L'AED, analyse exploratoire des données

Au cours des siècles, les soucis du statisticien sont restés sensiblement les mêmes : données manquantes ou aberrantes, précision des mesures, variables qualitatives, petits échantillons...

Autrefois, l'absence de moyens de calcul, et, dans une moindre mesure, de théorie, ont conduit le statisticien à développer des outils simples et pertinents, en particulier des graphiques, lui permettant de résoudre ses problèmes : tous les statisticiens faisaient de l'exploration. À partir du XIX^{e} siècle, la théorie statistique a élaboré des procédures d'analyse et d'inférence rigoureuses et optimales, au prix d'hypothèses sur la nature et la collecte des données : l'échantillonnage aléatoire simple, la moyenne et la loi normale devenaient des éléments centraux de l'analyse statistique. Mais, malgré des progrès considérables, il faut bien reconnaître que les données prennent toujours un malin plaisir à malmener, parfois même ridiculiser, les méthodes statistiques usuelles les plus sophistiquées (que les explorateurs qualifient de « confirmatoires »).

C'est ce constat qui a amené John Wilder Tukey, au début des années 1970, à créer *l'analyse exploratoire des données*[1]. Plongeant ses racines dans les pratiques traditionnelles, l'exploration a su profiter des formidables progrès de l'informatique pour progresser rapidement, tant sur le terrain du développement théorique que sur celui des applications. Aujourd'hui, tous les grands logiciels lui font une large part.

L'AED, c'est donc, pêle-mêle, des instruments efficaces et simples de mise en œuvre, des logiciels puissants et ergonomiques, la réhabilitation des graphiques comme outils d'analyse, mais aussi et surtout une attitude du statisticien face à son problème et ses données.

« It is better to be approximately right than exactly wrong » (Tukey)

La statistique confirmatoire repose sur une logique d'expérimentation : vous partez d'une hypothèse précise, vous collectez des données selon un plan d'échantillonnage

1. L'expression « analyse des données » est à prendre ici dans un sens très général, et non dans son acception plus réduite renvoyant à l'analyse factorielle et à la classification.

simple, vous calculez la ou les statistiques ad hoc, vous testez votre hypothèse et décidez enfin si elle est « statistiquement significative ». La pratique répond assez rarement à ce schéma : vous cherchez parfois à répondre à des questions peu précises, les données « préexistent » et ont été recueillies selon des plans de sondage complexes, il y a des observations aberrantes... Dans ces cas, leurs hypothèses fondamentales étant violées, les méthodes classiques ne sont plus optimales.

L'analyse exploratoire part au contraire des données et repose sur une logique d'observation. L'explorateur, grâce à une boîte à outils très fournie, va regarder ses données sous toutes les facettes, tenter de mettre en évidence des structures, et, le cas échéant, formuler des hypothèses plausibles. Il n'accorde pas une grande importance à « l'optimalité » de son outil, savoir que celui-ci se comporte « bien » dans la plupart des situations lui suffit.

Mais il serait absurde d'opposer statistique classique et statistique exploratoire, tout aussi absurde de vouloir les confondre. Elles occupent toutes deux une place de choix dans l'analyse statistique, et peuvent se révéler complémentaires. Ainsi, un explorateur peut très bien utiliser des outils classiques. De même, l'AED peut aider à la construction d'un modèle. Dans l'exemple simpliste de l'étude de la relation entre deux variables quantitatives, l'explorateur commencera par tracer le nuage de points ; puis il cherchera graphiquement les transformations des variables conduisant le plus près possible de la linéarité ; s'il ne détecte pas de contre-indication majeure, il optera pour une régression linéaire robuste ; enfin, un examen attentif des résidus lui permettra qualitativement de valider ou non sa démarche, et, si rien ne s'y oppose, il reviendra à des méthodes de régression classique, par les moindres carrés ordinaires.

« Je n'ai des idées que parce que j'ai des images » (Euler)

L'engouement pour la statistique confirmatoire a été tel que les méthodes anciennes ont été mises à l'écart. Soudain, on leur a reproché leur manque d'optimalité, de justification théorique... Ainsi, la médiane a été délaissée (il faudra des dizaines d'années pour reprendre conscience de ses propriétés de robustesse)[2]. De même, les méthodes graphiques ont été jugées « subjectives » et abandonnées.

2. Qui sait aujourd'hui que Boscovich, en 1755, a mis au point une méthode de régression linéaire robuste, basée sur les moindres écarts à la moyenne et qui met donc en vedette la médiane ? Il faudra attendre les années 1800 et les travaux de Legendre et Gauss pour voir apparaître la régression linéaire par moindres carrés !

L'AED s'inscrit au contraire dans cette tradition de statistique robuste et graphique. En même temps, elle intègre les plus récents progrès de la statistique et de l'informatique, en particulier de nombreux outils « non paramétriques », basés sur les statistiques d'ordre, et des procédures de représentation graphique à la fois simples et performantes, issues d'études spécifiques sur la perception visuelle. Grâce à l'animation, les méthodes classiques les plus élémentaires deviennent de puissants instruments d'analyse : celui qui n'a pas vu l'utilisation exploratoire d'un histogramme manque réellement quelque chose...

Bienvenue dans le monde de l'AED !

Sophie Destandau, chargée d'études à la division « conditions de vie des ménages » de l'Insee, et Monique Le Guen, ingénieur de recherche au CNRS (Matisse, Université Paris I, UMR CNRS 8595), vous entraînent dans les articles suivants à la découverte du monde de l'AED.

Il sera d'abord question d'images, réelles ou mentales, puis de l'oeuvre de Tukey. Difficile en effet d'aborder un sujet où la représentation graphique tient une place aussi importante sans commencer par rappeler le rôle essentiel que jouent les images dans le bon fonctionnement de notre cerveau. Tout aussi difficile de ne pas évoquer le riche itinéraire de l'inventeur de l'analyse exploratoire des données, qui a formé et orienté plus de cinquante thésards dont Brillinger, Hoaglin, Morgenthaler, Mosteller et Velleman, et à qui nous devons également et entre autres la technique de la *Median Polish*, le lissage par médianes mobiles, le *Jackknife*, le *Box Plot* et le *Stem and Leaf*.

Dans le troisième article, on trouvera un exemple frappant des erreurs d'interprétation auxquelles pourrait conduire un usage mal maîtrisé des pourtant si précieuses techniques de la statistique classique.

Seront ensuite passés en revue les fondamentaux de l'AED, visualisation interactive et réexpression des données, résistance/robustesse des indicateurs et des procédures, analyse des résidus, qui tous participent d'un même objectif : établir des hypothèses plausibles, en se gardant autant que faire se peut des présuppositions.

Les trois articles suivants sont dédiés à la présentation de SAS/Insight, à l'examen un bon logiciel d'introduction à l'analyse des données, offrant en outre l'avantage d'être facilement accessible à l'Insee et dans une majorité de SSM puisque partie intégrante de SAS-micro.

En conclusion, même si c'est par là que tout commence, il sera brièvement question de l'enseignement de la statistique. En la matière, beaucoup de voix s'élèvent aujourd'hui pour promouvoir un apprentissage basé sur la visualisation interactive de données réelles, en s'appuyant sur les principes de l'AED.

De très nombreuses références, bibliographiques mais pas seulement, sont données en annexe, concernant le domaine statistique mais également celui des sciences cognitives.

Dominique LADIRAY

Actuellement en fonction à Eurostat,
Dominique Ladiray est professeur d'analyse des données
à l'Ensae et à l'Ensai.

De l'importance de l'image

L'analyse exploratoire des données est au cerveau droit ce que l'analyse confirmatoire est au cerveau gauche.
Les deux doivent communiquer pour traiter l'information.

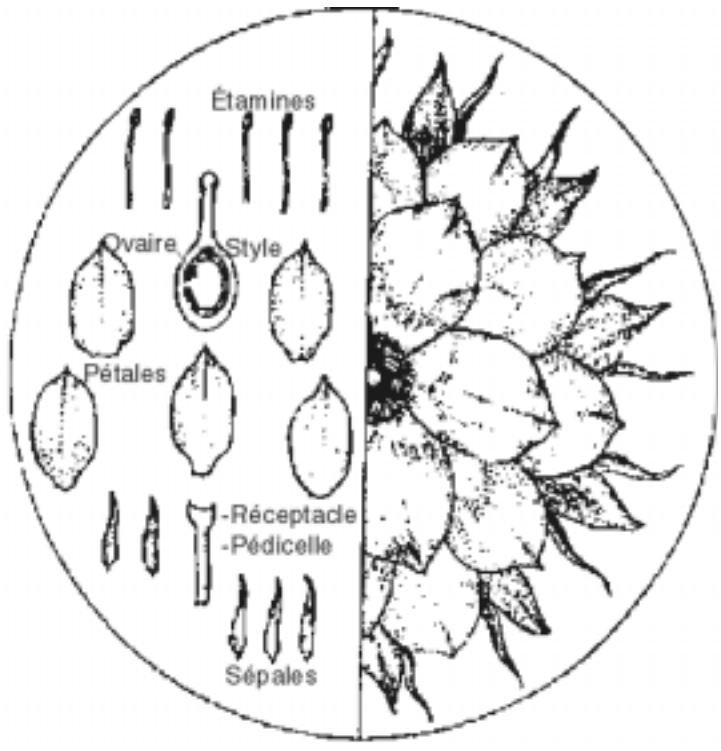
Les neuro-sciences (neuro-anatomie, neuro-physiologie, neuro-biochimie, neuro-biologie, et plus récemment neuro-pédagogie) ont ces dernières années considérablement avancé. Elles ont notamment montré que le traitement de l'information était affaire complémentaire des deux hémisphères cérébraux :

- plus analytique, l'hémisphère gauche est particulièrement apte à traiter l'information verbale ; il procède de façon linéaire et séquentielle, en décomposant un tout en ses différents éléments ;

- plus synthétique, l'hémisphère droit est très efficace pour le traitement visuel et spatial, c'est-à-dire celui des images ; il recherche et construit des structures, en reconnaissant les relations entre éléments séparés ; sa façon de traiter l'information est beaucoup plus globale ; il serait le siège de l'intuition créatrice.

Le mandala ci-dessous aidera votre « cerveau droit » à mémoriser ces différences de fonctionnement.

Si chacun des deux hémisphères joue sur des registres différents, c'est bien leur complémentarité qui donne à la pensée toute son efficacité et sa flexibilité (Linda Williams, 1997). L'un a besoin de l'autre, aucun des deux ne peut fonctionner seul de manière efficace. Ces conclusions confirment la nécessité des images, réelles ou mentales, dans la compréhension et la création, en particulier s'agissant de la chose mathématique et donc statistique.



Mandala de Linda WILLIAMS(1997),
extrait de « Deux cerveaux pour apprendre, le gauche et le droit ».
Avec l'aimable autorisation de reproduction des Éditions Organisation.

Images et nombres

Pour le neuro-biologiste Dehaene, qui s'est penché sur l'enseignement des mathématiques. Il y a quelque part dans nos circonvolutions cérébrales une représentation mentale des nombres associée aux symboles qui les matérialisent (1, 2, ...) et aux vocables qui les nomment (un, deux, ...). Cette représentation peut être naturelle, ou culturelle. Le lien entre les nombres naturels et les doigts de la main lui paraît inné, puisque c'est avec leurs doigts que tous les enfants de toutes les cultures apprennent à compter : *Il est plausible que le doigts et les nombres occupent des territoires cérébraux voisins et étroitement liés*. Il estime en revanche que la représentation ordonnancée selon un axe de moins l'infini à plus l'infini (avec le zéro au milieu) provient de notre culture, et plus particulièrement pour ce qui concerne les nombres négatifs, dont l'invention fut d'abord considérée comme une hérésie. Pour Pascal, la soustraction 0 moins 4 était un pur non-sens. Il est donc fort probable que Pascal et ses contemporains n'avaient aucune représentation mentale des nombres négatifs. Dehaene considère à ce sujet que seul le concept de température (il fait moins six degrés) est susceptible de conférer à l'enfant une image intuitive des nombres négatifs.

Dans le domaine des mathématiques un peu moins élémentaires, Il fallut attendre, pour que se développe l'usage des nombres complexes inventés en 1545 par l'Italien Jérôme Cardan, que le mathématicien anglais John Wallis en propose en 1685 une représentation géométrique sur plan cartésien (la partie réelle occupant l'axe horizontal et la partie imaginaire l'axe vertical).

Le même phénomène se reproduira à propos des fractales, inventées en 1920 par Fatou et Julia. Cette invention est longtemps demeurée sans suite, jusqu'à ce que Mandelbrot, chercheur à IBM, la sorte de l'oubli en montrant sur des images spectaculaires ce à quoi correspondaient les fractales et en baptisant « géométrie fractale » cette nouvelle branche des mathématiques.

Images et logique

Quand les propositions se multiplient, la pensée logique est vite débordée. Ainsi et pour reprendre un exemple (ici légèrement adapté) cher à Lewis Carroll, élaborer sans l'aide de la visualisation une conclusion logique valide à partir des trois propositions ci-dessous n'est pas un exercice si facile :

1. Les casseroles de Monique sont toutes inutiles.

2. Les seuls objets en fer blanc que possède Monique sont des casseroles.

3. Tous les cadeaux que Sophie a offerts à Monique sont utiles.

Si on s'appuie sur des diagrammes, comme nous l'ont enseigné Carroll puis Venn, la solution devient évidente : « Les cadeaux que Sophie a offerts à Monique ne sont pas en fer blanc ».

Images et mathématiques

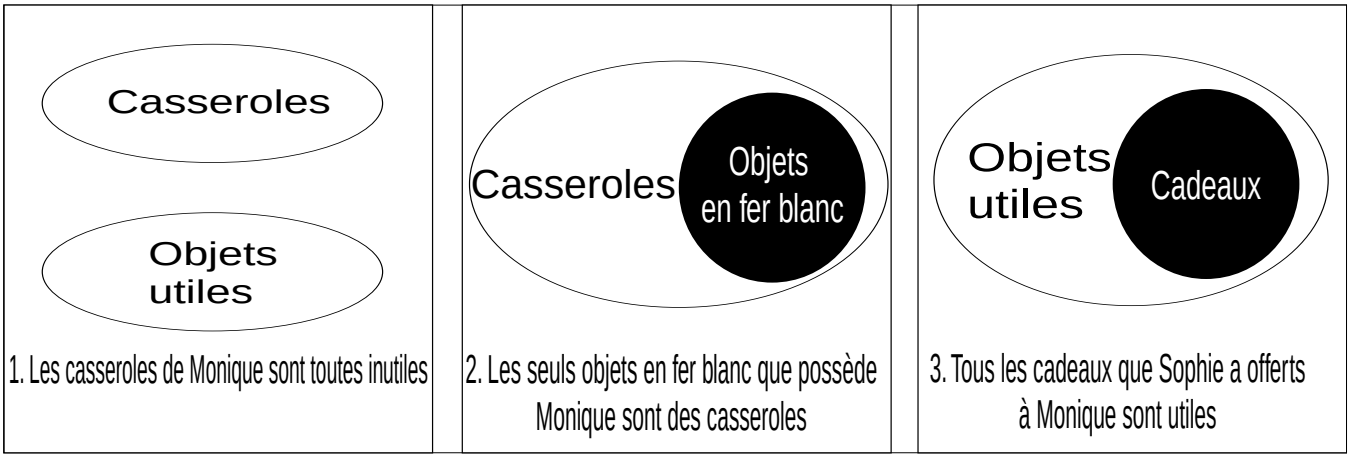
Poincaré l'affirmait, c'est par l'intuition qu'on invente et par la logique qu'on démontre. Il ajoutait que l'intuition faisait essentiellement appel à des images mentales, sans logique apparente, tandis que la démonstration réclamait au contraire un raisonnement logique se déroulant pas à pas.

Plus tard, Hadamard (« Essai sur la psychologie de l'invention dans le domaine mathématique », 1952) distinguera dans la création mathématique quatre étapes : préparation, incubation, illumination et vérification. Dans la phase de préparation, il faut acquérir des informations qui sont dissociées, parcellaires, il faut rassembler des indices sans cohérence apparente. Dans la phase d'incubation, l'esprit cherche à relier ces informations et indices pour trouver une cohérence, une logique, donner du sens. Dans cette deuxième phase, ce ne sont pas les mots qui sont traités mais des signes et des images. Puis soudain vient une illumination, quelque chose de très rapide, fugitif, une sensation d'avoir compris la cohérence. C'est l'instant où l'on dit « Eureka », ou « Insight » pour les Anglo-Saxons. On a alors une compréhension globale du problème à résoudre. La dernière étape va consister à vérifier, en démontrant par un raisonnement logique, linéaire, ce que l'on a pressenti ou ressenti. Cette étape de vérification est souvent très longue et empreinte de doutes, mais on est guidé par la presque certitude d'avoir découvert une vérité. À son issue seulement, on pourra porter un jugement raisonné.

Les travaux actuels de Dehaene nous enseignent que même en mathématiques l'intuition concrète joue un rôle crucial, tant au niveau de la compréhension qu'à celui de l'invention.

Les mathématiques modernes, discutables et discutées

Des avis très critiques ont été émis sur la réforme des maths modernes qui avait été décidée dans les années 1960-70, sous l'impulsion des Bourbakistes épaulés par le psychologue Piaget. À ce sujet, les réflexions de Dehaene (« La bosse des maths », 1997), sont directes, utiles et sans ménagements : *Nos écoles se contentent souvent d'inculquer une arithmétique mécanique et dépourvue de sens... En France, la fameuse réforme dite des "maths modernes" a ravagé le sens mathématique d'une génération d'écoliers en présentant, selon le pédagogue B. Charlot, "un enseignement formalisé à l'extrême, coupé de tout support intuitif ou présenté à partir de situations artificielles, et très sélectif"... Aux difficultés considérables que pose déjà l'arithmétique*



à tout cerveau normalement constitué vient alors s'ajouter un trouble affectif : la phobie des mathématiques. Nous pouvons lutter contre ces difficultés si nous bâtissons les connaissances mathématiques, dans le cerveau de nos enfants, sur le concret et non sur l'abstrait... Les enfants ne demandent qu'à aimer les mathématiques, pour peu qu'on leur en présente les aspects ludiques plutôt que la symbolique abstraite. Parents, jouez au jeu de l'oie avec vos enfants : vous leur donnerez un bon départ en arithmétique.

Dieudonné, membre influent de l'école Bourbakiste et donc ancien partisan de l'abstraction des mathématiques, était lui-même revenu, dans « Pour l'honneur de l'esprit humain, les mathématiques d'aujourd'hui » (1987), sur ses penchants au tout algèbre et à l'interdit de l'image en mathématiques : *Le remplacement du langage algébrique par le langage géométrique apporte des simplifications considérables et fait apparaître des propriétés qui restent insoupçonnées lorsqu'elles sont enfouies sous un fatras de calculs. Ainsi les modèles mathématiques de la programmation linéaire et de l'optimisation se comprennent et se traitent beaucoup mieux lorsqu'on interprète les inégalités qui y figurent sous forme géométrique.*

Pascal : La mémoire est nécessaire pour toutes les opérations de l'esprit.
Voltaire : Les Muses, filles de la Mémoire, nous enseignent que sans mémoire on n'a pas d'esprit.

En 1994, le vulgarisateur des mathématiques du chaos Ian Steward saluait le retour de la géométrie : *Après s'être recyclées dans le formalisme avec Bourbaki, les mathématiques actuelles reviennent en courant vers l'étage géométrique de la spirale, aussi vite que leurs jambes peuvent les porter.*

L'omniprésence de Mnémosyne

Les chercheurs impliqués dans les sciences du cerveau en sont convaincus, c'est de la compréhension du fonctionnement de la mémoire que se déduiront les règles de traduction entre la matière et l'esprit (Rose, « La mémoire, des molécules à l'esprit », 1992).

