OFFICE DE LA RECHERCHE SCIENTIFIQUE ET TECHNIQUE OUTRE-MER
(O.R.S.T.O.M.)

Section de Démographie

Document de travail n° 6

L'EXPLOITATION INFORMATIQUE EN DÉMOGRAPHIE

Jacques VAUGELADE

janvier 1978

Paris Mars 1978 ANT

Rectificatif de la NI n° 54

1°)- Ligne 9, page 1, lire: "Ce code carte est limité à 2 caractères" au lieu de 2 caractères.

2°)- Page de verso, 1e ligne: "Le programme prévoit une autre possibilité peu utilisée par les démographes, c'est un caractère spécial de l'indicatif pour indiquer la dernière unité, par exemple le dernier individu du ménage."

3°)- Page de verso, ligne 18: "des unités valides" au lieu de unités invalides.

J. Vangelade.

L'EXPLOITATION INFORMATIQUE EN DEMOGRAPHIE

Par J. VAUGELADE (*)

S O M M A I R E

Annexes

(*) Ce texte est l'édition révisée du chapitre "L'exploitation" du même auteur publié en 1973 dans Sources et Analyse des données démographiques, 1ère partie, ORSTOM, INED, INSEE, MIN-COOP.

INTRODUCTION

Les techniques de l'exploitation doivent être un outil pour le responsable d'une enquête. Toutefois, les impératifs d'une exploitation sont astreignants. La meilleure connaissance de ces contraintes laisse une plus grande liberté au responsable d'une enquête dans la définition de ses objectifs.

Que les renseignements viennent d'une enquête ou du dépouillement d'un système administratif (état civil...), nous parlerons d'un questionnaire qui est rempli pour chaque individu (ou unité statistique).

Ces renseignements standardisés peuvent être présentés en tableaux. La présente note a pour objet l'exploitation qui va du questionnaire au tableau en passant par la carte perforée et l'ordinateur qui sont les moyens actuellement les plus employés.

Dans une enquête, l'exploitation est une phase qui se situe après la conception et sa réalisation et avant l'analyse et la rédaction. Mais elle présente la particularité de faire appel à des connaissances qui sortent du domaine habituel d'un responsable d'enquête.

Cette situation conduit à faire de l'exploitation un maillon souvent faible alors que des données d'excellente qualité ont été recueillies. Or, c'est le maillon le plus faible qui donnera sa valeur à l'enquête. Cette note s'adresse donc plus particulièrement à des responsables d'enquêtes qui ont peu de connaissances en informatique. Les différentes phases de l'exploitation seront envisagées et, en tenant compte des contraintes de l'informatique, on essaiera d'indiquer les solutions qui paraissent les meilleures. Les techniques d'enquête sont supposées connues.

On se limitera au tableau statistique qui résulte directement de l'enquête et qui exclut donc les calculs d'analyse (table de mortalité...). De tels tableaux peuvent être obtenus :

. soit manuellement :
  . en faisant de petits bâtons dans des cases
  . par des fiches à perforations marginales qui présentent un intérêt certain pour des enquêtes portant sur quelques centaines d'unités
. soit de manière automatique,
  . avec des cartes perforées et des machines mécaniques (trieuse, interclasseuse, tabulatrice) qui ont aujourd'hui pratiquement disparu
  . en utilisant l'informatique dont le moyen central est l'ordinateur. Ce sera le seul sujet abordé ici.

Au niveau des principes, le nombre d'unités statistiques importe peu. Au stade de l'exploitation, il est fondamental et influe sur les solutions retenues. Pour fixer les idées, nous admettrons que 5 000 unités constituent un petit fichier et 500000 un grosfichier.

Une annexe aborde quelques problèmes d'informatique qu'il peut être utile de connaître. Dans ce texte, le mot carte désigne la carte perforée pour ordinateur (voir § 2, la saisie des données).

.../...

I. LA CODIFICATION

C'est l'opération qui consiste à transformer les réponses du questionnaire en caractères généralement numériques, des chiffres (1).

. Le plus souvent, chaque question constitue une variable. Le code est la correspondance entre les possibilités de réponse et les postes de la variable.

Exemple : question = variable : sexe
réponses possibles : masculin, féminin, non déclaré
code à trois postes : 1 = masculin, 2 = féminin, 3 = non déclaré (ou 9 = non déclaré).

Dans ce cas, un seul chiffre suffit pour la variable, mais une variable peut nécessiter plusieurs chiffres.

Il faut, autant que possible, limiter le nombre de codes différents. Si plusieurs questions ont pour réponse Oui-Non, il faut impérativement qu'elles aient le même code (par exemple 1 = oui, 2 = non). Pour "Inconnu", il est souvent commode de prendre un poste fixe. 9, par exemple, est souvent utilisé.

. Parfois, on fait correspondre une seule variable à plusieurs questions.

Exemple :

Dans le cas d'une question du type "Oui, Non, ne sait pas", suivie d'une question "Si Oui, pourquoi ?", il est préférable de réunir ces deux questions en une seule variable avec les postes :

1 = Non
3 = Oui, première raison
4 = Oui, deuxième raison
etc...
9 = ne sait pas.

L'exploitation en sera facilitée car on s'intéresse souvent aux deux questions simultanément.

. On peut également faire correspondre plusieurs variables à une seule question. On en verra un exemple dans le cas de réponses multiples.

.../...

---

(1) on peut concevoir des codes alphabétiques (par exemple pour le sexe : M, F). Mais, en l'absence d'assurances certaines de l'informaticien, il est plus prudent de se limiter à ces codes numériques, ce qui facilite la perforation et l'exploitation.

I.1. Les types de code. Deux catégories : qualitatifs et quantitatifs.

I.1.1. Codes qualitatifs : simple ou emboîté

Un code qualitatif peut être simple. C'est le cas de la variable sexe donnée comme exemple ci-dessus.

Un code peut être emboîté. Par exemple, pour un code sur la cause de décès à deux chiffres, le premier indique des grands groupes (accidents, maladies...), le deuxième permet de préciser la nature de l'accident dans le cas d'un accident, ou la maladie précise dans le cas d'une maladie. Cependant, le deuxième chiffre pris isolément n'a pas de sens.

Dans lescas décrits ci-dessus, les possibilités sont exclusives : la réponse ne peut être à la fois : masculin et féminin, oui et non. Ce n'est pas le cas d'une question portant sur les langues parlées, un individu pouvant connaître plusieurs langues.

Ce genre de question, avec des réponses multiples, peut être codifié de plusieurs façons :

a) les codes binaires sont quelques fois employés : chaque possibilité est repérée par une puissance de 2.

Exemple :

Si l'on a trois codes : 1 = anglais ; 2 = espagnol ; 4 = français, les combinaisons seront les suivantes :

0 = personne qui ne parle aucune langue
1 = anglais seul
2 = espagnol seul
3 = (1 + 2) pour anglais et espagnol
4 = français seul
5 = (4 + 1) pour français et anglais
6 = (4 + 2) pour français et espagnol
7 = (4 + 2 + 1) pour les trois langues
9 = inconnu.

C'est un système parfait (1), mais la codification est difficile et l'exploitation peu commode quand le nombre des possibilités augmente. L'exemple donné est un cas commode car le nombre de possibilités est faible et tous les croisements sont possibles avec une seule variable.

b) une autre solution est de réserver une variable pour chaque possibilité, chaque variable ayant les postes 1 = Oui, 2 = Non, 9 = inconnu. Cette solution convient également quand le nombre de possiblités n'est pas trop élevé; c'est encore un système parfait.

.../...

(1) on appelle système parfait celui qui ne perd aucune information.

c) les deux première solutions sont satisfaisantes quand il y a peu de possibilités. Lorsque les possiblités sont plus nombreuses, une troisième solution est donc de prévoir un nombre maximum de possibilités par individu et un nombre identique de variables.

En reprenant l'exemple a), si on estime que peu de personnes parlent les trois langues, on se limitera à deux variables : une pour la première langue et une autre pour la deuxième langue. Chaque variable aura le même code : 0 = néant ; 1 = anglais ; 2 = espagnol ; 3 = français ; 9 = inconnu. S'il arrive qu'un individu parle trois langues, on ne peut en codifier que deux et il y a une perte d'information (on peut, pour cet individu, tirer au hasard la langue qu'on ne codifie pas pour éviter de biaiser ce renseignement). C'est donc un système non parfait. Mais si le nombre maximum est bien choisi, la perte d'information doit se produire dans un très petit nombre de cas et il est possible de l'admettre. C'est une bonne solution quand les possibilités sont assez nombreuses mais que seul un petit nombre d'entre elles peuvent se produire simultanément.

Cette dernière solution peut être rendue parfaite par le recours à un nombre changeant de variables ; le traitement n'en est pas simple et cette zone changeante devra être placée en fin de carte perforée.

