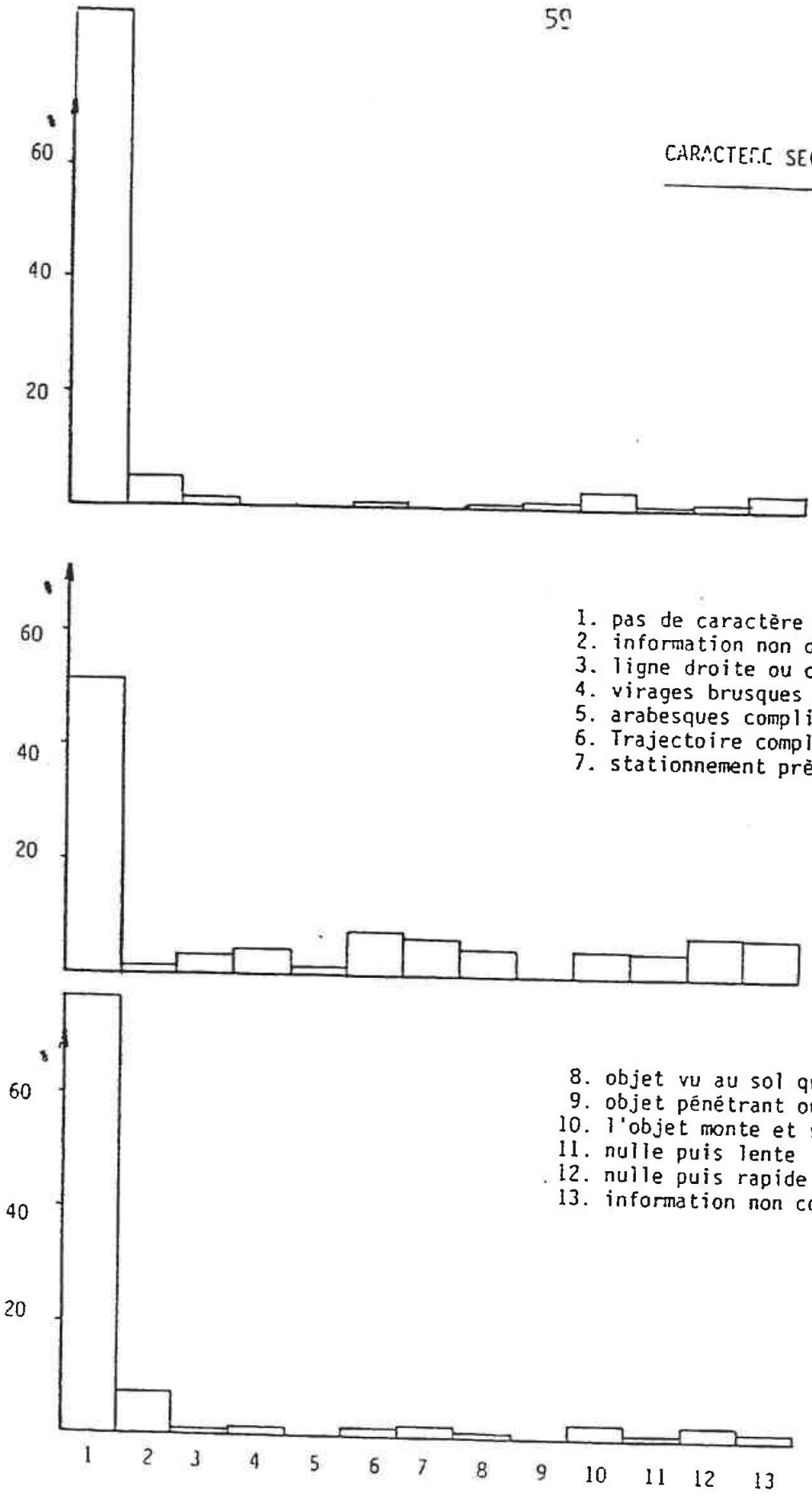


10. noir
11. gris
12. métallique (argent, aluminium poli)
13. plusieurs couleurs
14. couleur(s) changeante(s)
15. marron
16. or
17. information non codable
18. violet

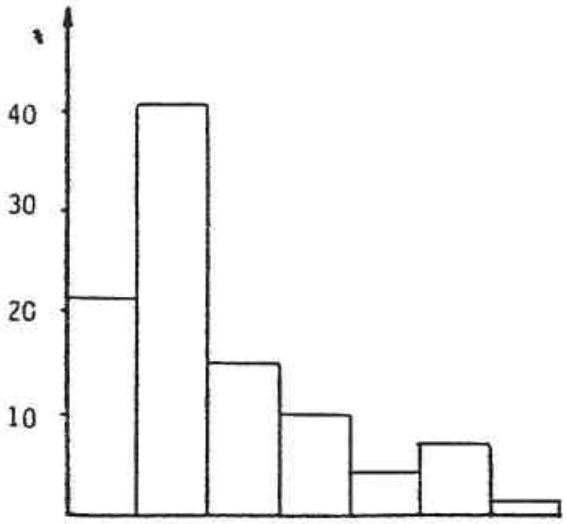
CARACTERE PRINCIPAL DE LA TRAJECTOIRE

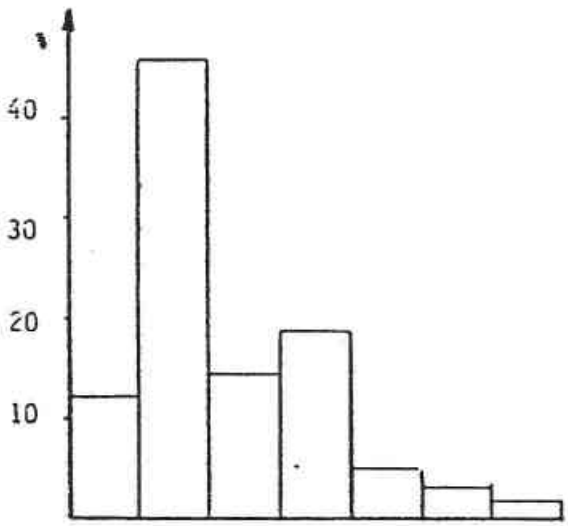


CARACTERE SECONDAIRE DE LA TRAJECTOIRE

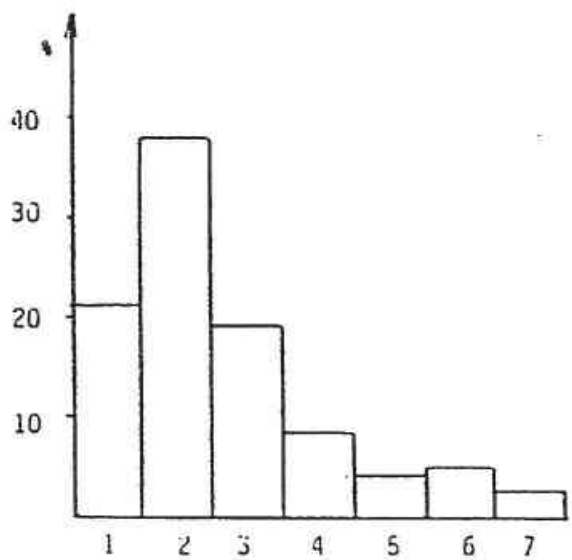


APPRECIATION DE LA VITESSE

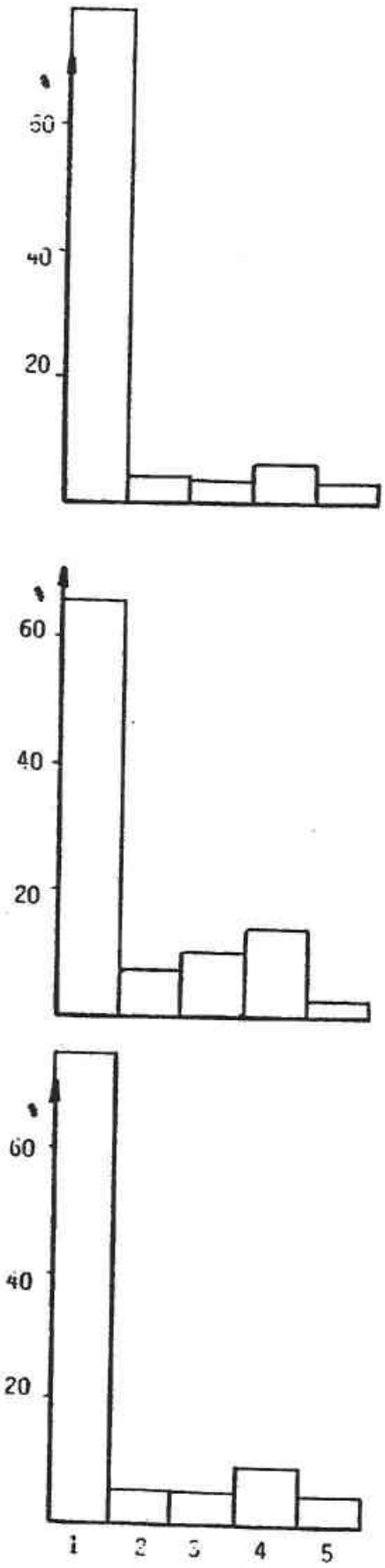




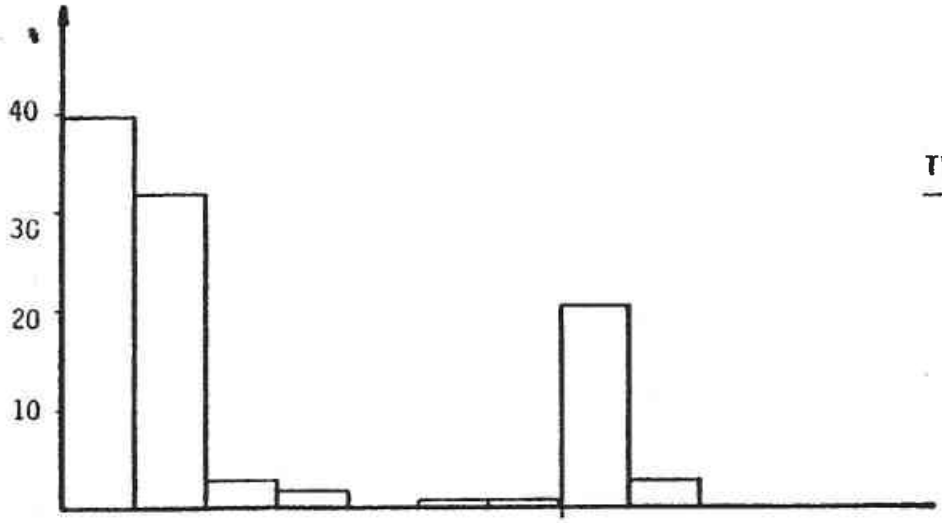
1. information non disponible
2. lente ou très lente ou immobile
3. très rapide
4. variable
5. fulgurante
6. vitesse d'un avion
7. information non codable