Mais pour ce qui concerne la pratique, on le sait depuis les Anciens, la clé d'une bonne mémorisation est le classement ordonné des choses dont on veut se souvenir, en les associant à des représentations, c'est-à-dire des images : *Les personnes désirant soumettre cette faculté à un entraînement doivent choisir en pensée des lieux distincts, puis former des images mentales des choses dont ils veulent se souvenir et ranger ces images dans les divers lieux. De cette façon, l'ordre des lieux conservera l'ordre des choses ; et les images des choses évoqueront les choses elles-mêmes* [Cicéron, « De oratore »].

Monique LE GUEN

John Wilder TUKEY

John Wilder Tukey est né en 1915 à New Bedford dans le Massachussets. Éduqué par ses parents, tous deux enseignants, il fréquente assidûment la bibliothèque publique, où il a l'occasion de découvrir le *Journal of the American Chemical Society* et les *Transactions of the American Mathematical Society*. Il obtient en 1937 un Master's Degree en chimie, à l'Université de Brown, puis, en 1939, un doctorat de mathématiques à l'Université de Princeton. Il débute sa carrière professionnelle dans cette même université, comme enseignant en mathématiques. Lorsque la deuxième guerre mondiale éclate, il rejoint le *Fire Control Research Office*, où il découvre la statistique. De retour à Princeton, il s'intègre à la communauté naissante des statisticiens (Wilks, Cochran, Moods, Winsor...).

Une carrière impressionnante

À partir de 1945 et tout au long de sa carrière, Tukey se partagera entre l'enseignement de la statistique, à l'Université de Princeton, et la recherche et le développement, au sein de la direction technique des laboratoires AT&T Bell Company (à Murray Hill). Membre, entre autres comités, de l'Académie des Sciences, de la Société américaine de Philosophie et du *President's Science Advisory Committee*, il sera également consultant en entreprise, en particulier à la Xerox Corporation Palo Alto CA, où seront conçues les premières interfaces graphiques ainsi que la souris.

Il part officiellement en retraite en 1985, à l'âge de 70 ans, mais continuera ses travaux : *When I formally retire (at the end of June 1985) I do not plan to stop thinking or working. I plan to continue to provide both new techniques and new annoying-but-true statements*. En 1995, la Fondation Minerva organise en son honneur, à l'Université de Princeton, un symposium réunissant ses principaux thésards et élèves. En 1997 encore il publiait un article (« More Honest Foundations For Data Analysis », *Journal of Statistical Planning and Inference*), toujours avec Princeton University comme adresse.

Une oeuvre de visionnaire

Auteur de plus de 500 articles, seul ou en collaboration, Tukey a touché à tous les domaines de la statistique classique : probabilités, séries temporelles, analyse de la variance, régression... Il a rassemblé et reformulé nombre d'anciens outils statistiques. Surtout, il en a imaginé de nouveaux, en s'attachant à clarifier les concepts. Sa plus grande contribution est sans doute d'avoir distingué deux étapes dans l'analyse de données, d'abord l'analyse exploratoire, ensuite l'analyse confirmatoire, en insistant sur la nécessité de l'une et l'autre démarches : *We need both exploratory and confirmatory data analysis*. Brillinger, de l'Université de Berkeley, le considère comme l'un des statisticiens les plus influents de ce siècle.

Dès 1967, dans le droit fil d'un article visionnaire sur l'avenir de l'analyse de données (« The future of Data Analysis », 1962), Tukey préconisait le développement de l'usage de l'informatique, la construction d'interfaces facilitant la relation homme-machine et la mise au point de logiciels adaptés au traitement de l'information. Dix ans plus tard, la parution de son livre « Exploratory Data Analysis », qui donnait une visibilité à son enseignement du début des années soixante-dix, révolutionnait la façon dont les statisticiens percevaient l'analyse de données.

Le progrès technologique (micro-ordinateurs, souris, interfaces graphiques, langages orientés-objet) a permis, à partir de 1985, la concrétisation informatique des idées de Tukey.

Princeton et AT&T Bell Company
Hauts lieux de la recherche en traitement de l'information et cybernétique

À Princeton comme à l'AT&T Bell Company, la personnalité remarquable de Tukey bénéficiera d'un environnement humain tout aussi exceptionnel.

À la fin des années 1940, début des années 1950, l'Université de Princeton rassemblait en effet de grands mathématiciens et logiciens, notamment Norbert Wiener (1894-1964), John Von Neumann (1903-1957) et Alan Turing (1912-1954) : à eux trois, ils jetteront les bases de l'informatique en concevant[1] *et réalisant les premiers calculateurs séquentiels et digitaux. Wiener, auteur de «* Cybernetics or Control and Communication in the Animal and the Machine *» (1948), et Von Neumann, auteur de «* The Computer and the Brain *» (1958), seront également les pionniers de ce qui deviendra, dans les années 1980, les sciences cognitives.*

À l'AT&T Bell Company , Tukey côtoiera l'ingénieur et mathématicien Claude Shannon (né en 1916), chercheur dans cette compagnie de 1941 à 1972 et autre grand nom des mathématiques et de la logique. Shannon soutint sa thèse au Massachussets Institute of Technology sur l'usage de l'algèbre de Boole pour analyser et optimiser des circuits de relais de communications. Il est l'auteur de « A Mathematical Theory of Communication », publié dans le Bell System Technical Journal *(1948).*

1. La fameuse « architecture Von Neumann ».

Actuellement, tous les nouveaux logiciels statistiques intègrent l'analyse exploratoire des données (AED). De même, les dernières versions des « anciens » logiciels statistiques ont été enrichies de nouvelles fonctions relevant de l'AED.

L'influence de Tukey est également patente dans l'enseignement de la statistique. À titre d'exemple, les ouvrages de Fox & Long « Modern Methods of Data Analysis » (1990) et d'Erickson & Nosanchuk « Understanding Data: An Introduction to Exploratory Data Analysis for Students in the Social Sciences » (1992) sont directement inspirés de ses idées.

Des idées-forces

Pour Tukey, qui a toujours accordé une grande importance à la réflexion épistémologique, la statistique est tout à la fois une science, un art, une philosophie, et une technique pour inférer du particulier au général. Elle n'est pas plus une branche des mathématiques que ne le sont la physique, la chimie ou l'économie. L'usage bien compris des modèles mathématiques est bien sûr nécessaire, mais laisser la statistique mathématique envahir toute la statistique serait une erreur fatale. En analyse des données, le praticien doit faire appel à sa propre intuition, son action doit être guidée et non déterminée par ce qui découle de la théorie ou ce qui a été établi par l'expérience.

Lors de son allocution de départ en retraite, Tukey réaffirmait quelques-unes de ses idées maîtresses, fondées sur 44 années de statistique et d'analyse de données :

- toute supposition doit être considérée comme un point de départ mais non comme une vérité ;

- la stratégie de traitement des données est plus importante que les tactiques d'exécution d'analyses statistiques spécifiques ;

- avec l'accumulation des données, les modèles fonctionnels, plus faciles à vérifier, doivent être préférés aux modèles stochastiques ;

- il faut, à l'adresse des utilisateurs, se focaliser sur les résultats les plus significatifs (comme de simples comparaisons), en les appuyant sur des techniques de représentation graphique toujours plus puissantes.

« *Il est du devoir des statisticiens d'informer leurs clients que ce qu'ils essaient de faire est plus difficile que ce qu'ils peuvent penser* » [Tukey, 1967].

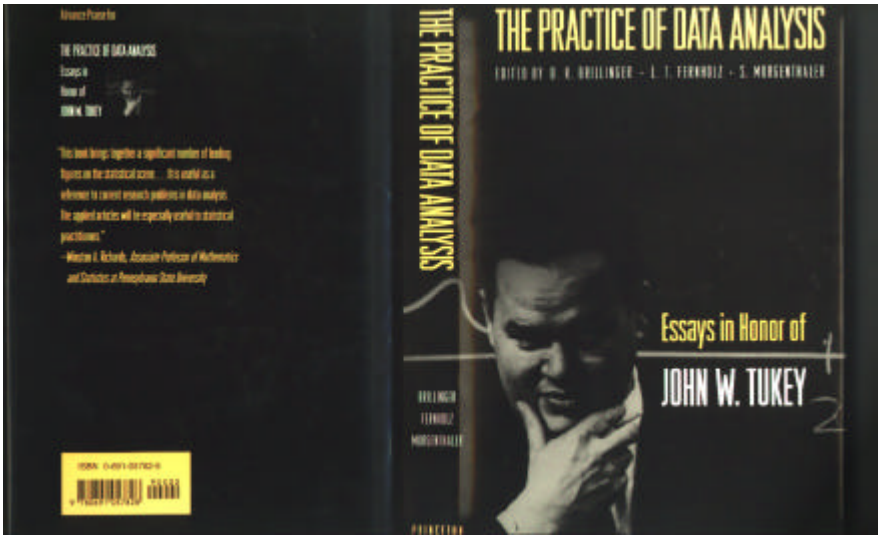
Monique LE GUEN

Tukey a encadré et orienté plus de 50 thésards, parmi lesquels Mosteller, Goodman, Wallace, Dempster, Brillinger, Hartigan, Wannacott, Hoaglin, Velleman et Morgenthaler, qui tous ont œuvré pour développer l'analyse exploratoire des données.

On lui doit, mais la liste n'est pas exhaustive, la technique de la Median Polish*, le lissage par médianes mobiles, l'algorithme de la transformée de Fourier rapide (FFT)*[1]*, quelques lois de probabilités, le* Jackknife *(qu'il a lui-même baptisé ainsi, du nom du couteau multi-usages du boy-scout), les graphiques* Stem and Leaf *et* Box Plot*, sans oublier, bien sûr, la* Tukey's Line*, le* Tukey's Quick Test*, le* Tukey's Test for Non-Additivity*, le test de* Siegel-Tukey *et le critère de* Tukey-Kramer*.*

Trimming*,* Winsorized Mean *et* Bit *(Binary digIT) sont autant de mots ou expressions qu'il a inventés.*

1. L'algorithme qui a le plus changé notre société selon B. Burke-Hubbard.



Brillinger D.R, Fernholtz L.T. & Morgenthaler S.
The Practice of Data Analysis, Essays in Honor of John W. Tukey
Princeton University Press, 1997

« We need both exploratory and confirmatory data analysis »

Dans le protocole d'analyse des données tel que le définit Tukey, exploration puis confirmation, l'approche confirmatoire suit une logique d'expérimentation. Elle met en oeuvre les techniques de la statistique mathématique classique, avec ses briques de base que sont la moyenne, la loi normale, le test de Student, l'égalité des variances, la linéarité, la modélisation. Ces techniques sont rigoureuses, bien formalisées, impressionnantes même. Mais reposant sur des présupposés, elles ne sont plus optimales si ces derniers ne sont pas vérifiés.

L'approche exploratoire suit au contraire une logique d'observation. L'explorateur va regarder ses données sous toutes les facettes, tenter de mettre en évidence les structures qu'elles recèlent, enfin et le cas échéant formuler des hypothèses plausibles.

Ne pas se laisser abuser par les mots

Pour Tukey, les tests ne doivent pas être considérés comme des témoignages infaillibles, et les modèles ne sont utiles que pour mieux comprendre les données.

Et si par exemple il reconnaît volontiers que le test de Student a constitué une véritable avancée, il en dénonce néanmoins quelques inconvénients cachés : il repose sur une supposition (la normalité), il met trop l'accent sur l'exactitude de la solution dans le cadre d'un problème « idéal », il détourne le statisticien théoricien d'un travail de recherche de solutions pour des problèmes « non idéaux », il empêche de s'attaquer aux problèmes de mélanges de populations.

Dans la même veine, il réfute le qualificatif de « normale » attaché à la loi de Gauss, selon lui très optimiste et uniquement lié aux formidables propriétés mathématiques de la loi en question.

Rester le plus possible attaché aux données

En statistique, l'objectif final est bien de comprendre, et de faire comprendre aux utilisateurs, les structures ou configurations que recèlent les données qu'on a recueillies. Comme le dit Velleman, *we must be clear that the underlying goal is understanding, not p-values*.

Aussi Tukey va-t-il mettre l'accent sur l'attitude à adopter face à des données réelles, et sur la compréhension des techniques utilisées pour traiter ces données. La philosophie du praticien doit être de rester le plus possible attaché aux données. Ce sont elles qui doivent le guider dans le choix du modèle, et les tests ne sauraient remplacer les idées : la statistique confirmatoire n'est pas une grande prêtresse, mais une servante (*not a high priestess but a handmaiden*).

Voir pour croire ou pour ne pas croire

Un exemple donné par Anscombe (« Graphs in Statistical Analysis »,1973), et que nous reprenons ici, est particulièrement frappant quant aux erreurs d'interprétation que pourrait entraîner un usage non contrôlé des techniques de la statistique mathématique, en l'occurrence la régression linéaire.

Bien sûr, la théorie du modèle linéaire nous donne aujourdhui de nombreux indicateurs numériques pour étudier les résidus, détecter les points influents... et porter ainsi un jugement sur la qualité du modèle. Dans de nombreux cas cependant, le graphique reste un outil irremplaçable (Cook & Weisberg, 1994).

Monique LE GUEN

Normale ?

La loi de Gauss, que Galton qualifiera de normale, est issue de la loi des erreurs utilisée par les astronomes. Bien que son usage soit devenu très courant, sa légitimité a souvent été mise en cause. Ainsi, Poincaré citait, à propos de la loi des erreurs, cette boutade d'un ami physicien : *Tout le monde y croit fermement parce que les mathématiciens s'imaginent que c'est un fait d'observation et les observateurs que c'est un théorème de mathématiques* (« La science et l'hypothèse », 1902). Tukey reprendra cette citation. Il rappellera également que Student lui-même disait que jamais les observations et mesures n'étaient distribuées selon les « courbes magiques en forme de cloche » (*magic bell-shaped curves*).

L'exemple d'Anscombe (revu par Tomassone et al.)

Effectuons une régression linéaire sur les 5 couples de variables ci-dessous (X,Ya), (X,Yb), (X,Yc), (X,Yd) et (Xe,Ye).

Obs	X	Ya	Yb	Yc	Yd	Xe	Ye
1	7	5,535	0,113	7,399	3,864	13,715	5,654
2	8	9,942	3,770	8,546	4,942	13,715	7,072
3	9	4,249	7,426	8,468	7,504	13,715	8,491
4	10	8,656	8,792	9,616	8,581	13,715	9,909
5	12	10,737	12,688	10,685	12,221	13,715	9,909
6	13	15,144	12,889	10,607	8,842	13,715	9,909
7	14	13,939	14,253	10,529	9,919	13,715	11,327
8	14	9,450	16,545	11,754	15,860	13,715	11,327
9	15	7,124	15,620	11,676	13,967	13,715	12,746
10	17	13,693	17,206	12,745	19,092	13,715	12,746
11	18	18,100	16,281	13,893	17,198	13,715	12,746
12	19	11,285	17,647	12,590	12,334	13,715	14,164
13	19	21,365	14,211	15,040	19,761	13,715	15,582
14	20	15,692	15,577	13,737	16,382	13,715	15,582
15	21	18,977	14,652	14,884	18,945	13,715	17,001
16	23	17,690	13,947	29,431	12,187	33,281	27,435

Surprise, les résultats sont strictement identiques (les incrédules peuvent vérifier) :
même droite de régression, même R2, mêmes erreurs-type sur les coefficients, mêmes Student et mêmes p-values.

Model Equation

YA = 0.52 + 0.81 X

Summary of Fit

Mean of Response	12.60	R-Square	0.62
Root MSE	3.22	Adj R-Sq	0.59

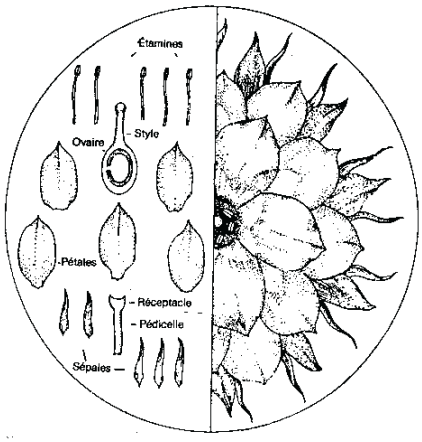
Analysis of Variance

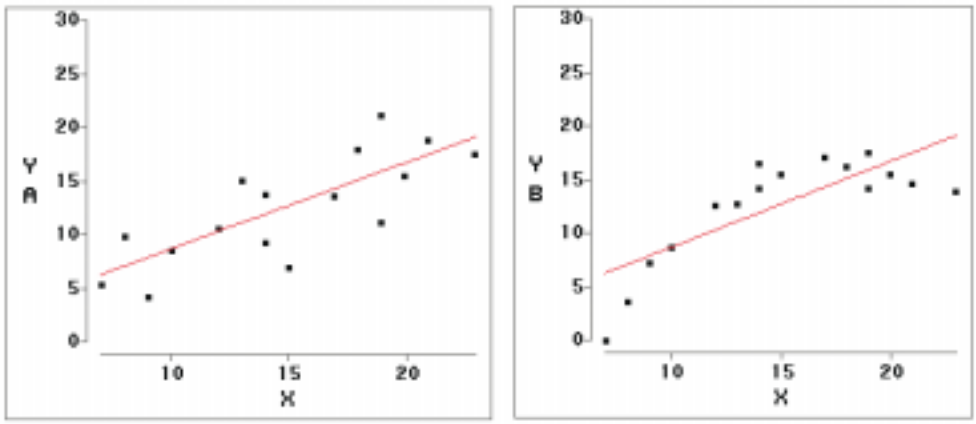
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
Model	1.0	234.6	234.6	22.6	0.0003
Error	14.0	145.4	10.4		
C Total	15.0	380.1			

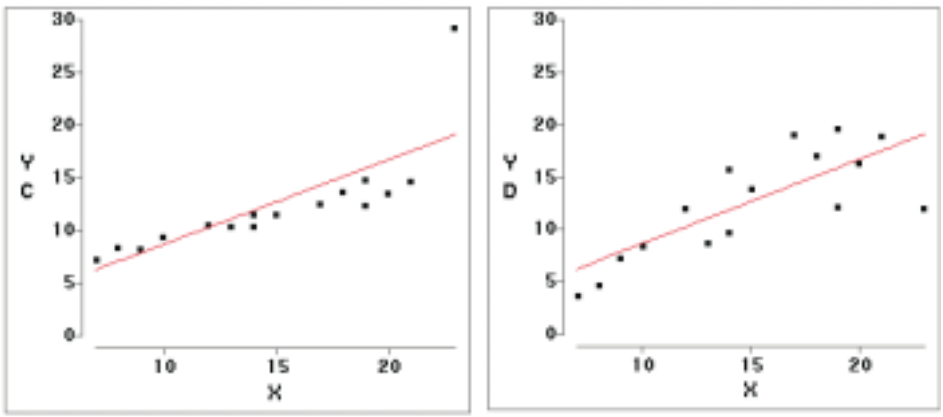
Parameter Estimates

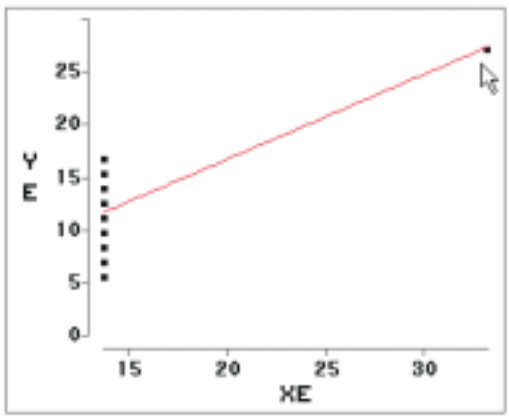
Variabl	DF	Estimat	Std Err	T Stat	Prob >|	Toleran	Var Inflat
INTERC	1.00	0.52	2.67	0.20	0.8476	.	0
X	1.00	0.81	0.17	4.75	0.0003	1.00	1.00

Notre cerveau gauche analytique a bien du mal
à s'en sortir !
Pourquoi ne pas faire appel à notre cerveau droit
en faisant de simples graphiques ?