I.1.2. Codes quantitatifs

Ils concernent les questions dont la réponse est déjà un nombre ; par exemple, l'âge ou le nombre d'enfants. En général, on prend les valeurs possibles comme postes de la variable. Un poste "inconnu" doit être prévu à moins d'imposer l'estimation des réponses inconnues. Il faut prévoir un poste maximum auquel sont ramenés tous les nombres supérieurs.

Exemple :

Si l'âge est codifié avec deux chiffres, il faut adopter une convention pour les âges des personnes de 100 ans et plus. En général, on se fixe un maximum auquel sont ramenés les âges supérieurs. Ainsi, on pourra codifier les 98, 99 et 100 ans et plus à 98 ; 99 étant réservé pour inconnu.

Si l'on s'intéresse à la différence d'âge entre époux, on peut soit calculer cette différence et la codifier, soit, ce qui est préférable, effectuer le calcul à l'ordinateur, à partir de chacun des âges (voir §1-4, possibilités de recodification).

Cependant, si l'on veut codifier une quantité négative, on peut ajouter un nombre assez grand (par exemple 10) pour que toutes les quantités deviennent positives. Dans ce cas, le poste 0 correspondra aux quantités égales à moins de 10 et aux quantités négatives plus petites que moins 10 (moins 11, moins 12...). On peut bien sur codifier le signe avec un caractère (+ ou -) ou un chiffre 1 = +, 2 = -) supplémentaire.

Plutôt que de codifier la réponse telle quelle sans transformation, il est possible d'effectuer un regroupement de classe (par exemple 1 = 0-4 ans ; 2 = 5-9 ans...) ou de faire tout calcul jugé souhaitable (par exemple exprimer un résultat en pourcentage plutôt qu'en valeur brute...). Mais, souvent, le procédé est peu employé puisque ces regroupements sont réalisables à l'ordinateur à partir des données brutes (§1.4).

.../...

I.2. Problèmes liés à l'identification des documents

L'identification permet de relier l'individu enquêté au questionnaire et ce dernier aux données informatiques. Si sur le questionnaire le nom et l'adresse de l'individu sont suffisants, pour les données informatiques, l'identifiant est une variable particulière (1)

Elle est nécessaire pour revenir au document de base pour certaines corrections et quand plusieurs questionnaires sont à relier entre eux.

Alors que pour les autres variables les corrections sont relativement faciles, pour les identifiants les erreurs sont graves par leurs conséquences. Elles peuvent conduire à considérer comme appartenant à un individu des cartes d'un autre individu, à la suite, par exemple, de la permutation de deux chiffres du numéro d'identification.

Que ces erreurs soient dues au chiffrement ou à la perforation, le principe pour résoudre ce problème est celui de la redondance de l'identification. En voici quatre exemples :

a) dans une zone d'enquête (ou grappe), les individus sont numérotés en séquence à partir de 1, les unités d'habitation également. Les unités d'habitation sont des renseignements **redontants** pour identifier les cartes d'un individu, le numéro d'individu suffirait. Pour toutes les cartes d'un individu, on répète les deux numéros; ainsi, s'il y a une erreur sur l'un des numéros, on a de grandes chances de s'en apercevoir. Il y a peu de chances d'avoir une erreur sur l'autre numéro qui fasse que les deux numéros erronés soient justement ceux d'un autre individu.

b) dans l'exemple ci-dessus, on peut en outre numéroter les individus dans chaque unité d'habitation. On a dans ce cas une double identification à l'intérieur de chaque zone. Mais on peut avoir des erreurs sur l'identification de la zone. Le problème est abordé à la fin de ce paragraphe.

c) on peut en plus de l'identification répéter pour chaque carte des caractéristiques essentielles de l'individu comme le sexe et l'année de naissance.

d) on peut enfin utiliser une clé. Par exemple, si l'identification est un code emboîté (région, commune, unité d'habitation, individu), on considère l'ensemble des chiffres comme un seul nombre et la clé est le reste de la division de ce nombre par, par exemple, 97. Soit un numéro d'identification 1874, le reste de la division par 97 est 04 (1874 = 20 x 97 + 4). On considèrera en fait 187 404 comme d'identifiant. S'il y a une erreur (écriture ou perforation), par exemple, en écrivant 787 404, la clé est fausse, ce devrait être 787 417.

L'utilisation d'une clé est commode pour le traitement; cela permet de rejeter à priori une carte, alors que l'utilisation d'une autre information redondante ne permet pas en cas de discordance de connaître la mauvaise carte. Le calcul des clés est le système le plus coûteux à établir et il se justifie mieux s'il s'agit d'un fichier permanent (enquête à passages répétés de longue durée) car dans ce cas une détérioration progressive et cumulative du fichier serait catastrophique. (L'utilisation du seul reste de la division par 9 est déconseillée car il ne permet pas de rendre compte de l'inversion de deux chiffres; pour une clé à un chiffre, il vaut

.../...

(1) Plus facilement que les autres variables, l'identifiant peut comprendre des lettres.

mieux choisir le reste de la division par 7).

Il est conseillé de classer les documents par grappe, ou zone géographique, surtout quand cette information intervient dans l'identification. Il suffit d'indiquer que les colonnes correspondantes (si possible consécutives) sont identiques sur toutes les cartes. Cela pourra être fait en série à la perforation et éliminera tout risque d'erreur. Du moins, s'il y a erreur, elle concernera tout un groupe ou toute une zone et la correction sera relativement aisée.

I.3. Le chiffrement

C'est l'opération manuelle qui assure la transformation des réponses aux questions en variables codées (en chiffres ou lettres). Dans les enquêtes bien préparées, le chiffrement se fait sur le questionnaire. Le contrôle est alors plus aisé. Mais si la préparation est insuffisante ou si l'on veut réduire la dimension du questionnaire pour des raisons pratiques, on doit préférer la codification sur des feuilles indépendantes du questionnaire (voir les exemples de feuilles de chiffrement).

L'ordre des variables est important. Cet ordre doit faciliter le travail des chiffreurs. Mais, quel que soit l'ordre retenu, il a peu d'influence sur la programmation.

Les postes de chaque variable et la correspondance avec les réponses aux questions sont décrits sur un code. Au cours du chiffrement, il est souvent nécessaire, pour des cas omis, de rajouter des postes à des variables ou de modifier la signification d'un poste en y incluant les réponses omises. Ces modifications doivent être répertoriées au fur et à mesure. Un document final reprenant les codes et leurs modifications doit nécessairement être réalisé sous peine de rendre l'exploitation périlleuse et l'interprétation des résultats hasardeuse.

Dans le cas de plusieurs cartes par individu, le chiffrement sur une feuille par individu permet de n'écrire qu'une seule fois l'identification. Lors de la perforation, on procèdera également à une seule perforation et à la reproduction de l'identifiant sur les cartes suivantes. Les risques d'erreur lors de l'écriture ou de la perforation sont ainsi minimisés.

Dans un atelier de chiffrement, l'affichage des différents codes sur des panneaux muraux permettra d'éviter au maximum au chiffreur la consultation répétée d'un document souvent épais.

Quand la codification est un peu lourde, il est préférable de diviser le travail. Cela permet de donner aux chiffreurs les plus compétents les parties les plus difficiles.

Il est inutile d'insister sur l'importance d'un contrôle systématique.

Quand des erreurs sont faites au chiffrement, il faut proscrire les surcharges, l'erreur doit être barrée et l'information exacte réécrite au-dessus. (Une autre solution est d'utiliser des morceaux de papier adhésif qui existe en plaques et en ruban).

.../...

I.4. La recodification

La recodification est l'opération qui consiste, à partir de variables codifiées, à calculer de nouvelles variables.

On peut, par exemple, recodifier des groupes d'âges à partir de l'âge, l'âge à un évènement à partir de la date de naissance et de la date de l'évènement, le nombre d'enfants si le numéro de la mère figure sur les cartes d'enfants...

Les variables de départ peuvent se trouver sur la même carte ou sur plusieurs cartes. Ce dernier cas peut être assez complexe et coûteux à l'ordinateur alors que c'est parfois si simple au chiffrement. L'avantage de l'ordinateur est de ne pas exclure de possibilités et de retarder ou de permettre la révision des décisions prises. La recodification par ordinateur est parfaitement fiable et ses possibilités sont illimitées pour autant que l'information de base soit codifiée. Il faut arbitrer entre le coût d'une codification supplémentaire et le coût d'une exploitation plus lourde. Le problème est à discuter avec l'informaticien qui s'occupe de l'enquête.

I.5. Conclusion

On trouve ci-joint plusieurs exemples de documents de codification.

