JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. PÈPE

L'automatisation dans la préparation des documents statistiques à l'exploitation

Journal de la société statistique de Paris, tome 100 (1959), p. 28-35

<http://www.numdam.org/item?id=JSFS_1959__100__28_0>

VI

VARIÉTÉS

L'automatisation dans la préparation des documents statistiques à l'exploitation

En présence des progrès rapides de l'électronique, l'esprit humain se laisse tenter par toutes les anticipations que l'on peut imaginer.

Science ou fiction, rêves d'aujourd'hui et réalités de demain? Telles sont les questions que pose la lecture de cet article.

***

A — LE GOULOT D'ÉTRANGLEMENT

Entre autres difficultés, l'exploitation des données statistiques rencontre un goulot d'étranglement dont la suppression ne ressort malheureusement pas de la théorie des queues, ni de la méthode de Montecarlo.

Les machines à cartes perforées ont, il y a déjà quelques décennies, pris efficacement la relève des méthodes manuelles qui ne mettaient en jeu que du papier, des crayons et quelques machines calculatrices de bureau fort simples; et, depuis peu, les calculatrices électroniques semblent prêtes à succéder aux machines à cartes perforées, déjà fort intéressantes et fort efficientes, pour améliorer encore la vitesse et l'économie des résultats.

Mais ces progrès ne concernent que l'exploitation proprement dite, c'est-à-dire :

— la sélection et le dénombrement des données;

— l'analyse, la tabulation et l'impression des résultats.

Ils sont beaucoup plus faibles, et parfois même inexistants en ce qui concerne trois catégories préalables d'opérations fastidieuses, monotones et interminables, avoir :

1° la vérification préliminaire des données, afin d'en éliminer les lacunes, les omissions, les incompatibilités, les invraisemblances;

2° la traduction des données initiales du document de base en un langage que comprennent les machines. Jusqu'ici ces dernières ne savent interpréter que les chiffres, et à la rigueur les lettres de l'alphabet (encore, ne faut-il pas leur demander des opérations compliquées sur des lettres). Un premier aspect de cette traduction est donc le remplacement des mots du langage courant par des nombres ayant un sens codifié.

3° Quant au deuxième aspect de la traduction, il réside dans l'incapacité des machines à comprendre les documents écrits à la main, dactylographiés ou imprimés, même s'ils ne comportent que des chiffres.

Il faut y substituer certains supports physiques (carte perforée, bande de papier perforé, bande magnétique) où les chiffres ont subi une deuxième traduction codifiée, sous forme de perforations, de taches coloriées lisibles

par une cellule photoélectrique, de spots magnétisés lisibles par une tête magnétique, etc...

Parfois même, cette traduction n'est pas directe et se fait par étages successifs : on substitue au chiffre initial une marque graphique, puis à la marque graphique une perforation, puis à la perforation une magnétisation de ruban. Toutefois le premier étage seul requiert l'intervention humaine; tous les autres sont assurés par des machines automatiques à grande vitesse.

De ces trois catégories d'opérations, les deux premières sont pratiquement inconnues dans les applications mécanographiques autres que la statistique, telles que comptabilité, stocks, facturation, etc... Les documents parviennent à l'atelier d'exploitation sous une forme à la fois simple et élaborée, les données qui y figurent ont déjà fait l'objet de vérifications par les comptables, les magasiniers, etc... Ils n'ont rien de commun avec ceux que les statisticiens obtiennent des personnes enquêtées, ne sachant pas ou ne voulant pas répondre, et n'accordant qu'une attention minime à la qualité de leurs réponses. De plus la grande majorité des documents commerciaux ou industriels ne comporte aucun chiffrement, ou alors il s'agit de codes très simples, souvent préimprimés et ne s'appliquant qu'à une ou deux rubriques.

Enfin, dans les ateliers des entreprises commerciales ou industrielles, l'établissement des cartes ou des bandes se fait à un rythme régulier, sans à-coups, avec un personnel stable et peu nombreux. Il n'a rien de commun avec les afflux de documents de recensements ou d'enquêtes, qu'il faut traiter très vite avec un renfort en main d'œuvre recruté à la hâte, connaissant le caractère temporaire de son embauchage, et que les errements budgétaires ne permettent guère de rémunérer convenablement. C'est toutefois cet aspect astreignant et lent de la perforation des cartes qui a provoqué diverses recherches dont nous parlerons plus loin (Cf. infra section D).