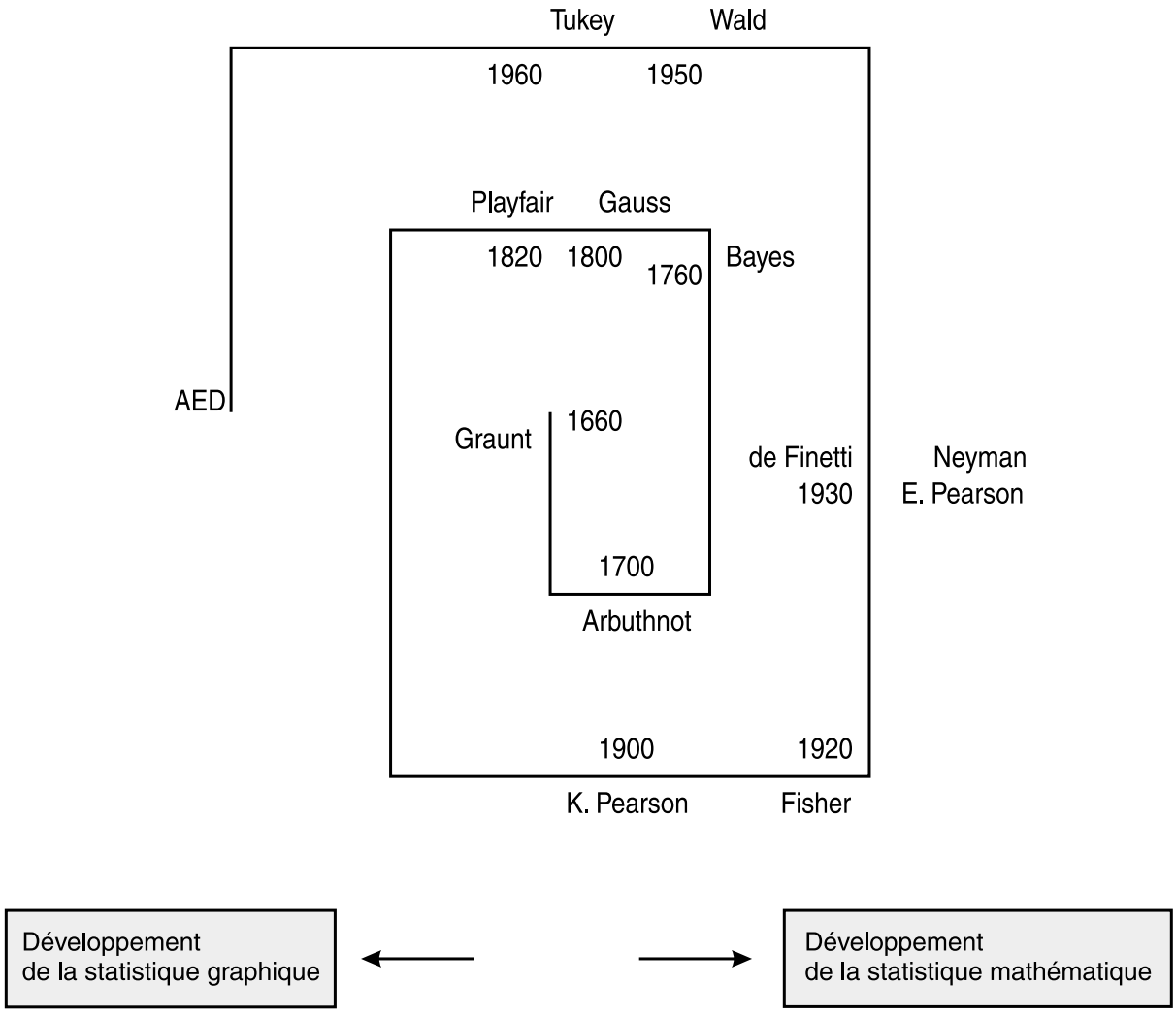








Le chemin de la statistique, d'après Peter J. HUBER



AED mode d'emploi

L'objectif de l'analyse exploratoire des données est de découvrir des structures, des groupes, des relations..., en se gardant autant que faire se peut des présupposés. Cette approche s'appuie sur un équipement adéquat, en l'occurrence un logiciel statistique perfectionné, intégrant des fonctions de représentation graphique adaptées. À tous les stades de l'analyse, l'explorateur devra en effet pouvoir visualiser ses données et l'effet des traitements qu'il leur aura appliqués, cela en mode interactif et sous différents points de vue. Réexpression des données, résistance et robustesse des indicateurs et des procédures et analyse fine des « restes » (en analyse exploratoire, on préfère parler de restes plutôt que de résidus) sont les trois autres fondamentaux de l'AED.

Mettre en scène les données

L'explorateur, pour se faire une idée des structures que recèlent ses données, doit pouvoir s'appuyer sur des représentations graphiques expressives, les plus complètes possibles, matérialisant la façon dont se distribuent les variables étudiées : *Ideas come from previous exploration more than from lightning strokes*.

À cette fin, Tukey a perfectionné des représentations graphiques déjà existantes, imaginant par exemple le *Stem and Leaf,* cousin de l'histogramme. Il en a également inventé de nouvelles, notamment le *Box Plot* (abréviation de *Box and Whiskers Plot,* qui peut se traduire par boîte à moustaches ou encore boîte à pattes). Parmi les autres représentations graphiques les plus couramment utilisées en analyse exploratoire des données, on citera en particulier les diagrammes de dispersion et matrices de diagrammes de dispersion (pour l'étude des liaisons pouvant exister entre les variables), les *QQplots* et *PPplots* (qui permettent de comparer des distributions observées à des distributions théoriques), les diagrammes de rotation en 3D (pour l'interprétation des plans factoriels).

Graphiques et statistique

Avec trois nombres vous faites une phrase, avec quatre à vingt nombres vous faites un tableau, au delà de vingt nombres vous faites un graphique (Tufte).

Dès l'origine (Quételet, Galton) et pendant très longtemps, la représentation graphique a joué dans l'analyse des données réelles un rôle essentiel. Cette pratique s'est toutefois diluée aux premiers temps de l'informatique (1950-1975), car peu adaptée aux sorties ligne à ligne des imprimantes de l'époque. Pendant cette période, on s'est donc essentiellement appliqué à produire du chiffre. Mais avec les nouveaux outils informatiques et sous l'impulsion de Tukey, la statistique graphique a aujourd'hui retrouvé toute sa respectabilité.

La représentation graphique n'est pas seulement un moyen de regarder les données. Couplée avec l'interactivité, elle devient en effet un puissant outil d'analyse. Par exemple, l'utilisation de marqueurs différents va permettre de discriminer des groupes de données sur un diagramme de dispersion, ajoutant ainsi une troisième dimension à la relation spatiale entre deux variables. On pourra même y faire apparaître une quatrième variable au moyen de la couleur.

Des exemples concrets et commentés sont donnés dans l'article « Visualisation interactive et réexpression des données avec Sas/Insight ».

Réexprimer les données

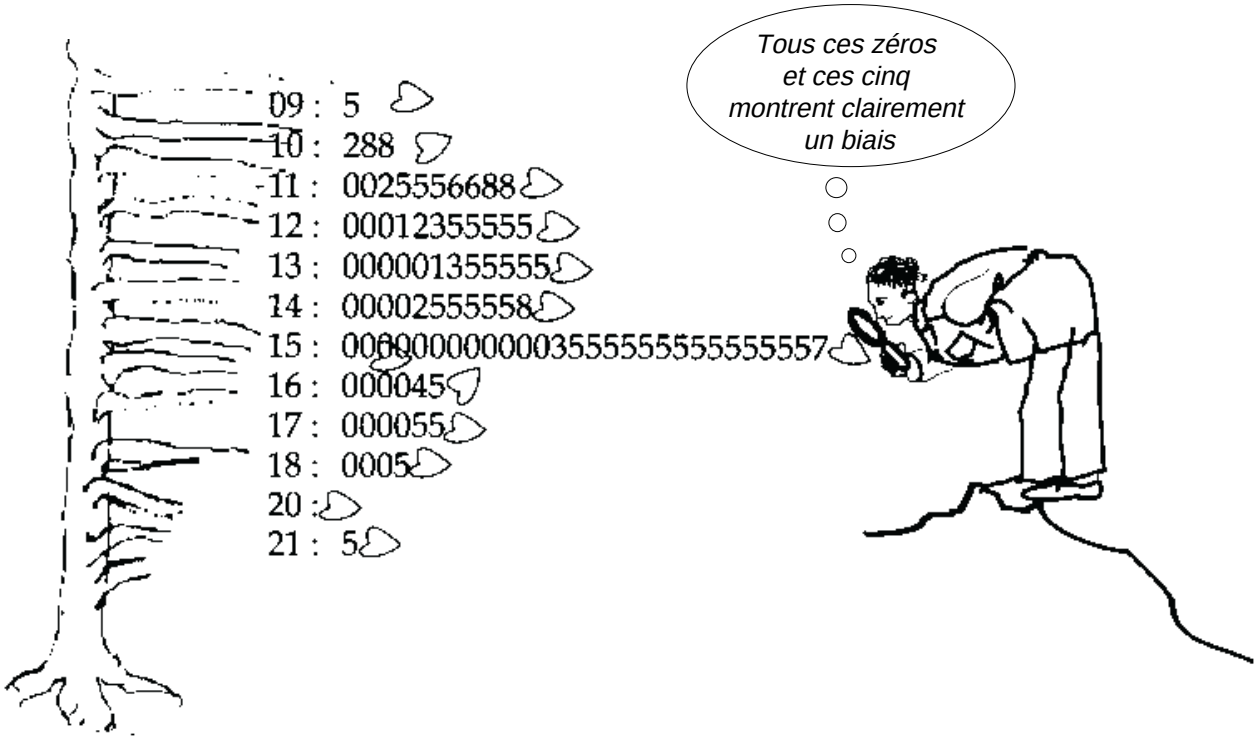
Le novice a souvent tendance à rester attaché à la forme première des données. Or celles-ci ne sont pas « données » (*data are not given*). On peut les adapter, afin d'y découvrir des structures cachées. Mais en la matière, il faut

	Puissance	Transformée	
	etc.	etc.	*Monter l'échelle*
	4	Y^4	
	3	Y^3	
	2	Y^2	
Variable brute	**1**	**Y**	
	1/2	$Y^{1/2}$	
	0	lnY	
	-1/2	$-1/\sqrt{Y}$	
	-1	-1/Y	
	-2	$-1/ Y^2$	
	-3	$-1/ Y^3$	
	etc.	etc.	*Descendre l'échelle*

Échelle de puissance de Tukey
(selon la représentation originale de Horber E.)

Le Stem and Leaf, cousin de l'histogramme

John Tukey a inventé la représentation *Stem and Leaf* (tige et feuille), qui permet de résumer les observations tout en conservant les données individuelles.



Lecture :
sur la 1re ligne, lire la valeur 95 ;
sur la 2e ligne, lire les 3 valeurs 102, 108 et 108 ;
sur la 3e ligne, lire les 10 valeurs 110, 110, 112, etc.

Réalisation d'après une idée de Larry Gonick & Woollcott Smith
(« The carton guide of statistics », Harper Perennial, 1993)

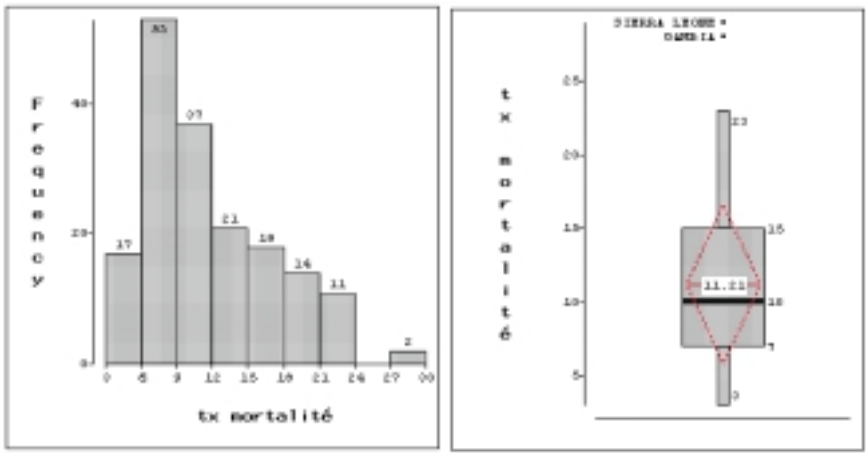
Les Box Plots, une mine d'informations

La variable ici étudiée est le taux de mortalité (pour mille habitants), dont on a relevé la valeur dans 173 pays (53 en Afrique, 39 en Amérique, 44 en Asie, 28 en Europe et 9 en Océanie).

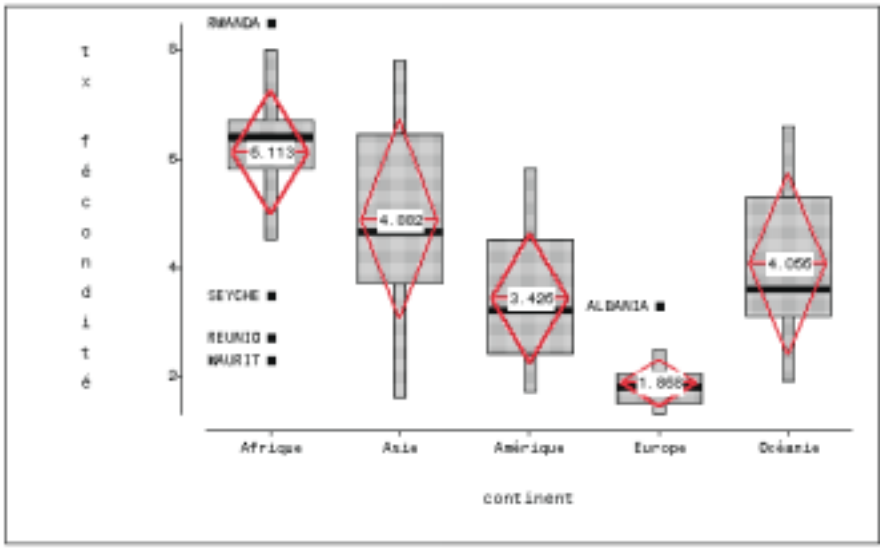
À gauche est donné un histogramme classique. À droite un Box Plot, qui apporte de nombreuses informations supplémentaires :

- la hauteur de la boîte (rectangle central) détermine l'étendue de la partie centrale de la distribution (taux de mortalité compris entre les quartiles Q1 et Q3, valeurs 7 et 15) ;
- la bande horizontale matérialisée à l'intérieur de cette même boîte indique la position de la médiane (valeur 10) ;
- l'étendue des queues de distribution hors données atypiques est déterminée par la hauteur des moustaches, étant précisé que l'extrémité de la moustache inférieure est ici fixée par la plus petite valeur (= 3) supérieure ou égale à Q1-1,5 (Q3 - Q1), celle de la moustache supérieure par la plus grande valeur (= 23) inférieure ou égale à Q3 +1,5 (Q3 - Q1) ;
- de part et d'autre des moustaches sont mises en évidence les observations atypiques, avec valeur inférieure à Q1 - 1,5 (Q3 - Q1) ou supérieure à Q3 + 1,5 (Q3 - Q1), ici la Gambie et la Sierra Leone, où le taux de mortalité est particulièrement élevé.

On peut également faire apparaître les valeurs de la moyenne et de l'écart-type. Dans la représentation ci-dessous, la petite diagonale du losange superposé à la boîte indique la position de la moyenne (valeur 11,21), sa grande diagonale, de longueur 2 σ, permet d'apprécier la valeur de l'écart-type.



Intéressons-nous à présent à la distribution du taux de fécondité (nombre d'enfants par femme). Par simple juxtaposition des Box Plots relatifs aux différents continents, on voit tout de suite que c'est en Europe que la concentration est de très loin la plus forte (autour d'une moyenne faible et très proche de la médiane), qu'il existe dans ce même continent un pays atypique (l'Albanie), que c'est en Asie que la distribution est la plus dispersée, et cetera, et cetera.
Ce type de représentation constitue une excellente introduction visuelle à l'analyse de la variance.



aussi savoir raison garder... Aussi Tukey a-t-il proposé à l'usage de l'explorateur une échelle de transformations raisonnable et raisonnée.

Dans la pratique, deux utilisations très répandues de la transformation (ou réexpression, selon le terme de Tukey) des données sont la symétrisation de la distribution d'une variable et la linéarisation d'une liaison entre deux variables.

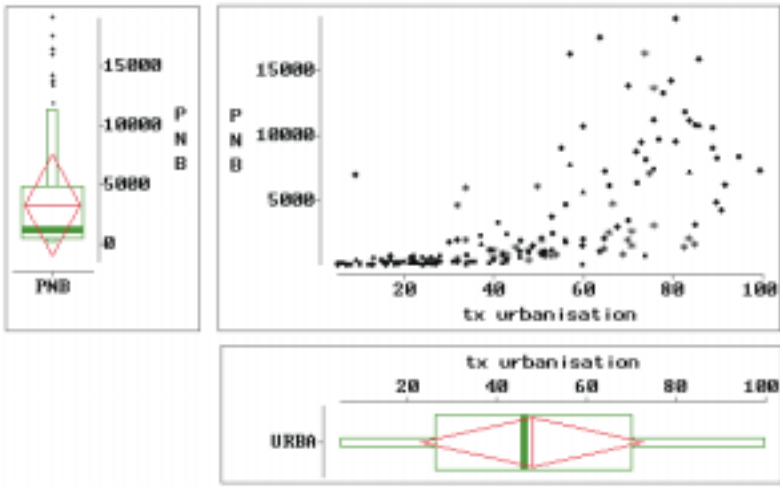
Si par exemple la courbe représentative de la distribution d'une variable est assymétrique avec une queue plus étalée à droite, peut-être pourra-t-on découvrir, en s'intéressant à son logarithme (puissance 0 dans l'échelle de Tukey), que cette variable suit la loi log-normale (un cas bien connu est celui de la distribution du *revenu*). De même, si l'on cherche à établir la liaison, a priori non linéaire, pouvant exister entre deux variables, on pourra utilement procéder à des essais de transformation successifs sur l'une ou l'autre de ces deux variables, en montant ou descendant l'échelle de Tukey. Peut-être mettra-t-on alors au jour une liaison linéaire entre la première variable et une transformée de la seconde.

Avec les nouveaux outils de la micro, réaliser ces transformations est devenu très facile et « compréhensible ».

De l'utilité de la réexpression des données

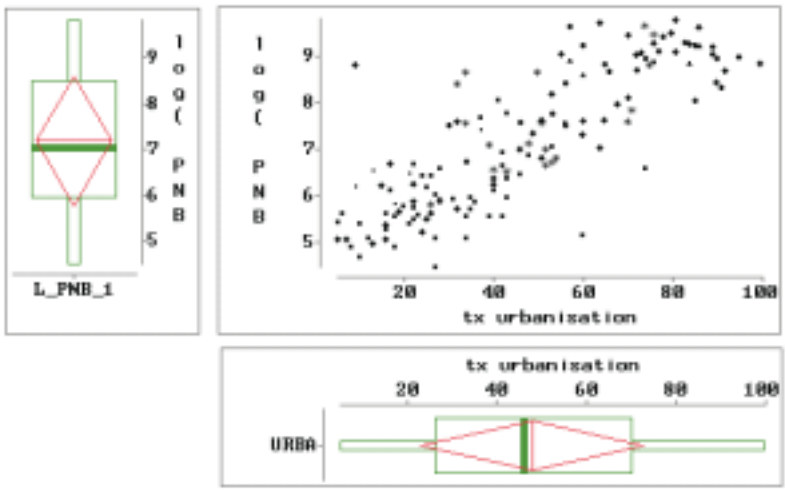
Les variables ici étudiées sont le PNB et le taux d'urbanisation, le nombre d'observations (nombre de pays observés) est égal à 173.