Le premier questionnaire, socio-démographique, avec chiffrement sur la partie droite du questionnaire de base. On remarquera que la lettre d'identification de la série de cartesest pré-imprimée ainsi que le code de carte. Les questions 11 et 12 sont des questions fermées avec une liste de réponses possibles. Dans ce cas, la réponse peut être directement notée dans la case correspondante. Cette précédure peut être une source d'erreur : l'enquêteur, après avoir posé la bonne question et entendu la bonne réponse, peut se tromper de numéro en reportant dans la case de chiffrement. Cette erreur peut également se produire avec des cases à cocher en face des bonnes réponses. Il est préférable d'obliger l'enquêteur à écrire la réponse, en abrégé pour les réponses les plus courantes ; les erreurs de report sur le questionnaire sont ainsi minimisées.

Le deuxième exemple (fiche résumé migration) concerne toutes les migrations d'un individu. Chaque ligne correspond à une migration et une carte est perforée pour chaque ligne. La zone d'identification (colonnes 1 à 16) est commune à toutes les cartes d'une même fiche et est reproduite sur chaque carte.

Le troisième exemple concerne une feuille de chiffrement indépendante du questionnaire. Chaque individu est représenté par une ligne. Les colonnes identiques pour tous les individus (colonnes 1 à 3) sont écrites une seule fois.

Le quatrième exemple est du même type que le troisième avec trois lignes pour chaque individu.

Dans les deux premiers exemples, la codification doit être mise au point en même temps que le questionnaire. Avec une feuille de chiffrement, la mise au point de la codification peut éventuellement être reportée. Pour les questions ouvertes dont on ne connaît pas la liste des réponses, un

.../...

dépouillement manuel sur un échantillon doit fournir une liste des réponses les plus fréquentes avant de commencer la codification. L'idéal est d'établir cette liste d'après les documents de la pré-enquête.

Dans tous les cas, il faut bien distinguer la codification des non-réponses ou non-déclarés de celle des non-concernés. Par exemple, il faut prévoir un poste du code (éventuellement le blanc) pour les hommes si une question concerne les seules femmes. Dans les enquêtes sociologiques, on distingue en général la réponse "ne-sait-pas" du refus de répondre et du non-concerné.

Questionnaire Socio-Démographique – Urbain-DAKAR

Nom enquêteur :

Date de l'interview :

Durée de l'interview :

PARTIE I

IDENTIFICATION DU MIGRANT SERER A DAKAR

1) Commune — U (1) — 0 2 (2 3)
2) Quartier — (4 5)
3) Parcelle n° — (6 7)
4) Chef de parcelle
5) Nom N° — (8 9 10)
6) Prénom
7) Sexe 1.M 2.F Situation matrimoniale — (11) (12)
8) Age — (13 14)
9) Tim (Matriclan)
10) Village d'origine dans Niakhar — (15 16)
11) Religion — (17)

1. Catholique
2. Musulman
3. Religion traditionnelle
4. Sans religion
5. Autres
6. Inconnu

12) Caste — (18)

1. Forgeron
2. Griot
3. Cordonnier
4. Bûcheron
5. Noble
6. Paysan
7. Autres
8. Inconnu

Description de la carte 02 : Questionnaire socio-démographique

Colonne	Variable	N° question	Code
1	Identification de la série de carte		U
2-3	Code de carte		02
4-5	Quartier et commune	1,2	Voir code quartier
6-7	N° parcelle		01 à 99
8-10	N° d'individu		001 à 310
11	Sexe (toujours connu)	7	1 Masculin 2 Féminin
12	Situation matrimoniale	7	0 Célibataire 1 Marié 1 ép. ou Mariée 2 Marié 2 ép. 3 Marié 3 ép. 4 Marié 4 ép. et + 5 Veuf divorcé 9 inconnu
13-14	Age en années révolues	8	01 à 97 98 = 98 et + 99 = inconnu
15-16	Village d'origine	10	Voir liste des villages
17	Religion	11	Voir questionnaire
18	Caste	12	Voir questionnaire

ORSTOM

Section de démographie

Enquête N GAYOKHEM

FEUILLE DE CODIFICATION

POPULATION INITIALE

Codifieur :

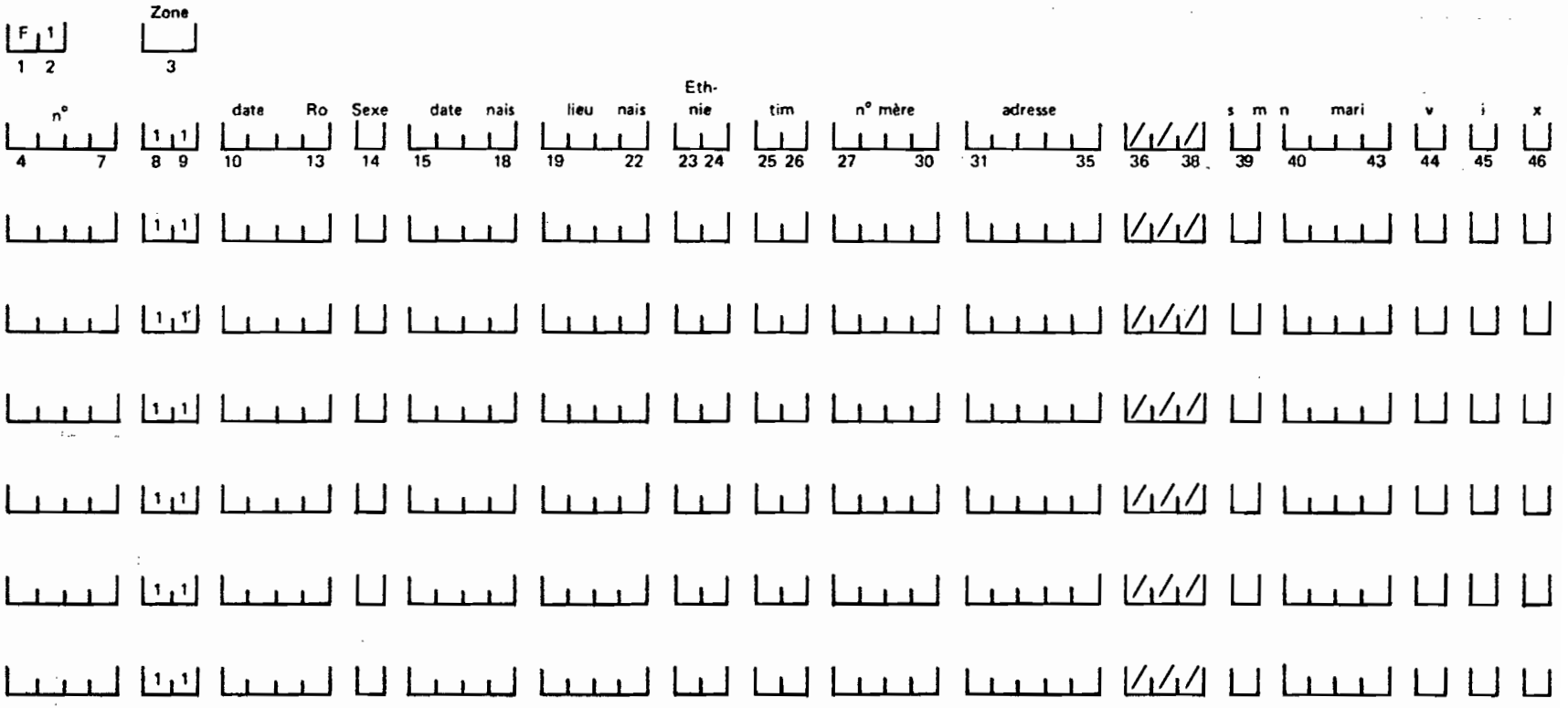
Contrôleur :

Mai 1971

Type 1

Village :

N° feuille :



FICHE RESUME MIGRATION C

Village .... N° V .... [3] [2 — 6]

CZ .... N° Z .... [7 — 11]

