I - LES PRINCIPES DE L'EXPLOITATION DU RECENSEMENT SUR L'ENSEMBLE ELECTRONIQUE DE L'ETABLISSEMENT CENTRAL

INTRODUCTION -

Les principes sommaires que nous exposons ci-dessous concernent seulement l'exploitation du recensement de la population de 1962 (sondage au 1/20 et exploitation exhaustive).

Ce dépouillement est en effet la première opération à grande échelle réalisée à l'I.N.S.E.E. sur matériel électronique. Il est vraisemblable que dans un avenir proche de nombreuses améliorations des techniques de codification et de programmation supprimeront certaines contraintes et introduiront plus de souplesse dans les procédures d'exploitation.

D'autre part étant donné l'ampleur du travail que représente le dépouillement du recensement et le peu de temps dont on disposait entre la réception de l'ordinateur et le début de l'exploitation du sondage au 1/20, il a été décidé de recourir à la programmation automatique : l'ordinateur engendre lui-même les programmes de dépouillement à partir d'un nombre restreint d'instructions qui définissent les tableaux désirés. Dans ces conditions, à la fois les données et les tableaux à produire doivent se présenter sous des formes déterminées, requises par la programmation automatique : la normalisation est la rançon des méthodes automatiques.

C'est pourquoi il ne faut pas accorder à l'exposé qui suit une portée générale et un caractère définitif en ce qui concerne la méthodologie du dépouillement sur ordinateur. Des exploitations d'envergure moindre, comme les enquêtes par sondage par exemple, correspondent à des traitements plus souples et plus proches du "sur mesure"; enfin, l'expérience aidant, surtout dans le domaine d'utilisation de l'électronique qui est le nôtre où peu d'organismes manipulent comme l'I.N.S.E.E. des masses énormes de données, il est certain que sur de nombreux points cet exposé sera rapidement dépassé.

1. La bande magnétique

Sur bande magnétique, par zones de longueur constante, sont reportées les données relatives aux diverses unités statistiques. Chaque zone est l'analogue d'une carte mécanographique : elle contient un certain nombre de critères comportant chacun un ou plusieurs caractères. Les caractères correspondent aux colonnes de la carte; ainsi, la catégorie socioprofessionnelle (CS) comporte deux caractères : la dizaine de CS (DCS) ou premier chiffre de la CS et l'unité de CS (UCS) ou deuxième chiffre de la CS.

Les diverses unités statistiques d'une même bande sont nécessairement (1) de même nature - comme les cartes d'un même fichier sont également de même nature : fichier "individu", fichier "logement", fichier "immeuble". Chacun des caractères est chiffré sur la bande entre 0 et 9 (2), à l'exclusion de tout autre chiffrement.

Même si pour certaines unités statistiques certains critères sont sans objet, il est nécessaire de prévoir un chiffrement. Ainsi la catégorie socioprofessionnelle est dépourvue de signification pour les enfants de moins de 15 ans. Dans ce cas, les bulletins individuels ne sont pas chiffrés, ou chiffrés "espace" (+) et les cartes n'ont pas de perforation dans les colonnes correspondantes. Sur bande, l'équivalent de l'espace est "zéro" et les enfants de moins de 15 ans recevraient la CS 00 (déjà réservée aux agriculteurs exploitants) si on ne prenait pas de précaution. On a donc décidé d'affecter la CS 99 (inactifs) aux enfants pour éviter toute confusion.

L'ordre de classement des unités statistiques sur la bande restera immuable au cours de toutes les exploitations : l'un des principes de dépouillement du recensement étant précisément de ne pas recourir au tri (sauf cas particuliers, comme nous le verrons par la suite). Cet ordre de classement est l'ordre géographique : district, commune, canton, département, région de programme. Ces critères de tri des unités statistiques jouent un rôle très important dans les opérations de dépouillement. On les appelle "échelons géographiques" (à l'exclusion des autres caractéristiques géographiques qui n'interviennent pas dans la séquence des unités statistiques reportées sur bandes, telles que la catégorie de commune, la région agricole, etc...).

Ainsi la bande magnétique comporte une succession d'unités statistiques de même nature, triées suivant les échelons géographiques.

2. Le dépouillement - Les tableaux bruts et définitifs

Lors du dépouillement, la bande magnétique défile devant une tête de lecture qui procède à deux opérations :

- sélection des unités statistiques qui font partie du tableau.
- ventilation de ces unités statistiques suivant les critères de croisement constituant le tableau.

a) Les critères d'appartenance

Les unités statistiques qui font partie du tableau envisagé sont définies par un ou plusieurs critères d'appartenance reliés par les opérateurs logiques "et" et "ou".

---

(1) du moins dans les bandes magnétiques du recensement.

(2) Cette condition est propre aux bandes du recensement de la population : on peut écrire les bandes magnétiques en caractères alphanumériques.

Ainsi, un tableau relatif à la population des ménages ou d
ploitations agricoles habitant les communes rurales correspond au
tenance :

- l'individu appartient à un logement siège d'une exploitation agricole (en code : EXPC ≠ 0)

ou

- le chef de ménage est agricole (en code CSCOM = 1).

} et la commune e est rura (en code : DCOM = 0)

D'où le champ du tableau défini par les critères d'appartenance :

$$\left[(EXPC \neq 0) \text{ ou } (CSCOM = 1)\right] \text{ et } \left[DCOM = 0 \right]$$

Les critères d'appartenance sont divisés en deux catégories :

- les critères d'appartenance géographique : ce sont les critères qui sont identiques pour toutes les unités d'un même territoire : ainsi la catégorie de commune, la zone de peuplement industriel ou urbain, la région agricole, ... sont des caractéristiques identiques pour toutes les unités statistiques d'une commune.

Ces caractéristiques ne sont pas placées dans la zone de chaque unité statistique mais mises en facteur en tête de chaque commune, constituant l'enregistrement spécial géographique qui comprend les données de la carte maitresse communale.

- les critères d'appartenance non géographiques sont les autres critères de la bande

Il n'y a aucune différence de nature entre les critères d'appartenance géographique et non géographique; seul le traitement interne à l'ordinateur diffère.

b) Les critères de ventilation

Au cours de la lecture de la bande magnétique, l'ordinateur enregistre le passage d'une unité statistique si elle appartient au tableau envisagé.

Si, par exemple, on croise le sexe et la CS à un chiffre, le calculateur prépare 20 cases de mémoire. Lorsque la zone relative à un salarié agricole du sexe masculin est lue par l'ordinateur, la case de mémoire portant l'adresse (S = 1, DCS = 1) enregistre "1" si l'individu satisfait aux critères d'appartenance. Si l'individu n'entre pas dans le champ du tableau, aucune mémoire ne conserve trace de son passage.

Dans les cases de mémoire il est possible de totaliser soit un nombre constant par unité statistique (1 dans le cas ci-dessus du compte carte, 20 pour un tableau du sondage au 1/20), soit une quantité figurant dans la zone de l'unité statistique (nombre d'enfants, nombre de pièces des logements, etc...), soit encore plusieurs données de nature différente : dans une case on totalise le nombre d'enfants, dans une autre le nombre de pièces, etc... Un même tableau peut comporter jusqu'à six critères de cette nature appelés critères de totalisation.

c) Les mémoires - la bande code

La préparation des cases de mémoire est assurée par consultation automatique d'une bande magnétique spéciale : la bande code qui porte la liste des postes possibles de chacun des critères de ventilation. Si dans un tableau on croise le sexe et la dizaine de CS, la bande code indique que le sexe a 2 postes et la dizaine de CS 10 postes. Le nombre de cases de mémoire préparées pour enregistrer les unités entrant dans le champ du tableau est le produit de ces deux nombres, soit 20 cases.

L'ensemble électronique IBM 7070 compte 50.000 positions de mémoire dont environ 20.000 sont utilisées pour la programmation. Chaque case de mémoire comporte dix positions c'est-à-dire peut enregistrer des nombres de 10 chiffres. Au cours d'un même passage de la bande magnétique, on peut établir entre 1.500 et 3.000 cases de tableaux réparties en un nombre quelconque de tableaux. Le nombre de cases ainsi produites dépend du nombre de tableaux, des critères d'appartenance définissant ces tableaux, du nombre d'individus figurant dans les tableaux.

Dans un même tableau, on peut croiser jusqu'à 8 critères (4 critères au maximum en ligne, 4 critères au maximum en colonne). Les tableaux établis au cours d'un même passage peuvent comporter des champs différents, être produits pour des échelons géographiques différents (certains à l'échelon canton, d'autres à l'échelon région de programme....). Un même individu peut appartenir au cours d'un même passage à un nombre quelconque de tableaux.

