LES METHODES MODERNES DE TRAITEMENT DE L'INFORMATION ET LEUR INCIDENCE SUR LA STATISTIQUE DE DEMAIN

par D. HIRSCHBERG,
Université Libre de Bruxelles.

Mon exposé s'intitule « Les méthodes modernes de traitement de l'information et leur incidence sur la statistique de demain » et c'est dans cet ordre que je me propose de discuter ces deux aspects. Je n'ai pas l'intention de rappeler ici les principes du traitement de l'information ou d'en retracer les origines, je me limiterai plutôt à un examen des tendances actuelles et à un pronostic d'environ dix ans quant à leur développement futur. A plus courte échéance il ne faut sans doute pas s'attendre à des transformations radicales des habitudes de la statistique et dix ans paraissent par ailleurs l'extrême limite de ce que l'on peut prévoir dans un domaine aussi mouvant que le traitement de l'information. Ce que je dirai sera de toute façon entaché d'un double incertitude puisque je serai obligé d'extrapoler très loin des tendances récentes du traitement de l'information et d'en prévoir l'incidence sur les procédés de la statistique, sans autre base que des arguments de plausibilité.

Qu'entend-on exactement par méthodes modernes de traitement de l'information ? Il est sans doute plus aisé de répondre à cette question aujourd'hui qu'il y a deux ou trois ans, lorsque l'évolution était graduelle : il ne peut en effet s'agir actuellement que de la « troisième génération des ordinateurs ». Dans cette terminologie la troisième génération est celle des microcircuits, comme la deuxième était celle des transistors et des ferrites et la première celle des tubes à vide. Mais cette désignation décrit imparfaitement la nouvelle génération, puisqu'elle insiste sur l'aspect technologique somme toute secondaire alors que le progrès se situe principalement sur le plan de la conception et de la structure et se traduit par le déplacement de l'accent, des ordinateurs proprement dits vers les systèmes à traiter l'information considérés comme un tout.

Une des conséquences de ce changement d'attitude est qu'il n'est pas possible de comparer les performances des nouveaux systèmes avec celles des plus anciens en se basant sur des caractéristiques envisagées une à une. Je citerai néanmoins quelques ordres de grandeur comme indication du

chemin parcouru dans les dix dernières années. Pendant cette période la vitesse interne des ordinateurs a grosso modo centuplé de même que celle d'entrée et de sortie, qui est passée de dix mille à environ un million de caractères par seconde. La capacité de la mémoire a également augmenté d'un facteur de l'ordre de la centaine, qu'il s'agisse de la mémoire rapide accessible en un temps de l'ordre de la microseconde, passée de cent mille à dix millions de caractères ou qu'il s'agisse de la mémoire dite « lente », accessible en une fraction de seconde, dont la capicté a progressé de dix millions à environ un milliard de caractères. Ce qui précède se rapporte aux caractéristiques maxima des ordinateurs mais l'indice performance/prix a, lui aussi, centuplé, à condition de mesurer la performance d'une façon réaliste en termes de problèmes traités par exemple par mois plutôt qu'en termes d'instructions exécutées par microseconde. Ce bilan impressionnant montre incidemment combien une extrapolation à dix ans est dangereuse.

Deux raisons ont motivé cette révision fondamentale : le désir d'améliorer le rendement des centres d'ordinateurs traditionnels et l'avènement de nouvelles applications, dites « en temps réel », auxquelles je reviendrai dans quelques instants.

Le rendement d'un centre d'ordinateurs est une caractéristique globale, relative à toute l'unité, y compris le personnel. La difficulté fondamentale à surmonter est le déséquilibre entre le travail humain et la vitesse sans cesse croissante des machines.

C'est dans le domaine de la programmation que ce déséquilibre s'est manifesté en premier lieu. Le remède a consisté dans l'emploi de langages de programmation bien adaptés aux problèmes traités, faciles à manier par les programmateurs et d'une forte densité, en ce sens que peu d'instructions suffisent à la description d'un problème donné. En calcul scientifique on utilise couramment des langages de programmation proches du formalisme algébrique, principalement le Fortran et l'Algol et on estime que l'on parvient de la sorte à ramener les temps de programmation et de mise au point au dixième de leurs valeurs originales. Un autre avantage est la possibilité pour les ingénieurs ou les physiciens qui posent les problèmes d'écrire eux-mêmes les programmes ou du moins à participer activement à leur élaboration. Il est vrai que dans les applications administratives les langages de programmation synthétiques sont moins répandus, mais ils ont tendance à se généraliser et ils s'imposeront au fur et à mesure que les équipements deviendront plus puissants et les applications plus nombreuses et plus complexes. Les pro-

grammes rédigés dans des langages artificiels doivent être traduits automatiquement vers le langage de la machine pour pouvoir être exécutés, cette opération est la « compilation » des programmes.