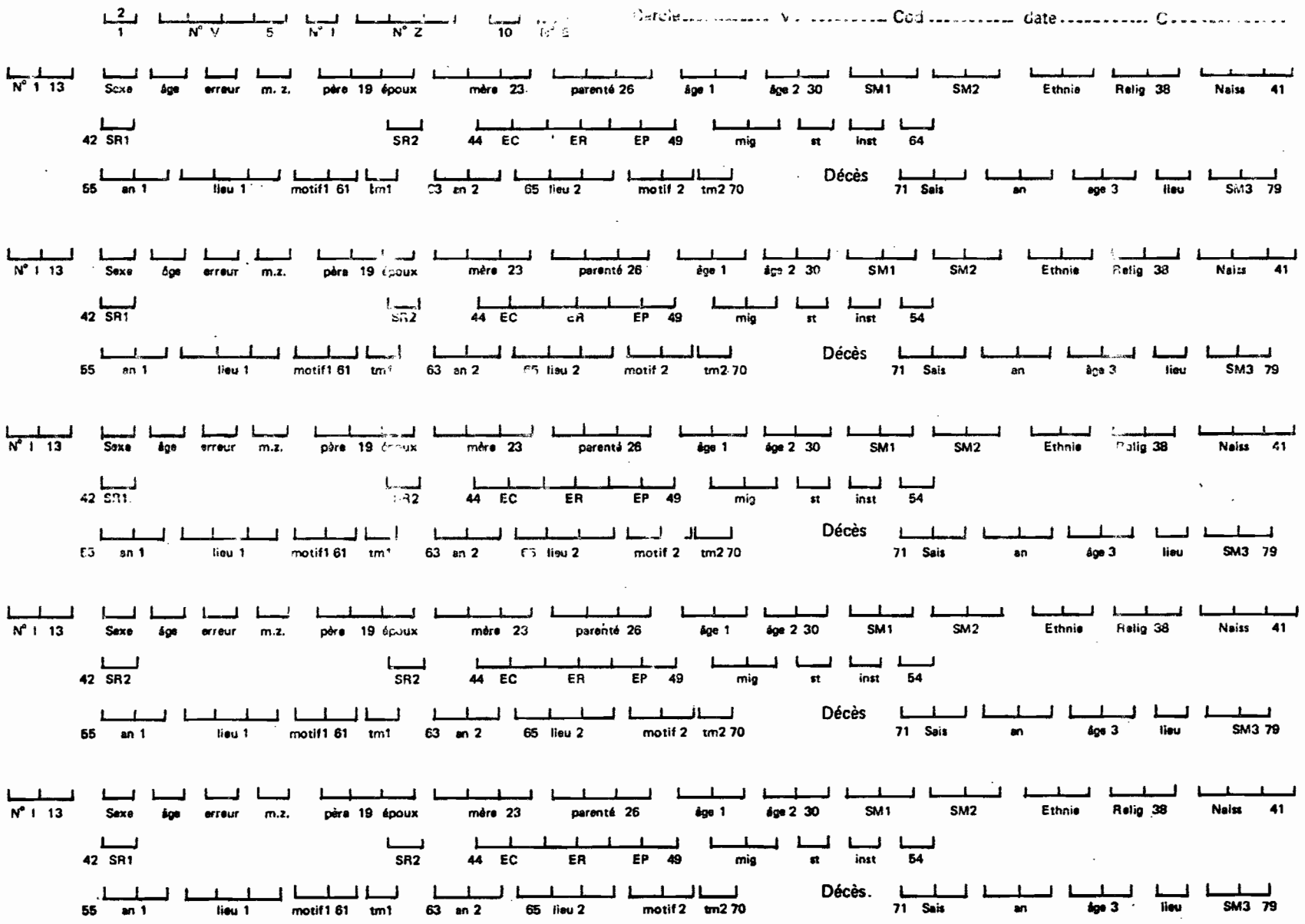
Nom .... N° I .... [12]

Sexe .... [14] Age en 1960 [15] [16]

Répondant lui-même
Autre nom ....
Parenté ....

date enquête ....
enquêteur ....
date contrôle ....
contrôleur ....
codifieur ....
date codification ....

N° MIG.	Départ					Séjour				Retour				Evénements matrimoniaux		Observations
	Nb d'années	Saison	SM	Accomp. Ep. Enf.	AF CE CZ	Lieu	V B	Durée	Emploi	Nb. d'années	Saison	SM	Accomp. Ep. Enf.	Nature	Date	
17	19			23	26	27		31	34	36		40	41	43		48
17	19			23	26	27		31	34	36		40	41	43		48
17	19			23	26	27		31	34	36		40	41	43		48
17	19			23	26	27		31	34	36		40	41	43		48
17	19			23	26	27		31	34	36		40	41	43		48


2
1
N° V
5
N° I
N° Z
10
Cod
date
N° I 13
Sexe
âge
erreur
m. z.
père 19
époux
mère 23
parenté 26
âge 1
âge 2 30
SM1
SM2
Ethnie
Relig 38
Naiss 41
42 SR1
SR2
44 EC
ER
EP 49
mig
st
inst
64
55 an 1
lieu 1
motif1 61
tm1
63 an 2
65 lieu 2
motif 2
tm2 70
Décès
71 Sais
an
âge 3
lieu
SM3 79
N° I 13
Sexe
âge
erreur
m.z.
père 19
époux
mère 23
parenté 26
âge 1
âge 2 30
SM1
SM2
Ethnie
Relig 38
Naiss 41
42 SR1
SR2
44 EC
ER
EP 49
mig
st
inst
54
55 an 1
lieu 1
motif1 61
tm1
63 an 2
65 lieu 2
motif 2
tm2 70
Décès
71 Sais
an
âge 3
lieu
SM3 79
N° I 13
Sexe
âge
erreur
m.z.
père 19
époux
mère 23
parenté 26
âge 1
âge 2 30
SM1
SM2
Ethnie
Relig 38
Naiss 41
42 SR1
SR2
44 EC
ER
EP 49
mig
st
inst
54
55 an 1
lieu 1
motif1 61
tm1
63 an 2
65 lieu 2
motif 2
tm2 70
Décès
71 Sais
an
âge 3
lieu
SM3 79
N° I 13
Sexe
âge
erreur
m.z.
père 19
époux
mère 23
parenté 26
âge 1
âge 2 30
SM1
SM2
Ethnie
Relig 38
Naiss 41
42 SR2
SR2
44 EC
ER
EP 49
mig
st
inst
54
55 an 1
lieu 1
motif1 61
tm1
63 an 2
65 lieu 2
motif 2
tm2 70
Décès
71 Sais
an
âge 3
lieu
SM3 79
N° I 13
Sexe
âge
erreur
m.z.
père 19
époux
mère 23
parenté 26
âge 1
âge 2 30
SM1
SM2
Ethnie
Relig 38
Naiss 41
42 SR1
SR2
44 EC
ER
EP 49
mig
st
inst
54
55 an 1
lieu 1
motif1 61
tm1
63 an 2
65 lieu 2
motif 2
tm2 70
Décès.
71 Sais
an
âge 3
lieu
SM3 79

II. LA SAISIE DES DONNEES

II.1. Méthodes récentes

La lecture optique qui consiste à lire des signes graphiques écrits dans des zones réservées est peu utilisée car difficile à mettre en oeuvre.

L'enregistrement direct sur un support magnétique est de plus en plus utilisé ; il est amené à remplacer la carte perforée. Les différences avec cette dernière sont :

a) la possibilité de dépasser les 80 colonnes qui constituent la limite de la carte perforée ;
b) l'enregistrement se fait en général en liaison avec un petit ordinateur qui peut effectuer certains contrôles (erreurs sur une variable, sur des variables croisées, ou sur la structure : voir § 4.1.).

Deux possibilités existent en cas d'erreur signalée :

. ou bien la personne qui effectue la saisie corrige ; elle doit alors être formée au chiffrement ;
. ou bien la personne n'a pas été formée au chiffrement et elle vérifie qu'elle n'a pas fait d'erreur de saisie.

II.2. La carte perforée

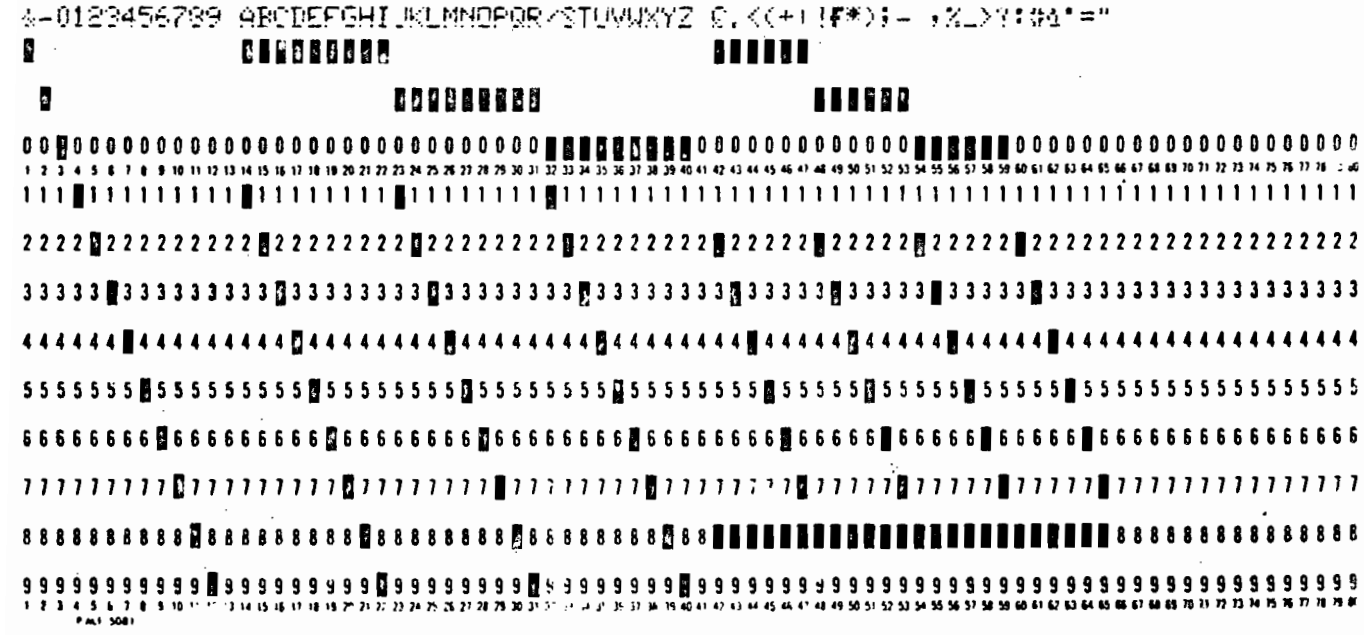
Dans le cas de saisie sur carte perforée, les étapes ne sont pas simultanées comme dans le mode de saisie précédent.