La vitesse à laquelle sont effectués les tableaux peut varier de 2.000 à 5.000 individus par minute, suivant la contexture des tableaux (nombre de tableaux, critères d'appartenance, nombre d'individus appartenant aux tableaux du passage).

Les tableaux bruts contiennent automatiquement - sans consommation supplémentaire de cases de mémoire - certains totaux partiels en ligne et en colonne (à condition, bien entendu, qu'on les ait demandés).

Ainsi un tableau croisant en ligne respectivement les trois critères x = Sexe, y = Etat matrimonial, z = Age : (S x M x AGE) peut fournir dans l'ordre suivant les seuls totaux partiels :

x y . : distributions marginales par sexe et état matrimonial

x . . : distributions marginales par sexe.

. . . : total général

. . z : distributions marginales par âge (regroupement suivant le critère mineur).

Par contre les totaux du genre . y . , x . z et . y z ne peuvent être obtenus au stade du tableau brut. C'est une remarque importante, valable de la même façon pour les croisements en colonne : l'ordre des critères de croisement en ligne ou en colonne n'est pas indifférent en ce qui concerne les totaux partiels. Si en ligne et en colonne on ne croise pas plus de deux critères, toutes les totalisations sont alors possibles. Dans le cas de quatre critères, x y z u les seules totalisations possibles sont au nombre de 5 : x y z . , x y . . , x . . . , ... et . . . u.

d) Les tableaux bruts

A la fin de chaque passage, les mémoires du 7070 qui contiennent les cases des tableaux du passage sont vidées sur une bande magnétique. Cette bande est ensuite introduite dans le 1401 dont l'imprimante fournit les tableaux sur papier accordéon de 42 cm de large à raison de 600 lignes par minute. Les tableaux peuvent éventuellement être perforés sur cartes mécanographiques (vitesse de perforation : 250 cartes par minute). Chaque carte comporte alors cinq colonnes de tableaux.

Les tableaux bruts se présentent sous forme standard : 14 colonnes de 8 chiffres chacune et 50 lignes par feuillet.

Si les critères ligne ou colonne comportent - y compris les totalisations - respectivement plus de 50 et 14 combinaisons, le reste du tableau figure sur les feuillets suivants, numérotés de 1 à n dans l'ordre vertical puis horizontal. Les totaux partiels sont isolés du corps du tableau par une ligne blanche et apparaissent après le détail correspondant c'est-à-dire en dessous pour les lignes et à droite pour les colonnes.

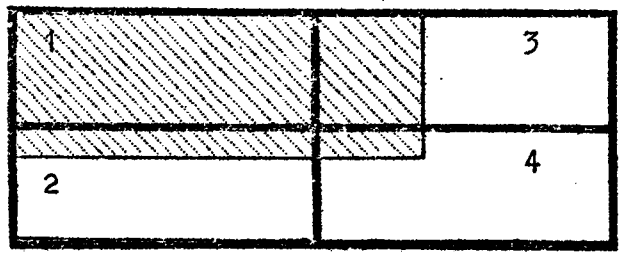
Exemple : Dans un tableau, on croise en ligne le sexe et l'âge quinquennal, en colonne la nationalité en 3 postes et l'état matrimonial :

$$\underset{(2+T)}{(\ S\ }\times\ \underset{(20+T)}{A\ )}\quad\times\quad\underset{(3+T)}{(\ N\ }\times\ \underset{(4+T)}{M\ )}$$

Ce tableau comporte - avec les totalisations toutes possibles ici puisque en ligne comme en colonne, on ne croise que 2 critères.

$(2 + 1) \times (20 + 1) = 63$ lignes et $(3 + 1)\ (4 + 1) = 20$ colonnes

Il se présente sur quatre feuilles :



sur les premières feuilles figurent les 63 lignes utiles et les 14 premières colonnes - sur les feuilles 3 et 4 figurent à nouveau les mêmes 63 lignes avec les 6 colonnes qui n'ont pu figurer sur les deux premières feuilles.

Sur chaque feuille d'un tableau les en têtes de lignes et de colonnes sont portées en code.

Sur la première feuille du tableau, apparaît le numéro et le titre en clair du tableau. On dispose de 60 caractères (y compris les blancs) alphanumériques pour écrire le titre.

Après leur sortie de l'imprimante sur papier accordéon, - les tableaux sont ensuite coupés par feuillet, éventuellement reproduits en un nombre variable d'exemplaires (offset) et brochés en liasses d'une centaine de feuilles format 27 x 42.

3. Les tableaux définitifs

Les bandes magnétiques qui contiennent les tableaux bruts sont reprises et introduites dans le 1401 pour conduire aux tableaux définitifs. Au cours de cette opération on peut procéder à de nombreux aménagements du tableau et à des calculs :

- addition, suppression, déplacement de certaines lignes ou colonnes. Obtention de sous-totaux. Aération du corps du tableau par introduction de lignes ou colonnes blanches.

- addition algébrique de deux ou plusieurs tableaux de même forme.

- division de tous les termes d'un tableau par une ligne ou une colonne de diviseurs (calculs de pourcentages).

- division des cases d'un tableau par les cases correspondantes d'un autre tableau de même forme.

  etc....

Les tableaux définitifs sont prêts pour la publication. Les en-têtes des lignes et des colonnes sont seuls réalisés à la main.

Malgré les avantages que procurent les tableaux définitifs, il ne faut pas perdre de vue que les tableaux bruts doivent être aussi proches que possible des tableaux définitifs. C'est sous la forme des tableaux bruts que l'information non publiée sera conservée et par conséquent il importe, dès le stade de leur conception, de penser les tableaux bruts en termes de tableaux définitifs. Demander un tableau brut croisant trop de critères revient à immobiliser l'ordinateur pendant longtemps, à gaspiller des mémoires inutiles et conduit à la fabrication de tableaux monstrueux par leur poids et leur coût et à peu près sûrement illisibles.

II - LES CONSEQUENCES DE LA METHODE DE DEPOUILLEMENT DU RECENSEMENT

De l'exposé rapide qui précède, il résulte un certain nombre de conséquences pratiques en ce qui concerne les procédures de codification et de dépouillement. Nous allons reprendre les points examinés plus haut :

1. Les unités statistiques d'une même bande sont de même nature et il est impossible (1) de croiser des données figurant sur deux bandes différentes.

Or les unités statistiques définies dans le recensement sont de nature différente et hiérarchisées :

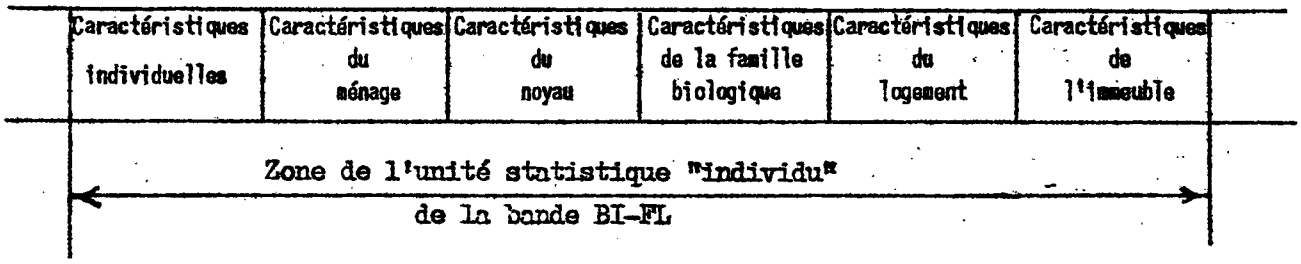
- individus groupés par logements groupés par immeubles
- individus d'un même logement constituant un ménage et groupés en noyaux familiaux et familles biologiques.

En outre, il est nécessaire de croiser des données relatives à une unité statistique avec des données relatives à une autre unité statistique : individus d'après les caractéristiques de leur ménage, de leur noyau, logements d'après les caractèristiques de l'immeuble, etc...

C'est pourquoi il a été décidé de créer deux bandes magnétiques : l'une où l'unité statistique est l'individu (bande BI-FL), l'autre où l'unité statistique est le logement (bande FL-BH).

a) La bande BI-FL

Sur la bande BI-FL sont reportées les données relatives à l'individu et aux unités auxquelles il appartient :



Ainsi figurent dans la zone de l'individu tous les critères qu'on envisage de croiser avec des caractéristiques individuelles. Les caractéristiques d'un même immeuble sont reportées dans la zone "individu" de chacun des habitants de l'immeuble (du moins celles d'entre elles qu'il a été jugé utile de reporter). De même dans la zone "individu" d'un chef de ménage sont reportés : sa CS propre, la CS de son chef de ménage, la CS de son chef de noyau c'est-à-dire trois fois la même chose puisqu'un chef de ménage est nécessairement chef de noyau.