Dans cette première fenêtre graphique sont regroupés deux Box Plots, le premier relatif à la distribution du PNB, le second à la distribution du taux d'urbanisation, ainsi qu'un diagramme de dispersion présentant la façon dont se répartissent les 173 pays observés suivant ces deux mêmes variables.



La même configuration est reprise ci-dessous, à cet important détail près : la variable PNB a été transformée en logPNB (puissance zéro sur l'échelle de Tukey).

On constate alors, d'une part que la distribution de logPNB est quasi symétrique, d'autre part qu'il semble y avoir une liaison linéaire entre cette variable et le taux d'urbanisation.



Résistance et robustesse

Dans le langage courant, résistance et robustesse désignent des qualités très voisines. S'agissant de la chose statistique, on admettra, comme Velleman, que le qualificatif de résistant doit plutôt s'appliquer à un indicateur, celui de robuste à une méthode.

Un indicateur sera dit résistant s'il est peu sensible à l'effet des données atypiques : ainsi la médiane est un indicateur de tendance centrale résistant, mais la moyenne ne l'est pas.

Une méthode sera dite robuste si elle est peu sensible à des déviations du corpus d'hypothèses qui en principe la sous-tend : ainsi, la pourtant si précieuse régression linéaire basée sur les moindres carrés n'est pas une méthode robuste.

Dispersion d'une distribution

Un principe général, couramment admis, est que « les données sont toujours normales en leur milieu ». Encore faut-il déterminer où se situe le milieu... En phase exploratoire, on évitera donc de se focaliser sur la moyenne, très sensible à l'effet des données atypiques. On préfèrera au contraire s'intéresser à des indicateurs de tendance centrale résistants, tels que la médiane, les moyennes tronquées ou winsorisées, la *midmean* ou encore la *trimean.*

Une autre règle d'or de l'AED est qu'une distribution doit être étudiée sur toute sa longueur. Confronté par exemple à une courbe en forme de cloche, l'explorateur ne se laissera pas obnubiler par l'hypothèse de normalité, souvent illusoire, et privilégiera la mise en œuvre de méthodes non paramétriques. Il se préoccupera bien sûr de la forme plus ou moins aplatie (Kurtosis) du milieu de la courbe, mais accordera une attention prioritaire aux aspects de symétrie. Les queues de distribution notamment doivent faire l'objet d'un examen approfondi, et c'est bien dans cette perspective que Tukey avait imaginé le *Box Plot.*

Ajustement d'une liaison linéaire

Robustesse toujours, Tukey a inventé, en matière d'ajustement linéaire, un substitut à la droite des moindres carrés, la *Tukey's Line*, à construire, dans un graphique cartésien (X, Y), selon le schéma ci-dessous :

- on découpe le nuage de points selon la valeur de X en trois groupes de taille identique (même nombre de points dans chacun) ;
- on calcule les valeurs médianes de X et de Y pour les premier et troisième groupes ; les valeurs ainsi calculées déterminent deux points (le point médian du premier groupe et celui du troisième) ;
- on trace la droite passant par ces deux points (droite des médianes croisées) ;
- la *Tukey's Line* est la parallèle à cette droite qui partage le nuage de points en deux groupes de taille identique.

Analyser les restes

La démarche exploratoire est par nature itérative. Prenons l'exemple de la distribution des taux de fécondité en Afrique, dont le *Box Plot* est présenté dans l'encadré de la page 19. On remarque instantanément que quatre pays se distinguent : le Rwanda, les Seychelles, la Réunion et l'Île Maurice. L'explorateur va essayer d'expliquer ces particularités en échafaudant des hypothèses : les trois pays aux taux de fécondité les plus bas sont des îles de l'océan Indien ; ce point commun est-il un élément d'explication ? Si oui, pourquoi Madagascar n'est-il pas atypique ? etc. Lorsqu'il aura trouvé une explication plausible à cette particularité, l'explorateur va s'intéresser aux autres pays, au « reste » de la distribution, qu'il représentera par un autre graphique, graphique qui révèlera peut être d'autres particularités qu'il expliquera, etc.

L'idée de base de l'exploration est en effet de décomposer les données selon une structure, les points atypiques dans notre exemple, et un reste, dont on recherchera à nouveau la décomposition en structure + reste, et ainsi de suite.

Pour transcrire cet enchaînement, on trouve dans la littérature la séquence d'« équations » suivante :

Data = Smooth_1 + Rough_1
Rough_1 = Smooth_2 + Rough_2
Etc.

Les données sont donc divisées en une structure souvent simple, interprétable, et qualifiée de « lisse » dans la mesure où elle ne pose plus de problème majeur, et en un reste, non encore étudié et interprété, qualifié de « rugueux ».

Remarquons que le reste n'a aucune raison d'être « petit ». Ainsi, dans notre exemple, la structure exhibée concerne quatre pays, et le reste, qui concerne 49 pays, contient sans doute la majeure partie de l'information. C'est pourquoi les explorateurs ne parlent pas de « résidu », ce terme faisant souvent référence, en statistique classique, à quelque chose de petit.

Cette stratégie itérative d'analyse est toujours utilisée en exploration, où les restes font l'objet d'une attention toute particulière.

Sophie DESTANDAU,
Dominique LADIRAY
et **Monique LE GUEN**

Les moyennes tronquées (*Trimmed Means*) sont des moyennes arithmétiques calculées après élimination des valeurs extrêmes (la plus petite et la plus grande, les 2 plus petites et les 2 plus grandes, etc.).

Les moyennes winsorisées (*Winsorized Means*) ont été introduites par Winsor dans les années 1940. Au lieu de supprimer les valeurs extrêmes, Winsor les remplace par la valeur la plus proche non suspecte.

La midmean est la moyenne des valeurs supérieures ou égales au premier quartile (Q1) et inférieures ou égales au troisième (Q3).

La **trimean** est égale à (Q1+2Q2+Q3) / 4, où Q2 désigne la médiane.

L’analyse de données à la française dans la typologie de Tukey

Sous la dénomination générique d’analyse de données à la française sont regroupées l’ACP (analyse en composantes principales), l’AFC (analyse factorielle des correspondances) et l’ACM (analyse des correspondances multiples) ainsi que les méthodes de classification, soit autant d’approches relevant de l’analyse exploratoire multidimensionnelle au sens de Tukey. C’est d’ailleurs bien ainsi que les statisticiens français les présentent aujourd’hui.

Cf. Lebart L., Morineau A., Piron M. : Statistique exploratoire multidimensionnelle, 1997.

À la découverte de l'AED avec SAS/Insight

Avec les récents progrès de la micro, les méthodes de représentation graphique, dont l'usage s'était partiellement dilué aux premiers temps de l'informatique, sont revenues en force sur le devant de la scène statistique.

Ces dernières années ont ainsi vu fleurir de nouveaux logiciels statistiques (Datadesk, S+, JMP, etc.) harmonisant procédures exploratoires et procédures confirmatoires, conformément au principe qu'avait affirmé John Tukey : *We need both exploratory and confirmatory data analysis.*

Dans le même temps, les logiciels « dinosaures », SAS en particulier avec son nouveau module SAS/Insight, se sont enrichis de fonctions spécifiques répondant aux préoccupations de l'AED.

Un bon logiciel d'introduction à l'AED...

Partie intégrante du système SAS, SAS/Insight réunit beaucoup des qualités attendues d'un logiciel d'AED :

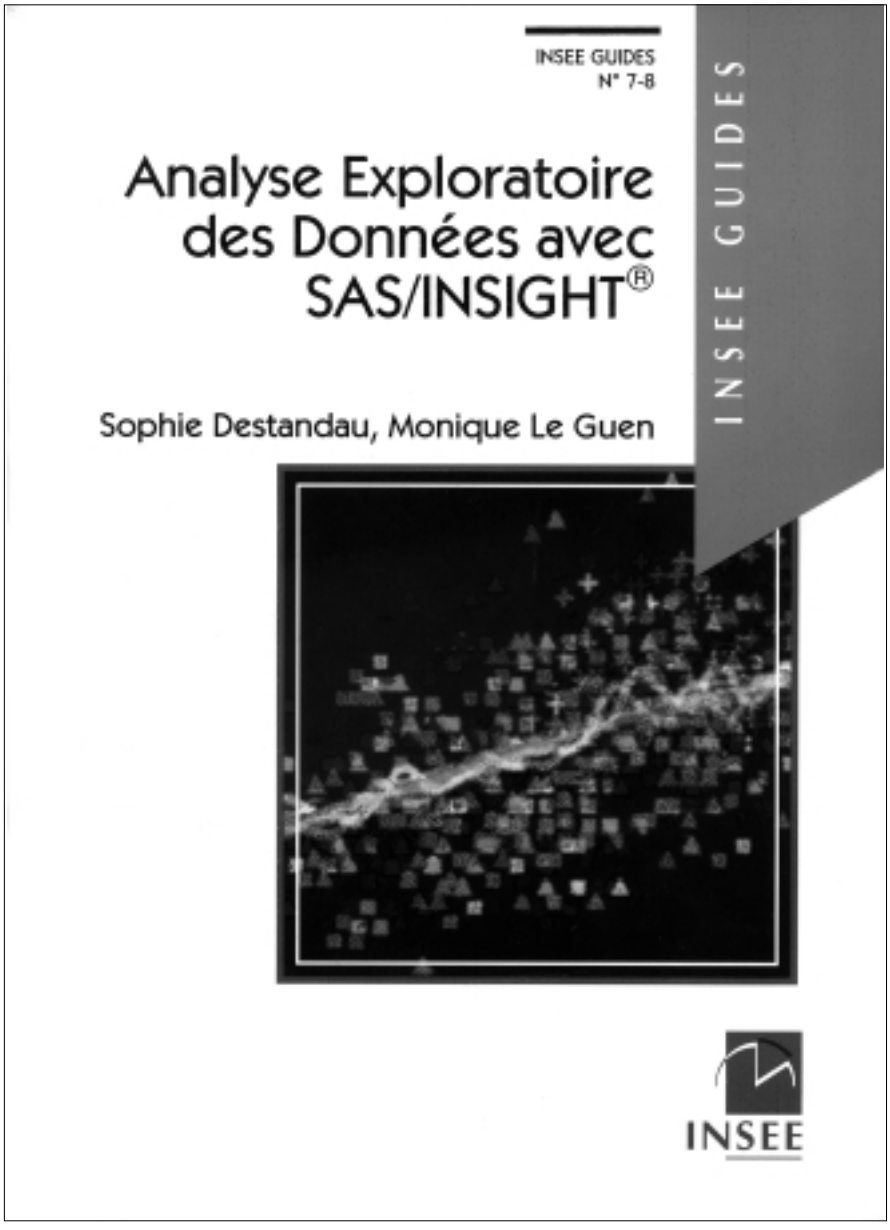
- il offre des possibilités étendues en matière de représentation graphique (histogramme, Bar Chart, Box Plot, Mosaic Plot, Scatter Plot, Line Plot, Rotating Plot et QQplot) et de réexpression des variables (plus de trente fonctions proposées, mathématiques et statistiques) ;

- il sait calculer des statistiques résistantes, par exemple des moyennes tronquées ou winsorisées ;

- il inclut des fonctions de mise en oeuvre de procédures d'ajustement paramétriques (ajustement de la distribution de la variable étudiée selon la loi normale, la loi log-normale, la loi exponentielle ou encore la loi de Weibull), mais également non paramétriques (estimations de probabilité selon la méthode du noyau) ;

- il intègre un menu de modélisation **Fit** très complet.

Bien sûr, SAS/Insight a aussi des défauts. Ainsi, ses fonctions de représentation graphique excluent la réalisation de graphiques de type Stem and Leaf ou encore PPplot ; plus généralement, elles ne sont pas réellement adaptées à des sorties « diffusion », car ne permettant pas de légender les graphiques. Par ailleurs, les possibilités offertes par ce logiciel en matière de transformation des variables pourront paraître insuffisantes à certains. Enfin, on pourra parfois lui reprocher un certain manque de souplesse : par exemple, deux manipulations sont nécessaires pour afficher le nom en clair d'une variable dans un tableau ou dans un graphique.



Statut statistique des variables

Les différentes options proposées par SAS/Insight en matière de représentation graphique et d'analyse de données sont fonction du statut statistique des variables étudiées. Deux types de variables y sont en effet distingués, d'une part les variables d'intervalle, d'autre part les variables nominales. Une variable à valeurs exclusivement numériques peut être définie comme d'intervalle ou nominale, au choix de l'explorateur. En revanche, une variable à valeurs non exclusivement numériques sera nécessairement considérée comme nominale.

... d'utilisation très facile

À ses nouveaux utilisateurs pratiquant déjà SAS-micro, SAS/Insight réservera une première bonne surprise. La table SAS appelée s'affiche en effet sous forme d'un tableau à deux dimensions, en ligne les observations, en colonne les variables, tout comme si une Viewtable avait été demandée. Ce tableur est bien sûr assorti de fonctions de gestion, certes beaucoup moins impressionnantes que celles que peuvent proposer Excel ou Quattro, mais néanmoins appréciables : création ou suppression de variables ou d'observations, extraction de sous-tables par sélection des observations selon les valeurs des variables, tri des observations selon les valeurs d'une ou plusieurs variables, dans l'ordre ascendant ou dans l'ordre inverse, reformatage des variables...

Deuxième très bonne surprise, l'apprenti explorateur de données va pouvoir immédiatement constater que l'utilisation de SAS/Insight ne nécessite aucun langage particulier : il lui suffira en effet de se laisser guider par les menus, et de sélectionner au moyen de la souris l'option qu'il aura retenue parmi celles qui lui sont proposées.

Le reste ne sera plus qu'affaire de pratique... Des exemples concrets de ce que l'on peut faire avec SAS/Insight, mais il ne s'agit que d'un premier aperçu, sont présentés dans les deux articles ci-après.

Précision importante, SAS/Insight inclut une aide en ligne en mode hypertexte, incorporant en particulier un véritable dictionnaire statistique. Si par exemple on veut se remémorer la signification de tel ou tel indicateur produit en résultat d'une analyse statistique (cf. l'article « Analyser une distribution avec SAS/Insight »), il suffira de « cliquer » sur l'intitulé de l'indicateur en question pour voir s'afficher sa définition à l'écran.

⇒ Si vous avez SAS-micro, essayez SAS/Insight.

⇒ Si vous butez sur une difficulté, reportez-vous au n° 7-8 d'Insee Guides « Analyse Exploratoire des Données avec SAS/Insight ».

⇒ Si vous voulez aller plus vite, suivez un stage SAS/Insight.

⇒ Si vous voulez en savoir plus sur l'AED, allez visiter le site Web http://www.unige.ch/ses/sococ/mirage :
vous y trouverez, entre autres, des informations sur l'Association MIRAGE, qui organise chaque deuxième semaine de septembre, à Carcassonne, une École d'été sur l'analyse exploratoire des données (la prochaine se tiendra du 11 au 16 septembre 2000).

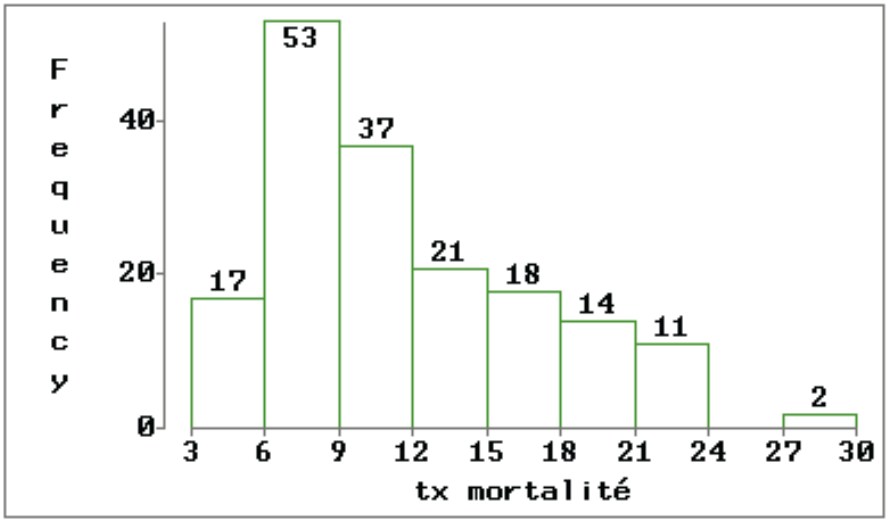
Sophie DESTANDAU

Visualisation interactive et réexpression des données avec SAS/Insight

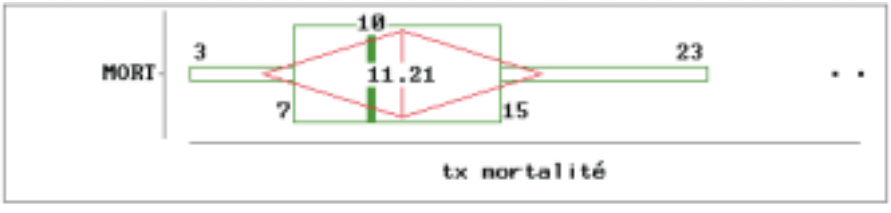
À l'exception des *Line Plots*, la totalité des graphiques présentés dans cet article ont été réalisés à partir d'une même table SAS incluant 173 observations, en l'occurrence 173 pays, et 25 variables, en particulier les taux de natalité (NAT) et de mortalité (MORT) pour 1 000 habitants, le taux d'accroissement naturel annuel de la population (ACCR) en %, le taux de fécondité ou nombre d'enfants par femme (FERTI), la part (en % de la population totale) des moins de 15 ans (AGE15) et des plus de 65 ans (AGE65), le produit national brut (PNB) et le taux d'urbanisation (URBA) en % de la population totale.

La première partie du menu *Analyze* de SAS/Insight propose sept types de graphiques différents. La façon dont se distribue une variable sera représentée sous forme d'un histogramme ou d'un Box Plot si la variable en question est une variable d'intervalle, par un Bar Chart ou un Mosaic Plot s'il s'agit d'une variable nominale. Le Line Plot et le Scatter Plot (diagramme de dispersion, ou nuage de points) permettent de visualiser la façon dont se répartissent les observations suivant deux variables distinctes, le Line Plot étant plus particulièrement adapté à la représentation d'une série temporelle. Enfin, la façon dont se répartissent les observations suivant trois variables distinctes peut être visualisée au moyen d'un Rotating Plot (diagramme de rotation).