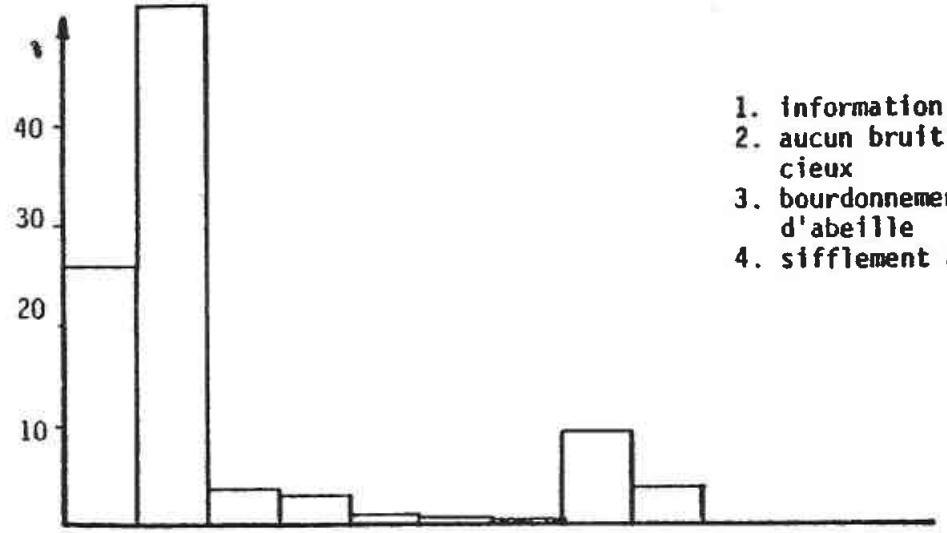
APPRECIATION DE L'ACCELERATION



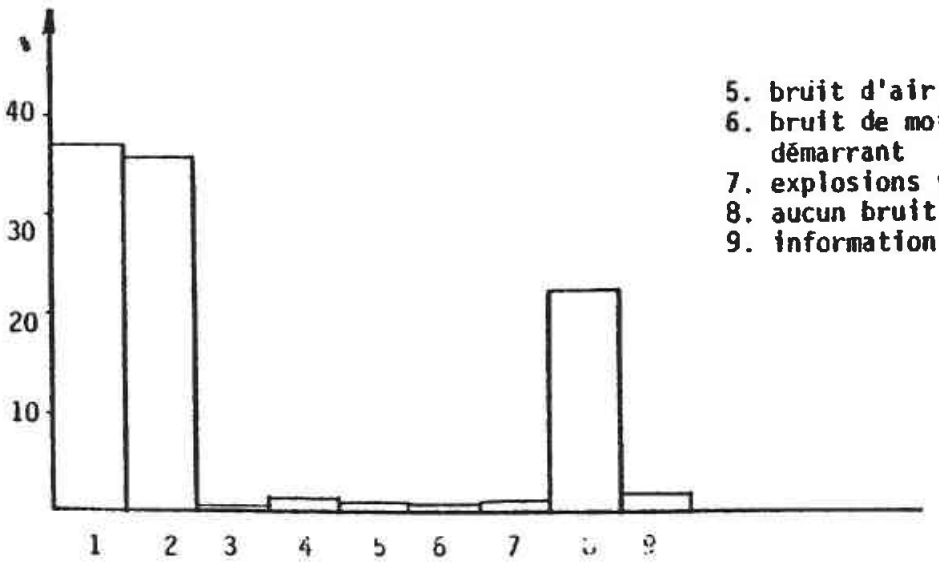
1. information non disponible
2. faible
3. variable
4. très élevée
5. information non codable



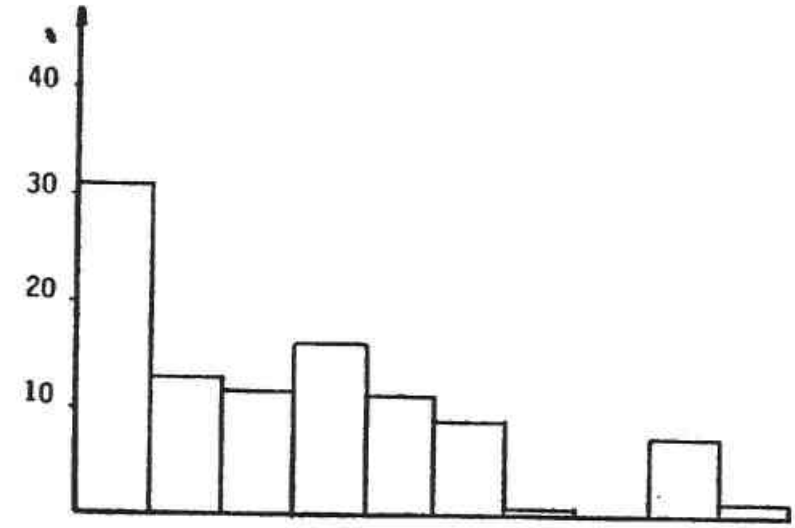
TYPE DE BRUIT



1. information non disponible
2. aucun bruit, silence total, objet silencieux
3. bourdonnement, vrombissement, bruit d'abeille
4. sifflement aigu

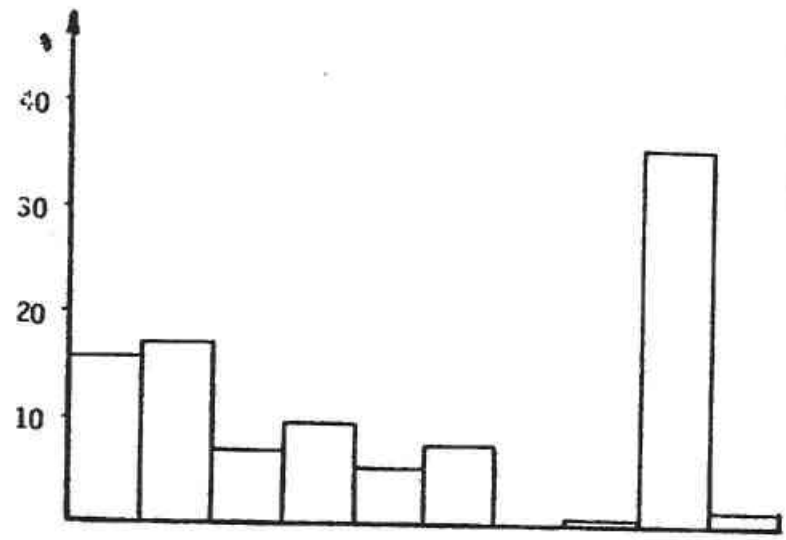


5. bruit d'air comprimé s'échappant
6. bruit de moteur électrique démarrant
7. explosions violentes
8. aucun bruit perceptible
9. information non codable

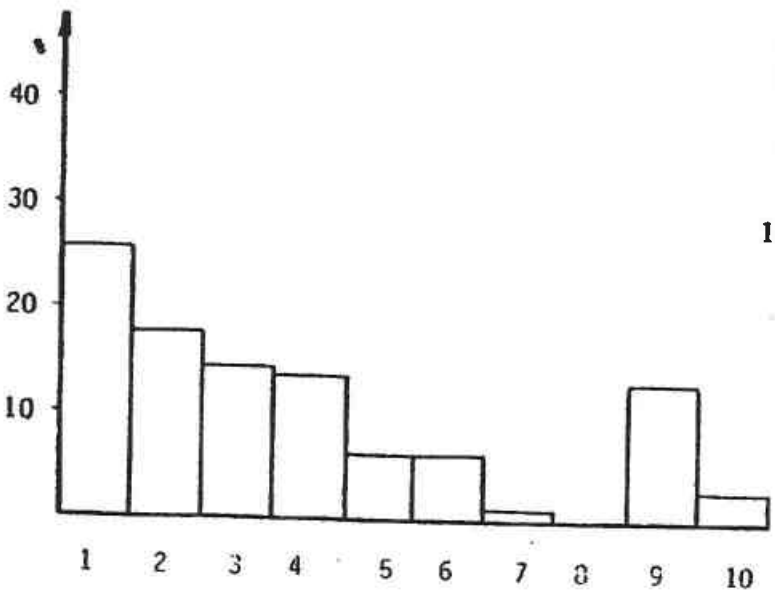


HAUTEUR ANGULAIRE

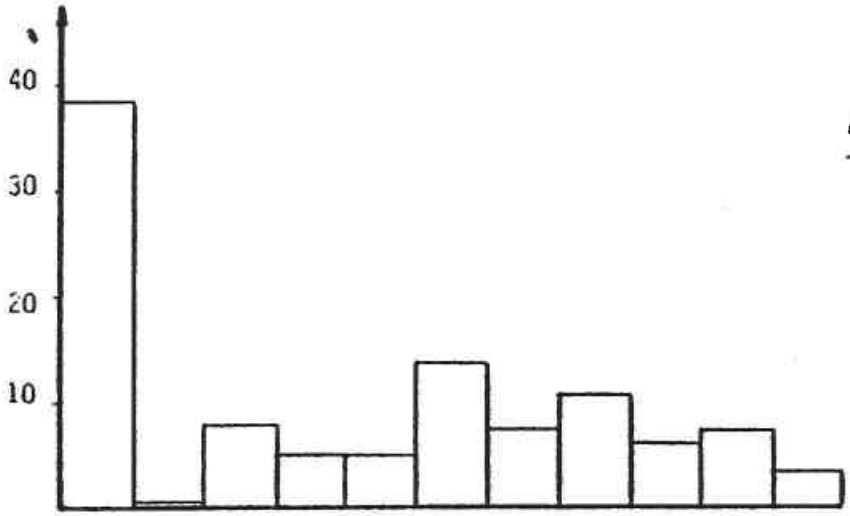
EN DEBUT D'OBSERVATION



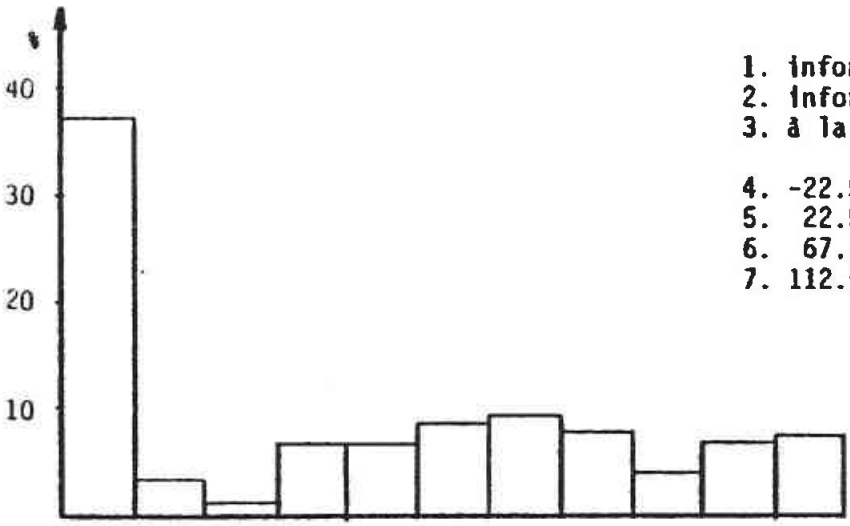
1. information non disponible
2. de 0 à 15 degrés (ou bas sur l'horizon)
3. de 15 à 30 degrés
4. de 30 à 40 degrés
5. de 40 à 60 degrés