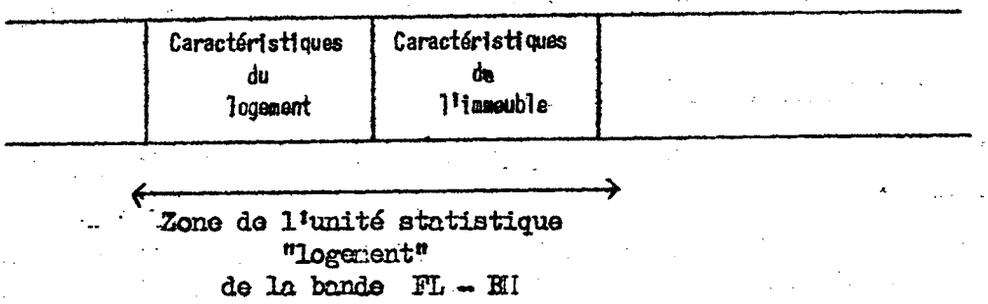
(1) Tout au moins, avec le programme général utilisé pour le dépouillement du recensement.

La bande BI-FL permet de faire la statistique des logements, ménages, noyaux, à condition d'introduire des critères d'appartenance. La statistique des ménages est effectuée à partir de l'exploitation des zones individus "chefs de ménage". De même la statistique des familles biologiques est obtenue à partir des données relatives aux individus à la fois "membres de famille biologique et chefs de noyau".

Cette procédure suppose - notons le - que chaque ménage comporte un et un seul chef de ménage, ce dont il conviendra de s'assurer au moment de la transcription des cartes sur bande.

b) La bande FL - BI

Sur la bande FL-BI sont portées les données relatives au logement et à l'immeuble :



Parmi les caractéristiques du logement on a porté également les caractéristiques du ménage qui occupe ce logement.

Cette bande permet de faire la statistique des logements ainsi que des immeubles. En effet on a affecté un lien de chaque logement avec l'immeuble (analogue au lien des individus avec le chef de ménage). La statistique des immeubles est effectuée à partir de l'exploitation des zones "logement" de lien 1 (il y a un logement et un seul ayant le lien 1 par immeuble).

c) Les unités statistiques reportées sur chacune des bandes.

- Sur la bande BI-FL sont reportés tous les individus recensés, qu'ils habitent un logement ou non. En revanche, les logements ou immeubles qui ne comportent aucun habitant ne figurent pas sur cette bande.

- Sur la bande FL-BI sont reportés tous les logements d'habitation et tous les immeubles - y compris les immeubles qui ne comportent aucun logement d'habitation.

Caractéristiques des individus	logements	immeubles	bande BI-FL	bande FL-BM
Population des résidences principales			×	×
	résidences principales		×	×
		immeubles comportant au moins une résidence principale	×	×
	logements vacants résidences secondaires			×
		immeubles ne comportant que des logements vacants ou secondaires		×
		immeubles ne comportant aucun logement d'habitation		×
Population des habitations mobiles			×	
	habitations mobiles		×	
Population des ménages collectifs Population comptée à part. Lot F.			×	

La bande BI-FL permet de faire la statistique des résidences principales et des habitations mobiles, c'est-à-dire des ménages ordinaires; la bande FL-BI permet de faire la statistique des résidences principales, secondaires ou vacantes c'est-à-dire des logements d'habitation. En ce qui concerne le logement, la première est la bande démographique, la seconde la bande de l'habitat.

2. Les unités statistiques correspondent à des zones de la bande magnétique de longueur constante, de dessin invariable, où tous les critères doivent être chiffrés.

Or pour certaines unités statistiques, certains critères sont dénués de signification :

. Caractéristiques du "logement" pour les individus qui n'habitent pas un logement : population des habitations mobiles, des ménages collectifs, de la population comptée à part.

. Caractéristiques du chef de ménage pour les individus qui n'appartiennent pas à un ménage (population comptée à part, ménages collectifs).

. Caractéristiques du ménage occupant le logement pour les logements vacants ou les résidences secondaires.

. etc....

Dans le cas des cartes mécanographiques, on ne chiffrait pas les critères correspondants - ce qui revenait à les chiffrer + (espace). Or sur bande, l'espace correspond au chiffrement 0.

D'où la nécessité dans ces différents cas de chiffrer 0 les caractéristiques sans objet. Mais certains codes accordent au chiffrement 0 une signification particulière. Il est donc nécessaire d'introduire un critère supplémentaire pour séparer le " "0" sans objet du "0" qui a une signification.

Ainsi sur la bande BI-FL pour la CS du chef de ménage : 00 signifie Agriculteur exploitant. La zone "CS du chef de ménage" est également chiffrée 00 pour les individus qui n'ont pas de chef de ménage.

Le critère supplémentaire introduit pour distinguer la signification de ces deux chiffrements identiques est NUSI : nature de l'unité statistique individu :

NUSI = 1 pour les individus qui n'appartiennent pas à un ménage, c'est-à-dire pour lesquels le type de carte est : 2 (population comptée à part)
3 (population des ménages collectifs)
4 (bulletins 2bis B non reclassés : lot F)

NUSI = 2 pour la population des habitations mobiles, c'est-à-dire habitant un logement de catégorie CL = 8.

NUSI = 3 pour la population habitant une résidence principale.

Ainsi ce code NUSI en 3 postes permet de définir, sur la bande BI-FL :

- la population qui ne vit pas en ménage ordinaire : NUSI = 1.
- la population des ménages ordinaires: NUSI $\neq$ 1
- la population des résidences principales: NUSI = 3

Si dans un tableau on croise la CS du chef de ménage avec d'autres caractéristiques, il conviendra d'ajouter le critère d'appartenance NUSI $\neq$ 1. De même pour un tableau croisant le nombre de pièces des logements occupés : NUSI = 3.

NOTA : Il existe un code analogue dans la bande FL-BM : le code NUSL.

3. La bande est triée une fois pour toutes suivant les caractéristiques géographiques

Cette procédure à l'avantage de limiter l'écriture des bandes magnétiques : les caractéristiques géographiques sont écrites une seule fois en tête de chaque commune pour toutes ses unités, mais surtout de faciliter le dépouillement lorsqu'un tableau doit être produit pour toutes les zones géographiques d'un même échelon (tableaux régionaux, départementaux, ...).

En effet, si un tableau de 100 cases doit être établi pour chaque département, le nombre de cases de mémoire utilisées ne sera que de 100, les mémoires étant vidées automatiquement à la fin de chaque département. Ainsi un tableau de 100 cases effectué pour chaque département n'est pas plus long à obtenir qu'un tableau de 100 cases pour la France Entière. Cette procédure sera particulièrement utile lorsqu'on établira des tableaux pour les échelons géographiques fins (tableaux communaux ou cantonaux dans l'exploitation exhaustive).

Enfin, pour des dépouillements locaux n'intéressant qu'une petite zone géographique, seule la bande du département concerné (ou des départements s'il y en a plusieurs) sera exploitée.

En revanche, si le découpage géographique retenu n'est pas un échelon géographique, c'est-à-dire n'est pas un critère utilisé pour constituer la séquence des unités sur la bande : grandes agglomérations (de 100.000 habitants et plus), agglomérations de 50.000 habitants et plus, zones de peuplement industriel ou urbain, régions agricoles.., il n'est plus possible d'utiliser la procédure indiquée ci-dessus : un tableau de 100 cases effectué pour chacune des 43 agglomérations de 100.000 habitants et plus serait équivalent à un tableau France Entière de 4.300 cases.

Dans ce cas, on procède de la façon suivante : on numérote chaque grande agglomération à l'intérieur du département (code CCRTC). Au cours d'un premier passage où on ne retient que les 30 départements ayant au moins une grande agglomération, on effectue le tableau pour la plus grande agglomération du département (critère d'appartenance portant sur CCRTC). L'échelon géographique est alors le département : les mémoires sont vidées à la fin de chaque département - c'est-à-dire à la fin de chaque agglomération. Au cours du 2ème passage, on ne retient que les 5 départements comportant au moins 2 agglomérations etc.... Ainsi au moyen de 7 passages de 43 départements au total (en fait : 52 départements à cause des agglomérations multi-départementales) on produit 2.000 cases de tableau pour chaque agglomération de 100.000 habitants et plus.

Dans le cas des régions agricoles cette procédure n'est plus applicable parce que le nombre de régions agricoles d'un même département peut être très important (26 en Gironde ...). On a prévu - dans le cas du sondage au 1/20 au moins - d'avoir recours à une bande annexe comportant les seules données relatives à l'agriculture (population agricole au sens le plus large).