Les cartes perforées les plus courantes ont 80 colonnes (voir modèle joint). Dans chaque colonne, 12 perforations sont possibles : les 10 perforations inférieures sont désignées par les chiffres de 0 à 9. Les deux supérieures sont appelées perforations 11 et 12. La perforation 11 représente le signe "moins", la perforation 12 représente le "et" commercial. Une colonne permet donc de représenter un chiffre avec une perforation. Par convention, on représente une lettre majuscule par la combinaison de deux perforations, et des signes divers : (), +, *, ..., avec deux ou trois perforations dans une même colonne. Chaque variable occupera autant de colonnes qu'elle a de chiffres (ou de caractères).

Les perforations 0 à 9 (situées dans les colonnes 3 à 12 dans le modèle joint) peuvent être lues par tous les ordinateurs. Pour les autres perforations -alphabétiques et caractères spéciaux-, la convention dépend des ordinateurs.

La perforation s'effectue sur des machines commandées par un clavier semblable à celui d'une machine à écrire. La vérification s'effectue sur une machine semblable mais elle ne perfore pas, mais vérifie la concordance entre la perforation déjà faite et ce qui est composé au clavier. Ces deux opérations doivent être faites par des personnes différentes car cela diminue la probabilité de faire la même erreur qui peut résulter de la lecture de chiffres mal formés. Les deux opérations bien faites doivent laisser très peu d'erreurs.

.../...



Carte perforée 80 colonnes (code IBM-EBCDIC)

II.3. Carte et unité d'information

Avant la généralisation de l'ordinateur, les cartes étaient traitées avec du matériel mécanique. Ce matériel ne permettait pas directement de relier des variables sur deux cartes différentes. Aussi, on cherchait à utiliser des codes astucieux pour que toutes les variables soient contenues sur une seule carte. Cela n'est plus nécessaire, mais s'il ne manque que quelques colonnes et que quelques astuces permettent de se limiter à une carte, l'exploitation sera simplifiée.

Il est toutefois exceptionnel de tomber juste sur 80 colonnes. Si l'information est plus courte, une partie de la carte reste inutilisée. On peut toutefois, si l'information est très courte, mettre plusieurs individus par carte. Dans ce cas, chaque moitié, tiers... de carte doit comprendre l'identification complète de l'individu. L'économie réalisée résulte uniquement du stockage des cartes. L'économie sur la perforation est nulle et le coût des cartes est faible. Le traitement à l'ordinateur s'en trouve compliqué. Cela ne se justifie donc que pour de gros fichiers. (Le traitement avec du matériel mécanique s'en trouve très compliqué).

Si l'information sur un individu nécessite plusieurs cartes, on prendra un nombre entier de cartes pour chaque individu : fichier multicartes. Cela pose le problème de l'identification (voir § 1.2.). Pour distinguer les cartes entre elles il faut un code-cartes (1 pour la première, 2 pour la deuxième... par exemple). L'identification doit être aux mêmes places sur toutes les cartes, de même le code-cartes. Ceci est nécessaire pour le tri qui permettra de placer consécutivement les cartes concernant un même individu.

III. LES FICHIERS

L'ensemble des cartes concernant la population enquêtée constitue un fichier-carte. Le fichier-carte est lu une fois par l'ordinateur et mis sur bande magnétique. Les cartes peuvent ensuite être archivées. La bande magnétique n'est pas limitée à 80 colonnes. Aussi toutes les variables des cartes concernant un individu peuvent être rassemblées en un enregistrement unique. Ce fichier-bande a une structure qui peut être simple ou hiérarchisée.

III.1. Fichier simple

C'est un fichier où l'information est la même pour tous les individus, par exemple une et une seule carte de chaque type par individu. Toutes les unités d'information ont une même description.

III.2. Fichier hiérarchisé

C'est un fichier qui comprend des unités d'information de plusieurs types. Par exemple, ménage et individu, créées à partir de cartes ménages et de cartes individus.

Il est possible de faire des tableaux sur les ménages ou sur les individus, mais chaque individu étant rattaché au ménage auquel il appartient, des tableaux croisant des variables du ménage avec celles des individus sont aussi possibles. Ainsi, on pourra obtenir non seulement une distribution des ménages selon le type de logement, mais aussi une distribution des individus selon le type de logement. On dit que c'est un fichier à plusieurs niveaux, avec, au niveau 1 (supérieur), les ménages et au niveau 2 (inférieur), les individus.

Un autre exemple pourrait être constitué des mères au niveau 1 et des enfants au niveau 2. Les deux niveaux sont des individus mais on leur fait jouer un rôle différent.

On peut construire des fichiers à un nombre quelconque de niveaux. Le traitement d'un fichier hiérarchisé, par rapport à un fichier simple, complique le travail de l'informaticien. A l'inverse, y échapper implique une codification plus lourde. On peut toujours se ramener à un fichier simple. Ainsi, dans l'exemple ci-dessus, les variables intéressantes du ménage peuvent-elles être codifiées à la fois pour le ménage et pour chacun des individus du ménage.

On distingue deux sortes de fichiers hiérarchisés selon que la structure est fixe ou variable.

III.2.1. Fichier hiérarchisé à structure fixe

Par exemple, mère-enfant. Le nombre d'enfants étant fixé au maximum. Il revient à un fichier simple avec des zones répétitives ; pour un nombre d'enfants inférieur à un maximum, certaines de ces zones ne sont pas utilisées. La plupart des programmes assignent une longueur maxima à cette structure.

.../...

III.2.2. Fichier hiérarchisé à structure variable

Pour ceux-ci, il n'est pas possible de fixer un maximum car il serait trop grand. Par exemple, dans le cas d'un fichier concession-individus, de certains pays africains, le nombre d'individus peut dépasser cent personnes. On est alors obligé d'avoir une structure variable.

III.3. Nature de l'information

L'information sur carte est dite en caractères ; chaque chiffre ou signe occupe une colonne. Quand on crée un fichier, on le copie, en général, sur une bande magnétique. L'information ainsi recueillie peut être conservée telle quelle en caractères ou être transformée en information numérique sous forme binaire. Cette dernière forme, mieux adaptée à l'ordinateur, permet un traitement plus rapide, donc plus économique.

IV. LE CONTROLE DES DONNEES

C'est une opération essentielle qui permet de tester la qualité de l'enquête de terrain, du chiffrement et de la perforation-vérification.

IV.1. Types d'erreurs

IV.1.1. **Erreur sur une seule variable**

Une variable peut avoir un ou plusieurs chiffres. Les valeurs possibles sont en nombre limité. Il y a erreur si on trouve des valeurs hors du domaine des possibles (liste des postes).

IV.1.2. **Erreurs sur des variables croisées**

La valeur prise par une ou plusieurs variables peut restreindre le domaine des possibles pour une autre variable. Par exemple, une femme à la naissance d'un enfant ne peut avoir un âge au-dessous de 10 ans ou au-dessus de 50 ans (deux variables : le sexe et la naissance d'enfant restreignent le domaine des possibles pour l'âge). Par exemple, il y aura erreur si une femme, à la naissance d'un enfant, a 65 ans.

IV.1.3. **Erreurs de structure**

Ce type d'erreur se rencontre dans deux cas. Pour les fichiers simples, si une carte est en double (deux cartes ont même identification et même code-carte), ou si une carte manque. On doit avoir une carte et une seule de chaque type. Pour un fichier hiérarchisé, en reprenant l'exemple mère-enfant, on ne peut avoir un enregistrement enfant sans enregistrement mère et les enregistrements enfants doivent être en nombre égal à celui déclaré dans l'enregistrement mère.