6. de 60 à 90 degrés
7. au-dessous de l'horizon, sous un avion
8. observé d'un avion (même altitude ou au-dessus)
9. objet vu au sol ou près du sol
10. information non codable

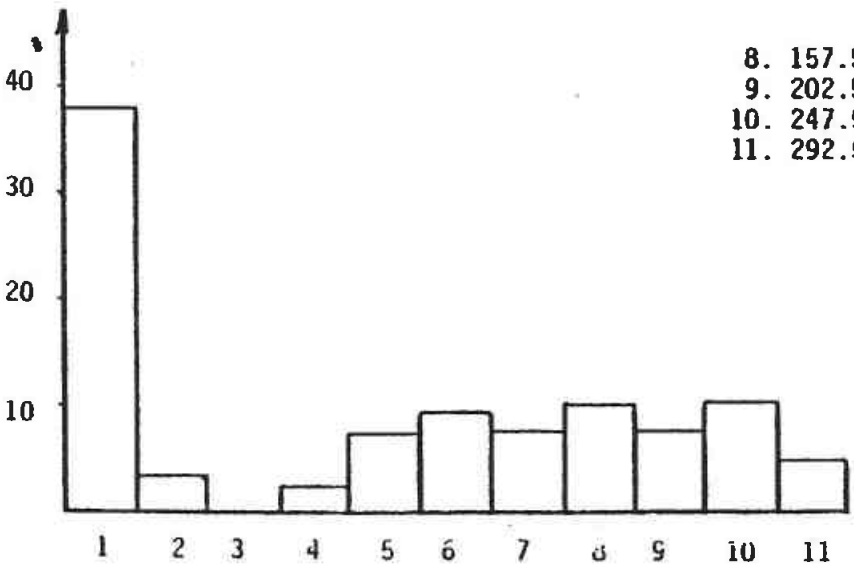


AZIMUT EN DEBUT D'OBSERVATION



1. information non disponible
2. information non codable
3. à la verticale (vers le zénith)

4. -22.5° à 22.5° vers le nord
5. 22.5° à 67.5° nord-est
6. 67.5° à 112.5° est
7. 112.5° à 157.5° sud-est



8. 157.5° à 202.5° sud
9. 202.5° à 247.5° sud-ouest
10. 247.5° à 292.5° ouest
11. 292.5° à 337.5° nord-ouest

ANNEXE 3

APPROCHE THÉORIQUE

Ph. BESSE Cl. VIDAL

I - INTRODUCTION

Il est fréquent, en pratique, d'avoir à étudier un phénomène décrit par une variable aléatoire (v.a.) Y inobservable (ou bien trop coûteuse à observer). Pour pallier à cet inconvénient l'étude de ce phénomène se fait au travers d'une v.a. X qui est -elle- observable et dont on peut estimer la liaison avec Y (en laboratoire ou sur un échantillon-test). Dans le cas où Y est une variable qualitative multidimensionnelle, une étude descriptive peut conduire à l'utilisation de l'Analyse Factorielle des Correspondances (A.F.C.) si Y est de la forme (Y_1,Y_2) où l'A.F.C. généralisée lorsque Y est un m-uple $(Y_1,Y_2,\ldots,Y_m)$. Pour déterminer cette analyse, il suffit de connaître (ou tout du moins d'estimer) la loi de Y. Celle-ci est déterminée (peut être estimée) dès que l'on connaît (sait estimer) la loi de X et la loi conditionnelle de Y à X.

Cette approche peut être utilisée en particulier dans le problème concret (à l'origine de cet article) exposé par P. BESSE [1] et que l'on peut résumer ainsi : des témoins observent un ensemble de phénomènes représentés par une variable qualitative multidimensionnelle Y dont ils font une estimation X (narration des témoins). Le problème est alors le suivant : comment rendre compte de Y à partir de l'étude faite sur X ?

En utilisant les outils introduits par J.F. MARTIN [8] sous l'appellation de codage "flou" (ou codage par une probabilité de transition), ce travail s'inscrit dans un cadre plus large visant à intégrer aux analyses factorielles les erreurs de mesure, de classement, les problèmes d'arrondi ou de lissage ... qui peuvent apparaître lors de l'observation des variables.

Ainsi, après avoir été sommairement rappelée, la notion de codage est étendue à des espaces produits de façon à pouvoir être appliquée à des couples (ou des m-uples) de variables. Ceci conduit alors à des problèmes d'indépendance (indépendance des erreurs entre elles, indépendance entre l'erreur sur une variable et les autres variables ...) qui sont discutés. Enfin, les problèmes d'estimation et donc de convergence sont résolus.

Remarque : Ce travail se rapportant à l'A.F.C. classique, toutes les tribus considérées par la suite sont de cardinal fini.

2 - T-CODAGE OU CODAGE PAR UNE TRANSITION

En pratique, les analyses factorielles non-linéaires utilisent essentiellement des fonctions indicatrices de modalités (pour des variables qualitatives) ou d'intervalles (pour des variables quantitatives), pour modéliser les situations à analyser ; ceci revenant à employer le codage disjonctif complet. La simplicité d'emploi et de mise en oeuvre de cet outil (qui permet d'exhiber simplement une base orthonormée) est aussi cause de sa rigidité, de son inadaptabilité à certains problèmes spécifiques. C'est pourquoi certains auteurs ont introduit d'autres formes de codage faisant appel à des fonctions spline (D. LAFAYE de MICHEAUX [6], J.O. RAMSAY - S. WINSBERG [11] et [12], J. VAN RIJCKEVORSEL [13]) ou bien à la notion de codage flou (J.P. BORDET [2], J.F. MARTIN [8]) - notions reprises très récemment par J.L. MALLET [7] et J.M. GAUTIER, G. SAPORTA [5] -. On peut encore citer D.M. TITTERINGTON [14] qui a développé des outils similaires - estimation de densités de probabilités discrètes par la méthode des noyaux - pour l'étude de données catégorielles.

Parmi ces divers modèles, l'approche probabiliste due à J.F. MARTIN [8] (codage par une probabilité de transition) a été choisie, car elle semble la plus adaptée, et ce pour différentes raisons :

* elle est synthétique et englobe les autres approches.

* le cadre précis dans lequel elle est développée permet de donner une signification au codage, et donc d'interpréter les résultats.

* enfin et surtout, ce n'est que dans ce cadre qu'il est possible de parler d'indépendance de codages (des différentes variables à analyser cf. 4.1.).

2.1 - rappels et notations

On rappelle ici sommairement les notions introduites par J.F. MARTIN [8], auquel on renvoie pour ce qui est des démonstrations.

Le codage par une (probabilité de) transition permet de tenir compte de certains problèmes parasites qui peuvent survenir lors de la saisie des données : erreurs de mesure, erreurs de classement, "bruit" qu'il faut lisser, erreurs systématiques d'arrondi ou encore, comme dans le cas qui nous intéresse, l'analyse d'une variable Y à l'aide d'une autre variable X.

Le principe adopté est alors d'associer à une observation $x = X(\omega)$ une probabilité (i.e. un codage)

P_x sur un espace $(F,\mathcal{F})$ servant à apprécier l'imprécision de la mesure. S'il n'y a pas d'imprécision (ou si l'on ne veut pas en tenir compte), il suffit d'employer le codage disjonctif classique.

Définitions

Soient $(E,\mathcal{E})$ et $(F,\mathcal{F})$ deux espaces probabilisables et P une transition sur $(E,\mathcal{E}) \times \mathcal{F}$, c'est-à-dire une application.

$$(E,\mathcal{E}) \times \mathcal{F} \longrightarrow ([0,1], \mathcal{B}_{[0,1]})$$

$$(x,B) \longrightarrow P_x(B)$$

telle que :

(i) $\forall x \in E \quad B \longrightarrow P_x(B)$ est une probabilité sur $(F,\mathcal{F})$

(ii) $\forall B \in \mathcal{F} \quad x \longrightarrow P_x(B)$ est mesurable de $(E,\mathcal{E})$ dans $([0,1],\mathcal{B}_{[0,1]})$.

Si X est une v.a. définie sur $(\Omega,\mathcal{A})$ à valeurs dans $(E,\mathcal{E})$, alors l'application P_X :

$$(\Omega,\mathcal{A}) \times \mathcal{B} \longrightarrow ([0,1], \mathcal{B}_{[0,1]})$$

$$(\omega,B) \longrightarrow P_{X(\omega)}(B)$$

définit une transition sur $(\Omega,\mathcal{A}) \times \mathcal{F}$.

Définition 1 :

On appelle T-codage sur $(E,\mathcal{E})$ relatif à $\mathcal{F}$ (resp. T-codage de X relatif à $\mathcal{F}$) une transition P sur $(E,\mathcal{E}) \times \mathcal{F}$ (resp. une transition P_X sur $(\Omega,\mathcal{A}) \times \mathcal{F}$).