Cette bande annexe sera ensuite triée par région agricole et exploitée à part, la région agricole devenant un échelon géographique de cette bande annexe.

Il en est de même pour l'étude des migrations alternantes ou des migrations définitives entre 1954 et 1962. Pour étudier la population active au lieu de travail ou la population suivant la résidence de 1954, on a recours à une bande supplémentaire extraite de la bande générale et triée suivant les caractéristiques de lieu de

travail ou de résidence antérieure. Cette bande n'intéresse que la fraction des personnes qui ont effectivement migré : actifs travaillant hors de leur commune de résidence, personnes habitant en 1962 une autre commune qu'en 1954. D'autre part on ne reporte sur cette bande que les éléments utiles : caractéristiques professionnelles, sexe, âge, etc... Ainsi, la bande qui servira à l'étude de ces problèmes particuliers est à la fois plus courte que la bande générale qui lui a donné naissance puisqu'elle ne contient que certaines caractéristiques de certains individus et plus maniable puisqu'elle sera triée suivant les critères géographiques utilisés.

4. La bande ne sera pas triée au cours de l'exploitation.

Cette condition oblige à passer l'ensemble des bandes pour obtenir un tableau quelconque, même si celui ci ne concerne qu'une population très restreinte (disséminée sur les bandes). Ainsi pour les tableaux portant sur les musulmans algériens, les rapatriés, les chômeurs, les apprentis, etc....

Pour accélérer le dépouillement il est prévu dans ces différents cas de recopier sur des bandes annexes ces sous populations particulières de faible effectif.

Pour la même raison, du fait qu'on ne trie pas la bande, on est amené à prévoir des procédures spéciales pour produire des tableaux qui comportent un nombre élevé de cases. Ainsi, le tableau croisant le sexe, la profession, la qualification et la catégorie d'activité économique contient environ 50.000 cases. Un tel tableau est produit de la façon suivante : on exploite la bande générale au cours d'un premier passage qui fournit 2.000 cases de tableau en même temps qu'on transcrit sur une bande annexe les quatre critères (croisés dans le tableau) pour les seuls individus dont l'adresse ne figurait pas parmi les 2.000 adresses disponibles. Cette bande annexe -beaucoup plus courte que la bande générale - est ensuite exploitée donnant naissance à son tour à 2.000 autres cases et à une seconde bande annexe etc... Les bandes annexes successives sont de plus en plus courtes et la procédure assez rapide.

5. Chaque caractère doit être chiffré entre 0 et 9

Les postes possibles de chaque caractère sur la bande sont seulement 0, 1, ..., 9, 0 correspondant à l'espace (+) des cartes mécanographiques. Or les cartes à partir desquelles est établie la bande magnétique peut recevoir 13 sortes de chiffrements : 0 à 9, X, Z, + (1).

Il est donc nécessaire de transformer les chiffrements X, Z et + en chiffrement uniquement numériques. Il s'ensuit que les codes de la bande peuvent différer des codes de la carte.

La procédure de transformation des codes en codes uniquement numériques est exposée dans l'annexe pages 208 à 228.

---

(1) Sans compter les perforations alphabétiques, non utilisées dans le recensement démographique.

6. La bande code comporte la liste des seuls postes possibles de chaque critère.

Il importe que la bande magnétique ne comporte que les seuls chiffrements inscrits sur la bande code. En effet si un individu, au moment du dépouillement, présentait un chiffrement non prévu, il ne pourrait être enregistré dans aucune case de mémoire et le calculateur se bloquerait.

Or, par suite d'erreurs de chiffrement ou de perforation, la carte peut comporter de tels chiffrements imprévus. Il est donc indispensable de vérifier intégralement le chiffrement des cartes et de transformer les chiffrements impossibles en chiffrements possibles. On profite d'ailleurs de cette vérification pour supprimer certaines incompatibilités de chiffrement relatives à deux ou plusieurs critères non indépendants. Ainsi, l'état matrimonial et l'âge ne sont pas indépendants : en-dessous d'un certain âge (on a retenu 17 ans pour les garçons et 15 ans pour les filles) seul l'état matrimonial "célibataire" (M = 1) est possible. Même si la carte d'un enfant comportait un chiffrement possible (M = 1 à 4, 9) on a inscrit sur bande M = 1. Outre les chiffrements impossibles, seules les incompatibilités "criantes" ont été éliminées.

L'intérêt de corriger automatiquement les incompatibilités au niveau des bandes est extrêmement grand : il ne sera plus nécessaire de corriger manuellement les tableaux et l'égalité comptable des effectifs ainsi assurée d'un tableau à l'autre est tout à fait précieuse pour identifier la nature exacte d'une sous-population concernée par un tableau particulier.

7. La bande code considère comme indépendants deux critères croisés et prépare un nombre de mémoires égal au produit des nombres possibles de postes.

Il est fréquent que les critères d'appartenance limitent le nombre des postes possibles des critères de ventilation (si les actifs sont ventilés par âge, les âges possibles ne commencent qu'à 15 ans par définition) ou que toutes les combinaisons de deux critères ne soient pas possibles (la CS a 30 postes d'actifs et le statut 10 postes alors que le croisement CS × Statut ne comporte que 136 postes).

Or le calculateur prépare un nombre de cases, de mémoire égal au produit des nombres de postes possibles. Pour éviter de gaspiller des cases de mémoire inutiles qui se traduiraient sur les tableaux bruts par des lignes ou des colonnes vierges, on a recours aux procédés suivants :

a) si le nombre de postes est réduit du fait des critères d'appartenance, on prévoit dans la bande code seulement - et non dans la bande magnétique définitive qui contient les données - un code spécial comportant moins de postes que le code initial.

Exemple : Actifs par sexe et âge : Comme un actif a nécessairement 15 ans ou plus, on utilise comme code de l'âge un code qui ne comprend que les postes 15 à 99. L'utilisation de ce code spécial dit "Age actif" est du ressort de la Division Electronique. Il est possible de créer ainsi à tout moment au cours de l'exploitation des codes de ce genre sans recourir à une codification nouvelle de la bande magnétique définitive.

b) Si le nombre de postes est réduit du fait des critères de ventilation, on constitue des super codes :

Ainsi, si dans un tableau on croise en ligne la CS et le statut, on utilisera le code unique à 3 chiffres: (CS × ST) qui comporte 136 postes possibles : 001 à 829. Ce nouveau critère est tout à fait analogue aux autres critères. En outre, chacun de ses chiffres a une appellation : dizaine de CS, unité de CS, statut et peut être utilisé indépendamment des deux autres. Le groupe des deux premiers chiffres est la CS.

L'utilisation des super codes est particulièrement précieuse dans le cas de la description des ménages suivant le nombre de personnes des divers liens: Nombre de personnes × Présence d'un conjoint × Nombre d'enfants × Nombre d'ascendants × Nombre de parents × Nombre d'autres personnes.

En l'absence d'un super code on aurait 10 × 2 × 10 × 10 × 10 × 10 = 200.000 combinaisons possibles alors que 100 combinaisons suffisent pour décrire 99,5 % des ménages ....

Signalons cependant un inconvénient des super codes : ils ne permettent pas au stade des tableaux bruts d'obtenir des totalisations marginales ..

Si le critère ligne est (CS × ST), on n'obtiendra pas les sous totalisations CS à 1 chiffre × statut, CS à 2 chiffres sans le statut, statut seul. Si on veut les obtenir, on peut, soit attendre le tableau définitif, soit établir des tableaux distincts où on croiserait en ligne successivement : (DCS × ST), CS, ST.

8. <u>Au stade des tableaux bruts la ventilation relativement à un critère est effectuée suivant l'ensemble des postes de ce critère.</u>

Il est impossible au niveau des tableaux bruts de regrouper certains postes de certains critères. Si on désire un tableau fournissant la pyramide des âges et si, seul, l'âge à deux chiffres a été porté sur la bande magnétique, les trois seuls découpages possibles de l'âge sont les suivants :

- découpage annuel depuis 00 jusqu'à 99 ans.
- découpage décennal en ventilant suivant la dizaine d'âge.
- découpage, qui n'aurait aucun sens en l'absence de critère d'appartenance portant sur la dizaine d'âge, où on ventilerait suivant l'unité d'âge.

Sans doute, au niveau des tableaux définitifs peut on effectuer tous les regroupements souhaitables. Mais pour que les tableaux bruts soient aussi lisibles et peu encombrants que possible et pour ne pas gaspiller des cases de mémoire inutilement on a été amené à recourir aux codes condensés et articulés.