Pour en revenir aux ateliers d'exploitation statistique, les tâches de préparation des documents à l'exploitation y sont particulièrement délicates, dispendieuses et extrêmement lentes, en comparaison de la vitesse des dispositifs électromécaniques ou électroniques qui sélectionnent et dénombrent les données, ou qui analysent, tabulent et impriment les résultats. Pour un recensement ou une enquête importante, la traduction des documents de base initiaux en langage acceptable par la machine représente couramment de 70 à 90 pour cent du temps et de la main d'œuvre consacrés à l'exploitation.

A notre époque, où l'automatisation est la nouvelle déesse du progrès, il faut donc espérer que l'homme arrivera à rendre ces trois tâches automatiques, et ne laissera à l'opérateur humain que son double rôle normal :

— fixer le programme du travail;
— contrôler la conformité de son exécution.

Les chercheurs sont à l'œuvre. Ils ont peu progressé jusqu'ici, et cette progression varie avec chacun des trois problèmes énoncés. Selon certains signes précurseurs, ils aboutiront dans un avenir qu'il serait vain d'exprimer par des délais.

Les renseignements sur leurs travaux sont peu nombreux, et ne sont

guère diffusés. Voici cependant ce que nous avons appris, et qui intéressera certainement nos lecteurs.

B — VÉRIFICATION PRÉALABLE DE L'INFORMATION

La révision individuelle des documents de base n'avait pas, au temps jadis, le caractère astreignant qu'elle a pris maintenant. A l'époque des exploitations manuelles, il était facile, en lisant chaque document, de constater que des âges ne concordaient pas avec la profession ou la situation matrimoniale, qu'une réponse manquait, etc... C'était l'occasion de revoir tout le document et d'y apporter les corrections jugées nécessaires pour la vraisemblance. Si même des anomalies avaient échappé, elles ressortaient des tableaux de résultats, peu nombreux, qu'on ajustait par suppression ou regroupements, en veillant aux recoupements. Cette méthode artisanale est devenue impossible avec le volume des documents et des tableaux. Or non seulement il faut un personnel nombreux et expérimenté pour déceler des anomalies, souvent peu apparentes, mais on n'aura jamais la certitude, quels que soient le temps et le prix qu'on y mette, d'avoir fait disparaître *toutes* les anomalies.

C'est peut-être dans cette voie que la mécanisation et l'automatisation sont les plus avancées. Les machines savent comparer des chiffres, constater des présences ou des absences insolites dans certaines zones du document chiffré qu'on leur fait lire. On peut donc leur confier cette opération de vérification préalable, appelée « editing » par les anglo-saxons (terme difficile à traduire, sinon peut-être par « révision » ou par « mise en forme »).

La trieuse-compteuse-imprimante utilisée en France pour le recensement de population de 1954 présentait des possibilités intéressantes et assez poussées dans ce domaine, mais elle se bornait à éjecter par rebut les documents défectueux. C'est déjà beaucoup.

Les calculatrices électroniques ont des possibilités beaucoup plus développées. On peut les programmer pour qu'elles détectent les anomalies, et ensuite qu'elles prennent une décision d'après l'importance de cette anomalie. Cette décision peut être très variable; par exemple : imprimer l'anomalie et s'arrêter, la laisser passer si elle se trouve entre deux limites fixées par le programme, ou la corriger et substituer au nombre testé un nombre vraisemblable, valeur moyenne ou valeur limite fixée par le programme.

On trouvera un exemple de confrontation de données avec des ratios préétablis et un plan de révision pour une enquête britannique sur la distribution dans la communication de Mr. J. Stafford au 30me Congrès de l'Institut International de statistique — Stockholm — 1957.

Cette révision automatique des données comporte un corollaire : une machine ne peut faire que ce que son programmeur lui a prescrit. Elle laissera passer les bourdes les plus extraordinaires et les plus apparentes, si on a oublié de l'instruire sur ce point particulier. Le programmeur lui-même ne pourra lui donner des ordres que dans la mesure où des spécialistes du domaine exploité lui auront énuméré les anomalies possibles et les conditions de leur redressement. Il y aura donc substitution, à une intellectualité de niveau relativement bas, dispersée entre les anciens exécutants manuels, d'une « phosphoration » intense préalable de quelques responsables supérieurs d'un rang très élevé.

C — LE CHIFFREMENT AUTOMATIQUE

Jusqu'ici, le chiffrement reste une opération mi-intellectuelle, mi-manuelle, et sa mécanisation est très limitée : application de quelques constantes de grands groupes par timbre humide, par composteur, ou par préperforation en série dans quelques zones de carte perforée.