IV.2. Les erreurs et leurs sources

Si on suppose que l'enquête a été correctement contrôlée, les erreurs proviennent en général de la codification. La perforation-vérification est une opération en principe sûre.

Les erreurs peuvent provenir d'un code mal fait, imprécis ou incomplet et/ou d'une incompréhension du chiffreur. C'est plus souvent une faute d'inattention ou le résultat d'un chiffre mal formé qui entraînent une lecture erronée à la perforation, lecture erronée qui doit être confirmée (par une personne différente à la vérification).

Une autre source d'erreur est la présence de blancs (absence de perforation) consécutifs en trop grand nombre. Dans ce cas, il peut y avoir une erreur sur le nombre de blancs et un décalage de toutes les colonnes qui suivent. On peut y remédier en limitant à 4 ou 5 le nombre maximum de blancs consécutifs sur une carte. On peut aussi revoir l'ordre des variables. Ou encore, après l'éventuelle série de blancs consécutifs, mettre une perforation fixe qu'il est facile de vérifier en trieuse ou à l'ordinateur. (Si possible, choisir une perforation fixe qui a peu de chances d'apparaître dans la colonne qui suit).

.../...

IV.3. Tabulation sommaire

Avant de faire un programme de vérification et de contrôle très lourd, il est conseillé de faire par exemple sur un échantillon, ou à la suite d'une préenquête, une tabulation sommaire dans le but de mettre en évidence les erreurs principales.

Une tabulation ultra-sommaire consiste à établir la fréquence de chaque perforation pour les diverses colonnes. Elle permettra de voir toutes les erreurs simples sur les variables à un chiffre et donnera des indications sur les autres erreurs. On peut vérifier que l'effectif d'hommes est égal à l'effectif des non-concernés à une question spécifique aux femmes... (il peut bien sûr y avoir des compensations mais cela est peu probable).

IV.4. Méthodes de correction

Une solution consiste à ne rien corriger. On accepte les erreurs et on pourra corriger les tableaux finaux. Il faut, toutefois, savoir où les postes erronés seront placés dans les tableaux; ce peut être dans une ligne et une colonne rebut.

On peut distinguer trois méthodes de correction avant la tabulation : par élimination, par retour au document de base, automatique.

IV.4.1. Correction par élimination

Ce procédé est dangereux et risque d'entraîner des biais. S'il y a, par exemple, des contrôles croisés liés au sexe, les erreurs n'étant pas aléatoires, on peut être amené à éliminer plus d'individus d'un sexe que de l'autre. Une solution est de vérifier la distribution de quelques variables essentielles sur les éliminés. S'il y a discordance avec la distribution sur les non-éliminés, cette méthode est à rejeter. Dans le cas d'un sondage, l'élimination modifie le taux de sondage ; il faut en tenir compte.

Dans le cas d'une erreur de structure, la correction par élimination peut entraîner des biais. Ainsi, soit un fichier de population dont les décès sont portés sur des cartes spéciales : c'est un fichier hiérarchisé, niveau 1 = individu, niveau 2 = décès. L'élimination des individus ayant deux cartes de décès peut conduire à une sous-estimation de la mortalité si la carte en double provient du décès d'un autre individu. Conserver les deux décès peut conduire à une surestimation s'il s'agit vraiment d'une carte en double. Dans le cas des erreurs de structure, le retour au document de base est préférable.

IV.4.2. Correction par retour au document de base

C'est une méthode sûre, mais éventuellement coûteuse. La correction peut se faire de deux manières; soit la carte entière est reperforée, soit une carte spéciale est perforée pour corriger chaque code erroné.

La première méthode permet, en général, de réutiliser le même programme ; la méthode de correction des codes nécessite un programme spécial. Le plus souvent, on crée un petit fichier des erronés corrigés qui est interclassé avec le gros fichier par un programme standard. Pour de très

.../...

petits fichiers, on peut classer les cartes perforées à la trieuse, enlever les mauvaises cartes, les remplacer par les bonnes et repasser tout le fichier.

IV.4.3. Correction automatique

C'est une correction faite par l'ordinateur ; elle réclame une étude particulière et implique une programmation supplémentaire. Elle consiste, par exemple, à donner un âge plausible à une femme enceinte dont l'âge codifié est trop jeune. Le principe de correction le plus satisfaisant est de retenir pour la femme d'âge erroné l'âge de la précédente femme du fichier de caractéristiques voisines (les caractéristiques retenues seront le sexe bien sûr et, par exemple, le nombre d'enfants ou l'âge de l'aîné...).

Si, comme on l'a supposé, l'âge était erroné, cela ne crée pas de biais. Par contre, si l'âge était exact et la situation de grossesse erronée, on est conduit à vieillir systématiquement une partie de la population. Un autre risque est de créer de nouvelles incohérences par exemple entre l'âge de la femme et celui de son conjoint ; le risque peut toujours être éliminé en augmentant la liste des caractéristiques retenues dans le choix de l'individu précédent qui sert à donner une valeur à la variable erronée. Si le fichier est classé géographiquement, le fait de prendre la valeur de la variable pour un individu précédent respecte la distribution régionale de la variable. Au début du fichier, il n'y a pas d'individu précédent, on doit donc prendre pour chacun des cas des valeurs moyennes.

Une autre tentation serait de corriger les biais d'enquête par des corrections automatiques contrecarrant le biais. Cela ne permet plus l'analyse et la critique des biais qui devraient conduire à une correction analytique. Avant de faire une correction systématique qui ne s'impose que pour une population importante, il faut au minimum faire une correction par retour au document de base sur un échantillon afin de connaître les informations les plus sûres qui permettent de corriger celles qui le sont moins.

Un cas intéressant de correction automatique, parce qu'il est simple, est de considérer un poste supplémentaire pour chaque variable. Les variables erronées sont mises automatiquement à ce poste supplémentaire. Il n'y a donc pas confusion avec les non-déclarés qui viennent de l'enquête. Pour un petit fichier, la correction automatique, qui nécessite un programme spécial, est trop coûteuse.

IV.4.4. Conclusion sur les méthodes de correction

Pratiquement, il est possible de combiner les diverses méthodes. Pour les erreurs de structure, on peut juger préférable de ne pas faire de correction automatique. Pour les autres erreurs, le traitement peut dépendre de l'erreur particulière ; ce peut être, soit un retour au document de base, soit une correction automatique, soit plus rarement une élimination.

Malgré les critiques que nous avons portées, la correction par élimination peut être utile pour obtenir les premiers résultats provisoires.

Dans tous les cas, il faut prévoir une statistique par type d'erreur et nombre d'individus erronés. En plus, dans le cas de correction automatique, il faut prévoir une statistique des types de correction afin de connaître les modifications apportées au fichier.

.../...

V. LA TABULATION

C'est l'objectif que nous nous sommes assignés pour l'exploitation et une étape obligatoire, même si d'autres analyses sont réalisées.

V.1. Les programmes généraux

La tabulation est une opération relativement standard au point de vue informatique ; c'est une opération fréquemment répétée ; aussi, pour éviter une programmation particulière, donc coûteuse, pour chaque tableau et chaque enquête, des informaticiens ont écrit des programmes généraux. Certains sont disponibles gratuitement. D'autres sont vendus par des sociétés de "softwares". D'autres ne sont utilisables que sur l'ordinateur d'une société qui vend du temps d'ordinateur.

Théoriquement, un programme général est une solution idéale ; pour produire un tableau, il suffit d'indiquer la position des variables : les libellés, le mode de présentation, quelle variable en ligne, quelle variable en colonne, calcul des pourcentages... En général, ces programmes permettent de recodifier, de filtrer... La programmation est ainsi presque nulle.

Cette solution idéale présente des inconvénients. Il y a un apprentissage à faire, donc un investissement à réaliser, d'autant plus lourd que le programme est plus général et offre plus de possibilités. D'un autre côté, il faut choisir un programme adapté à l'ordinateur dont on dispose, mais deux ordinateurs sont rarement absolument identifiques et on peut avoir des problèmes de passage d'un ordinateur à l'autre. A l'utilisation, on peut avoir des déboires par suite d'une mauvaise compréhension des ordres à donner au programme qui vient d'un apprentissage insuffisant ou d'une notice d'utilisation mal faite.

Si l'on s'oriente vers cette solution, il faut réunir des conditions minima : une notice bien faite -c'est essentiel- une assistance par un informaticien qui connaît le programme pour les premiers essais et les premiers tableaux.

Une fois qu'un programme est rendu opérationnel, beaucoup de travail de programmation est économisé et l'on peut répondre rapidement à la demande de nouveaux tableaux.

Un programme dit "général" ne peut pas faire n'importe quoi ; il y aura toujours des cas où on aura recours à une programmation complémentaire. On peut aussi, quand cela est possible, modifier légèrement la demande de tableaux pour la rendre acceptable par le programme général.