Ainsi le code condensé de l'âge : AC qui constitue un code entièrement distinct de l'âge détaillé est un code à 3 postes :

AC = 1 pour les personnes de moins de 20 ans

AC = 2 pour les personnes de 20 à 64 ans

AC = 3 pour les personnes de 65 ans et plus

Pour éviter de prévoir à l'avance tous les découpages qui seront utilisés et d'allonger démesurément la bande magnétique en constituant un grand nombre de codes condensés, on a d'autre part mis au point les codes articulés dont on trouvera le principe page 29 à l'occasion du code A : âge articulé.

9 - Les critères intervenant dans un même tableau sont nécessairement croisés.

Un tableau de la forme : (Sexe × Age) × (CS + ST) ne peut être réalisé directement au niveau des tableaux bruts. Il doit être décomposé en deux parties :

1ère partie : (Sexe × Age) × (CS)

2ème partie : (Sexe × Age) × (ST)

Au niveau des tableaux définitifs il sera possible de fusionner les deux parties et de présenter l'ensemble sous la forme désirée.

III - LE TRAITEMENT DES DONNEES PREALABLEMENT A L'EXPLOITATION

LA CONFECTION DES BANDES MAGNETIQUES DEFINITIVES

Nous avons évoqué certaines opérations nécessitées par la procédure de dépouillement : vérification intégrale du chiffrement en vue de la suppression des postes impossibles de chaque critère et de l'élimination des incompatibilités les plus gênantes, fabrication de codes condensés et articulés en vue de croiser, dès le stade des tableaux bruts, des critères suivant le découpage le plus adéquat, transformation des chiffrements de la carte en chiffrements uniquement numériques.

Il est d'autres opérations qu'il a été possible d'entreprendre grâce à l'utilisation de l'ordinateur : suppression du poste non déclaré de certains critères ce qui a l'avantage de réduire d'une unité le nombre de postes de ces critères mais surtout d'améliorer la présentation et même la qualité de l'information recueillie, fabrication de codes calculés qui résulte d'une analyse souvent complexe des unités statistiques et qui n'aurait pu être menée à bien qu'au prix d'un travail de chiffrement long et coûteux.

Nous précisons dans ce chapitre la suite des opérations concernant le traitement des données avant tout travail de dépouillement. La procédure décrite se rapporte à l'échantillon au 1/20 mais il est probable que l'ensemble des données du recensement sera traité de la même façon. (ou de façon analogue).

1. L'information de base : les cartes mécanographiques

L'information de base est constituée par les cartes mécanographiques établies par les directions régionales.

Ces cartes sont de nature et de dessin différents :

carte maîtresse communale	(type x)
carte "individu-logement"	(type 1 à 4)
carte "immeuble"	(type 0)

Ces cartes sont mises en séquence, c'est-à-dire triées dans l'ordre des échelons géographiques (département, canton, commune) et à l'intérieur de la commune suivant les numéros de district, d'immeuble et de logement (caractéristiques communes aux divers dessins de cartes) :

Séquence des cartes à l'introduction :

Numéro du district	Numéro du immeuble	Numéro du logement	Type de carte	Catégorie de logement	Unité statistique		LOT
col. 11-12-13	col. 14-15-16	col. 17-18	col. 1	col. 71			
000	000	00	X	+	Carte maitresse communale		
000	900	00	4	+	BI 2bis B des personnes qui n'ont pu être reclassées dans un logement		F
001 à 999	001 à 699	00	0	+	Immeubles		A
		01 à 99	1	1 à 5	Population des résidences principales	Chefs de ménage et logements	A
				+		Autres personnes	A
				6	Résidences secondaires		A
				7	Logements vacants		A
	700 ou 750	01 à 99	1	8	Population des habitations mobiles	Chefs de ménage	E
				+		Autres personnes	E
	800	01 à 99	3	+	Ménages collectifs		D
	900	01 à 99	2	+	Population comptée à part		B_1 et C

2. Le report des cartes sur bande - L'introduction

L'ordinateur 1401 effectue la transcription sur bande, en contrôlant la séquence des cartes et en vérifiant que :

- le nombre de résidences principales de chaque immeuble est correct : le nombre de cartes de type 1 et de catégorie de logement 1 à 5 est égal au nombre de résidences principales figurant sur la carte immeuble correspondante.

- le nombre de résidences secondaires (et de logements vacants) est également correct : le nombre de cartes de type 1 et de catégorie de logement 6 (et 7) est égal au nombre de résidences secondaires (et vacantes) figurant sur la carte immeuble correspondante

- le nombre de personnes de chaque logement ou habitation mobile est correct : le nombre de cartes de type 1 et de CL + est égal au nombre annoncé dans la carte chef de ménage correspondante.

En cas de désaccord il est établi une liste des cartes litigieuses ; si l'erreur peut être corrigée facilement, la bande reçoit les chiffrements corrigés. Dans le cas contraire - en particulier lorsque certaines cartes manquent dans le fichier - il est fait appel à la Direction Régionale concernée qui, remontant aux questionnaires, perfore de nouvelles cartes.

Après ce contrôle des données, on aboutit à certaines définitions :

ménage : ensemble des cartes de type 1 ayant le même numéro de logement : NPT (nombre de personnes au total).

résidence principale : carte de type 1 de CL = 1 à 5.

logements d'un immeuble : ensemble des cartes de type 1 et de CL = 1 à 7 ayant le même numéro d'immeuble

population d'un immeuble: ensemble des cartes de type 1 et de CL = 1 à 5 ou + ayant le même numéro d'immeuble

chef de ménage : carte de type 1 de CL = 1 à 5 (définition identique à celle des résidences principales) ou de CL = 8 (habitations mobiles)

A ce stade, la bande de travail sur laquelle sont recopiées les cartes comporte les mêmes chiffrements que celles-ci (caractères alpha-numériques).

3. Vérification du chiffrement :

Correction des incompatibilités - Suppression des chiffrements impossibles

Cette bande de travail est ensuite reprise en vue de la vérification du chiffrement : correction des incompatibilités et des chiffrements erronés.

Le principe de la méthode de correction est le suivant :

a) si pour certains individus un seul poste est possible, on porte automatiquement le seul chiffrement possible.

Ainsi lorque la CS est 67, le statut est 4 nécessairement. Tout autre statut que 4 est corrigé en 4.

b) s'il existe un poste "non déclaré" le poste impossible est corrigé en "non déclaré"

c) s'il n'existe pas de poste "non déclaré", on corrige

- en mettant à la place du poste erroné le poste de loin le plus fréquent, s'il en existe un.

- en tenant compte d'un autre critère supposé correctement chiffré et en corrigeant automatiquement de façon unique le poste erroné.

- en affectant au hasard l'un des postes possibles à la place du poste impossible.

Chaque critère est examiné après celui ou ceux qui servent à le contrôler, pour éviter de remettre en question un critère déjà vérifié. De ce fait, l'ordre des corrections est le suivant : immeuble, logement, individu.

Il est établi une statistique des corrections de façon à déceler toute erreur systématique de chiffrement ou même de programmation : il serait pour le moins fâcheux de corriger à tort des chiffrements exacts en chiffrements inexacts ....

On trouvera en annexe, à propos de chacun des codes de la bande définitive, la procédure de correction utilisée.

Le schéma de correction est le suivant :