Histogramme



Box Plot

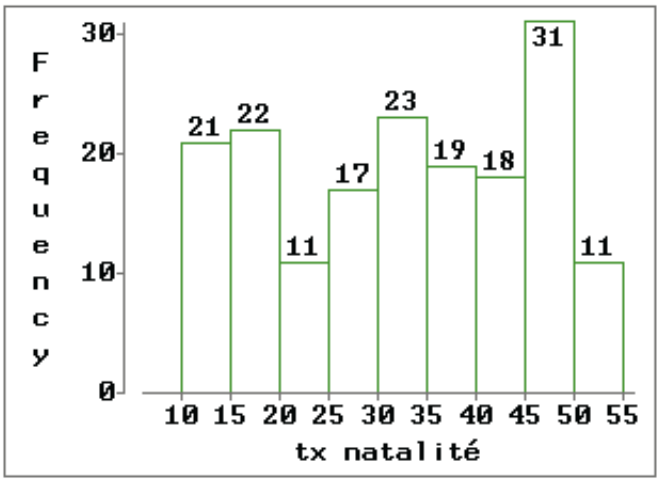


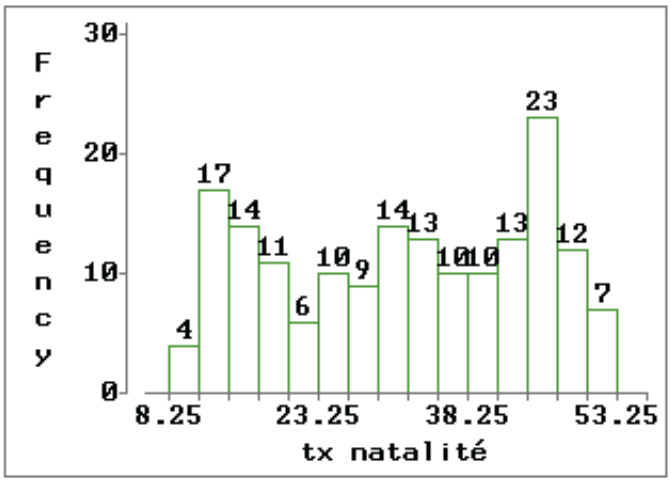
Histogramme et Box Plot

L'histogramme révèle la forme, ou plutôt une forme, de la distribution étudiée, ici celle du taux de mortalité.

Le Box Plot apporte de nombreuses informations supplémentaires :
- la longueur de la boîte détermine l'étendue de la partie centrale de la distribution (taux de mortalité compris entre les quartiles Q1 et Q3, valeurs 7 et 15) ;
- la bande verticale matérialisée à l'intérieur de la boîte indique la position de la médiane (valeur 10) ;
- l'étendue des queues de distribution hors points atypiques est déterminée par la longueur des moustaches, étant précisé que l'extrémité de la moustache de gauche est ici[1] fixée par la plus petite valeur (= 3) supérieure ou égale à Q1 - 1,5 (Q3 - Q1), celle de la moustache de droite par la plus grande valeur (= 23) inférieure ou égale à Q3 + 1,5 (Q3 - Q1) ;
- de part et d'autre des moustaches sont mises en évidence les observations atypiques, avec valeur inférieure à Q1 - 1,5 (Q3 -Q1) ou supérieure à Q3 + 1,5 (Q3 - Q1), ici deux pays avec taux de mortalité > 23 ;
- la petite diagonale du losange superposé à la boîte indique la position de la moyenne (valeur 11,21) ; sa grande diagonale, de longueur 2σ, permet d'apprécier la valeur de l'écart-type.

1. Le coefficient appliqué à (Q3 - Q1) est paramétrable.





Redécoupage d'un histogramme

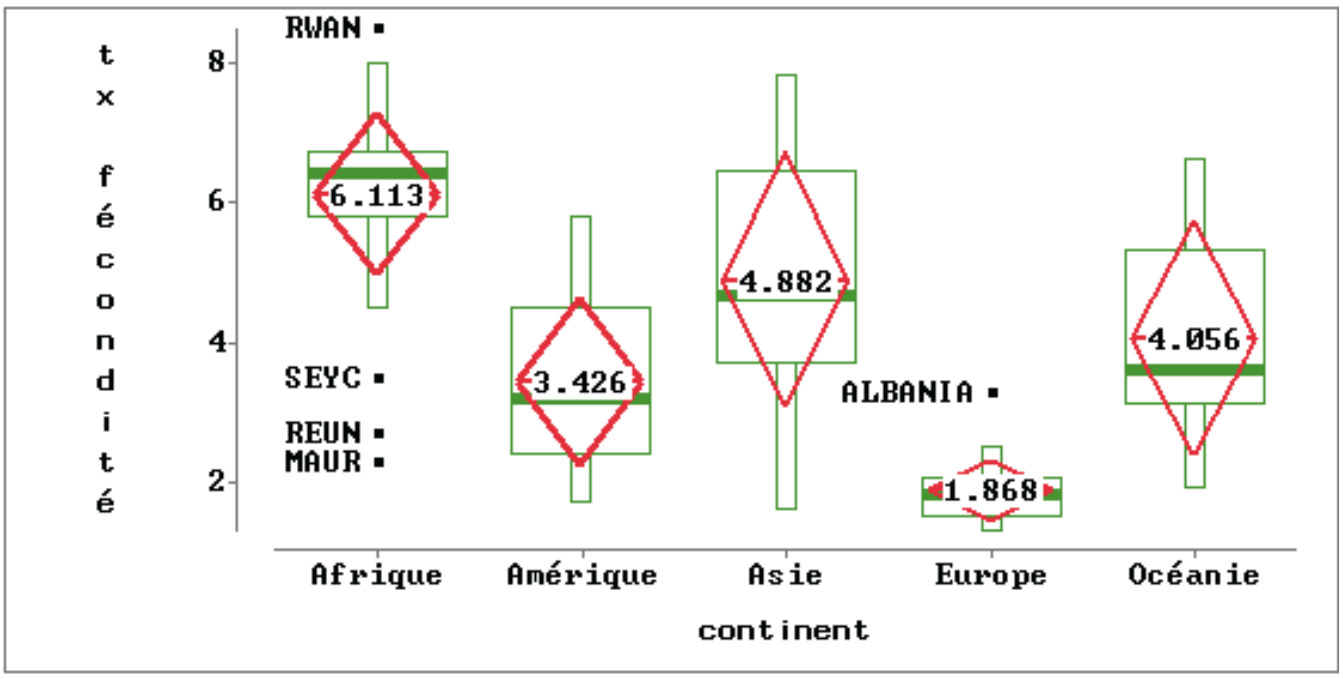
Avec SAS/Insight, il est très simple et très rapide, grâce à la souris, de modifier le découpage en classes de la variable étudiée. L'effet sur l'allure de l'histogramme peut être surprenant, comme le montre l'exemple ci-dessus relatif à la distribution du taux de natalité (pour 1 000 habitants).

Juxtaposition de Box Plots

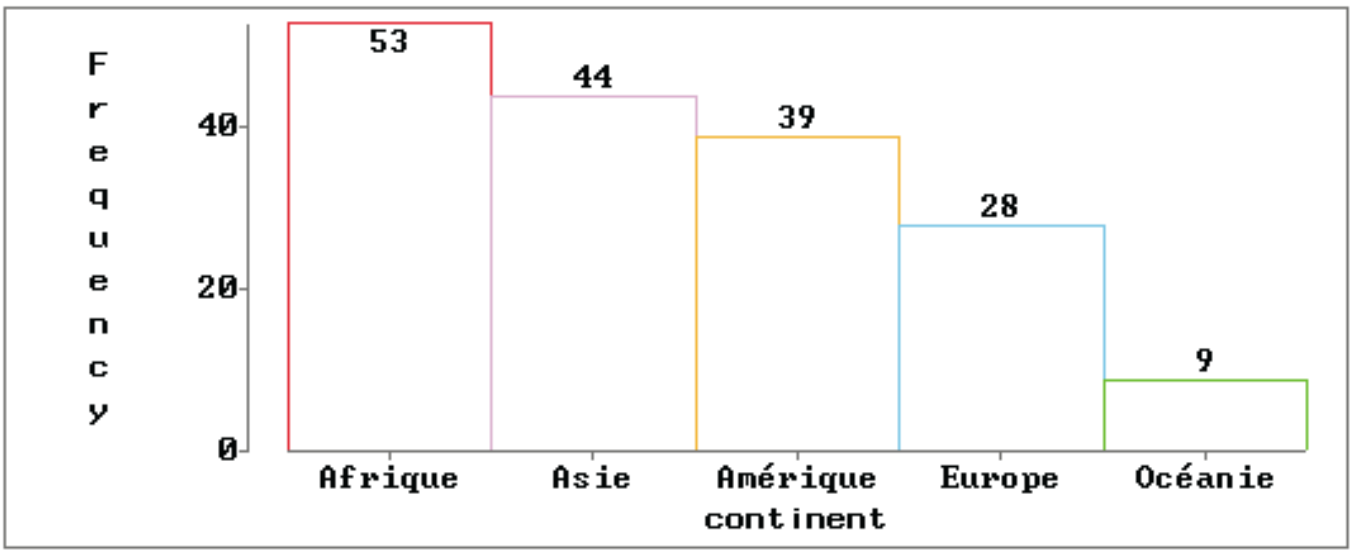
Les 173 pays observés ont été répartis en 5 groupes, selon le continent.
Les 5 Box Plots juxtaposés ci-dessous, ici en mode vertical, représentent la distribution du taux de fécondité (nombre d'enfants par femme) dans les différents continents.

Ce type de présentation est particulièrement efficace :

- il permet d'apprécier en un coup d'oeil la façon dont se distribue une variable d'intervalle en fonction des modalités d'une variable nominale ;
- en outre, il constitue une excellente introduction visuelle à l'analyse de la variance.



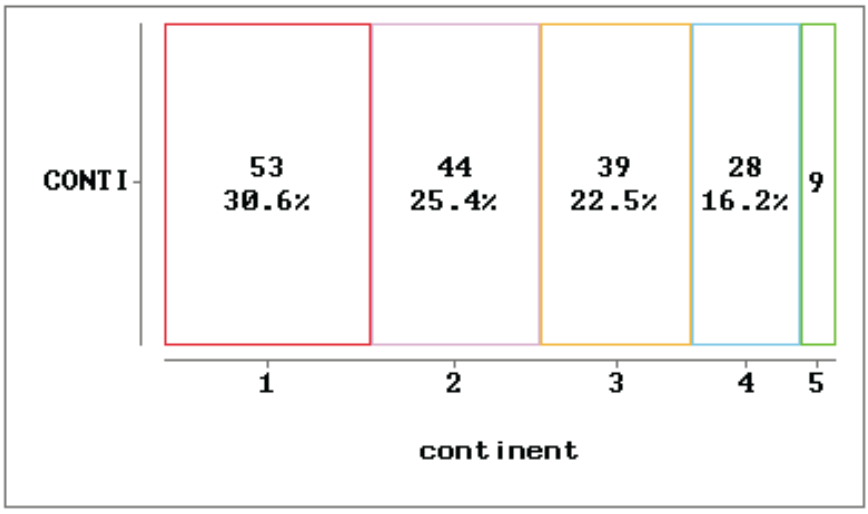
Bar Chart



Bar Chart et Mosaic Plot

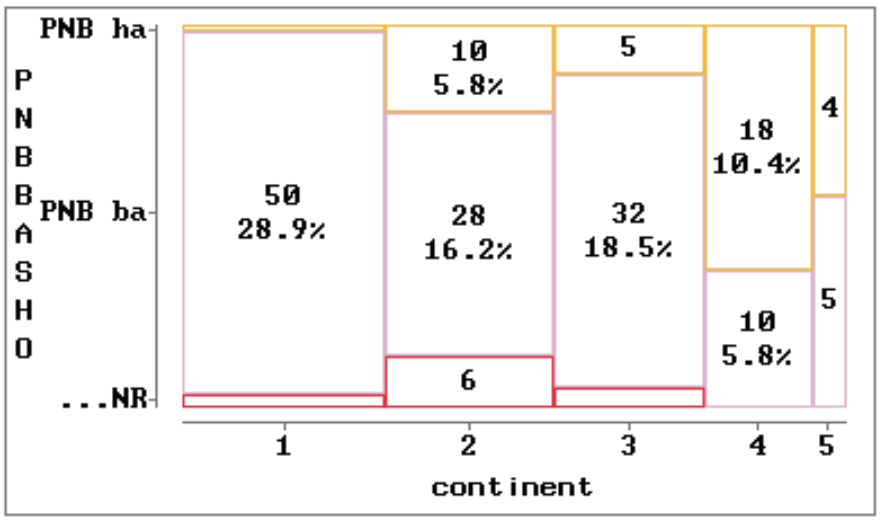
Mosaic Plot

La façon dont se répartissent les observations selon les valeurs d'une variable nominale, ici la répartition des pays suivant le continent, peut être représentée au moyen d'un Bar Chart ou d'un Mosaic Plot.

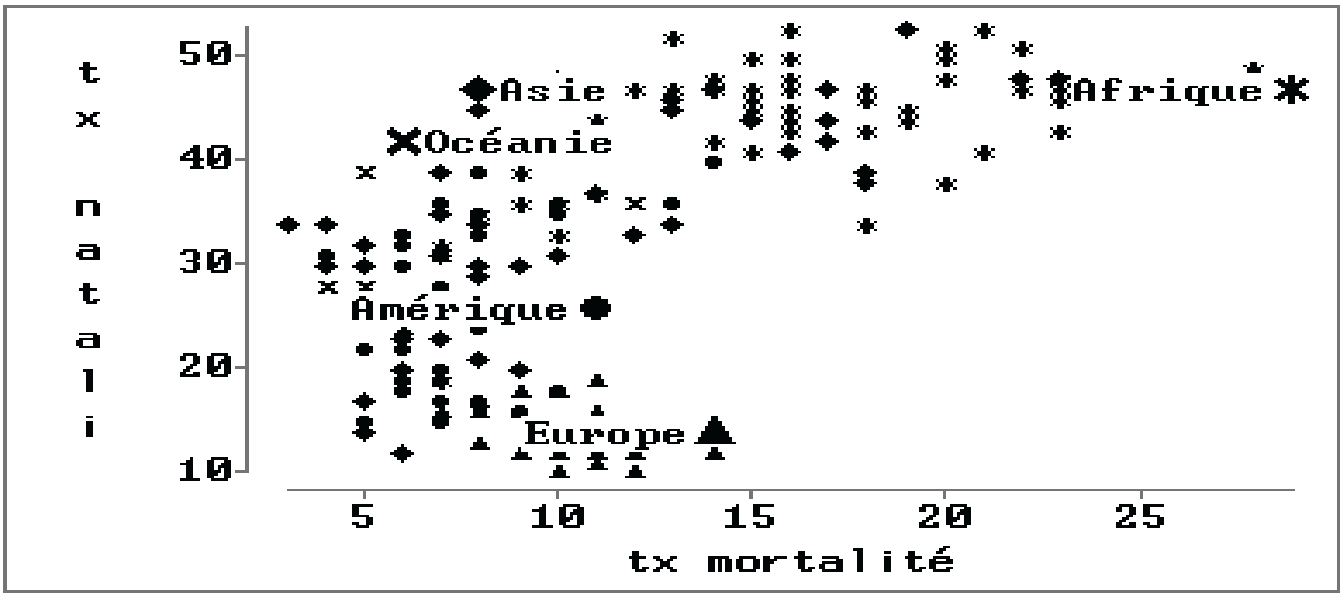


Une autre utilisation possible du Mosaic Plot, équivalent visuel de la procédure *Proc Freq* de SAS[2], est la représentation des croisements entre deux variables nominales.

Ainsi, le deuxième graphique ci-contre permet de visualiser la façon dont se répartissent les 173 pays observés selon le continent et le PNB (grossièrement découpé en 3 modalités : haut, bas, non-réponse).



2. À ceci près toutefois que la *Proc Freq* calcule également les pourcentages en ligne et en colonne, en sus des pourcentages par rapport à l'effectif total.

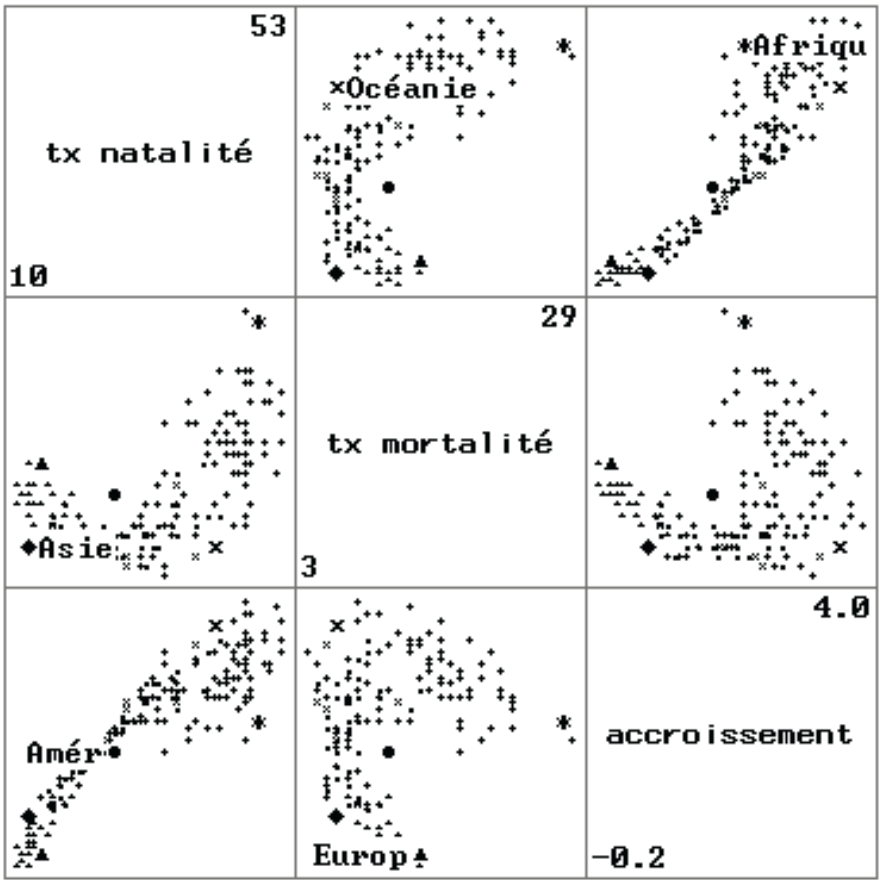


Le Scatter Plot

Le Scatter Plot, ou diagramme de dispersion, permet tout à la fois d'apprécier le type de liaison (linéaire ou autre) pouvant exister entre deux variables, de diagnostiquer une éventuelle hétéroscédasticité, de repérer les groupes (clusters) et les observations atypiques (outliers).

Le Scatter Plot ci-dessus visualise la façon dont se répartissent les 173 pays observés suivant le taux de natalité et le taux de mortalité. Y a été ajoutée une troisième dimension, via l'utilisation de cinq marqueurs différents repérant le continent d'appartenance des pays (on peut ainsi remarquer une forte concentration de pays africains dans la partie supérieure droite du diagramme). On aurait tout aussi bien pu en en ajouter une quatrième au moyen de la couleur.

Les liaisons 2 à 2 entre plusieurs variables d'intervalle, ici les taux de natalité et de mortalité (pour 1 000 habitants) et le taux d'accroissement naturel de la population (en %), peuvent être visualisées au moyen d'une matrice.



Guide de lecture de gauche à droite et de haut en bas

Cadre 1 : Les valeurs extrêmes du taux de natalité sont 10 et 53.

Cadre 2 : Répartition des pays suivant le taux de natalité (en ordonnées) et le taux de mortalité (en abscisses).

Cadre 3 : Répartiton des pays suivant le taux de natalité (en ordonnées) et le taux d'accroissement (en abscisses).

Cadre 4 (symétrique du cadre 2) : Répartition des pays suivant le taux de mortalité (en ordonnées) et le taux de natalité (en abscisses).

Cadre 5 : Les valeurs extrêmes du taux de mortalité sont 3 et 29.

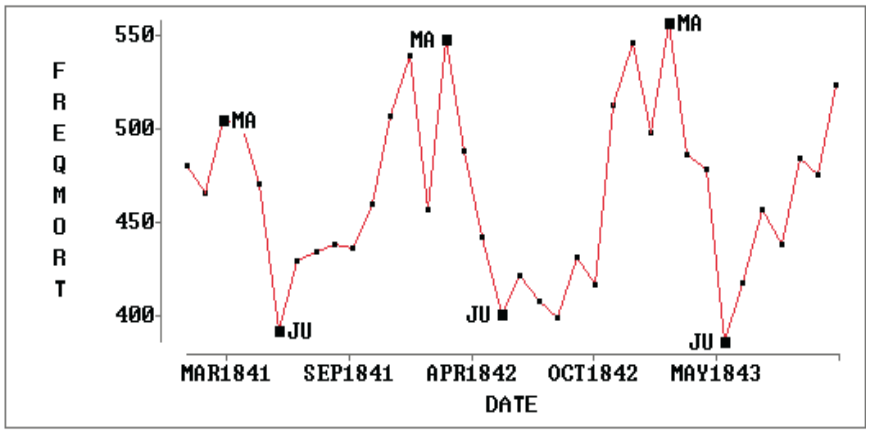
Cadre 6 : Répartition des pays suivant le taux de mortalité (en ordonnées) et le taux d'accroissement (en abscisses).