V.2. Le langage de l'exploitation d'enquête : lexique

Pour utiliser un programme d'exploitation d'enquête, il faut en connaître le vocabulaire. Les termes employés dépendent de chaque programme ; cependant, quels que soient ceux-ci, ils correspondent à des fonctions semblables d'un programme à l'autre. La liste ci-après donne les termes le plus couramment utilisés pour les principales fonctions.

.../...

Code : c'est la correspondance entre les postes d'une variable et les réponses à une question (voir § 1).

Edition : les règles d'édition sont celles qui gouvernent la présentation des tables ; on peut avoir une présentation avec des textes (voir ce mot) ou parfois au choix une présentation standard où les postes, les variables et les tableaux ne sont répérés que par des nombres.

Filtre : c'est la combinaison de postes d'une ou plusieurs variables pour sélectionner une partie des individus pour faire un tableau sur une sous-population. Les individus qui ne répondent pas aux conditions du filtre ne sont pas comptabilisés dans le tableau.

Niveau de ventilation : dans un fichier hiérarchisé, pour chaque ventilation, il faut en préciser le niveau. Par exemple, pour un fichier mère-enfant, un tableau sur le nombre d'enfants selon l'âge de la mère entraîne un comptage des mères (niveau 1). Un tableau sur l'âge des mères à la naissance est un comptage des enfants (niveau 2).

Postes : ce sont les différentes possiblités d'une variable (voir §1). Des termes différents peuvent être employés pour ce concept.

Rebuts : la liste des postes d'une variable est limitée. Normalement, le contrôle a pour but de ramener tous les postes dans ces limites. Si cela n'a pas été fait systématiquement, il est utile de comptabiliser les rebuts dans un poste supplémentaire pour chaque variable. Le traitement des rebuts varie avec les programmes.

Recodification : calcul d'une nouvelle variable à partir des variables existantes (voir §1.4.).

Tableau : c'est le produit final de l'exploitation et le résultat d'une ou plusieurs ventilations combinées (voir plus bas un exemple de ventilation).

Textes dans un tableau : on peut avoir trois sortes de textes, un texte qui sert de titre au tableau, un texte qui est le nom en clair de la variable, un texte qui est le nom en clair pour chaque poste d'une variable. Pour ces trois catégories de textes les mots titres, libellés, intitulés sont indifféremment employés selon les auteurs.

Variable : c'est la codification d'une question (voir §1).

Variable somme : une ventilation est une répartition des individus selon les variables de ventilation. Les cases de la ventilation sont mises à zéro au début du fichier et à chaque rupture. Pour chaque individu (plus généralement pour chaque unité du niveau de ventilation), on ajoute une unité dans la case définie par les valeurs des variables de cet individu.

Plus généralement, on peut totaliser n'importe quelle variable : ce peut être un coefficient de pondération (inverse du taux de sondage), une variable dont on veut calculer la moyenne.....

Ventilation : la ventilation est l'opération qui répartit tous les individus dans les cases d'un tableau. Chaque case d'un tableau est définie par la valeur de chacune des variables qui définissent ce tableau.

.../...

Un tableau peut comprendre deux, trois variables, ou plus.

Une ventilation se définit par :

. un filtre
. une ou des variables de ventilation
. éventuellement, une variable de rupture
. un niveau de ventilation
. une variable sommée.

Exemple :

Fichier logement (niveau 1), individu (niveau 2)
Au niveau 1 est porté le type de logement
Au niveau 2 sont portés l'âge, le sexe pour chaque individu
Un tableau donnant les effectifs d'individus, par sexe et type de logement est une ventilation de niveau 2, comprenant trois variables de ventilation : âge, sexe et type de logement. Il n'y a pas de filtre, ni de variable de rupture; la variable sommée est la variable unité.

Un tableau donnant les salaires moyens des salariés selon le type de logement et l'âge de l'individu nécessite deux ventilations : celle ci-dessus avec un filtre sur les salariés qui donne les effectifs; une autre ventilation identique qui donne le total des salaires au lieu de l'effectif ; la variable sommée est le salaire. Le tableau sur les salaires moyens est obtenu en divisant case à case les résultats des deux ventilations (salaire global divisé par l'effectif des salariés correspondant).

V.3. La demande de tableaux

Au niveau du vocabulaire, les termes définis ci-dessus doivent suffire. Pour chaque tableau demandé, il est bon de diviser la population totale du tableau par le nombre de cases du tableau. Si le nombre moyen d'individus par case est trop faible, la ventilation est trop poussée. Le nombre de cases s'obtient en faisant le produit du nombre de postes des variables de ventilation et de rupture.

La population totale du tableau est le nombre d'individus dans le fichier. Dans un fichier hiérarchisé, c'est le nombre d'unités du niveau de ventilation. Ceci dans le cas où il n'y a pas de filtre ; s'il y a un filtre, il ne faut prendre en compte que la population qui satisfait aux conditions du filtre.

VI. PLACE DE L'INFORMATICIEN

VI.1. Langage

Le responsable d'une enquête a en général en face de lui un informaticien (analyste de préférence) ; il leur faudra se créer un langage commun. Ce problème se complique car souvent le responsable de l'enquête ne sait pas ce qu'il veut obtenir de l'utilisation de l'ordinateur, ne sachant pas ce qui est possible. De même, l'informaticien ne comprend pas toujours les préoccupations du responsable d'enquête (concernant par exemple la spécificité des calculs démographiques). Il en résulte que les premières séances de travail sont déroutantes et peu productives ; cette situation doit normalement évoluer favorablement.

VI.2. Enchaînement

Un enchaînement idéal des opérations est représenté par l'organigramme ci-contre.

On remarquera que l'informaticien doit intervenir très tôt ; le mieux est qu'il intervienne dès la mise au point des questionnaires. Il arrive souvent qu'on lui demande d'intervenir, une fois l'enquête réalisée, au stade de la conception de la codification. C'est encore une solution acceptable. Par contre, l'intervention de l'informaticien après la codification complique en général singulièrement la programmation. En effet, nous avons vu plusieurs fois qu'il y avait le choix entre des opérations effectuées à l'ordinateur et des opérations manuelles de codification. Il est souhaitable que l'informaticien apporte sa contribution à ces choix.

VI.3. Mise au point des programmes

Cette question ne concerne pas le seul informaticien. Le responsable de l'enquête doit regarder de très près les résultats pour s'assurer que les résultats de la programmation répondent à ce qu'il a défini. En effet, des incompréhensions sont toujours possibles.

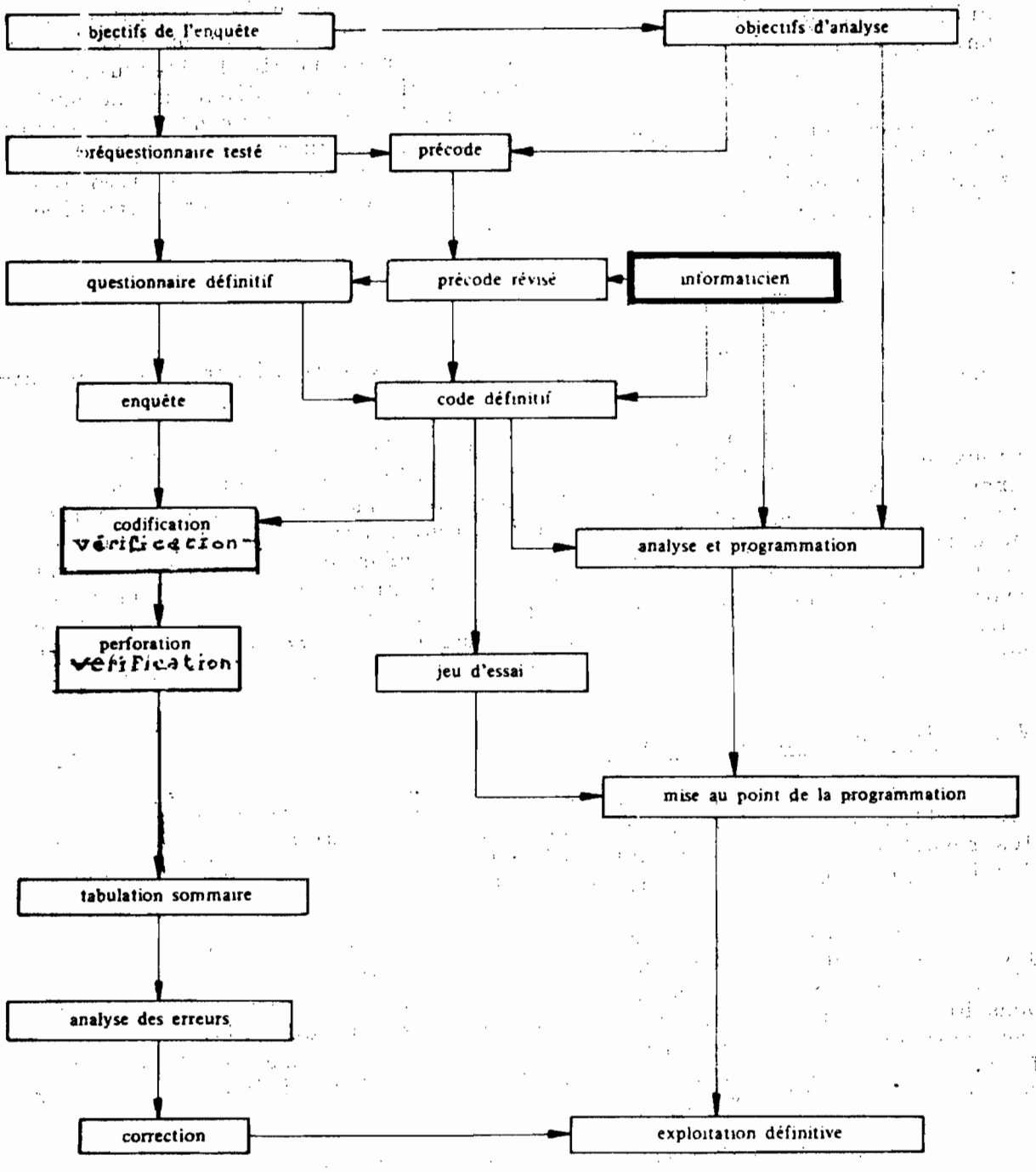
Une solution à ce problème est de constituer un jeu d'essai. Ce jeu d'essai peut être sommaire. 10 à 20 cas peuvent suffire selon la complexité des situations possibles. Le jeu d'essai doit être constitué par le responsable de l'enquête qui prévoira quelques cas normaux, mais surtout des cas anormaux qui pourront être de véritables pièges tendus à l'informaticien. Tous les coups sont permis ; ce sera l'assurance d'un minimum d'ennuis lors du passage au fichier réel.