Le deuxième aspect du déséquilibre homme-machine concerne l'exécution des programmes. Plus les machines sont rapides et plus la configuration de leurs unités est complexe, plus il devient difficile et finalement impossible à un opérateur humain de les gérer avec un rendement acceptable, en évitant l'inaction de certaines unités par suite de l'absence de données ou de consignes. Cette surveillance continuelle ainsi que l'alimentation et la répartition permanente des tâches doit être confiée à un programme de gestion, appelé moniteur ou programme opérationnel.

Dans les deux cas l'ordinateur est chargé de tâches qui initialement incombaient au personnel du centre de traitement de l'information : gestion de la machine ou écriture des programmes en langage machine, cette dernière activité étant remplacée par la « compilation ». La structure des premiers ordinateurs était déterminée par leur tâche primaire, à savoir le traitement des données. Les tâches secondaires se sont progressivement greffées sur cette structure sans beaucoup l'affecter. La nouvelle génération, en revanche, a été conçue en fonction de tout l'éventail de ces travaux primaires et secondaires, avec le souci de maximiser le rendement global.

L'autre considération qui a conduit à repenser la structure des systèmes à traiter l'information est l'avènement des applications « en temps réel », applications qui se situent dans le contexte de l'activité de production des entreprises. Jusqu'ici les ordinateurs assuraient en effet la mécanisation des bureaux d'étude et des départements administratifs mais le flux d'informations qui conditionne la production à tout instant leur échappait presque totalement. A présent les constructeurs des systèmes à traiter l'information s'attaquent énergiquement à ces problèmes, aussi bien dans le domaine de la production industrielle que dans celui de la production de services.

Ces problèmes de production présentent la particularité de se passer « en temps réel », autrement dit dans le temps physique, qu'il faut bien distinguer de la variable « t » des mathématiciens dont l'échelle et l'origine peuvent être modifiées à souhait. Un système en temps réel opère en liaison directe avec un processus dynamique et il doit réagir avant que celui-ci ne franchisse certaines limites. Un tel système est soumis à un flux aléatoire d'informations et de demandes de service provenant d'un grand nombre de terminaux dispersés à travers l'entreprise, ces demandes devant souvent être satisfaites quasi instantanément, avec des délais de l'ordre de la seconde. Le

traitement de l'information se complique ici de problèmes de télécommunications et de trafic et un programme opérationnel perfectionné est indispensable pour gérer le système.

Les premières applications en temps réel étaient d'ordre militaire, la première application civile a été la réservation à distance de places dans les avions, mais à l'heure actuelle presque tous les secteurs économiques sont représentés dans les études en cours. Trois systèmes en temps réel au moins fonctionneront en Belgique en 1966, deux dans la sidérurgie et un système dans l'industrie du verre. Plusieurs études se trouvent à l'état de projets ou d'avant-projets et dans quelques années cette évolution englobera les banques, les grands magasins, les hôpitaux, les services publics, etc. Il va de soi que les systèmes en temps réel assimilent des flux d'informations beaucoup plus importants que les systèmes traditionnels. Ce fait, ajouté à l'accroissement « naturel » du nombre d'ordinateurs permet d'estimer que vers 1975 le flux d'informations soumis au traitement automatique aura au moins centuplé par rapport à sa valeur actuelle.

Quelle sera l'incidence de cette intensification et de cette généralisation sur les méthodes de travail de la statistique, c'est ce que je voudrais examiner à présent. Je ne parlerai pas de statistique mathématique non pas que j'en sous-estime l'intérêt, mais par ce qu'elle ne présente pas de difficultés particulières du point de vue du traitement de l'information. Les calculs auxquels on est conduit, d'ailleurs souvent des variantes d'algorithmes classiques d'algèbre linéaire, peuvent être résolus par presque n'importe que ordinateur actuel, compte tenu de leurs dimensions relativement modestes. Je me limiterai à la statistique dite « descriptive » et je m'intéresserai surtout aux problèmes que rencontrent les instituts de statistique nationaux, confrontés avec le rassemblement et le traitement de données très nombreuses.