C'est dans un chiffrement entièrement automatique que les progrès semblent les plus faibles. En effet, si cette opération comporte pour beaucoup la consultation machinale de codes, elle demande souvent une interprétation intellectuelle. Certains caractères statistiques, comme l'activité individuelle ou collective, sont généralement écrits sous des appellations très diverses ou par des sigles; souvent la désignation du métier, de l'entreprise ou de l'établissement est incorrecte, ou non conforme. L'opérateur humain peut alors apporter la correction voulue, mais elle échappe évidemment à l'opération précédente de révision automatique.

La solution automatisée existe : c'est la machine électronique à traduire. On sait qu'il y a aux U.S.A. un ordinateur qui, en échange d'un texte anglais (simple) restitue un texte russe, et vice-versa. Les Russes ont aussi leur traductrice électronique russo-américaine et sino-russe. Le vocabulaire de ces machines est encore limité, et il s'exprime parfois assez incorrectement, mais ce sont des prototypes qui se développeront. Le principe n'a rien de mystérieux. Le dispositif de lecture de l'ordinateur lit sur carte perforée un mot anglais en clair, et le traduit en signes binaires; puis un dispositif chercheur fouille dans la mémoire de la machine où est enregistré un lexique anglo-russe. Dès qu'il a trouvé le mot lu (c'est-à-dire quand il y a comparaison égalité entre le nombre binaire correspondant aux lettres du mot lu par la machine et le même nombre binaire contenu dans la mémoire), la traduction correspondante voisine est extraite par la machine de la mémoire-lexique, et va temporairement dans une mémoire intermédiaire. Quand la phrase est terminée, le programme déclanche quelques règles de grammaire (telles que les accords) et de position relative des mots (telles que la place du verbe dans la phrase), puis imprime le résultat après conversion en langage normal.

Pour appliquer ce système au chiffrement statistique, il faut donc que la machine électronique ait une mémoire assez vaste pour contenir tous les codes nécessaires à un recensement. Le développement récent des mémoires sur bandes magnétiques, qui sont illimitées en capacité, constitue à cet égard une solution immédiatement applicable. Si nous avons alors un dispositif capable de lire directement le langage en clair (voir infra section D) la machine ira chercher sur la bande-code voulue, la traduction en code du mot lu, et la placera à l'endroit approprié d'une autre bande magnétique qui correspond à l'information du questionnaire où ce mot à été lu. Ces nouvelles bandes magnétiques, comportant toute l'information numérique ou codée relevée sur les questionnaires successifs seront la source de documentation utilisée pour le dépouillement du recensement. Les bandes-codes, établies une fois pour toutes, pourront servir à nouveau tant qu'on le voudra, l'information qu'elles contiennent sous forme de spots magnétiques s'y conservant tant

qu'on ne procède pas à une opération d'effaçage spéciale, et pouvant faire l'objet de mises à jour.

Le principe est donc simple, et la machine sera même moins compliquée qu'une machine de traduction bilingue. Toutefois, l'application est assez difficile, puisqu'elle présuppose :

D'une part *la lecture en clair* (dont nous parlons ensuite), et qui n'est pas encore réalisable. En attendant, nous serons contraints de traduire intégralement en langage machine, sur carte perforée ou bande perforée, l'information telle qu'elle figure sur le questionnaire. Nous déplacerons le goulot d'étranglement sans le supprimer, puisque, si nous éliminons le chiffrement (opération intellectuelle), nous accroissons la durée de l'opération matérielle de perforation, en ayant à perforer intégralement l'information alphabétique au lieu de sa codification chiffrée.

D'autre part, la consultation *au hasard* de l'information d'une bande qui est classée *en séquence*, (en principe dans l'ordre de la nomenclature). Cette consultation au hasard s'obtient par une programmation appropriée, mais, comme elle implique des recherches en séquence normale, puis en séquence inverse, elle entraîne des opérations de déroulement et de réenroulement de la bande qui, mettant en œuvre des systèmes mécaniques, ont une vitesse limitée. Cette lenteur relative freine considérablement le rendement de fonctionnement et elle constitue un handicap très sérieux de l'efficacité de la méthode. Le prix de revient peut, dans le cas d'un assortiment de codes nombreux et longs, devenir inadmissible. Toutefois les constructeurs de machines électroniques développent des mémoires à consultation au hasard dont la capacité peut se trouver suffisante pour résoudre mieux le problème. A ce point de vue, des systèmes commes les mémoires à disques et les mémoires à noyaux magnétiques, auxquelles l'accès peut se faire au hasard, sont intéressantes, mais leur adaptation aux problèmes des statistiques n'a pas encore été étudiée.