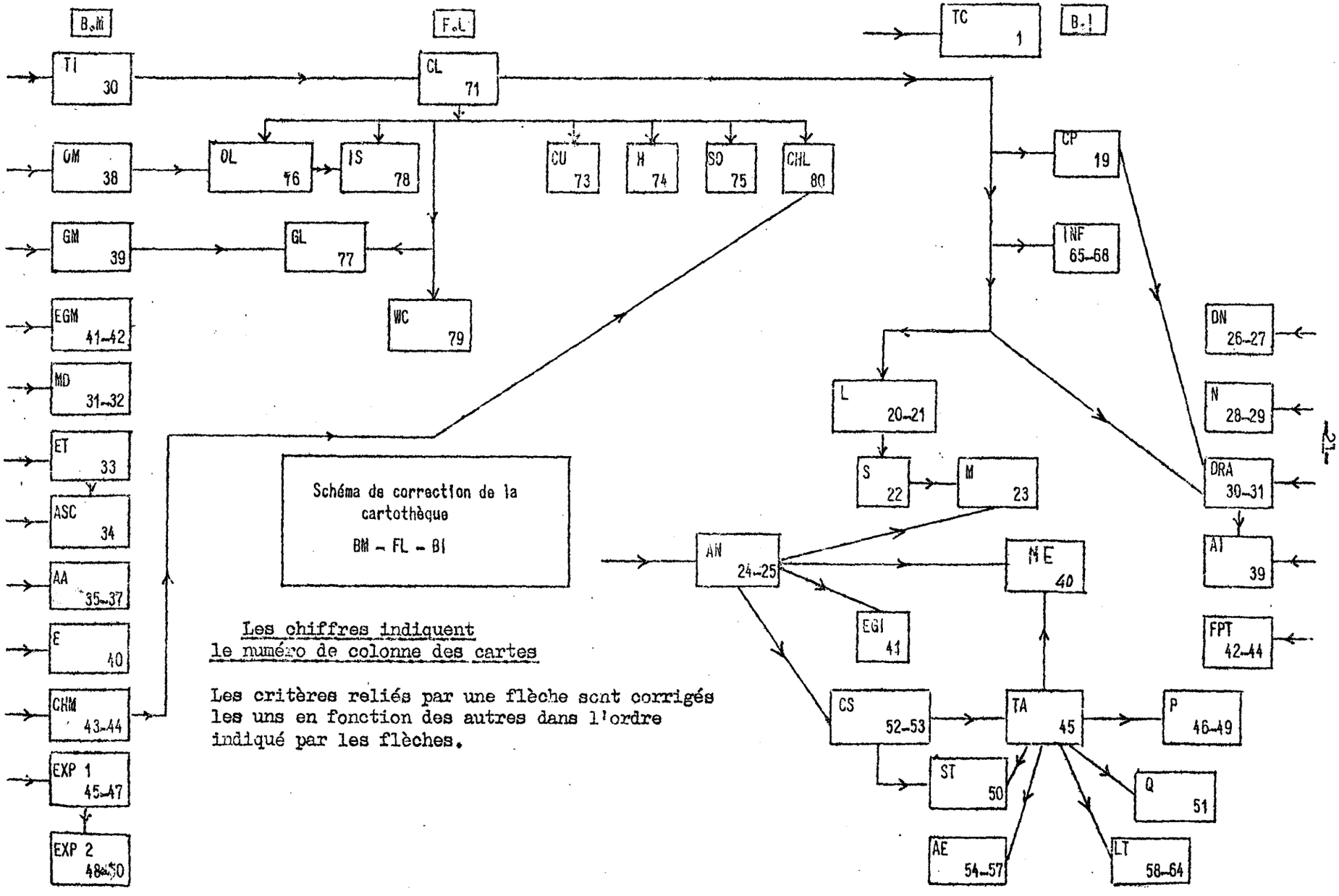


Schéma de correction de la cartothèque

BM - FL - BI

Les chiffres indiquent le numéro de colonne des cartes

Les critères reliés par une flèche sont corrigés les uns en fonction des autres dans l'ordre indiqué par les flèches.

4. Transformation des codes en codes uniquement numériques

Nous avons vu ci-dessus que la bande magnétique définitive est écrite en caractères numériques seulement. Or la carte peut comporter les perforations 0 à 9, X, +. Il importe donc de transformer les X ou + éventuels en chiffrements numériques.

Méthodes utilisées

Plusieurs méthodes sont possibles suivant le critère envisagé :

a) Si le code du critère comporte (y compris X et +) 10 postes au plus, on transformera X et + en l'un des chiffrements numériques non utilisés :

Exemple :

. Le lien avec le chef de ménage peut être chiffré L = 1 à 9, +.

On transformera + en 0.

. L'année d'installation comporte les postes AI = +, 0 à 2, 4 à 9.

On transformera le + en 3.

b) Si le code du critère comporte, outre le poste + un poste sans objet chiffré entre 0 et 9, on transforme + en ce poste sans objet, la sélection étant obtenue par critère d'appartenance :

Exemple :

. Le type d'activité est chiffré + pour les enfants de 14 ans et moins. Pour les personnes de 15 ans et plus TA = 0 correspond aux inactifs. On transformera donc TA = + en TA = 0, la sélection entre TA = 0 et TA = + pouvant être obtenue par critère d'appartenance portant sur l'âge.

c) S'il n'est pas possible d'utiliser l'une des deux méthodes précédentes, on peut soit envisager l'utilisation d'un caractère supplémentaire, soit se résoudre à une légère perte d'information.

Exemples : . le pourcentage d'invalidité est chiffré INF = 0 à 9, X, +

La transformation est la suivante :

Code carte	Code bande
0 à 9	00 à 09
+	10
X	11

. le nombre d'étages au dessus du rez de chaussée comporte les postes ET = 0 à 9, X.

On fusionne les postes "8"et "9" pour réserver le chiffrement "9" aux non déclarés.

Code carte	Code bande
0 à 8	0 à 8
9	8
X	9

Cette dernière procédure ne sera utilisée que pour ce critère.

5. Elimination du poste "non déclaré" de certains critères

La présence, dans le code de certains critères, d'un poste "Non déclaré" gêne considérablement la présentation des résultats du recensement :

Lorsque le critère envisagé est combiné en ligne ou en colonne, dans un tableau statistique, on est conduit - pour assurer les concordances d'un tableau récapitulatif à l'autre - à ajouter une ligne ou une colonne "Non déclaré", ce qui peut augmenter considérablement la taille du tableau. Ainsi le tableau donnant la répartition de la population par sexe, âge quinquennal, état matrimonial comporte - outre les récapitulations - 2 × 20 × 4 = 160 cases, si tous les critères sont déclarés. Si l'état matrimonial comporte seul un poste "non déclaré", le tableau a 200 cases et si les trois critères comportent chacun un poste non déclaré, le tableau a 3 × 21 × 5 = 315 cases.

Lorsque, le critère envisagé est un critère d'appartenance, c'est-à-dire que chacune de ses modalités conduit à un tableau différent, on est nécessairement amené - toujours pour assurer la concordance d'un tableau récapitulatif à l'autre - à créer un tableau correspondant à la modalité "Non déclaré". Ainsi la description des ménages suivant le sexe et l'état matrimonial du chef de ménage en deux postes (marié - non marié) est assurée par 2 × 2 = 4 tableaux statistiques. Si le sexe et l'état matrimonial comportent un poste "non déclaré", ce n'est plus 4 mais 3 × 3 = 9 tableaux statistiques qui deviennent nécessaires - et parmi ces 9 tableaux, 5 n'ont guère de sens : "Le chef de ménage est de sexe indéterminé - mais marié", ... "Le chef de ménage est de sexe indéterminé et d'état matrimonial indéterminé" .... Notons qu'après la suppression des chiffrements impossibles pour chacun des critères de la bande, seul l'état matrimonial comporte dans l'exemple présent un poste "non déclaré".

Au stade de la publication, on peut envisager plusieurs solutions, en ce qui concerne le (ou les) poste(s) "non déclaré" d'un tableau :

- faute de pouvoir corriger ce poste, on publie le tableau tel quel, avec cette ligne ou cette colonne "non déclaré".

- on juge inadmissible ce poste, surtout dans le cas où le critère est un critère d'appartenance, et on "bloque" le poste non déclaré avec l'un des postes possibles. Ainsi, dans la description des ménages suivant l'état matrimonial du chef de ménage : les deux tableaux sont "marié, "non marié", le poste "non marié" comprenant la modalité "non déclaré".

- on juge inadmissible ce poste "non déclaré" et on ventile, sur les tableaux, proportionnellement aux réponses "déclarées" les effectifs du poste "non déclaré". Cette procédure - peu justifiable dans certains cas par ce que les "non déclarés" proviennent souvent d'une modalité spécifique et sont loin de se répartir proportionnellement aux réponses déclarées - est désastreuse en ce qui concerne les concordances de tableau à tableau. D'autre part, cette procédure ne permet pas de profiter pleinement des tableaux définitifs qu'il faudrait remanier avant publication.

a) Choix des critères sur lesquels porte l'élimination des "non déclaré".

De même que c'est au niveau des bandes qu'il convient de corriger les incompatibilités, pour éviter d'avoir à faire des "raccommodages" pénibles sur les tableaux, de même c'est au moment de l'introduction de l'information sur bandes qu'il convient de corriger les postes "non déclaré" jugés inadmissibles. Le choix des critères ainsi corrigés peut être guidé par les considérations suivantes :

- si le poste "non déclaré" représente une faible fraction de l'effectif total, il est utile de l'éliminer : c'est l'optique "nettoyage des tableaux".

- si le critère comportant un poste "non déclaré" est un critère fondamental apparaissant dans de nombreux tableaux combiné à d'autres critères, il est utile d'éliminer le poste "non déclaré" pour simplifier la présentation des résultats et éviter des calculs pénibles ultérieurs. Toutefois, il faut faire en sorte que la correction de ce poste n'introduise pas des incompatibilités avec d'autres postes correctement chiffrés. C'est ainsi qu'il sera difficile de corriger les données du bordereau de maison, qui peuvent créer des incompatibilités avec celles des feuilles de logement correspondantes.