Définition 2 :

Un T-codage sur $(E,\mathcal{E})$ relatif à $\mathcal{F}$ définit pour tout B de $\mathcal{F}$ une application mesurable ϕ_B définie par :

$$\phi_B : (E,\mathcal{E}) \longrightarrow ([0,1], \mathcal{B}_{[0,1]})$$

$$x \longrightarrow \phi_B(x) = P_x(B)$$

est appelée fonction de codage (associée à B).

Exemples

* Au codage disjonctif classique est associée la transition

$P = \delta$; $(x,B) \longrightarrow P_x(B) = \delta_x(B)$ (où δ_x désigne la distribution de Dirac au point x) et, dans ce cas, les fonctions de codage sont les indicatrices :

$$\phi_B(x) = \mathbb{1}_B(x).$$

Le codage conditionnel : Soit (X,Y) un couple de v.a. définies sur $(\Omega,\mathcal{A},\mu)$ à valeurs dans $(E \times F, \mathcal{E} \otimes \mathcal{F})$; pour tout B de $\mathcal{F}$, la fonction :

$$\phi_B : (E,\mathcal{E}) \longrightarrow ([0,1],\mathcal{B}_{[0,1]})$$
$$x \longrightarrow \phi_B(x) = E_\mu^{X=x}(\mathbb{1}_B \circ Y)$$

(où $E_\mu^{X=x}(\mathbb{1}_B \circ Y)$ est encore la probabilité conditionnelle $\mu[Y \in B/_{X=x}])$ est une fonction mesurable.

Si, de plus, il existe une version régulière de la probabilité conditionnelle de Y à X, l'application :

$$(E,\mathcal{E}) \times \mathcal{F} \longrightarrow ([0,1],\mathcal{B}_{[0,1]})$$
$$(x,B) \longrightarrow P_x(B) = \phi_B(x) = E_\mu^{X=x}(\mathbb{1}_B \circ Y)$$

est une transition sur $(E,\mathcal{E}) \times \mathcal{F}$ et on appelle T-codage conditionnel de Y en X la transition sur $(\Omega,\mathcal{A}) \times \mathcal{F}$:

$$P_X : (\omega,B) \longrightarrow P_{X(\omega)}(B) = [E_\mu^X (\mathbb{1}_B \circ Y)](\omega).$$

Remarque : La fonction de codage ϕ_B n'est autre que la fonction de régression de $\mathbb{1}_B \circ Y$ en X.

T-codage d'une probabilité

Lorsque $(E,\mathcal{E})$ est muni d'une probabilité ν, le T-codage définit une probabilité sur $(F,\mathcal{F})$ par :

$$\mathcal{F} \longrightarrow [0,1]$$
$$B \longrightarrow E_\nu[P.(B)] = \int_E P_x(B)\, d\nu(x) = E_\nu(\phi_B).$$

Elle est appelée probabilité codée de ν par P et notée $E_\nu(P)$ (ceci correspond à la notation νP de J. NEVEU [10], image à gauche de la probabilité ν par P considéré comme un opérateur sous-markovien).

Avec les notations des paragraphes précédents, si μ_X désigne la loi de X, la probabilité codée de μ_X par P est égale à la probabilité codée de μ par P_X :

$$E_{\mu_X}(P) = E_\mu(P_X) .$$

Exemples ;

* Si P est la transition associée au codage disjonctif, on a :

$$E_{\mu_X}(P) = E_{\mu_X}(\delta) = \mu_X .$$

Si P est la transition associée au codage conditionnel, on trouve :

$$\star\ \forall B \in \mathcal{F} \quad [E_{\mu_X}(P)](B) = E_{\mu_X}[P_\cdot(B)]$$
$$= \int_E P_X(B)\, d\mu_X(x)$$
$$= \int_\Omega P_X(\omega)(B)\, d\mu(\omega) = \mu_Y(B).$$

2.2 - T-codage sur un espace produit

Position du problème

La variable X à laquelle est appliqué le T-codage peut être un couple (X_1, X_2) ou même un m-uple $(X_1, \ldots, X_m)$ prenant ses valeurs sur un espace produit $(E_1 \times \ldots \times E_m, \mathcal{E}_1 \otimes \ldots \otimes \mathcal{E}_m)$. Comment peut-on construire le codage de X à partir de celui des X_i $(i=1,\ldots,m)$?

Soit donc $(E, \mathcal{E}, \nu)$ un espace probabilisé et $(F, \mathcal{F})$ un espace probabilisable tels que :

$$E = E_1 \times E_2\ ,\ F = F_1 \times F_2 \quad \text{et} \quad \mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2\ ,\ \mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2\ .$$

On note P une transition sur $(E, \mathcal{E}) \times \mathcal{F}$.

T-codage marge

Pour tout x de E, P_x est une probabilité sur $(F, \mathcal{F})$ et on peut considérer ses marges sur chacun des espaces $(F_i, \mathcal{F}_i)_{i=1,2}$; soit ${}^iP_x$ ces probabilités. On a :

$$\forall x \in E\ ,\ \forall B_1 \in \mathcal{F}_1\ ,\ {}^1P_x(B_1) = \sum_{\ell=1}^{s} P_x(B_1 \times B_2^\ell)$$
$$\forall B_2 \in \mathcal{F}_2\ ,\ {}^2P_x(B_2) = \sum_{k=1}^{r} P_x(B_1^k \times B_2)$$

où $\{B_1^k\}_{k=1,\ldots,r}$ (resp. $\{B_2^\ell\}_{\ell=1,\ldots,s}$) désigne une partition de F_1 (resp. de F_2) engendrant $\mathcal{F}_1$ (resp. $\mathcal{F}_2$).

$P_\cdot(B)$ étant une application mesurable pour tout B de $\mathcal{F}$, les applications ${}^iP_\cdot(B_i)$ $(i=1,2)$ le sont aussi, comme sommes finies d'applications mesurables. Ainsi , ${}^iP$ est une transition sur $(E, \mathcal{E}) \times \mathcal{F}_i$.

Définition 3 :

On appelle T-codages marge de P, les transitions ${}^iP$ sur $(E, \mathcal{E}) \times \mathcal{F}_i$, telles que pour tout x de E, les lois ${}^iP_x$ sur $(F_i, \mathcal{F}_i)$ soient les marges de P_x.

Proposition 1

La loi ν codée par un T-codage marge ${}^iP$ de P, est la marge de la probabilité codée de ν par P, soit :

$$ {}^i[E_\nu(P)] = E_\nu({}^iP). $$

La démonstration est évidente par interversion des signes de sommation.

Codage produit

Produit simple

Si P^i désigne un T-codage sur $(E_i, \mathcal{E}_i)$ relatif à $\mathcal{F}_i$ et si $\mathcal{S}$ est le semi-anneau des "rectangles" :

$\mathcal{S} = \{B_1 \times B_2 \;/\; B_1 \in \mathcal{F}_1 \text{ et } B_2 \in \mathcal{F}_2\}$, on définit l'application

$P = P^1 \otimes P^2$ sur $(E, \mathcal{E}) \times \mathcal{S}$ par :

$$ P : (E,\mathcal{E}) \times \mathcal{S} \longrightarrow ([0,1], \mathcal{B}_{[0,1]}) $$
$$ (x = (x_1,x_2), B_1 \times B_2) \longrightarrow P_x(B_1 \times B_2) = P^1_{x_1}(B_1) \times P^2_{x_2}(B_2). $$

Proposition 2

P est un T-codage sur $(E, \mathcal{E})$ relatif à $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$.

Démonstration :

$\mathcal{F}_1$ et $\mathcal{F}_2$ étant finies, on peut décomposer tout élément de $\mathcal{F}_1 \otimes \mathcal{F}_2$ en une réunion finie d'ensembles minimaux deux à deux disjoints et ainsi définir P sur $E \times \mathcal{S}$.

Soit :

$$ P_x(B) = P_x\Big(\bigcup_{i=1}^{n} B_1^i \times B_2^i\Big) = \sum_{i=1}^{n} P_x(B_1^i \times B_2^i) $$

$$ = \sum_{i=1}^{n} P^1_{x_1}(B_1^i).P^2_{x_2}(B_2^i). $$

D'autre part, pour tout B de $\mathcal{F}$, l'application : $x \to P_x(B)$ est mesurable comme somme finie d'applications mesurables. □

Produit des marges

Si ${}^iP$ est un T-codage sur E relatif à $\mathcal{F}_i$, on définit :

$$ \forall x \in E,\ \forall B_1 \times B_2 \in \mathcal{S},\ P_x(B_1 \times B_2) = {}^1P_x(B_1) \,.\, {}^2P_x(B_2) $$
$$ = ({}^1P \otimes {}^2P)_x\,(B_1 \times B_2). $$

$^1P \otimes {}^2P$ est de même un T-codage appelé codage produit de ses marges.

Remarque

Si Card $E_1 = p$, Card $E_2 = q$ et si les tribus $\mathcal{F}_1$ et $\mathcal{F}_2$ sont engendrées par des partitions de cardinaux respectifs r et s, on voit alors que la définition d'un T-codage sur E relatif à $\mathcal{F}$ nécessite la connaissance (ou du moins dans la pratique l'estimation) de p q r s termes.

On aura donc tout intérêt à se ramener, si cela est possible, à des codages produits plus simples, puisque :

* pour le produit simple, il suffit de connaître pr + qs termes.
* pour le produit des marges, il suffit de connaître pq(r+s) termes.

3 - ANALYSES ET T-CODAGES

3.1 - Rappels et notations

L'A.F.C. peut être introduite de différentes manières équivalentes entre elles. Nous en utiliserons deux au cours de cet exposé :

Analyse canonique (A.C.) d'une probabilité sur un espace produit

Soit (X_1,X_2) un couple de v.a. défini sur $(\Omega, \mathcal{A}, \mu)$ et à valeurs dans $(E_1 \times E_2)$, $\mathcal{E}_1 \otimes \mathcal{E}_2$). On désigne par $\mathcal{A}_k$ (k=1,2) les tribus complétées de $X_k^{-1}(\mathcal{E}_k)$ et, respectivement, par $\{\alpha_1^i\}_{i=1,\ldots,p}$ et $\{\alpha_2^j\}_{j=1,\ldots,q}$ les partitions finies de Ω qui les engendrent. On pose :

$$p_{ij} = \mu(\alpha_1^i \cap \alpha_2^j) \quad , \quad p_{i.} = \sum_{j=1}^{q} p_{ij} = \mu(\alpha_1^i)$$

$$p_{.j} = \sum_{i=1}^{p} p_{ij} = \mu(\alpha_2^j).$$

On note $\mu_X = \mu_{(X_1,X_2)}$ l'image de μ par (X_1,X_2) sur $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$, μ_{X_1} et μ_{X_2} ses marges et λ le produit $\mu_{X_1} \otimes \mu_{X_2}$.