Quelle a été jusqu'ici l'influence réciproque de la statistique et du traitement automatique de l'information ? On sait que la carte perforée a été inventée par Hollerith en vue de recensement de 1890 de la population des Etats-Unis mais on constate que la statistique a assez rapidement perdu ce caractère de moteur du traitement de l'information qu'elle avait possédé initialement. Ce rôle a été repris par la gestion administrative, le calcul scientifique et récemment par les problèmes de production, comme je viens de l'indiquer. La statistique a en revanche bénéficié des progrès généraux du traitement de l'information. Les ordinateurs électroniques notamment lui ont permis de réduire le coût des travaux, d'améliorer leur qualité, de les accélérer et d'en accroître le nombre et l'importance.

La liste des travaux réalisés actuellement par exemple à l'INS est éloquente à cet égard. Elle comprend des recensements généraux extrêmement volumineux et des applications périodiques à grande échelle comme les statistiques du commerce extérieur et des transports, qui exigent le dépouillement mensuel d'un demi-million de documents douaniers.

A cela s'ajoutent des statistiques diverses, industrielles, agricoles, sociales, démographiques, financières ,etc. La même impression se dégage de l'examen des travaux des instituts d'autres pays ou de ceux des organismes internationaux. Pour fixer les idées je prendrai comme exemple l'application périodique la plus volumineuse de l'office de statistique des Communautés Européennes, à savoir l'élaboration trimestrielle des tableaux analytiques du commerce extérieur.

Il s'agit non seulement de la publication qui porte ce nom mais d'un ensemble de travaux plus vaste. Dans cette application on s'est principalement attaché à réduire le délai total et on a pu le ramener à environ six mois. Les données de base correspondent à environ 50 millions de caractères par trimestre. Près de 30 % arrivent sur cartes, le reste sur bandes magnétiques. Ces données sont conformes aux normes de chaque pays et ne sont pas standardisées. La sortie finale comporte de l'ordre de 150 millions de caractères dont 10 % environ sont publiés, le reste sert à des fins de vérifications éventuelles. Les opérations comprennent des contrôles d'exactitude des données, la traduction des codes nationaux en codes standards, des calculs de conversion d'unités et de totalisation et finalement l'impression de listes et de tableaux. La publication elle-même se fait par reproduction des listes en off-set.

Pour réaliser ces tâches les Instituts de Statistiques disposent d'équipements comparables à ceux des services mécanographiques des grandes entreprises, mais leurs problèmes sont plus volumineux et les données plus difficilement accessibles. Il ne faut donc pas s'étonner de ce que les applications statistiques soient souvent plus lentes et moins intégrées que celles qui ont trait à la gestion des entreprises. Ceci préoccupe beaucoup les responsables : au delà d'un délai critique qui dépend du phénomène envisagé les résultats statistiques cessent en effet d'être des éléments de décision et deviennent de simples enregistrements historiques. Mais il ne suffit pas d'accélérer le traitement proprement dit, il faut également agir sur l'acheminement des données et c'est là que l'évolution future du traitement de l'information jouera sans doute un rôle essentiel.

La statistique n'a certainement pas besoin des temps de réponse de l'ordre de la seconde, que permettent d'atteindre les systèmes en temps réel. Le béné-

fice que la statistique tirera de la généralisation de ces systèmes sera surtout indirect. Comme toutes les informations auront été enregistrées à la source à des fins d'exploitation immédiate, il sera aisé de se procurer celles qui intéressent l'institut de statistique, rapidement, à peu de frais et avec un taux d'erreurs très bas.

Des sources particulièrement importantes pour les instituts de statistique sont les services administratifs de l'Etat. Là aussi on cherche à mécaniser au maximum la gestion quotidienne. Les moyens techniques actuels permettent d'envisager une solution en temps réel pour l'ensemble d'un pays. Mais au préalable, il faut créer un fichier national contenant la liste de tous les habitants et de toutes les entreprises, codés de façon à ne laisser subsister aucune ambiguité de désignation. Quand un tel système fonctionnera, les services de statistique disposeront d'un inventaire démographique et économique permanent et détaillé. Il leur sera possible de structurer automatiquement ces données selon des critères particuliers en fonction des problèmes qu'ils auront à résoudre. La procédure lente et coûteuse qu'est la prise d'informations nouvelles par voie d'enquête s'imposera de moins en moins et lorsque de telles enquêtes s'avéreront indispensables, le fichier national sera utile, comme système de référence.

Voilà comment par l'accélération des travaux et la constitution de véritables banques d'informations avec mise à jour permanente les progrès du traitement de l'information permettront à la statistique de vaincre graduellement les obstacles matériels qui l'empêchent d'être un instrument de décision parfaitement adapté à la gestion des affaires publiques. Il serait présomptueux de prédire que cette évolution aboutira avant 1975, mais on peut estimer qu'elle aura considérablement progressé à cette époque.