- enfin, compte tenu de la méthode de correction exposée ci-dessous, il ne sera possible d'éliminer les "non déclaré" que si le coût de l'opération n'est pas exagéré : il importe en effet de ne pas encombrer inconsidérément les mémoires de l'ordinateur au cours de la correction.

Il faut remarquer toutefois, qu'étant donnés les avantages de l'opération, il n'est pas inutile d'y consacrer quelques moyens. Par ailleurs, nous allons voir que la méthode utilisée améliore non seulement la présentation mais sans doute également la qualité de l'information.

Après examen des critères comportant un poste "non déclaré", il a été décidé de corriger :

- sur la carte "immeuble"

AA : Année d'achèvement (colonne 35 seulement)

- sur la carte "logement" :

CU : Existence et nature de la cuisine.
H : Nombre de pièces d'habitation.
SO : Statut d'occupation.

- sur la carte "individu" :

M : Etat matrimonial
AE : Activité économique

b) La méthode de correction

La méthode de correction utilisée est empruntée au "Bureau of the Census" qui l'a mise au point pour le dépouillement du recensement américain de 1960. On en trouvera la description dans un article de TORE DALENIUS (Statistical Review n° 7 - 1962, pages 395 à 400).

Cette méthode automatique est la suivante :

Considérons trois critères du bulletin individuel et, pour fixer les idées : le sexe, l'état matrimonial (en deux postes) et l'âge (annuel) et supposons que sur chaque carte le sexe et l'état matrimonial soient toujours déclarés alors que l'âge peut ne pas l'être.

Les individus sont donc caractérisés par un ensemble de trois chiffrements

(S, M, A) s'ils ont tout déclaré

(S, M, -) s'ils n'ont pas déclaré leur âge

Associons à l'opération de redressement le tableau à double entrée S × M dans lequel nous portons des âges quelconques et a priori possibles : ces âges sont par exemple des médianes de distributions connues (recensement antérieur, enquête antérieure) ou même des âges déterminés au jugé. Le choix du tableau initial n'a pas trop d'importance pour la suite (sauf s'il comportait un nombre très élevé de cases).

Sexe / Etat matrimonial	MASCULIN	FEMININ
Marié	45	40
Non marié	18	(21)

Ce tableau initial est introduit dans une mémoire de l'ordinateur. Les individus dont les caractéristiques sont portées sur la bande magnétique sont examinés un à un. Supposons que le premier individu soit une femme non mariée d'âge déclaré 30 ans. Alors, à la place de 21 dans la case "Femme non mariée" on porte 30 et le tableau devient :

Sexe / Etat matrimonial	MASCULIN	FEMININ
Marié	45	40
Non marié	18	(30)

Et ainsi de suite ..... Chaque individu déclaré remplace, sur le tableau, l'individu antérieur qui possèdait les mêmes caractéristiques de sexe et d'état matrimonial que lui.

Le tableau initial se déforme au fur et à mesure que défilent sur bande les individus ayant déclaré leur âge (ce qui explique que l'amorçage de la procédure ne conditionne guère la suite des opérations).

Au bout d'un certain temps, le tableau est, par exemple, le suivant :

Sexe / Etat matrimonial	MASCULIN	FEMININ
Marié	61	(42)
Non marié	27	12

Supposons que sur la bande se présente ensuite un individu d'âge non déclaré mais de sexe et d'état matrimonial déclaré : il s'agit par exemple d'une femme mariée; on convient de lui affecter l'âge 42 ans, c'est-à-dire l'âge de la dernière femme mariée précédemment examinée.

Il est facile de voir que cette correction revient à une correction aléatoire qui en moyenne est une répartition proportionnelle. La distribution des âges attribués aux femmes mariées d'âge non déclaré est en moyenne la distribution des âges des femmes mariées qui ont déclaré leur âge. Par conséquent, la correction est en moyenne celle qu'on ferait à partir des tableaux pour évaluer à la date du recensement la répartition de la population par sexe, âge, état matrimonial. L'intérêt de la méthode est que la correction est effectuée au niveau des bandes une fois pour toutes et non au niveau de chacun des tableaux.

Généralisation de la méthode

L'exemple que nous venons de citer est volontairement schématique. D'ailleurs dans notre recensement, l'âge est toujours déclaré (la chiffreuse ayant dû, éventuellement, en attribuer un d'après l'examen des bulletins figurant dans la feuille de ménage) alors que l'état matrimonial peut être non déclaré (lorsque la chiffreuse n'a pu l'établir de façon précise : H = 9 col. 23).

La méthode peut être généralisée dans plusieurs directions :

- on peut considérer non pas deux (ici le sexe et l'état matrimonial) mais un nombre quelconque de critères de contrôle comprenant chacun un nombre quelconque de postes. Le choix des critères de contrôle peut être délicat, il convient naturellement de retenir :

- des critères fortement en corrélation avec le critère qu'on se propose de corriger.

- des critères de contrôle qui ne risquent pas d'introduire d'incompatibilités avec d'autres critères non pris en considération (voir ci-dessous).

- des critères en petit nombre et ayant chacun un nombre restreint de postes (au besoin en regroupant certains des postes possibles) pour éviter d'alourdir exagérément l'opération (encombrement de la mémoire du calculateur, qui, malgré tout, a une capacité limitée). On peut corriger, de façon indépendante ou non, à la fois le critère envisagé et certains critères de contrôle. On a supposé plus haut que sexe et état matrimonial étaient toujours déclarés. Il se peut également qu'ils ne le soient pas.

Remarques sur la méthode

- Du fait qu'on peut utiliser simultanément plusieurs critères de contrôle, la répartition proportionnelle à laquelle aboutit - en moyenne - la méthode est meilleure que celle qu'on peut effectuer à partir des tableaux où on corrige souvent avec 0, 1 ou au maximum 2 critères de contrôle.

- Du fait que les individus sont inscrits en séquence sur la bande, l'attribution d'un poste "déclaré" aux critères qui présentent des "non déclaré" est stratifiée géographiquement ; l'individu "déclaré" qui permet d'attribuer le chiffrement à un autre individu "non déclaré" est voisin sur la bande, mais aussi dans la réalité. Les individus d'une région dont l'activité économique n'est pas déclarée sont ventilés proportionnellement à la répartition suivant l'activité économique des individus de la même région qui ont déclaré leur activité économique. Cette propriété de la méthode est d'un très grand intérêt.

Néanmoins, comme nous l'avons déjà souligné, la méthode a des limitations :

- Le nombre de critères de contrôle retenus - ou plus précisément le nombre de cases, éventuellement multidimensionnelles, du tableau de correction introduit en mémoire - ne doit pas être exagéré.

- La correction ne doit pas introduire des incompatibilités avec d'autres critères qui ne sont pas pris en considération.

Supposons que dans l'exemple précédent, on ait retenu, pour déterminer l'âge, le seul critère de contrôle : sexe. Le tableau est à un certain moment le suivant :

MASCULIN	FEMININ
61	12

L'individu suivant qui se présente sur la bande est une femme mariée d'âge non déclaré : la méthode de correction conduit à attribuer l'âge 12 ans. Or 12 ans est incompatible avec l'état matrimonial "marié" - A vouloir corriger un "non déclaré", on introduit une incompatibilité.

Par conséquent les critères de contrôle retenus doivent être assez nombreux pour ne pas créer d'incompatibilités avec les critères non retenus comme critères de contrôle (ceci est lié avec la condition de forte corrélation déjà indiquée plus haut) et pas trop nombreux pour ne pas alourdir l'opération. Pratiquement on utilisera deux critères de contrôle et parfois trois.

Il est établi une statistique des redressements de "non déclarés".

Cette précaution est indispensable pour s'assurer de la méthode de correction mais aussi pour connaître la répartition des "déclarés" et des "non déclarés" suivant le critère corrigé.

Exemple : Correction de SO (Statut d'occupation) en fonction de CL × TI (catégorie de logement et type d'immeuble) :

- Nombre de "non déclarés" sur SO par case du tableau de redressement (quel que soit le chiffrement attribué à la place de SO = 9).

Répartition des chiffrements attribués à ces "non déclarés" (quelle que soit la case CL × TI ayant permis le redressement).

C'est-à-dire :

- Répartition des non déclarés en 21 postes : CL × TI.
- Répartition des non déclarés en 6 postes : SO attribué

6. Elaboration des codes condensés, articulés et calculés

a) Les codes condensés

La nécessité de construire des codes condensés tient à la procédure de dépouillement : lorsqu'un critère est utilisé comme critère de ventilation, les unités statistiques appartenant au champ du tableau sont ventilées suivant tous les postes possibles de ce critère.