Cadre 7 (symétrique du cadre 3) : Répartition des pays suivant le taux d'accroissement (en ordonnées) et le taux de natalité (en abscisses).

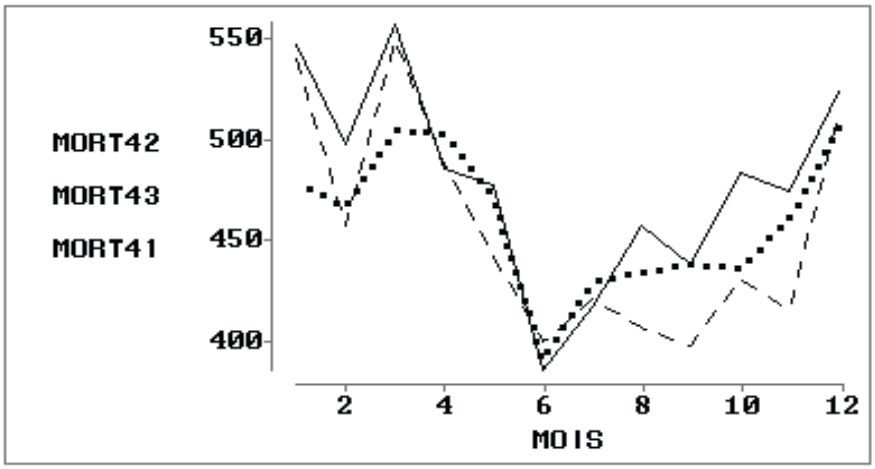
Cadre 8 (symétrique du cadre 6) : Répartition des pays suivant le taux d'accroissement (en ordonnées) et le taux de mortalité (en abscisses).

Cadre 9 : Les valeurs extrêmes du taux d'accroissement sont -0,2 et 4,0.

Nombre d'enfants mort-nés en Belgique
de janvier 1841 à décembre 1843



Nombre mensuel d'enfants mort-nés en Belgique,
années 1841 à 1843



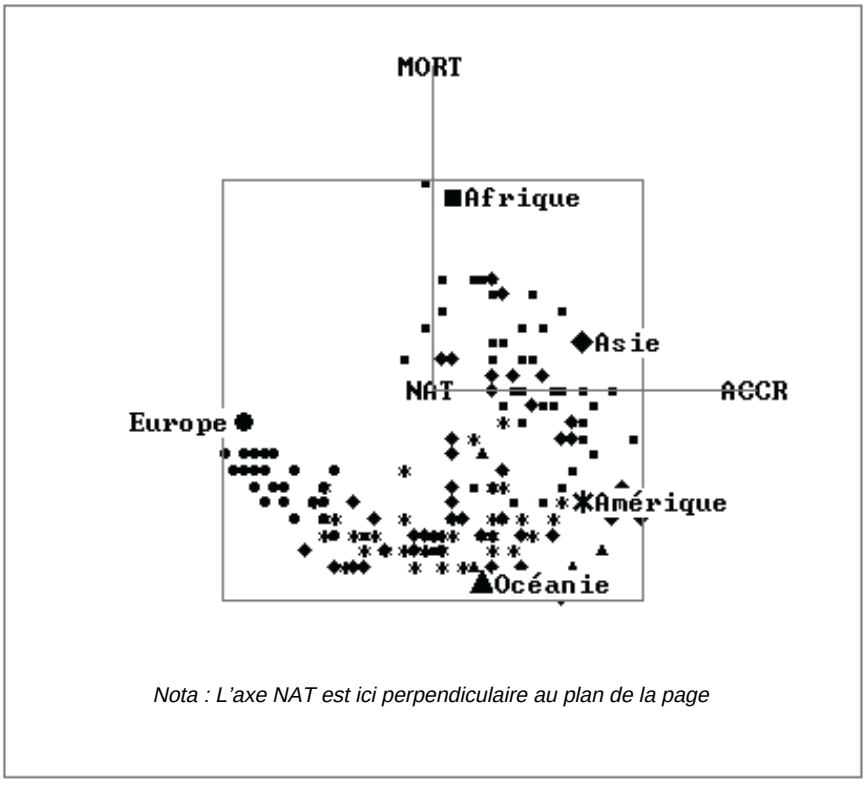
Le Line Plot

Le Line Plot, ou tracé de données reliées par des lignes, est adapté à la représentation des séries temporelles. L'utilisation de la couleur ou de tracés différents facilitera la visualisation en cas de prise en compte d'une troisième dimension déterminée par une variable nominale (cas du deuxième graphique).

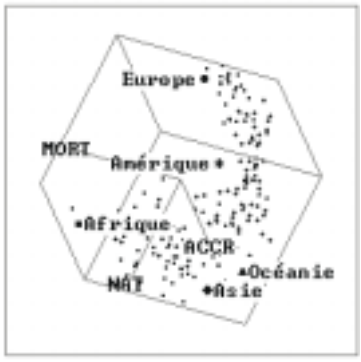
Le Rotating Plot

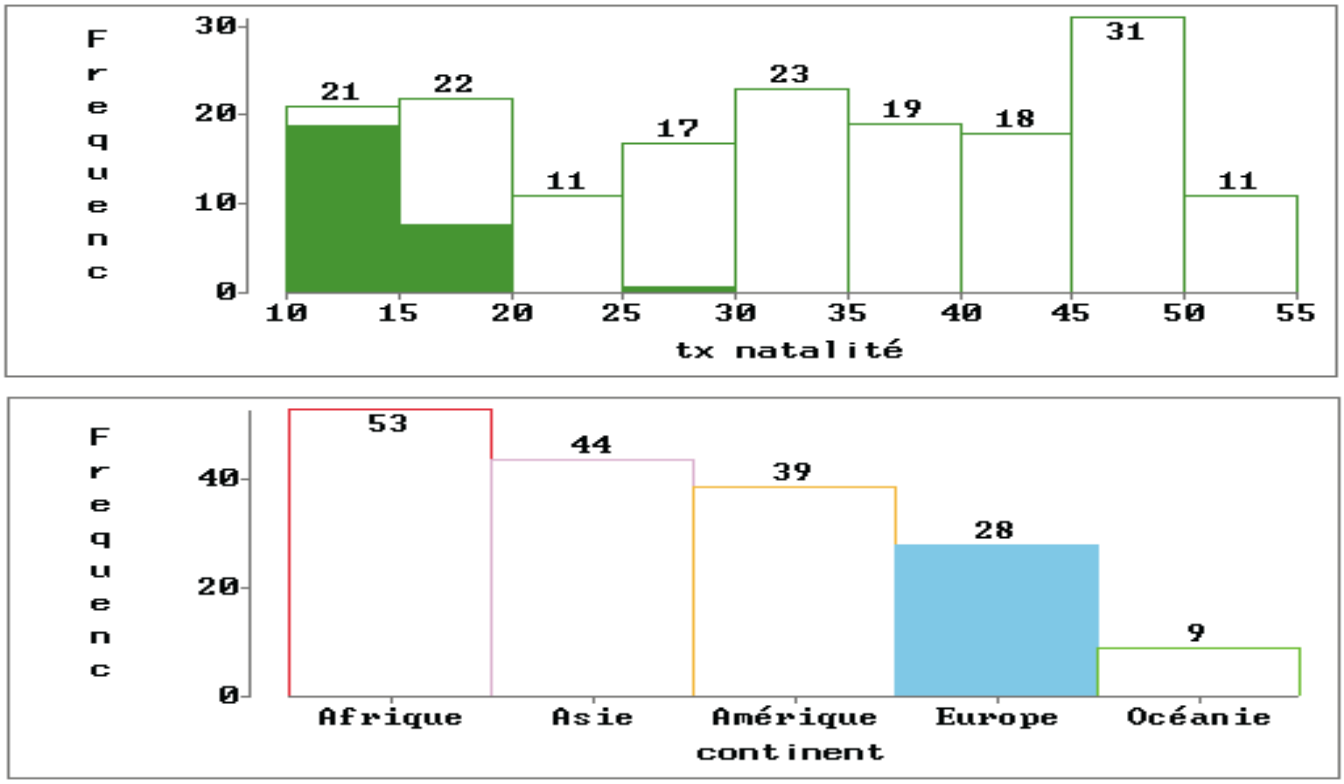
Le Rotating Plot, ou graphique dynamique de rotation en 3 D, permet de repérer d'éventuelles structures qui ne sont ni visibles sur des graphiques statiques, ni détectables par des méthodes analytiques.

De ce point de vue, l'exemple ci-dessous pourrait paraître assez mal choisi, puisque l'on s'y intéresse à la répartition de nos 173 pays suivant le taux de natalité, le taux de mortalité et le taux d'accroissement naturel de la population. En faisant tourner le nuage de points, on met en évidence la relation liant ces trois variables, ici ACCR = (NAT - MORT) /10 puisque les taux de natalité et de mortalité sont exprimés en ‰ et le taux d'accroissement en %.



Nota : L'axe NAT est ici perpendiculaire au plan de la page





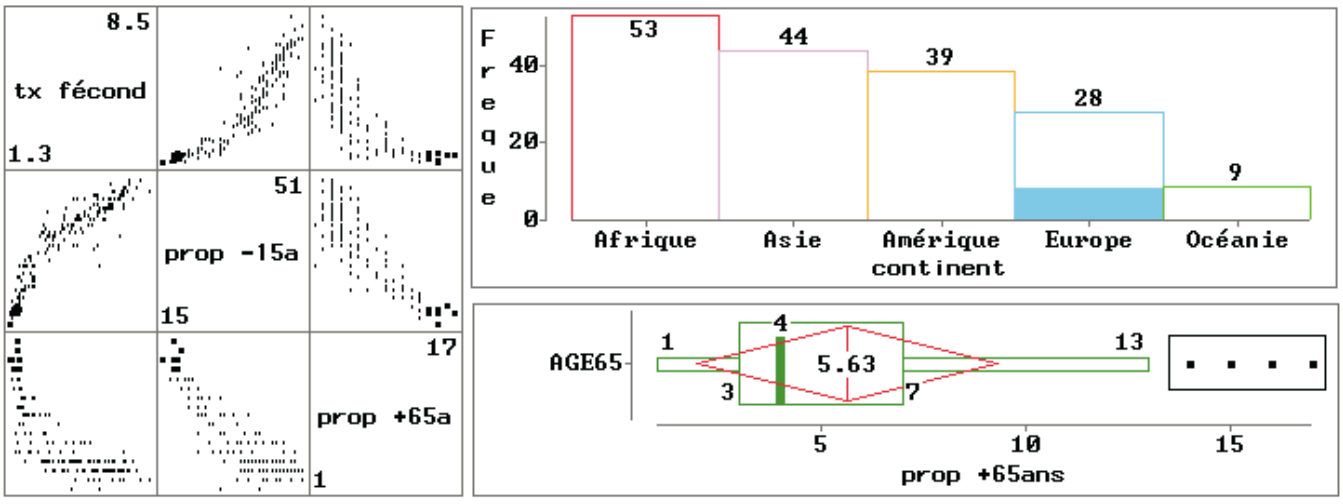
Interactivité

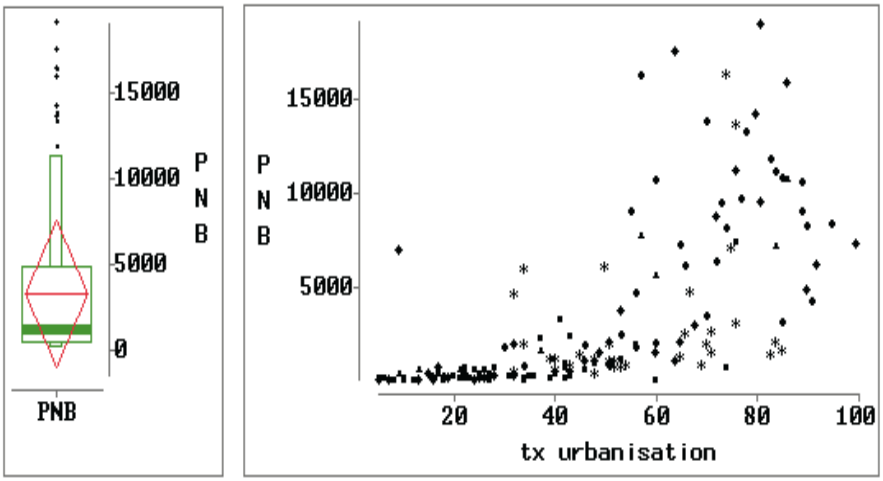
Il est tout à fait possible de regrouper plusieurs représentations graphiques dans une même fenêtre et d'animer l'ensemble grâce à l'interactivité.

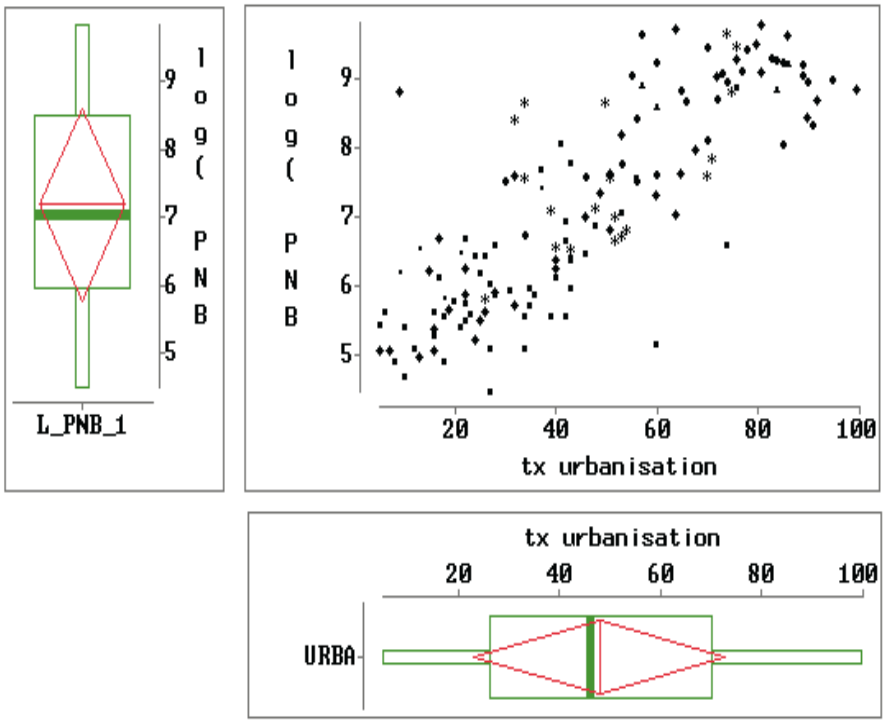
Regroupons par exemple dans une même fenêtre un histogramme relatif à la distribution du taux de natalité, et un Bar Chart donnant la répartition par continent des 173 pays observés. Si l'on sélectionne la barre Europe du Bar Chart, on verra alors instantanément se superposer à l'histogramme global un deuxième histogramme limité au cas de ce continent.

Plus spectaculaire encore, regroupons dans une même fenêtre ce même Bar Chart, le Box Plot relatif à la distribution de la proportion des plus de 65 ans ainsi qu'une matrice de diagrammes de dispersion mettant en oeuvre la proportion des moins de 15 ans, celle des plus de 65 ans et le taux de fécondité (nombre d'enfants par femme).

Si l'on sélectionne (en les encadrant) les quatre observations atypiques du Box Plot, on voit immédiatement se surimprimer la position des pays en question sur le Bar Chart et les diagrammes de dispersion (effet loupe) : ils sont situés en Europe, le taux de fécondité et la proportion des moins de 15 ans y sont particulièrement faibles, la proportion des plus de 65 ans y est particulièrement élevée.







Réexpression des données

On l'a vu dans l'article « AED mode d'emploi », l'un des fondamentaux de l'analyse exploratoire des données est la transformation ou réexpression des variables, l'opération pouvant en particulier permettre de symétriser une distribution ou encore de linéariser une liaison.

Avec SAS/Insight, il est possible de directement transformer une variable à partir d'une représentation graphique, sans qu'il soit nécessaire de revenir au tableur.

Soit par exemple la première fenêtre ci-contre, dans laquelle nous avons regroupé deux Box Plots, le premier relatif à la distribution du PNB, le second à la distribution du taux d'urbanisation, et le Scatter Plot présentant la façon dont se répartissent les 173 pays observés suivant ces deux mêmes variables.

Sélectionnons le libellé de variable « PNB » sur le premier Box Plot ou le Scatter Plot, et transformons cette variable en son logarithme (logPNB) par l'intermédiaire du menu *Edit # Variables*. On obtient immédiatement la représentation transformée résultante, sur laquelle on va pouvoir constater, d'une part que la distribution de logPNB est quasi symétrique, d'autre part qu'il semble y avoir une liaison linéaire entre cette variable et le taux d'urbanisation.

Sophie DESTANDAU
et **Monique LE GUEN**

Les sous-menus de Analyse # Distribution

Tables, Graphs, Curves

Tables

Moments Quantiles Frequency Table C.I for Mean >>> Location Tests...
Gini's Mean Difference
Trimmed Mean, (1/2)N >>> Trimmed Mean, (1/2)Percent >>>
Winsorized Mean, (1/2)N >>> Winsorized Mean, (1/2)Percent >>>

Graphs

Histogram/Bar Chart Box Plot/Mosaic Plot QQ Plot

Curves

Parametric Density... Kernel Density...
Empirical CDF CDF Confidence Band >>>
Parametric CDF... Test for a Specific Distribution... Test for Distribution ...
QQ Ref Line...

Analyser une distribution avec SAS/Insight

Le menu *Analyze # Distribution* de SAS/Insight est organisé en trois sous-menus, *Graphs*, *Tables* et *Curv*es. Dans le sous-menu *Graphs,* on retrouve quatre des sept types de graphiques déjà passés en revue dans l'article précédent (histogramme, Bar Chart, Box Plot et Mosaic Plot), sur lesquels nous ne reviendrons donc pas. Mais bien sûr, nous nous attarderons sur le QQplot.

Autre précision liminaire, il ne sera ici question que de variables d'intervalle. S'agissant d'une variable nominale, le menu *Analyze # Distribution* ne propose en effet rien de plus que les représentations graphiques de type Bar Chart ou Mosaic Plot, à l'exception toutefois d'un tableau de comptage des observations selon les valeurs de la variable en question avec indication des pourcentages afférents, autrement dit une simple *frequency table*.

La totalité des exemples donnés dans cet article ont été élaborés à partir d'une même table SAS incluant 173 observations, en l'occurrence 173 pays, et 25 variables, en particulier le taux de natalité pour 1 000 habitants (NAT), la population totale en millions d'habitants (POP87) et le taux de fécondité ou nombre d'enfants par femme (FERTI).

Indicateurs statistiques

En la matière, le premier volet du sous-menu *Tables* inclut l'ensemble des options proposées par la *Proc Univariate* de SAS, soit la production des tables statistiques relatives aux moments et quantiles et la mise en oeuvre de trois tests de position spécifiques, les deux derniers non paramétriques : test T de Student pour tester l'hypothèse « moyenne = Mu », test des rangs signés de Wilcoxon et test du signe pour tester l'hypothèse « médiane = Mu ».

Il incorpore également une option de détermination de l'intervalle de confiance de la moyenne

Un exemple concret est donné ci-dessous, relatif à la distribution de la variable FERTI (taux de fécondité ou nombre d'enfants par femme).