Le jeu d'essai doit entraîner un contrôle manuel de l'exploitation complète, variable par variable, individu par individu. C'est pour cela qu'il ne faut pas multiplier le nombre de cas, ce contrôle étant vite fastidieux.

Ensuite, si le fichier est volumineux, il est prudent de réaliser une exploitation sur un échantillon et de vérifier la cohérence statistique des distributions obtenues.

.../...

ORGANIGRAMME D'ENCHAINEMENT



VI.4. La production de tableaux

Grâce à l'utilisation d'un programme général, la production de tableaux est facilitée. Avec de l'habitude, le responsable d'une enquête peut écrire lui-même la demande de tableaux sans recourir à l'informaticien. L'analyse peut se faire par étapes en produisant les tableaux en plusieurs séries en fonction des premiers résultats de l'analyse.

ANNEXE I
NOTES TECHNIQUES SUR LES ORDINATEURS, LES BANDES MAGNETIQUES ET LES PROGRAMMES UTILITAIRES

Caractéristiques d'un ordinateur

Un ordinateur est d'une marque et d'un modèle donnés. Pour un même modèle on peut avoir des tailles de mémoire différentes. Enfin la configuration peut être variable, elle concerne les unités périphériques : lecteur de cartes, perforateur de cartes, imprimante, dérouleur de bandes magnétiques, disques... Les périphériques ont eux-mêmes des caractéristiques qui peuvent varier.

Un ordinateur peut accepter ou non différents langages. Les plus usuels sont :
- le COBOL qui est bien adapté au traitement des fichiers et par suite à l'exploitation d'enquête
- le FOTRAN qui est bien adapté au calcul scientifique mais qui peut servir à l'exploitation d'enquête
- le PL1 qui représente une synthèse de COBOL et FORTRAN.

En outre un ordinateur est régi par un système d'exploitation (pour IBM les plus courants sont le DOS sur les moyens ordinateurs, l'O.S. sur les gros ordinateurs). L'ASSEMBLEUR, langage très proche de celui de la machine est plus long à programmer et peu utilisé.

L'ensemble de ces informations sert à caractériser un ordinateur.

Les bandes magnétiques.

Elles peuvent être de plusieurs types, les plus répandus sont :
-7pistes et 800 BPI ("bits per inches").
-9 pistes et 800 BPI
-9 pistes et 1600 BPI (on dit aussi densité 1 pour 800 BPI et densité 2 pour 1600 BPI).

Une bande est caractérisée par sa longueur qui conditionne la quantité d'informations qu'elle peut recevoir. Cette quantité se mesure en nombre d'enregistrements et longueur des enregistrements. En fait une bande n'est pas une suite continue d'enregistrements, les enregistrements sont séparés par les sauts où il n'y a rien. Ces sauts sont très importants en longueur et on a un moyen d'en minimiser le nombre. En effet on peut grouper un nombre entier d'enregistrements entre deux sauts : c'est un block. Pour minimiser la place sur la bande on doit prendre le plus long block possible, mais à la lecture et à l'écriture cela mobilise de la place en mémoire et on doit se limiter. Une autre limitation est liée à l'ordinateur, elle est variable mais toujours supérieure à 4096 caractères (pour les ordinateurs IBM). C'est donc un nombre à ne pas dépasser si on peut être amené à changer d'ordinateur.

Une autre caractéristique d'une bande est le langage dans lequel elle a été créée. Il peut y avoir des incompatibilités entre langages. Si une bande est insuffisante pour un fichier il faut plusieurs bandes (ou volumes) on parle d'un fichier multivolume. Si on met plusieurs fichiers sur une même bande on parle d'un volume multi-fichier.

Tableau
Longueur (en mètres) nécessaire pour 100 blocks selon la longueur du block et le type de bande (source IBM)

Nombre de caractères (ou octets d'un block	Type de bande		
	7 pistes (1) 800 BPI	9 pistes (2) 800 BPI	9 pistes (3) 1600 BPI
80	2.19	1.80	1.81
120	2.32	1.93	1.87
160	2.45	2.06	1.93
200	2.58	2.19	2.00
400	3.22	2.83	2.32
600	3.86	3.48	2.64
800	4.51	4.12	2.97
1000	5.15	4.77	3.29
1500	6.76	6.37	4.09
2000	8.37	7.98	4.90
2500	9.98	9.59	5.70
3000	11.59	11.20	6.51

Formules données par ailleurs : (résultats sensiblement identiques)
Longueur en cm de 100 blocks de N caractères
(1) = 190,5 + 0,3145 N
(2) = 152,4 + 0,3175 N
(3) = 152,5 + 0,15875 (N + 82)

Une bande est caractérisée par un nom, dit nom de volume, et chaque fichier reçoit un nom ou label. La longueur de l'enregistrement et le nombre d'enregistrements par block définissent complètement la bande et ses fichiers. Ensuite pour utiliser la bande, il faut comme pour les cartes un code qui indique les postes de chaque variable, la position de la varialbe ;il faut en outre indiquer si la variable est en caractères ou en binaire.

Les programmes utilitaires

Ce sont des programmes standard qui effectuent des opérations courantes comme chargement d'un fichier carte sur une bande : tri d'un fichier ; impression du contenu de tout ou d'une pattie de bande ; fusion de deux fichiers préalablement triés ; ... Les programmes utilitaires n'exigent que peu de cartes de spécification pour être utilisés et évitent toute programmation. Ce sont des programmes généraux mais comme tous les informaticiens les utilisent, ils sont bien connus et ne posent normalement pas les problèmes rencontrés avec les programmes généraux d'exploitation d'enquête.

Quelques programmes généraux d'exploitation d'enquête

	MINITAB	SPSS	SAFE	LEDA	CENTS
Type de fichier traitable : simple	+	+	+	+	+
hiérarchisé fixe	0	+	+	+	+
hiérarchisé variable	0	0	+	+	+
Variables alphabétiques	0	+	?	+	+
Contrôle fichier simple	+	?	+	+	0
Recodification	+	+	0	+	+
Tabulation	+	+	0	+	+
Moyenne, écart-type	+	+	0	0	+
Régression	0	+	0	0	0
Test d'association	+	+	0	0	0
Analyse factorielle	0	+	0	0	0
Distribution des variables	+	+	0	0	+
Mise à jour par - correction des codes	0	?	0	0	0
- correction de l'unité	+		+	0	0
Contrôle de structure	0	?	+	+	0
Constitution de la structure	0		0	opérations + simultanées	0

Contrôle entre unités	0	?	?	+	?
Manuel A : anglais, F : Français	A,F bon	A bon	?	F bon	A bon
Utilisable petits ordinateurs (100 k)	+	+	+	0	+
Disponibilité	facilement transportable	facilement transportable (?)	facilement transportable	installation lourde	transportable