μ_X admet relativement à λ une densité qu'il suffit d'expliciter sur les pq modalités $(\{x_1^i\} \times \{x_2^j\})$ de (X_1,X_2) :

$$\frac{d\mu_{(X_1,X_2)}}{d\lambda}(\{x_1^i\} \times \{x_2^j\}) = \frac{\mu(\alpha_1^i \cap \alpha_2^j)}{\mu(\alpha_1^i).\mu(\alpha_2^j)} = \frac{p_{ij}}{p_{i.} \times p_{.j}} \quad .$$

On sait (cf. J. DAUXOIS et A. POUSSE [4]) que l'A.F.C. de (X_1, X_2) peut alors être définie comme l'A.C. de μ_X. Elle est obtenue par l'analyse spectrale de l'opérateur $E^{X_1} \circ E^{X_2}/L^2(\Omega_1)$ où l'on a :

$$E^{X_1} : g \in L^2(\Omega_2) \longrightarrow h = E^{X_1}(g) \quad \text{tq} : h(i) = \sum_{j=1}^{q} \frac{p_{ij}}{p_{i.}} g(j)$$

$$E^{X_2} : f \in L^2(\Omega_1) \longrightarrow k = E^{X_2}(f) \quad \text{tq} : k(j) = \sum_{i=1}^{p} \frac{p_{ij}}{p_{.j}} f(i)$$

où $f(i)$ (resp. $g(j)$) désigne la valeur prise sur α_1^i (resp. α_2^j) pour $i=1,\ldots,p$ (resp. $j=1,\ldots,q$).

Ces opérateurs ont pour matrices, relativement aux bases $(\mathbb{1}_{\alpha_1^i})_{i=1,\ldots,p}$ et $(\mathbb{1}_{\alpha_2^j})_{j=1,\ldots,q}$ de $L^2(\Omega_1)$ et $L^2(\Omega_2)$, respectivement $D_1^{-1}T$ et $D_2^{-1}{}^tT$, où T est la matrice de terme général p_{ij} et D_1 (resp. D_2) la matrice $\text{diag}(p_{i.})_{i=1,\ldots,p}$ (resp. $\text{diag}(p_{.j})_{j=1,\ldots,q}$).

L'A.F.C. est alors obtenue par l'analyse spectrale de la matrice : $D_1^{-1} T D_2^{-1} {}^tT$.

A.F.C. en tant qu'A.C.P.

Si on note encore (on confond les notations par souci de simplification) : $X_1 = (\mathbb{1}_{\alpha_1^i})_{i=1,\ldots,p}$ et $X_2 = (\mathbb{1}_{\alpha_2^j})_{j=1,\ldots,q}$ les paquets d'indicatrices, une façon plus classique d'introduire l'A.F.C. de (X_1, X_2) est de la définir (cf. par exemple CAZES et all. [3]) comme l'Analyse en Composantes Principales (A.C.P.) du tableau $Z = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ obtenu en superposant les deux paquets d'indicatrices. Le schéma de dualité associé s'écrit :

$$\begin{array}{ccc} \mathbb{R}^{p+q} & \xleftarrow{\quad Z = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad} & L^2(\Omega, \quad, \mu) \simeq \mathbb{R}^n \\ M \downarrow \uparrow V & & W \downarrow \uparrow D_n = \frac{1}{n} I_n \\ \mathbb{R}^{p+q} & \xrightarrow{\quad {}^tZ \quad} & L^2(\Omega, \quad, \mu) \end{array}$$

où $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$ et $M = \begin{bmatrix} V_{11}^{-1} & 0 \\ 0 & V_{22}^{-1} \end{bmatrix}$ avec $V_{ij} = X_i \, D_n \, {}^tX_j$.

Cette A.C.P. est obtenue par l'analyse spectrale de l'opérateur $W \circ D_n$ représenté par la matrice :

$$ {}^tX_1 \, V_{11}^{-1} \, X_1 \, D_n + {}^tX_2 \, V_{22}^{-1} \, X_2 \, D_n \, . $$

3.2 - A.F.C. codée

Introduction

J.F. MARTIN [8] a déjà décrit quelques applications du T-codage (ou codage flou) aux analyses factorielles (A.C.P. d'une probabilité codée, A.C. non linéaire d'un couple de v.a. codées ...) permettant entre autre d'introduire dans le modèle des notions d'erreurs de mesure, d'arrondi, ou encore de lissage de variables quantitatives. Un autre type d'application est proposé ici, concernant cette fois les variables qualitatives et donc l'A.F.C. . Il permet de tenir compte des erreurs de classement des individus à l'intérieur de chaque modalité ou encore d'estimer l'analyse de la v.a. Y "coûteuse" à observer par l'intermédiaire d'une variable X plus facilement accessible.

A.C. *T-codée d'une probabilité*

Définition 4 :

Soit ν *une probabilité sur* $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ *et* P_m *un T-codage relatif à* $\mathcal{F}_1 \otimes \mathcal{F}_2$ *, on appelle analyse canonique T-codée de* ν*, l'A.C. de la loi* $E_\nu(P)$ *sur l'espace* $(F_1 \times F_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$.

A.F.C. *codée d'un couple de v.a.*

De même, si $X = (X_1, X_2)$ est une v.a. définie sur $(\Omega, \mathcal{A}, \mu)$ à valeurs dans $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ muni de la loi image μ_X de μ par X, on a :

Définition 5 :

Si P est un T-codage sur $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ *relatif à* $\mathcal{F}_1 \otimes \mathcal{F}_2$*, on appelle A.F.C. T-codée de X, l'A.C. de la loi* $E_{\mu_X}(P)$ *sur l'espace produit* $(F_1 \times F_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$.

Remarque :

Si ${}^1P$ et ${}^2P$ désignent les T-codages marges de P, cette analyse est obtenue à partir de l'explicitation de la densité :

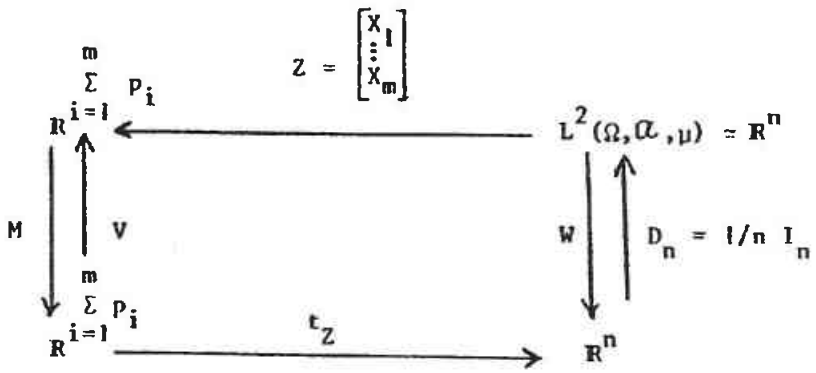
$$\frac{dE_{\mu}(P_X)}{d[{}^1E_{\mu}(P_X) \otimes {}^2E_{\mu}(P_X)]} = \frac{dE_{\mu_X}(P)}{d[E_{\mu_X}({}^1P) \otimes E_{\mu_X}({}^2P)]}$$

3.3 - m-A.F.C. codée

Définition

Une généralisation possible de l'A.F.C. à plus de deux variables qualitatives s'établit à partir de la définition donnée en 3.1.

Considérons m variables ayant chacune p_i modalités, on est conduit au schéma de dualité suivant :



où $V = \left[V_{ij}\right]_{\substack{i=1,\ldots,m \\ j=1,\ldots,m}}$ avec $V_{ij} = X_i\, D_n\, {}^tX_j$ et $M = \mathrm{diag}(V_{ii}^{-1})_{i=1,\ldots,m}$.

L'analyse cherchée est alors obtenue par l'analyse spectrale de l'opérateur ${}^tZMZD_n$ ou encore par celle de l'opérateur $M \circ V$.

Remarque :

Dans ce type de généralisation de l'A.F.C., seuls interviennent les tableaux V_{ij} croisant toutes les variables deux à deux ; c'est-à-dire qu'il n'est pas nécessaire de connaître la loi de $X = (X_1,\ldots,X_m)$ sur toute la tribu $\mathcal{E}_1 \otimes \ldots \otimes \mathcal{E}_m$, mais seulement ses "marges d'ordre deux".

m-A.F.C. codée

La notion de T-codage sur un espace produit introduite en 2.2 - se généralise sans difficulté à plus de deux espaces ; on obtient encore une transition sur $(E = E_1 \times \ldots \times E_m\ ,\ \mathcal{E} = (\mathcal{E}_1 \otimes \ldots \otimes \mathcal{E}_m)) \times (\mathcal{F}_1 \otimes \ldots \otimes \mathcal{F}_m)$.

D'autre part, compte tenu de la remarque précédente, la m-analyse ne nécessite que la connaissance des "marges d'ordre deux" de la loi $E_{\mu_X}(P)$ pour le calcul de la densité, soit :

$$\forall (i,j) \in [1,\ldots,m]^2 \quad , \quad {}^{ij}[E_{\mu_X}(P)] = E_{\mu_X}({}^{ij}P)$$

(propriété analogue à la proposition 1)
et donc il suffit de connaître les T-codages marges ${}^{ij}P$ sur $(E,\mathcal{E})$ relatifs à $\mathcal{F}_i \otimes \mathcal{F}_j$.

4 - APPLICATION AU PROBLEME CONCRET

4.1 - Codage conditionnel sur un espace produit

Position du problème

On souhaite étudier une variable Y inobservable, ou trop coûteuse à observer, à l'aide d'une variable X facilement observable mais entachée d'erreurs aléatoires (non nécessairement indépendantes ou additives). Il est alors impossible de déterminer une fonction de transfert de X en Y, mais on suppose connue (i.e., par exemple, estimable sur un échantillon-test) la loi de probabilité conditionnelle de Y à X qui permet de déterminer, par l'intermédiaire du codage, la loi de Y connaissant celle de X ($E_{\mu_X}(P) = \mu_Y$ cf. 2.1). On peut ainsi calculer l'A.F.C. de $Y = (Y_1,Y_2)$ qui ne dépend que de μ_Y.