Enfin, la parfaite *concordance* des termes lus sur les questionnaires avec les vocables enregistrés dans la machine; sans quoi, après des tentatives inutiles et prolongées, cette dernière devra déclarer forfait, laissant le statisticien dans un cruel embarras.

Notons ici que la révision dont nous parlions précédemment devra être postérieure au chiffrement, tout au moins en ce qui concerne les rubriques codifiées.

D — LA LECTURE AUTOMATIQUE DES DONNÉES

Nous avons vu jusqu'ici que la révision automatique était un problème résolu de façon très satisfaisante, mais partielle, dans la double mesure où le programmeur pensait à tout, et où il pouvait ramener la révision à des comparaisons de valeurs numériques ou à des confrontations de zones réceptrices d'informations. Par contre, pour le moment, le chiffrement automatique relève un peu du « il-n'y-a-qu'à », sa réalisation commercialisée, et surtout son efficacité, restent un peu aléatoires. Arrivons maintenant au problème de la lecture automatique des données.

1° La méthode classique d'établissement d'un document que la machine

puisse comprendre est la machine à clavier. Sa réalisation est très perfectionnée pour la carte perforée et pour la bande perforée; elle est encore au stade du laboratoire pour la bande magnétique. Ses inconvénients sont sa lenteur (au plus 8 000 caractères—chiffres ou lettres — à l'heure) et la nécessité d'avoir sous les yeux un document de base où figurent en clair les caractères chiffrés à transcrire.

2° On a cherché à l'accélérer par la lecture graphique de la carte à perforer. Elle substitue au tracé des chiffres arabes des bâtons ou des croix au crayon à des emplacements codifiés. Ces marques sont ensuite lues à grande vitesse par une machine d'établissement de cartes perforées.

La lecture graphique est parfaitement adaptable aux travaux comptables, où elle n'a pour objet que l'inscription de quelques chiffres, faite à loisir et épisodiquement. Son application à la traduction continue de grosses masses d'informations statistiques a été tentée. Tout ce qu'on peut conclure avec les réalisations commercialisées actuelles, c'est que le temps gagné à la perforation est amplement reperdu au traçage des marques, et que ce traçage est horriblement fatigant pour le personnel qui en est chargé. Aussi n'apparaît-il applicable qu'aux exploitations de volume limité.

3° De même, la comptabilité et les travaux de bureau s'accommodent fort bien de machines connectées qui établissent une carte perforée ou une bande de papier perforée comme sous-produit d'une facturation dactylographiée ou d'un calcul effectué sur calculatrice de bureau. Mais on imagine mal un recensement ou une enquête où les documents de base seraient spécialement établis par les personnes enquêtées ou par les enquêteurs sous forme dactylographiée et sur machine connectée, surtout que ces documents deviendraient inutiles aussitôt établis.

D'ailleurs cette solution ne résoud pas le problème du chiffrement automatique, qui reste entier.

4° Toutefois, faute de mieux, la lecture graphique, qui a une importante clientèle dans les entreprises privées, s'est développée et perfectionnée, et on a tenté de l'adapter au dépouillement statistique.

Pour cela, on a d'abord cherché à s'évader du format de la carte perforée, trop petite pour les documents de base statistiques.

La firme I.B.M. a réalisé des documents plus grands que la carte, où l'on trace de petits traits de crayon dans des ovales, ce travail étant confié à l'enquêteur (recensement canadien de 1950) qui est spécialement instruit et ne graphite que sporadiquement, surtout dans les campagnes. Le Bureau du Census en a fait le support de certaines de ses enquêtes permanentes par sondage, où l'on retrouve les mêmes conditions : graphitage par enquêteur itinérant spécialisé, et surtout document très aéré, comportant peu de rubriques. Cette méthode est coûteuse : le questionnaire sur carton spécial coûte cher, et il faut le transcrire dans une carte perforée.

Pour s'évader, lui aussi, du format de la carte, Bull a proposé la fiche bien connue de l'I.N.S.E.E. (recensement de population et recensement des fonctionnaires). Cette fiche est au point maintenant, elle est assez facile à marquer (taches de crayon graphité ultérieurement magnétisées). Mais ses inconvénients sont son prix de revient élevé (usinage de précision de la fiche pour un

bon entraînement dans les machines), et la compacité des pâtés de chiffrement qui ne peuvent être graphités qu'à l'atelier mécanographique, donc en travail continu, lent et délicat.

Plus récemment, on a proposé des documents beaucoup plus grands, sur simple papier, donc moins coûteux. L'information reste peu abondante et très aérée, et est graphitée directement par les enquêteurs spécialisés. Des pistes de taches noires permettent un entraînement régulier et bien centré sous le dispositif d'analyse, réglé par des cellules photoélectriques. Le document est de très grand format, fragile, et se prête mal à des questionnaires importants.