Or, pour éviter un détail excessif ou résumer l'information, on a souvent besoin de condensations des codes en quelques postes. Par exemple :

- AC (âge condensé) en 3 postes : moins de 20 ans, 20-64 ans, 65 ans et plus.
- MC (état matrimonial condensé) en 2 postes : marié, non marié.
- CSCO (CS condensé) en 2 postes : agricole, non agricole
- NC (nationalité condensé) en 4 postes : Français, devenu Français, Musulman algérien, étranger.
- etc. ...

Ces nouveaux codes sont entièrement distincts des codes qui leur ont donné naissance et qu'ils résument. Ils peuvent naturellement être utilisés comme critères de ventilation ou comme critère d'appartenance.

Le choix des codes qu'il est utile de condenser est fonction des tableaux que l'on compte produire. Ainsi le code CSCO a été introduit en vue de l'étude des populations agricoles : population des ménages agricoles, population des exploitations agricoles, etc....

Notons que si pour un tableau donné, on ne dispose pas du code condensé nécessaire, on peut éviter grâce à des critères d'appartenance de multiplier inutilement les cases du tableau et d'effectuer de nombreux regroupements sur le tableau définitif. Si on veut obtenir par exemple une ventilation de la CS en deux postes (agricole et non agricole) en l'absence du code condensé CSCO, on effectue le tableau en deux parties : la première partie pour la population agricole (critère d'appartenance : CS = 00, 10 ou 93) et la seconde partie pour la population non agricole. Au stade du tableau définitif, on fusionne les deux parties.

Cette procédure est néanmoins assez lourde et rend les tableaux bruts peu lisibles. D'autre part, chaque tableau nécessite un temps d'assemblage de programmes à peu près constant de 10 minutes sur 7070. Diviser ainsi un tableau en deux parties conduit à effectuer deux tableaux - et même 3 si on veut dès le stade des tableaux bruts la récapitulation toutes CS - ce qui double ou triple les temps d'assemblage.

b) Les codes articulés

Les codes articulés permettent, au moyen d'un nombre restreint de caractères sur la bande magnétique définitive, la ventilation de certains critères suivant un découpage variable adapté au tableau demandé.

Ainsi l'âge A est décomposé en 3 tranches ou articulations : I = 00 à 24 ans, II = 25 à 54 ans, III = 55 à 99 ans (voir page 134). Dans chaque articulation on définit 6 découpages différents : par exemple, le découpage n° 5 comporte 11 postes dans l'articulation n° I : 00 à 14 ans, 15 ans, 16 ans, ... 24 ans. Dans l'articulation n° II, le découpage n° 4 est annuel : 25 ans, 26 ans, ... 54 ans.

Dans l'articulation n° III le découpage n° 5 comprend 21 postes : découpage annuel de 55 à 74 ans puis l'ensemble 75-99 ans en un seul poste.

Pour définir un découpage de l'âge, on précise le numéro de découpage de chaque articulation. Ainsi le découpage utilisé pour l'étude fine de la population active : 0-14 ans, 15 ans, 16 ans, ... 74 ans, 75 ans et plus porte le nom

A 545 puisqu'il correspond au 5e découpage de l'articulation I,
au 4e " " II,
au 5e " " III.

Chaque découpage de l'ensemble des âges constitue un code condensé indépendant. L'âge 545 est un code distinct de l'âge 111 qui correspond au découpage en 3 postes seulement : 0 à 24 ans, 25 à 54 ans, 55 ans et plus.

On a prévu dans chaque articulation le découpage 0 qui signifie "hors champ" : le code A040 par exemple correspond au découpage de la seule tranche 25-54 ans en classes annuelles.

Grâce au code articulé de l'âge on peut aussi définir $7^3-1 = 342$ découpages différents constituant en fait 342 codes condensés distincts dont le chiffrement sur bande est réalisé au moyen de 18 caractères seulement.

On a articulé quelques-uns des critères les plus fréquemment utilisés dans les dépouillements : âge individuel (A), âge du chef de ménage (AM), âge du chef de noyau (AN), activité économique individuelle (AE), la catégorie d'activité économique individuelle (CAE), du chef de ménage (CAEM), du chef de noyau (CAEN), la catégorie socioprofessionnelle individuelle (CS), du chef de ménage (CSM), du chef de noyau (CSN), le statut individuel (ST), du chef de ménage (STM) et du chef de noyau (STN).

Les codes articulés permettent ainsi beaucoup de souplesse dans les découpages, sans pourtant être universels. Au stade des tableaux définitifs, quelques regroupements seulement doivent suffire à présenter le tableau sous la forme exactement désirée. Il convient de noter cependant que sur ordinateur la ventilation suivant un code articulé est un peu plus longue que suivant un code normal et qu'il est préférable lorsqu'il est possible de recourir au code détaillé correspondant. Ainsi le découpage de l'âge en classes annuelles peut aussi bien être défini par A444 que par AD (âge détaillé). Il est préférable d'utiliser le second.

C'est pourquoi, malgré la mise au point des codes articulés, on a conservé sur bande les codes initiaux de la carte, désignés par la lettre supplémentaire D (détaillé):

AD : âge détaillé (1er chiffre DAD, 2è chiffre UAD)

CSD : CS détaillé (DCSD, UCSD), sous code du supercode CSSTC : CS-Statut calculé (cf page 15/)

STD : Statut détaillé.

De même :

ADM : âge détaillé du chef de ménage (DADM, UADM)

CSDN : CS détaillé du chef de noyau (DCSDN, UCSDN)

etc ....

D'autre part - et c'est une raison supplémentairede reporter sur bande, en plus des codes articulés, les codes détaillés - un code articulé est seulement un critère de ventilation et ne peut être utilisé comme critère d'appartenance.

c) Les codes calculés

Alors qu'un code condensé est établi à partir d'un critère seulement, les codes calculés sont élaborés à partir d'un nombre quelconque de critères.

Par exemple le code (CSSTA) "CS-Statut agricole" résume les critères CS et statut en vue des tableaux sur l'agriculture :

CSSTA = 0 pour la population qui ne vit pas dans un ménage ordinaire.

CSSTA = 1 pour les agriculteurs exploitants patrons (CS = 00 et ST = 1 ou 2)

etc.... suivant le tableau ci-dessous :

Si à la fois			alors CSSTA =	Signification en clair :
NUSI =	CS =	ST =		
1	Quelconques		0	Ménages collectifs, population comptée à part, lot F
2 ou 3	00	1 ou 2	1	Agriculteurs exploitants (patrons)
		3	2	Agriculteurs exploitants (aides familiaux)
	10	Quelconque	3	Salariés agricoles
	21 à 82	0 à 3	4	Actifs non agricoles (patrons et aides familiaux)
		4 à 9	5	Salariés non agricoles
	92	0	6	Militaires du contingent
	93		7	Ancien agriculteur
	91, 94 à 99		8	Autres inactifs

Ainsi, les codes calculés réalisent des condensations parmi les croisements possibles d'un nombre quelconque de critères.

On trouvera pages 96 et 231 le code CEL qui résume en 38 postes seulement le croisement des critères suivants :

- Catégorie de logement - Nature et existence de la cuisine - Nombre de pièces - Matériaux dominants - Alimentation en eau du logement et de l'immeuble - Cabinets d'aisance - Installation sanitaire - Année d'achèvement de l'immeuble.

Cette analyse des caractéristiques des logements n'aurait pu être effectuée sur les tableaux où ces critères auraient été croisés du fait de leur ampleur. Par ailleurs de nombreux croisements auraient été sans grande signification.

D'autres codes calculés résultent d'une analyse des données au niveau du ménage ou du noyau : nombre des personnes des divers liens, nombre d'enfants de 16 ans et moins suivant leur âge, etc... L'analyse des noyaux familiaux qui n'avait pu être réalisée en 1954 que sur un échantillon au 1/100 est étendue à l'ensemble de la population. De plus, grâce à l'ordinateur, on a pu procéder à l'analyse des familles biologiques, subdivisions des noyaux familiaux. En ce domaine, l'utilisation de l'électronique a été un progrès considérable.

7. CONCLUSION

Tels sont les divers traitements que subissent les données portées dans les cartes mécanographiques en vue de la préparation de la bande magnétique définitive :

- Contrôle de la séquence.
- Suppression des chiffrements impossibles et des incompatibilités.
- Transformation des codes en codes uniquement numériques.
- Elimination du poste "non déclaré" de certains critères.
- Codification proprement dite : codes condensés, calculés, articulés.

Ces différentes opérations qui mettent en oeuvre l'ordinateur 7070 et l'ordinateur 1401 nécessitent environ 8 heures de 1401 et 7 heures de 7070 pour 100.000 personnes