Le T-codage conditionnel

Dans le cas qui nous intéresse, les variables qualitatives X et Y sont des couples et ont un nombre fini de modalités. Soit donc $E = E_1 \times E_2$ et $F = F_1 \times F_2$ avec :

$$E_1 = \{x_1^i / i \in I = \{1,\ldots,p\}\} \quad , \quad E_2 = \{x_2^j / j \in J = \{1,\ldots,q\}\}$$

$$F_1 = \{y_1^k / k \in K = \{1,\ldots,r\}\} \quad , \quad F_2 = \{x_2^\ell / \ell \in L = \{1,\ldots,s\}\}.$$

On définit alors le T-codage conditionnel de Y en X par :

$$\forall (i,j,k,\ell) \in I \times J \times K \times L \qquad \phi_{ijk\ell} = P_{(x_1^i,x_2^j)}(\{y_1^k\} \times \{y_2^\ell\})$$

$$= \mu[(Y_1,Y_2)=(y_1^k,y_2^\ell)/(X_1,X_2)=(x_1^i,x_2^j)]$$

et la fonction de codage se met sous la forme d'une matrice :

$$\Phi = [\phi_{ijk\ell}]_{\substack{(i,j)\in I\times J \\ (k,\ell)\in K\times L}} .$$

$\phi_{ijk\ell}$ exprime encore la probabilité de bon ou de mauvais classement : c'est la probabilité que $Y(\omega)$ prenne la modalité (y_1^k, y_2^ℓ) sachant que $X(\omega)$ est observée en (x_1^i, x_2^j).

Hypothèses de simplification

On a vu en 2.2 - que l'on pouvait diminuer le nombre de termes à estimer pour déterminer le T-codage P en l'exprimant sous la forme d'un produit. Afin de simplifier les notations on pose :

$$\alpha_1^i = \overset{-1}{X_1}(\{x_1^i\}),\ \alpha_2^j = \overset{-1}{X_2}(\{x_2^j\}) ,\quad \beta_1^k = \overset{-1}{Y_1}(\{y_1^k\}) ;\quad \beta_2^\ell = \overset{-1}{Y_2}(\{y_2^\ell\}) ,$$

et ainsi :

$$\phi_{ijk\ell} = \mu[\beta_1^k \cap \beta_2^\ell \,/\, \alpha_1^i \cap \alpha_2^j].$$

Considérons alors les hypothèses suivantes (valables pour tout (i,j,k,ℓ) de $I \times J \times K \times L$) :

$$(H_1) \qquad \mu[\beta_1^k \cap \beta_2^\ell \,/\, \alpha_1^i \cap \alpha_2^j] = \mu[\beta_1^k \,/\, \alpha_1^i \cap \alpha_2^j] \cdot \mu[\beta_2^\ell \,/\, \alpha_1^i \cap \alpha_2^j]$$

$$(H_2) \quad \begin{cases} \mu[\beta_1^k \,/\, \alpha_1^i \cap \alpha_2^j] = \mu[\beta_1^k \,/\, \alpha_1^i] \\ \text{et} \\ \mu[\beta_2^\ell \,/\, \alpha_1^i \cap \alpha_2^j] = \mu[\beta_2^\ell \,/\, \alpha_2^j] \end{cases}$$

Proposition 3

(i) $(H_1) \Longleftrightarrow P = {}^1P \otimes {}^2P$

(ii) (H_1) *et* $(H_2) \Longleftrightarrow P = P^1 \otimes P^2$

Démonstration :

(i) En explicitant,

$$P = {}^1P \otimes {}^2P \iff \forall x \in E,\ \forall B \in \mathcal{S} \quad P_x(B) = {}^1P_x(B_1) \cdot {}^2P_x(B_2)$$

$$\iff \forall (i,j,k,\ell) \in I\times J\times K\times L \quad \mu[\beta_1^k \cap \beta_{2/\alpha_1^i\cap\alpha_2^j}^{\ell}] = \mu[\beta_{1/\alpha_1^i\cap\alpha_2^j}^{k}] \cdot \mu[\beta_{2/\alpha_1^i\cap\alpha_2^j}^{\ell}]$$

$$\iff (H_1)$$

(ii) Supposons (H_1) et (H_2), alors : $\forall (i,j,k,\ell) \in I\times J\times K\times L$

$$P_{(x_1^i,x_2^j)}(\{y_1^k\}\times\{y_2^\ell\}) = \mu[\beta_1^k\cap\beta_{2/\alpha_1^i\cap\alpha_2^j}^{\ell}] = \mu[\beta_{1/\alpha_1^i}^{k}] \cdot \mu[\beta_{2/\alpha_2^j}^{\ell}]$$

$$= P^1_{x_1^i}(\{y_1^k\}) \cdot P^2_{x_2^j}(\{y_2^\ell\})$$

Réciproquement, si P est égal à $P^1 \otimes P^2$ on a :

$$\forall x \in E,\ \forall B \in \mathcal{S} \qquad P_x(B) = P^1_{x_1}(B_1) \cdot P^2_{x_2}(B_2)$$

en prenant $x = (x_1^i, x_2^j)$ et successivement $B = \{y_1^k\} \times F_2$ et $B = F_1 \times \{y_2^\ell\}$ on obtient (H_2),

puis en prenant $x = (x_1^i, x_2^j)$ et $B = \{y_1^k\} \times \{y_2^\ell\}$ et en utilisant (H_2), on obtient (H_1). □

Signification des hypothèses

(H_1) ou [(H_1) et (H_2)] sont des hypothèses d'indépendance conditionnelle permettant de simplifier l'expression de P.

En écrivant les variables sous la forme :

$Y_i = \psi_i(X_i, \varepsilon_i)$ pour $i=1,2$,

où ψ_i est une fonction mesurable et ε_i une v.a. "mesurant" l'erreur, on remarque que les hypothèses signifient :

*(H_1) : X_1 et X_2 étant connues, Y_1 et Y_2 sont indépendantes ou encore, les erreurs de mesure (ou de classement) sont indépendantes entre elles.

* (H_1) et (H_2) : X_1 et X_2 étant connues ; pour i=1,2 , Y_i ne dépend que de X_i ou encore, ε_i est indépendante de la valeur de X_{3-i} et, ε_1 et ε_2 sont indépendantes entre elles.

4.2 - Détermination pratique de l'analyse

Explicitation de la densité

Comme il est décrit en 3.1 - , le calcul de l'analyse de Y ne nécessite pas l'observation de Y, mais la connaissance de la loi de Y et, plus précisément, de la densité de cette loi par rapport au produit de ses marges dont l'existence est assurée car E et F sont de dimension finie.

L'expression de cette densité est plus ou moins complexe suivant les hypothèses qu'il est possible d'admettre.

Posons :

$$p_{ij} = \mu_X(\{x_1^i\} \times \{x_2^j\}) = \mu(\alpha_1^i \cap \alpha_2^j).$$

avec pour marges $p_{i.}$ et $p_{.j}$ et :

$$q_{k\ell} = \mu_Y(\{y_1^k\} \times \{y_2^\ell\}) = \mu(\beta_1^k \cap \beta_2^\ell)$$

avec pour marges $q_{k.}$ et $q_{.\ell}$.

La densité s'exprime alors :

$$\frac{dE\mu_X(P)}{d[E_{\mu_X}(^1P)\otimes E_{\mu_X}(^2P)]}(\{y_1^k\}\times\{y_2^\ell\}) = \frac{q_{k\ell}}{q_{k.}\times q_{.\ell}} = \frac{\sum_{i,j}\phi_{ijk\ell}\,p_{ij}}{\sum_{\ell=1}^{s}(\sum_{i,j}\phi_{ijk\ell}p_{ij})\times\sum_{k=1}^{r}(\sum_{i,j}\phi_{ijk\ell}p_{ij})}$$

$$= \frac{\sum_{i,j}\phi_{ijk\ell}p_{ij}}{(\sum_{i,j}\phi_{ijk.\ell}p_{ij})\times(\sum_{i,j}\phi_{ij.\ell}p_{ij})}$$

où, par analogie,

$$\phi_{ijk.} = \sum_{\ell=1}^{s}\phi_{ijk\ell} \quad \text{et} \quad \phi_{ij.\ell} = \sum_{k=1}^{r}\phi_{ijk\ell}$$

Avec les mêmes notations on a la :

Proposition 4

$$(i) \qquad P = {}^1P \otimes {}^2P \Rightarrow \frac{dE_{\mu_X}(P)}{d[E_{\mu_X}({}^1P) \otimes E_{\mu_X}({}^2P)]}\,(\{y_1^k\}\times\{y_2^\ell\}) = \frac{\sum_{i,j} \phi_{ijk.}\,\phi_{ij.\ell}\,p_{ij}}{(\sum_{i,j} \phi_{ijk.}\,p_{ij})\times(\sum_{i,j} \phi_{ij.\ell}\,p_{ij})}$$

$$(ii) \qquad P = P^1 \otimes P^2 \Rightarrow \frac{dE_{\mu_X}(P)}{d[E_{\mu_X}({}^1P) \otimes E_{\mu_X}({}^2P)]} = \frac{dE_{\mu_X}(P)}{d[E_{\mu_X}(P^1) \otimes E_{\mu_X}(P^2)]} \cdot \frac{dE_{\mu_X}(P)}{dE_{\mu_{X_1 \otimes X_2}}(P)} \,.$$

Démonstration :

(i) Evident

(ii) Si P est de la forme $P^1 \otimes P^2$ on peut poser :

$$\phi_{ijk\ell} = \phi^1_{ik} \cdot \phi^2_{j\ell}$$

et remarquant que :

$$\forall (i,j) \in I \times J \quad , \quad \sum_{k=1}^{r} \phi^1_{ik} = 1 \quad \text{et} \quad \sum_{\ell=1}^{s} \phi^2_{j\ell} = 1,$$

il vient alors :

$$\frac{dE_{\mu_X}(P)}{d[E_{\mu_X}({}^1P) \otimes E_{\mu_X}({}^2P)]}\,(\{y_1^k\}\times\{y_2^\ell\}) = \frac{\sum_{i,j} \phi^1_{ik}\,\phi^2_{j\ell}\,p_{ij}}{(\sum_i \phi^1_{ik}\,p_{i.})\times(\sum_j \phi^2_{j\ell}\,p_{.j})} = \frac{\sum_{i,j} \phi^1_{ik}\,\phi^2_{j\ell}\,p_{ij}}{\sum_{i,j} \phi^1_{ik}\,\phi^2_{j\ell}\,p_{i.}\times p_{.j}}$$

□

A.F.C. de (Y_1, Y_2)

On cherche donc l'A.F.C. de Y connaissant X. Pour ce faire il faut calculer l'A.C. de $E_{\mu_X}(P)$ qui est la loi de Y si P est le T-codage conditionnel. En fait les calculs présentés ci-dessous sont valables pour tout T-codage P pourvu qu'il se mette sous la forme élémentaire : $P = P^1 \otimes P^2$ (le codage de X est égal au produit des codages de X_1 et X_2) ; c'est-à-dire, dans le cas du codage conditionnel, si les hypothèses (H_1) et (H_2) sont vérifiées.