FERTI

Moments			
N	173.0000	Sum Wgts	173.0000
Mean	4.4000	Sum	761.2000
Std Dev	1.9846	Variance	3.9386
Skewness	0.0095	Kurtosis	-1.3668
USS	4026.7200	CSS	677.4400
CV	45.1044	Std Mean	0.1509

Quantiles			
100% Max	8.5000	99.0%	8.0000
75% Q3	6.2000	97.5%	7.4000
50% Med	4.5000	95.0%	7.2000
25% Q1	2.4000	90.0%	7.0000
0% Min	1.3000	10.0%	1.8000
Range	7.2000	5.0%	1.6000
Q3-Q1	3.8000	2.5%	1.5000
Mode	1.8000	1.0%	1.4000

Student's T Test for Mean		
Parameter (Mu)	Statistic	P-Value
4.0000	2.6510	0.0088

L'hypothèse ici testée (et rejetée) est moyenne = 4.

Signed Rank Test for Location				
Parameter (Mu)	N ^= Mu	N > Mu	Statistic	P-Value
4.0000	169	93	1733.0000	0.0062

L'hypothèse ici testée (et rejetée) est médiane = 4.

Sign Test for Location				
Parameter (Mu)	N ^= Mu	N > Mu	Statistic	P-Value
5.0000	172	73	-13.0000	0.0563

L'hypothèse ici testée (et non rejetée) est médiane = 5.

Confidence Interval for Mean			
Mean	Level (%)	Lower Limit	Upper Limit
4.4000	95.0000	4.1022	4.6978

Intervalle de confiance (ici à 95 %) de la moyenne.
Dans le cas précis, il n'y a bien sûr pas grand sens à déterminer un intervalle de confiance de la moyenne, non plus qu'à tester des hypothèses de type moyenne ou médiane = Mu, les indicateurs en question ayant été calculés à partir d'une information censément exhaustive.

Les trois autres volets du sous-menu *Tables* recouvrent la détermination d'indicateurs de dispersion ou de tendance centrale plus résistants que l'écart-type ou que la moyenne.

FERTI

Gini's Mean Difference	
G Statistic	Normal Std Dev
2.2876	2.0274

L'indicateur de dispersion proposé est calculé à partir de l'écart à la moyenne de Gini (2,2876), que l'on multiplie par 1/2 de racine de π. Il est ici égal à 2,0274, à rapprocher de l'écart-type 1,9846 (cf. table des moments).

Les indicateurs de tendance centrale proposés sont des moyennes tronquées ou winsorisées.

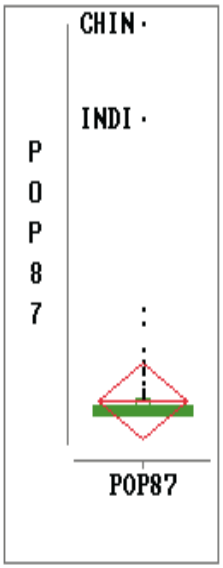
L'éloquent exemple ci-dessous porte cette fois sur la distribution de la variable POP87 (population en millions d'habitants).

Toutes observations confondues, la population moyenne par pays est de 29,05 millions d'habitants. Mais si l'on ne tient pas compte des deux pays les plus grands (Chine et Inde, cf. le Box Plot) ni des deux plus petits, la population moyenne par pays chute alors à 18,71 millions d'habitants, la moyenne winsorisée s'établissant quant à elle à 21,57 millions d'habitants.

POP87

Moments			
N	173.0000	Sum Wgts	173.0000
Mean	29.0477	Sum	5025.2500
Std Dev	105.8003	Variance	11193.7042
Skewness	7.9090	Kurtosis	68.9952
USS	2071289.02	CSS	1925317.13
CV	364.2297	Std Mean	8.0438

Quantiles			
100% Max	1062.00	99.0%	800.30
75% Q3	17.00	97.5%	174.90
50% Med	6.40	95.0%	107.10
25% Q1	1.20	90.0%	53.60
0% Min	0.05	10.0%	0.20
Range	1061.95	5.0%	0.20
Q3-Q1	15.80	2.5%	0.10
Mode	0.20	1.0%	0.10



Trimmed Mean						
(1/2)Percent	(1/2)N	Mean	Std Mean	DF	T: Mean=0	Prob >|T|
1.16	2	18.7148	3.6767	168	5.0901	0.0001

Les moyennes tronquées (Trimmed Means)
sont des moyennes arithmétiques
calculées après élimination des valeurs extrêmes
(la plus petite et la plus grande, les 2 plus petites et les 2 plus grandes, etc.).

Winsorized Mean						
(1/2)Percent	(1/2)N	Mean	Std Mean	DF	T: Mean=0	Prob >|T|
1.16	2	21.5665	3.6770	168	5.8653	0.0001

Les moyennes winsorisées (Winsorized Means)
ont été introduites par Winsor dans les années 1940.
Au lieu de supprimer les valeurs extrêmes,
Winsor les remplace par la valeur la plus proche non suspecte.

Ajustement à une loi paramétrique théorique

Le sous-menu *Curves* inclut des outils de contrôle spécifiques, au moyen desquels on va « regarder » si la distribution observée pourrait s'ajuster à une loi paramétrique théorique. Quatre types d'ajustement - aux lois normale, log-normale ou exponentielle, ou encore à la loi de Weibull - sont proposés dans ce sous-menu.

L'option *Parametric Density* permet de superposer dans une même fenêtre graphique la courbe, en forme d'histogramme, de la fonction de densité de probabilité de la distribution observée, et la courbe de la fonction de densité de probabilité de la distribution théorique, avec détermination automatique des classes de l'histogramme et des paramètres de la loi théorique retenue.

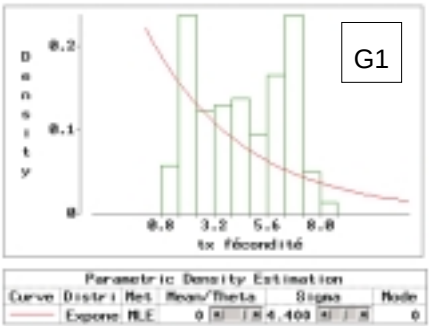
Dans l'exemple ci-contre (G1), la variable étudiée est le taux de fécondité ou nombre d'enfants par femme (variable FERTI), et la loi testée est la loi exponentielle. Bien sûr, l'inadéquation apparaît clairement.

On aurait tout aussi bien pu faire apparaître sur le même graphique les courbes relatives aux trois autres lois (G2).

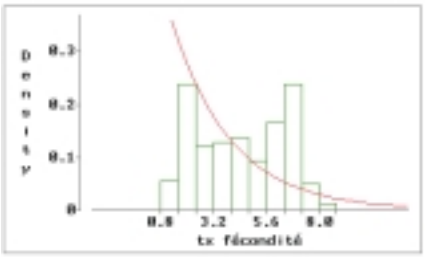
L'option *Parametric CDF* permet quant à elle de superposer dans une même fenêtre graphique la courbe cumulée empirique et la courbe cumulée théorique, avec détermination automatique des paramètres de la loi théorique retenue.

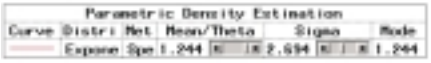
Dans l'exemple ci-contre (G3), la variable étudiée est toujours FERTI, et la loi testée est celle de Weibull. La courbe cumulée théorique apparaît en pointillés, la courbe cumulée empirique, beaucoup plus saccadée, en trait plein.

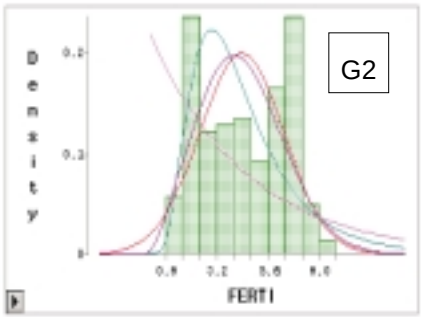
En jouant sur les paramètres de la loi, on pourra essayer d'obtenir un ajustement plus satisfaisant concernant telle ou telle partie de la distribution, ici son milieu.

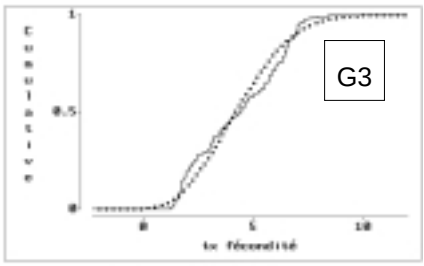


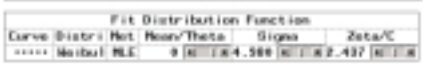
On peut faire varier les paramètres de la loi testée (ici Theta et Sigma) grâce aux curseurs prévus à cet effet et observer tout en même temps les conséquences de ces modifications sur la représentation graphique :

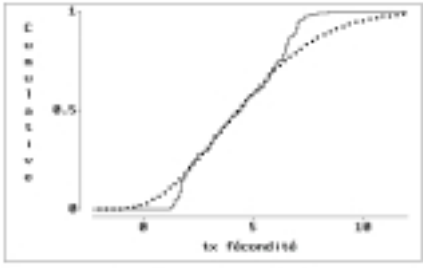


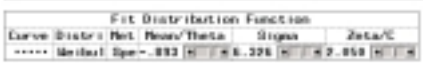












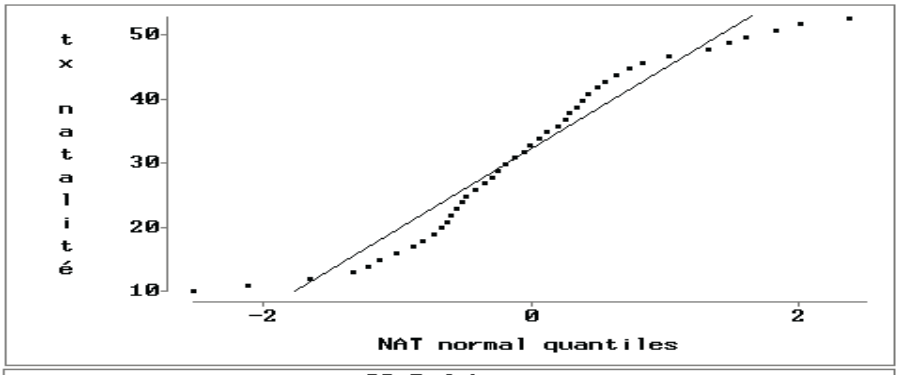
QQplot

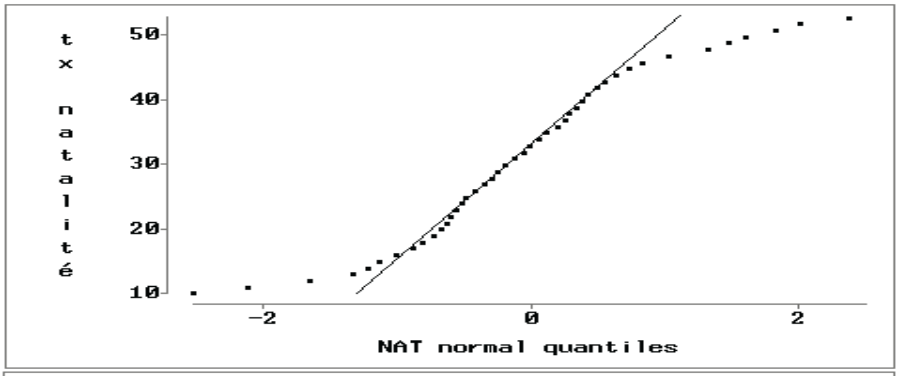
Le QQplot ou graphique quantile-quantile (cf. sous-menu *Graphs*) fait partie intégrante des outils que propose SAS/Insight s'agissant d'établir la plausibilité de l'ajustement d'une distribution empirique à une loi paramétrique théorique[1].

Les points positionnés sur le QQplot sont déterminés par, en ordonnées les quantiles observés (valeurs observées de la variable étudiée), en abscisses les quantiles théoriques (valeurs théoriques de la variable étudiée[2]). La plausibilité de l'ajustement est ainsi fonction de la linéarité de la représentation obtenue. Il est possible de superposer au QQplot une droite de référence (cf. sous-menu *Curves*), qui permettra d'apprécier plus facilement cette linéarité.

Dans l'exemple ci-dessous, la variable étudiée est le taux de natalité pour 1 000 habitants (variable NAT), le nombre d'observations (nombre de pays observés) est égal à 173, la loi testée est la loi normale. L'ordonnée à l'origine (intercept = 32,02) et la pente (slope = 12,60) de la droite de référence (ou droite de Henri) sont des estimateurs de la moyenne et de l'écart-type.

Là encore, on peut faire varier les paramètres grâce à des curseurs, avec visualisation instantanée de l'effet résultant sur la position de la droite.



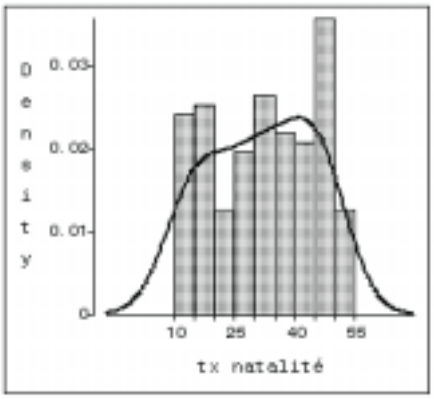


1. Un graphique cousin du QQplot, mais moins utilisé et non prévu dans SAS/Insight, est le PPplot ou graphique probabilité-probabilité, en ordonnées les probabilités cumulées observées, en abscisses les probabilités cumulées théoriques.
2. Les valeurs théoriques sont fixées par les fréquences cumulées des valeurs observées, ordonnées selon une échelle adaptée à la loi en question (par exemple l'échelle gaussienne pour la loi normale).

Ajustement non paramétrique

L'estimation de la densité de probabilité par la méthode du noyau (*Kernel Estimation*) est une méthode non paramétrique de lissage d'un histogramme utilisée depuis plus de 30 ans par les praticiens de l'analyse exploratoire des données.

Pour la mise en oeuvre de cette méthode, SAS/Insight offre le choix entre trois formes de noyau, gaussienne, triangulaire ou encore quadratique. Dans l'exemple ci-contre relatif à la distribution de la variable NAT (taux de natalité pour 1 000 habitants), c'est la première (noyau gaussien ou normal) qui a été retenue.



Kernel Density Estimation						
Curv	We	Me	C Value	Band	Mode	AMISE
——	No	AM	.785	6.72	41.6	.0003

Sophie DESTANDAU
et **Monique LE GUEN**

Analyses multivariées avec SAS/Insight

Là encore, SAS/Insight offre des possibilités étendues, depuis la simple superposition d'ellipses de confiance sur des diagrammes de dispersion (cf. l'article ci-avant) jusqu'à la modélisation au moyen d'un modèle linéaire généralisé, en passant par des analyses en composantes principales (ACP) à partir de matrices de corrélations ou de covariances.

Tout comme en dimension 1, les nouvelles variables générées dans le cadre d'une analyse multivariée sont enregistrées dans la table SAS de travail. Elles pourront donc à leur tour faire l'objet d'analyses détaillées.

Voir, apprendre, comprendre autrement

Pour être bien comprise, la statistique devrait être vue par les apprenants comme le prolongement d'un raisonnement naturel, commun à toute intelligence humaine. Or, elle est principalement enseignée de manière formelle, comme une branche des mathématiques pures. Cette voie n'est certainement pas la meilleure. Ainsi, l'Américaine Joan Garfied, spécialiste de l'enseignement des statistiques, rapporte que les recherches des psychologues ont montré que pour une grande majorité d'apprenants, même les notions de base (représentativité, indépendance, probabilité, moyenne pondérée, corrélation, causalité...) restaient floues. Les enseignants ont tendance à surestimer ce que comprennent leurs élèves. Dans le même temps, ceux-ci ont bien du mal à saisir la finalité des problèmes. En particulier, les étudiants de formation non mathématique ne semblent voir dans la statistique que des recettes de cuisine.

Du concret !

En accord avec les connaissances accumulées ces vingt dernières années en sciences cognitives, et plus spécifiquement en neuro-biologie et neuro-pédagogie, de plus en plus de voix s'élèvent aujourd'hui pour promouvoir l'enseignement de la statistique à partir de la visualisation sur données réelles et en utilisant les nouvelles technologies.

Regarder des données réelles sous des formes graphiques différentes, qui chacune apporte un nouveau point de vue, est en effet un moyen extrêmement puissant pour appréhender les concepts de la statistique. Tout un chacun peut toucher du doigt ce qu'est une distribution au vu des expressions graphiques proposées par les logiciels d'AED (analyse exploratoire des données). Ces mêmes représentations confèrent une réalité tangible aux caractéristiques de tendance centrale et de dispersion, comme au concept de variable aléatoire.

Bien sûr, tout graphique demande un apprentissage. Qui par exemple, parmi les profanes, imaginerait que l'allure d'un histogramme est fixée par l'amplitude des classes selon lesquelles on a découpé la variable étudiée ? Même un étudiant en statistique, tant qu'il n'aura pas fait l'expérience, ne pourra réaliser cette non-robustesse.

Expérience, simulation, EIAO

Toutes les études récentes montrent que les élèves apprennent plus facilement lorsqu'ils font eux-mêmes et sur données réelles.

Pour reprendre l'exemple de l'histogramme, on ne trouvera sans doute meilleur professeur qu'un logiciel statistique interactif permettant de modifier l'amplitude des classes et de visualiser en temps réel l'influence de ces modifications sur la forme de la distribution. Cette expérience réalisée par l'apprenant restera gravée dans sa mémoire à long terme. De la même façon, il percevra aisément ce que recouvre le concept de contribution, en visualisant les mouvements de la droite de régression résultant de la suppression d'observations situées à l'extérieur du nuage. Il aura alors intégré la nécessité de quantifier la contribution d'un point-observation. Dans la même veine, il a été constaté que simuler sur micro le théorème central limite était un excellent moyen d'en faire comprendre la signification ainsi que la portée.

Issues des nouvelles technologies, la visualisation et l'interactivité sont ainsi des moyens d'apprentissage très efficaces. Un autre avantage, fondamental, de la simulation sur ordinateur est qu'elle permet à la curiosité de l'apprenant de pleinement s'exprimer. Or, on le sait bien, c'est avec la curiosité et le questionnement que débute la connaissance, et c'est aussi en se trompant qu'on apprend, même si dans la tradition pédagogique française, la valeur formatrice de l'erreur n'est que modérément reconnue.

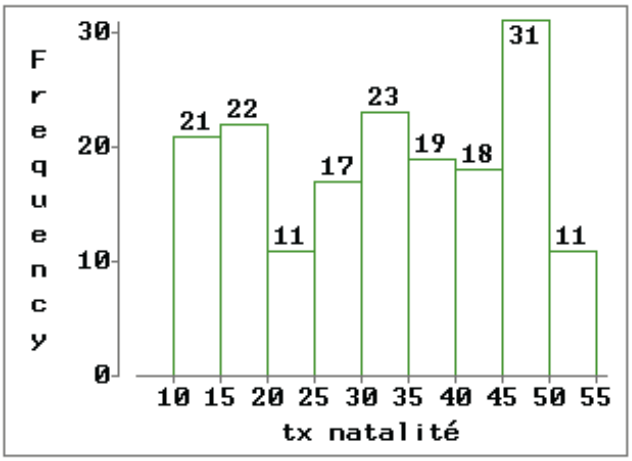
L'ordinateur, lui, ne porte pas de jugement ni ne fait de commentaires négatifs. Et comme l'informatique est appelée à devenir un levier de l'apprentissage, tous les espoirs sont permis. Bientôt, l'EIAO (enseignement interactif d'apprentissage par ordinateur) sera une réalité que tous les enseignants, les professeurs de statistique en particulier, devront assimiler. Bientôt, l'AED fera partie intégrante de la formation de base à la statistique.