Ces solutions présentent un intérêt certain, mais ne donnent qu'une satisfaction très partielle aux besoins de la préparation des données destinées aux exploitations statistiques.

Dans le même état d'esprit, deux administrations américaines, le Bureau du Census et le Bureau des Standards sont en train de mettre au point un procédé appelé Fosdic (initiales de Film Optical Sensing Device for Input to Computers — Dispositifs d'analyse optique de film pour l'entrée dans les calculatrices).

Il n'existe que très peu de renseignements sur son fonctionnement qui en reste au stade du laboratoire. Tout au plus en connaissons-nous le principe. Les documents de base sont marqués de traits noirs, (vraisemblablement par les enquêteurs et non par le public), et ne comportent que peu d'information jusqu'ici. La nouveauté est qu'ils sont microfilmés, ce qui substitue à leur transport dans les ateliers d'exploitation celui de petits rouleaux de films. Les microfilms sont ensuite lus par cellules photoélectriques et l'information est reportée directement sur bande magnétique.

Jusqu'à plus ample informé, cette solution semble donc ne pas comporter la révision des données, ni leur chiffrement automatique.

5° Il fallait une certaine audace pour s'attaquer au problème direct : concevoir une machine qui lise directement l'écriture sans intervention humaine. La Société britannique Solartron l'a tenté.

Cette machine électronique de lecture des documents se compose de trois parties principales : introduction des documents, explorateur, reconnaisseur. En réalité ces trois parties forment un complexe unique.

Le dispositif d'introduction des documents comporte des rouleaux pour l'introduction de rubans de papier et de microfilms, et une courroie transporteuse pour l'introduction de documents de travail, tels que factures, chèques ou questionnaires. La machine peut continuer son travail avec une introduction constante de documents, et il n'est pas nécessaire d'arrêter leur mouvement pendant la lecture.

L'exploration vient ensuite, par le moyen d'un « faisceau de balayage » explorateur, sorte de tube à rayons cathodiques, du type utilisé dans les techniques de la télévision pour analyser la séquence d'un film. Enfin, c'est l'unité de reconnaissance, le véritable cœur de la machine, qui a nécessité de très longues études pour sa réalisation.

Le prototype qui a été présenté pouvait lire à la vitesse de 120 caractères par seconde, mais il est prévu pour les machines de série une vitesse de lecture

de 200/300 caractères par seconde, et on peut même envisager ensuite la possibilité de porter ce chiffre à 600.

Ses inventeurs font remarquer que cette performance peut se comparer au travail de 150 à 300 opérateurs habiles, représentant un coût, en salaires, impôts, etc... de 300 000 livres sterling par an. La nouvelle machine E.R.A. coûtera paraît-il, de 20 000 à 50 000 livres, selon modèle.

Il faut souligner que l'E.R.A. est seulement un « œil »; en fait il n'enregistre pas, ni n'élabore autrement l'information lue. Son domaine est seulement d'envoyer l'information lue à plusiers types de machines déjà connues et utilisées (perforatrices de carte ou de bande de papier, enregistreur de bande magnétique, etc.) et d'accomplir cette opération de lecture à une vitesse très supérieure à celle de l'œil humain.

On estime, dans le cas des cartes perforées, qu'un Solartron E.R.A. permettra de perforer les cartes 144 fois plus rapidement que par le moyen d'opératrices ayant une rapidité de production de 300 cartes à l'heure.

Pour le moment cette machine ne lit que des caractères dactylographiés ou imprimés à des emplacements prédéterminés avec précision, et qui ont une forme, une dimension et un écartement donnés. Nous en sommes pas encore à la lecture de toutes les écritures manuelles, que les hommes savent si bien varier, et si souvent rendre à peu prés illisibles. C'est toutefois un début qui devait être signalé et salué avec intérêt.

***

Rêvons donc à loisir d'un « complexe combiné » de plusieurs unités où les opérateurs humains se borneront à introduire les document de base dans une première unité de lecture directe. Les impulsions électroniques émises passeront dans une unité traductrice pour chiffrement, puis dans une unité calculatrice qui procédera à la révision des données. Ces dernières seront alors enregistrées sur bande magnétique (ou autre support à inventer) et seront prêtes à l'exploitation. Souhaitons néanmoins que ce rêve ne nous force pas, comme dans la légende de Barberousse sous sa montagne, à nous rendormir trop de fois après des réveils prématurés.

P. Pèpe.

***