On note :

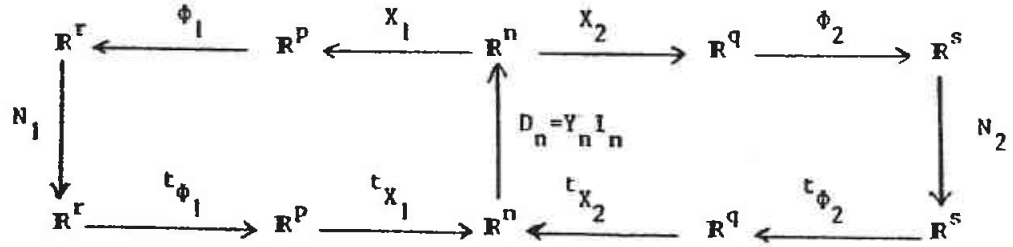
$${}^t\Phi_1 = \left[\phi^1_{ik}\right]_{\substack{i=1,\ldots,p \\ k=1,\ldots,r}} \qquad \text{et} \qquad {}^t\Phi_2 = \left[\phi^2_{j\ell}\right]_{\substack{j=1,\ldots,q \\ \ell=1,\ldots,s}}$$

$$C = X_1 D_n {}^t X_2 = \left[p_{ij}\right]_{\substack{i=1,\ldots,p \\ j=1,\ldots,q}} \quad \text{et} \quad C' = \left[q_{k\ell}\right]_{\substack{k=1,\ldots,r \\ \ell=1,\ldots,s}} .$$

C étant la table de contingence associée à X, C' est celle associée à Y et s'écrit aussi :

$$C' = \Phi_1 C {}^t\Phi_2 .$$

Le schéma de dualité associé à l'A.C. de $E_{\mu_X}(P)$ (équivalente à l'A.F.C. du couple (Y_1, Y_2)) est le suivant :



où $N_1 = \mathrm{diag}[(E_{\mu_X}({}^1P)(\{y_1^k\}))^{-1}]_{k=1,\ldots,r}$

et $N_2 = \mathrm{diag}[(E_{\mu_X}({}^2P)(\{y_2^\ell\}))^{-1}]_{\ell=1,\ldots,s}$.

L'A.F.C. cherchée est obtenue par l'analyse spectrale d'un opérateur ayant pour représentation matricielle :

$$N_1 \Phi_1 X_1 D_n {}^t X_2 {}^t\Phi_2 N_2 \Phi_2 X_2 D_n {}^t X_1 {}^t\Phi_1 = N_1 C' N_2 {}^t C'.$$

En pratique, et ce quelles que soient les hypothèses admises, il suffit donc de faire exécuter un programme classique d'A.F.C. sur le tableau C' de terme général $q_{k\ell}$.

Sous les hypothèses (H_1) et (H_2), cette matrice s'exprime simplement par : $\Phi_1 C {}^t\Phi_2$. L'analyse obtenue est équivalente à l'A.C.P. du tableau :

$$\begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

les espaces étant munis des métriques N_1 et N_2.

Remarque 1 :

Dans le cas très particulier où ${}^t[\mu(\{x_1^i\})]_{i=1,\ldots,p}$ et ${}^t[\mu(\{x_2^j\})]_{j=1,\ldots,q}$ sont respectivement vecteurs propres associés à la valeur propre 1 des matrices ${}^t\Phi_1$ et ${}^t\Phi_2$, J.F. MARTIN [8] a montré que μ_X était invariante par P, i.e. :

$$\mu_Y = E_{\mu_X}(P) = \mu_X .$$

Dans ce cas, et dans ce cas seulement, les analyses de X et de Y sont les mêmes, et nous dirons que l'A.F.C. de X est invariante par P.

Remarque 2 :

On peut appliquer cette méthode aux processus prévisionnels.

Dans l'hypothèse où X_1 et X_2 sont des chaînes de Markov de matrices de transition Φ_1 et Φ_2 , le point de vue adopté ici rejoint celui de A. YOUSFATE [13] : connaissant l'état initial de (X_1,X_2), on peut calculer les A.F.C. aux différents instants.

4.3 - Cas de l'A.F.C. généralisée

Position du problème

On se propose de généraliser ce qui précède à la situation suivante : soit $Y = (Y_1,\ldots,Y_m)$ une v.a. définie sur $(\Omega,\mathcal{A},\mu)$ à valeurs dans $(F = F_1 \times \ldots \times F_m$, $\mathcal{F} = \mathcal{F}_1 \otimes \ldots \otimes \mathcal{F}_m)$ et $X = (X_1,\ldots,X_m)$ à valeurs dans $(E = E_1 \times \ldots \times E_m$, $\mathcal{E} = \mathcal{E}_1 \otimes \ldots \otimes \mathcal{E}_m)$. Soit pour tout i, Card $E_i = p_i$ et nous utiliserons la définition de la m-A.F.C. codée donnée en 3.3 .

Cette analyse ne pose pas de problème théorique spécifique mais rend encore plus délicate la mise en oeuvre pratique à cause de la multiplication des hypothèses à considérer.

Hypothèses de simplification

Comme l'analyse de Y ne tient pas compte des relations (interactions) intervenant entre plus de deux variables, elle ne nécessite pas la connaissance du T-codage conditionnel P, mais seulement celle des marges ${}^{ij}P$ et donc une hypothèse du type de (H_1) n'est plus nécessaire.

On cherche donc à simplifier l'expression des m^2 transitions :

$$ {}^{ij}P : (E,\mathcal{E}) \times (\mathcal{F}_i \otimes \mathcal{F}_j) \longrightarrow ([0,1] , \mathcal{B}_{[0,1]})$$

$$(x,B_{ij}) \longrightarrow {}^{ij}P_x(B_{ij}) = \mu[Y \in B_{ij/X=x}].$$

5 - ESTIMATION DU CODAGE ET CONVERGENCE

5.1 - Estimation

En pratique, la loi conditionnelle de Y à X est inconnue et il est alors nécessaire de l'estimer sur un échantillon-test. Pour simplifier les notations nous considérons X et Y au lieu des couples (X_1,X_2), (Y_1,Y_2) et nous noterons $\Phi = [\phi_{ik}]$ (au lieu de $[\phi_{ijk\ell}]$) la matrice de codage.

On suppose que l'on observe un échantillon-test $(X_k,Y_k)_{k=1,\ldots,m}$ de (X,Y), puis un échantillon $(X_{k'})_{k'=m+1,\ldots,m+n}$ de X, indépendant du précédent.

Si M_{ik} désigne l'effectif de la modalité (x_i,y_k) de la variable (X,Y) pour l'échantillon de taille m et N_i l'effectif de la modalité x_i pour l'échantillon de taille m+n, l'estimateur du maximum de vraisemblance de ϕ_{ik} (voir J.F. MARTIN [8]) s'exprime :

$$\hat{\phi}_{ik} = \phi^m_{ik} = P^m_{x_i}(\{y_k\}) \begin{cases} = \dfrac{\sum_{h=1}^{m} \mathbb{1}_{\{x_i\}\times\{y_k\}} \circ (X_h,Y_h)}{\sum_{h=1}^{m} \mathbb{1}_{\{x_i\}} \circ X_h} & \text{si } \sum_{h=1}^{m} \mathbb{1}_{\{x_i\}} \circ X_h \neq 0 \\ = 0 \text{ sinon} \end{cases}$$

c'est-à-dire $\hat{\phi}_{ik} = \dfrac{M_{ik}}{M_{i.}}$.

L'estimateur de μ_X est alors : $\hat{\mu}_X(\{x_i\}) = \dfrac{M_{i.}+N_i}{m+n}$.

Ces deux estimateurs permettant de construire celui de la loi de Y :

$$\hat{\mu}_Y(\{y_k\}) = \sum_{i=1}^{p} \hat{\phi}_{ik}\, \hat{\mu}_X(\{x_i\}) = \frac{1}{m+n} \sum_{i=1}^{p} M_{ik} \frac{M_{i.}+N_i}{M_{i.}} .$$

5.2 - Convergence de T-codages

Position du problème

Soit P un T-codage sur $(E,\mathcal{E})$ relatif à $\mathcal{F}$ et $(X_n)_{n\in\mathbb{N}^*}$ une suite de v.a. définies sur $(\Omega,\mathcal{A},\mu)$, à valeurs dans $(E,\mathcal{E})$, indépendantes et de même loi que X. Pour tout n de $\mathbb{N}^*$, $(X_1,\ldots,X_n)$ est un n-échantillon de X auquel on peut associer le n-uplet de probabilités de transition $(P_{X_1},\ldots,P_{X_n})$ appelé échantillon T-codé.

On s'intéresse donc à la convergence, en un sens à préciser, de la suite $(\frac{1}{n}\sum_{i=1}^{n} P_{X_i})_{n\in\mathbb{N}^*}$ des moyennes empiriques des échantillons T-codés vers la probabilité T-codée $E_\mu(P_X)$.

De plus, comme le T-codage P est lui-même estimé, il faut étudier la double convergence de la suite

$$\left[\frac{1}{n}\left(\sum_{i=1}^{n} P_{X_i}^{m}\right)\right]_{(m,n)\in\mathbb{N}_*^2} .$$

Convergence vague et convergence faible

Soient P et $(P^m)_{m\in\mathbb{N}^*}$ des T-codages sur $(E,\mathcal{E},\nu)$ relatifs à $\mathcal{F}$. On peut définir comme le fait J.F. MARTIN [8] une "convergence vague ν-presque sûre (ν-p.s.)".

Définition 6 :

On dit que la suite $(P^m)_{m\in\mathbb{N}^*}$ converge vaguement ν-p.s. vers P si pour tout B de $\mathcal{F}$ la suite de v.a. $(P^m(B))_{m\in\mathbb{N}^*}$ converge ν-p.s. vers la v.a. P(B), i.e. :

(1) $\quad \forall B \in \mathcal{F} \qquad \nu\{x \in E / \lim_{m\to+\infty} P_x^m(B) = P_x(B)\} = 1.$

Généralement (1) est moins exigeante que la propriété :

(2) $\quad \nu\{x \in E / \forall B \in \mathcal{F} \quad \lim_{m\to+\infty} P_x^m(B) = P_x(B)\} = 1.$

Cependant, ici $\mathcal{F}$ étant une tribu finie, il est clair que (1) et (2) sont équivalentes. On utilisera donc (2) et on a la :

Définition 7 :

$(P^m)_{m\in\mathbb{N}^*}$ converge faiblement ν-p.s. vers P si pour presque tout x de E la suite de probabilités $(P_x^m)_{m\in\mathbb{N}^*}$ converge faiblement vers P_x (i.e. (2)).

Remarque :

Nous utilisons la notion de convergence faible puisque, $\mathcal{F}$ étant finie, les indicatrices $\mathbb{I}_B$ sont continues.

5.3 - Convergence de l'analyse

Convergence du T-codage conditionnel

En reprenant les notations de 5.1 -, le T-codage conditionnel est déterminé par :

$$P_{x_i}(\{y_k\}) = \frac{1}{\mu_X(\{x_i\})} \int_{[X=x_i]} \mathbb{I}_{\{y_k\}} \circ Y \; d\mu \; ;$$

Soit (H_{ij}) : ${}^{ij}P_x(B_{ij}) = \mu[Y \in B_{ij}/(X_i,X_j)=(x_i,x_j)]$

Ce sont des hypothèses du même type que (H_2) permettant de définir la transition ${}^{ij}P$ seulement sur l'espace $E_i \times E_j$ au lieu de E tout entier. Sommairement (H_{ij}) signifie que l'erreur commise sur la "mesure" de (X_i,X_j) est indépendante des valeurs prises par les autres variables X_k $(k\neq i$ et $k\neq j)$.

Ainsi si pour tout couple (i,j), l'hypothèse (H_{ij}) est vérifiée, le problème se ramène alors à m^2 fois la situation précédente de l'A.F.C. d'un couple de v.a..

Pour chacun des couples (X_i,X_j), on peut considérer les hypothèses (H_1^{ij}) et (H_2^{ij}) comme définies en 4.1 - et évaluer alors :

$$V'_{ij} = E_{\mu_X}({}^{ij}P)$$

(dont la simplicité relative dépend de l'admissibilité de (H_1^{ij}) et (H_2^{ij})).

m-analyse

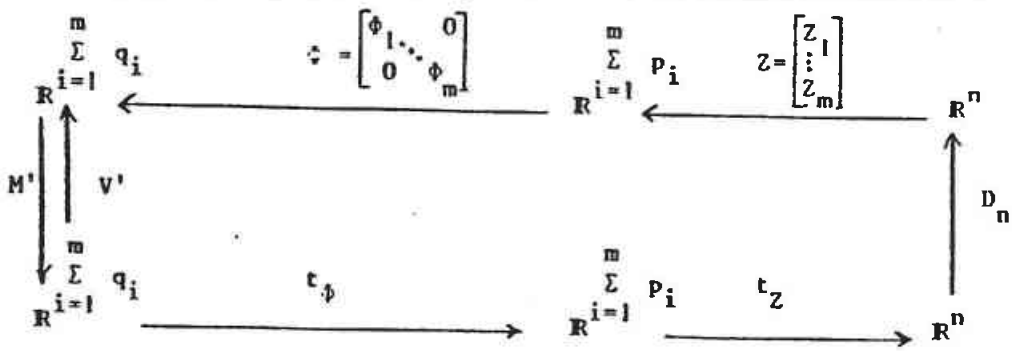
L'ensemble de ces blocs V_{ij} permet de construire la matrice à diagonaliser :

$$M' \circ V' = \begin{bmatrix} V'^{-1}_{11} & & 0 \\ & \ddots & \\ 0 & & V'^{-1}_{mm} \end{bmatrix} \times \left[V'_{ij}\right]_{(i,j)\in I^2}$$

Sous le jeu complet des hypothèses :

$$\forall(i,j) \in I^2 \quad (H_{ij}) \wedge (H_1^{ij}) \wedge (H_2^{ij})$$

et en se référent au paragraphe 3.3 - , on obtient le schéma de dualité :



où $V' = \Phi\, V\, {}^t\Phi$.

Ce schéma correspond à l'A.C.P. des variables $(\phi_1 X_1, \ldots, \phi_m X_m)$ lorsque l'espace des "individus" est muni de la métrique M'.

codage que l'on peut estimer sur l'échantillon-test $(X_i,Y_i)_{i=1,\dots,m}$ par :

$$P^m_{x_i}(\{y_k\}) = \begin{cases} \dfrac{\sum_{j=1}^{m} \mathbb{1}_{\{x_i\}\times\{y_k\}}\circ(X_j,Y_j)}{\sum_{j=1}^{m} \mathbb{1}_{\{x_i\}}\circ X_j} & \text{si } \sum_{j=1}^{m} \mathbb{1}_{\{x_i\}}\circ X_j \neq 0 \\ 0 \quad \text{sinon.} \end{cases}$$

Les tribus $\mathcal{E}$ et $\mathcal{F}$ étant finies, la suite de probabilités $(P^m_x)_{m\in\mathbb{N}^*}$ converge faiblement vers P_x uniformément en x sur $\mathcal{E}$ et on peut donc appliquer la proposition 6 p. 74 de J.F. MARTIN [8] montrant ainsi que :

$$\lim_{\substack{m\to+\infty \\ n\to+\infty}} \frac{1}{n} \sum_{i=1}^{n} P^m_{x_i} = E_\mu(P_X) \text{ faiblement } \mu\text{-p.s.}$$

Convergence de l'analyse

En utilisant le résultat ci-dessus pour $X = (X_1,X_2)$ et $Y = (Y_1,Y_2)$, on peut appliquer la proposition 9 p. 291 de J. DAUXOIS et A. POUSSE [4]. On en déduit la convergence uniforme p.s. de la suite des analyses obtenues par échantillonnage vers celles de Y et ainsi , la convergence des éléments propres (valeurs propres et facteurs canoniques de l'analyse) .

6 - CONCLUSION

Les outils développés permettent donc bien de résoudre théoriquement le problème concret posé en [1]. En pratique il faut être conscient que les difficultés dues à l'estimation de la probabilité de transition sont délicates à traitées. On a vu que pour simplifier ces questions il fallait pouvoir décomposer le codage P en un produit de ses marges, ceci ne peut se justifier alors que par la considération d'hypothèses de type probabiliste. Ainsi apparaît la nécessité d'introduire des outils adéquats (probabilités de transition) pour modéliser la notion de codage dans le cadre des analyses factorielles non-linéaires.

D'autres types d'application que celle proposée en [1] peuvent être suggérées. Il est fréquent de vouloir réactualiser une enquête ou un sondage sans pour autant interroger un échantillon complet. Il suffirait alors de ne réenquêter qu'un sous-échantillon de l'échantillon initial permettant d'estimer une probabilité de transition et donc, d'actualiser les représentations factorielles.

Dans un autre ordre d'idée, on considère les hypothèses proposées par YOUSFATE 15 (chaîne de Markov homogène d'ordre 1) dans le but de prévoir un processus qualitatif. Alors, connaissant les transitions de 2 ou plusieurs de ces processus on peut prévoir les représentations factorielles (A.F.C. ou m-A.F.C.) associées à ces processus.

REFERENCES

[1] - BESSE P. : "*Recherche statistique d'une typologie des descriptions de phénomènes aérospatiaux non identifiés*".
CNES/GEPAN - note technique n° 4 - TOULOUSE (1981)

[2] - BORDET J.P. : "*Etude de données géophysiques. Modélisations statistiques par régression factorielle*".
Thèse de 3° cycle - Université de PARIS VI,(1973).

[3] - CAZES P. - BAUMEDER A. - BONNEFOUS S. - PAGES J.P. : "*Codage et tableaux des Données Logiques - Introduction à la pratique des variables qualitatives*".
Cahiers du B.U.R.O. n° 27 - PARIS (1977).

[4] - DAUXOIS J. - POUSSE A. : "*Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique*".
Thèse - Université de TOULOUSE III (1976).

[5] - GAUTIER J.M. - SAPORTA G. : "*About fuzzy discrimination*".
COMPSTAT 82 - Proceedings in Computational Statistics, Physica-Verlag, WIEN (1982).

[6] - LAFAYE DE MICHEAUX D. : "*Approximation d'analyses canoniques non-linéaires de variables aléatoires et analyses factorielles privilégiantes*".
Thèse de docteur ingénieur - Université de NICE (1978).

[7] - MALLET J.L. : "*Propositions for fuzzy characteristic functions in data analysis*".
COMPSTAT 82 - Proceedings in Computational Statistics - Physica-Verlag, WIEN (1982).

[8] - MARTIN J.F. : "*Le codage flou et ses applications en statistique*".
Thèse de 3° cycle - Université de PAU et des pays de l'Adour (1980).

[9] - MARTIN J.F. : "*Le codage aléatoire - Utilisation en statistique*".
Journées de Statistiques - NANCY (1981).

[10] - NEVEU J. : "*Bases mathématiques du calcul des probabilités*".
MASSON - PARIS (1964).

[11] - RAMSAY J.O. - WINSBERG S. : "*Monotonic transformations to additivity using splines*".
Biométrika - 1980, 67, 3, pp. 669-74.

[12] - RAMSAY J.O. - WINSBERG S. ; *"Analysis of pairwise preference data using B-splines"*.
Psychometrika - vol. 46, N° 2, june, 1981. pp 171-186

[13] - RIJCKEVORSEL (VAN) J. : *"Canonical analysis with B-splines"*.
COMPSTAT 82 - Proceedings in Computational Statistics, Physica-Verlag, WIEN (1982).

[14] - TITTERINGTON D.M. : *"A comparative study of kernel-based density estimates of categorical data"*.
TECHNOMETRICS - vol. 22, n° 2, mai 1980. pp 259-268

[15] - YOUSFATE A. : *"Analyses factorielles des processus qualitatifs de type Markovien. Description et prévision"*.
Thèse de 3° cycle - Université de TOULOUSE III (1981).