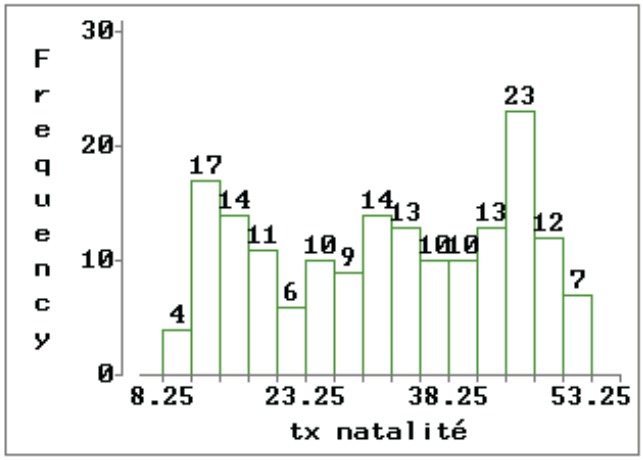
Monique LE GUEN

Deux exemples de ce que l'on peut découvrir avec un logiciel interactif d'AED

De la non-robustesse de l'histogramme

Avec un logiciel d'AED, il est très simple et très rapide de modifier le découpage en classes de la variable étudiée, avec visualisation instantanée de l'effet obtenu. Un exemple est donné ci-dessous, relatif à la distribution du taux de natalité (pour 1 000 habitants) dont a été observée la valeur dans 173 pays.



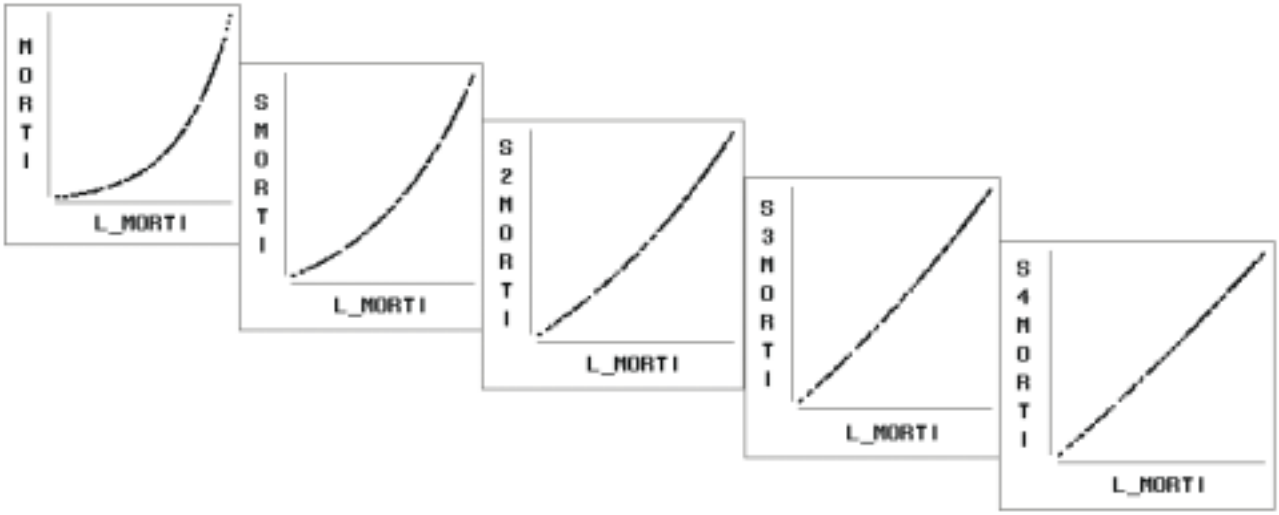


Des effets de la transformation des données (« to bend the data »)

La variable ici étudiée (variable MORTI) est le taux de mortalité infantile (pour mille naissances), dont on a relevé la valeur dans 173 pays.

Transformons cette variable en son logarithme, que nous noterons LMORTI, puis en sa racine carrée, notée SMORTI, puis en la racine de sa racine, S2MORTI, puis en la racine de la racine de sa racine, S3MORTI, enfin en la racine de la racine de la racine de sa racine, S4MORTI.

Il n'est pas très facile d'imaginer les effets de ces transformations mathématiques. Mais si l'on considère les graphiques cartésiens (MORTI, LMORTI), (SMORTI, LMORTI), (S2MORTI, LMORTI), (S3MORTI, LMORTI) et (S4MORTI, LMORTI), on constate qu'il existe entre LMORTI et S4MORTI une relation linéaire. Autrement dit, quand x est compris entre 5,5 et 183 (plage de valeurs occupée par MORTI), il existe une relation quasi linéaire entre lnx et $x^{\frac{1}{16}}$.



Références

Les trois ouvrages selon nous majeurs sur l'analyse exploratoire des données

ERICKSON B.H. & NOSANCHUK T.A. - *Understanding Data: An Introduction to Exploratory and Confirmatory Data Analysis for Students in the Social Sciences,* Milton Keynes, Open University Press, 1977 (réédité en 1992).

HOAGLIN D.C., MOSTELLER F. & TUKEY J.W. - *Understanding Robust and Exploratory Data Analysis,* New York, Wiley, 1985.

FOX J. & LONG J.S. - *Modern Methods of Data Analysis,* Sage Publications, 1990.

Un aperçu de l'oeuvre de Tukey

Nota 1 : Lorsque l'année d'édition est suivie d'une lettre, le code ainsi composé correspond à la numérotation des articles de Tukey telle qu'elle est notée dans « The Collected Works of John W. Tukey ».

Nota 2 : L'accès aux articles et ouvrages de Tukey sera plus facile si l'on commence par lire les écrits de ses élèves. Son style d'écriture est en effet particulier, et parfois très imagé. Sous sa plume, les quartiles peuvent devenir des « hinges » (littéralement « gonds ou charnières »), les valeurs extrêmes des « ones », la transformation d'une variable une « re-expression ». Lorsqu'il compare l'aplatissement d'une distribution observée à la loi normale, il pourra parler, plutôt que de Kurtosis, de « sharpness » ou de « spikyness » ce qui est plus compréhensible par le novice anglophone. Et les exemples de même nature sont foison ! Pour les francophones, la traduction n'est donc pas toujours évidente...

HOAGLIN D.C., MOSTELLER F. & TUKEY J.W. - *Understanding Robust and Exploratory Data Analysis,* New York, Wiley, 1985.

HOAGLIN D.C., MOSTELLER F. & TUKEY J.W. - *Fundamentals of Exploratory Analysis of Variance,*
Wiley, Series in Probability, 1971.

MOSTELLER F. & TUKEY J.W. - *Data Analysis for Regression,* Addison-Wesley Publishing Company, 1977.

TUKEY J.W. - *The Growth of Experimental Design in a Research Laboratory,*
Research Operations in Industry, Kings Crown Press, New York, pp. 303-313,1953c.
Également publié dans The Collected Works of J.W. Tukey, volume III.

TUKEY J.W. - *The Future of Data Analysis,* Annals of Mathematical Statistics, 33, pp.1-67, 1962a.
Également publié dans The Collected Works of J.W. Tukey, volume III.

TUKEY J.W. - *Analyzing Data: Sanctification or Detective Work,* American Psychologist, 24, pp. 83-91, 1969f.
Également publié dans The Collected Works of J.W. Tukey, volume IV.

TUKEY J.W. - *Exploratory Data Analysis*[1], Addison-Wesley, 1977.

TUKEY J.W. - *We Need Both Exploratory and Confirmatory Statistics,*
The American Statistician, vol 34, n° 1, pp. 23-25, 1980a.
Également publié dans The Collected Works of J.W. Tukey, volume III.

TUKEY J.W. - *Do Derivations Come from Heaven,*
manuscrit de 1981 publié dans The Collected Works of John W. Tukey, volume IV.

TUKEY J.W. - *Choosing Techniques for Analysis of Data,*
manuscrit de 1981 publié dans The Collected Works of John W. Tukey, volume IV.

TUKEY J.W. - *Sunset Salvo,* The American Statistician, 40, pp. 72-76, 1986a.
Également publié dans The Collected Works of John W. Tukey, volume IV.

TUKEY J.W. - *More Honest Foundations for Data Analysis,*
Journal of Statistical Planning and Inference, 57, pp. 21-28, 1997.

1. Cet ouvrage serait en cours d'actualisation.

The Collected Works of John W. Tukey
(Wadworth & Brooks / Monterey / Californie pour les volumes I à VII, Chapman and Hall depuis le volume VIII) :

Vol I : *Time Series 1949-1964* (ed. Brillinger D.R.), 1984
Vol II : *Time Series 1965-1984* (ed. Brillinger D.R.), 1984
Vol III : *Philosophy and Principes of Data Analysis 1949-1964* (ed. Jones L.V.),1986
Vol IV : *Philosophy and Principes of Data Analysis 1965-1986* (ed. Jones L.V.),1986
Vol V : *Graphics: 1965-1985* (ed. Cleveland W.S.), 1988
Vol VI : *More Mathematical 1938-1984* (ed. Mallows C.L.), 1990
Vol VII : *Factorials and ANOVA 1949-1962* (ed. Coc D.R.), 1992
Vol VIII : *Multiple Comparisons 1949-1983,* 1994

Autres références statistiques

En anglais

BEHRENS J.T. - *Principles and Procedures of Exploratory and Data Analysis,*
Psychological Methods, vol 2, n° 2, 131-160, 1997.

BELSEY D.A., KUH E. & WELSH R.E. - *Regression Diagnostics,* Wiley, 1980.

BRILLINGER D.R., FERNHOLTZ L.T. & MORGENTHALER S. - *The Pratice of Data Analysis,*
Essays in Honor of John W. Tukey, Princeton University Press, 1997.

CHAMBERS J.M., CLEVELAND W.S., KLEINER B. & TUKEY P.A. - *Graphical Methods for Data Analysis,*
Wadsworth International Group, Monterey, Californie, 1983.

CLEVELAND W.S. - *Visualizing Data,* Hobart Press, Summit, New Jersey, USA, 1993.

CLEVELAND W.S. - *The Elements of Graphing Data,* Hobart Press, Summit, New Jersey, USA, 1994.

CLEVELAND W.S. & MCGILL R. - *Graphical Perception:*
The Visual Decoding of Quantitative Information on Graphical Displays of Data,
J.R. Statist. Social A, 150, Part 3, pp. 192-229, 1987.

COOK R.D. & WEISBERG S. - *An Introduction to Regression Graphics,* New-York, Wiley, 1994.

DESJARDINS D. - *New Graphical Techniques for the Analysis of Census Data,*
Statistics Conference Symposium 97 Proceedings, Ottawa, Canada, 1997.

DEVILLE J.-C. & MALINVAUD E. - *Data Analysis in Official Socio-Economic Statistics,*
JRSS, 146, part 4, pp. 335-361, 1983.

ERICKSON B.H. & NOSANCHUK T.A. - *Understanding Data: An Introduction to Exploratory and Confirmatory Data Analysis for Students in the Social Sciences,* Milton Keynes, Open University Press, 1977 (réédité en 1992).

FOX J. & LONG J.S. - *Modern Methods of Data Analysis,* Sage Publications, 1990.

VELLEMAN P.F. & HOAGLIN D.C. - *Applications, Basics and Computing of Exploratory Data Analysis,*
Boston, Mass., Duxbury Press,1981.

WAINER H. - *Visual Revelations, Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot,*
Copernicus, Springer-Verlag, 1997.

En français

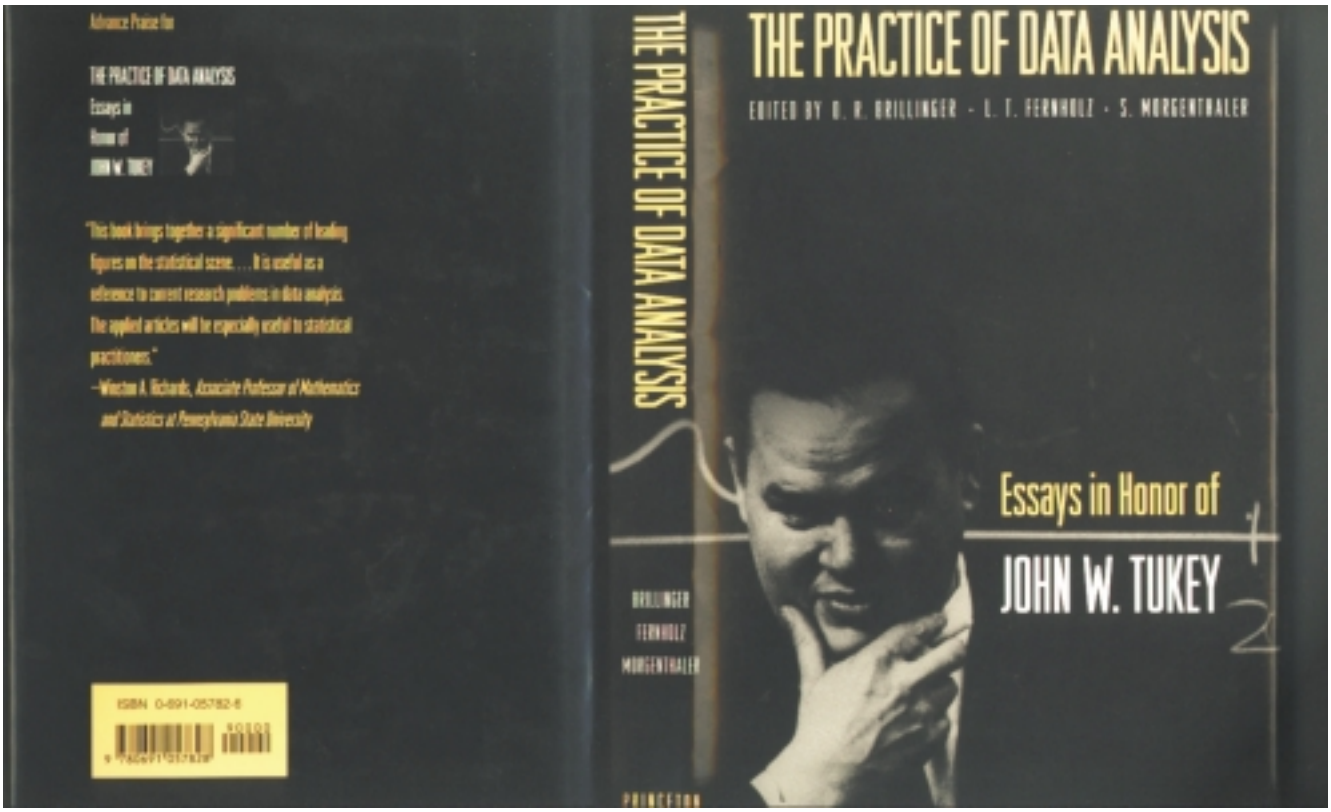
DESTANDAU S. & LE GUEN M. - *Analyse exploratoire des données avec SAS/Insight,*
Insee Guides, n° 7-8, 1998.

HORBER E. - *Visualisation graphique et semi-graphique,*
in Actes des Journées de méthodologie statistique des 18 et 19 octobre 1995,
Insee Méthodes, n° 59-60-61, pp. 15-17, 1995.

LADIRAY D. - *L'analyse exploratoire des données »,* Insee, Lettre du SSE, n° 30, 1998.

LADIRAY D. & ROTH N. - *Lissage robuste des séries chronologiques, une étude expérimentale,*
Annales d'Économie et Statistique, n° 5, pp. 147-181, 1987.

LEBART L., MORINEAU A. & FÉNELON J.-P. - *Traitement des données statistiques,* Dunod, 1979.


Advance Praise for
THE PRACTICE OF DATA ANALYSIS
Essays in
Honor of
JOHN W. TUKEY
"This book brings together a significant number of leading figures on the statistical scene. . . . It is useful as a reference to current research problems in data analysis. The applied articles will be especially useful to statistical practitioners."
—Winston A. Richards, Associate Professor of Mathematics and Statistics at Pennsylvania State University
ISBN 0-691-05782-6
THE PRACTICE OF DATA ANALYSIS
BRILLINGER
FERNHOLZ
MORGENTHALER
PRINCETON
THE PRACTICE OF DATA ANALYSIS
EDITED BY D. R. BRILLINGER · L. T. FERNHOLZ · S. MORGENTHALER
Essays in Honor of
JOHN W. TUKEY

LEBART L., MORINEAU A. & PIRON M. - *Statistique exploratoire multidimensionnelle,* CISIA-CERESTA, 1997.

SPRENT P. - *Pratique des statistiques non paramétriques,* INRA Éditions, Techniques et pratiques, 1992.

TOMASSONE R., AUDRAIN S., LESQUOY-DE TURKEIM E. & MILLIER C. - *La régression, nouveaux regards sur une ancienne méthode statistique,* Masson, 2e édition, 1992.

VALOIS J.-P. - *Analyse exploratoire et régression sur série chronologique : application à l'affaiblissement d'un champ pétrolier,* XXXe Journées statistiques ASU-SFdS, 1998.

Références en sciences cognitives et enseignement

Écrits

Annual Meeting of The American Statistical Association - *Proceedings of the Section on Statistical Education,* 1995.

ANDLER D. (sous la direction de) - *Introduction aux sciences cognitives,* Folio, Essais, Éditions Gallimard, 1992.

ANSCOMBE F. J. - *Graphs in Statistical Analysis,* The American Statistician, vol 27, 1973, pp. 17-21, 1973.

BERRONDO-AGRELL M. & AGRELL P. - *Vers une syntaxe des diagrammes de Venn : lutte contre un mythe,* SSP, tome 133, n° 1, pp. 140-150, 1992.

BERRONDO-AGRELL M. & FOURASTIÉ J. - *Pour comprendre les probabilités,* Hachette, Collection Les Fondamentaux, 1995.

BERRONDO-AGRELL M. & FOURASTIÉ J. - *Le calcul des probabilités compréhensible pour tous, exercices avec corrigés,* Gaëtan Morin Éditeur, 1998.

BERTIN J. - *La graphique et le traitement graphique de l'information,* Flammarion, 1974.

BRISSARD F. - *Développez l'intelligence de votre enfant par la méthode « de la Garanderie »,* Éditions Le Rocher, 1989.

BUZAN T. - *Une tête bien faite,* Éditions Organisations, 1999.

CARROLL L. - *Logique sans peine,* traduction et présentation de GATTÉGNO J. & COUMET E., Hermann, 1992.
Cet ouvrage reprend « Symbolic Logic » et « The Game of Logic » (1887),
« What the Tortoise Said to Achille » et « A logical Paradox » (1894).

CHANGEUX J.-P. - *L'homme neuronal,* Paris, Fayard, 1983.

CHANGEUX J.-P. & CONNES A. - *Matière à pensée,* Éditions O. Jacob, 1989.

CHANGEUX J.P. - *Raison et plaisir,* Éditions O. Jacob, 1994.

Courrier du CNRS, n° 66-67-68 (*spécial Imagerie scientifique*), 1987.

Courrier du CNRS, n° 79 (*Dossiers scientifiques - Sciences cognitives*), 1992.

CYRULNIK B. - *De la parole comme de la molécule,* Point Seuil, 1997.

DAMASIO A.R. - *L'erreur de Descartes, la raison des émotions,* Éditions O. Jacob, 1994.

DE LA GARANDERIE A. - *Pédagogie des moyens d'apprendre,* Le Centurion, 1982.

DEHAENE S. - *La bosse des maths,* Éditions O. Jacob, 1997.

DENIS M. - *Image et cognition,* PUF, Psychologie d'aujourd'hui, 1989.

DIEUDONNÉ J. - *Pour l'honneur de l'esprit humain, les mathématiques d'aujourd'hui,* Hachette, 1987.

DUPUY J.-P. - *Aux origines des sciences cognitives,* Éditions La Découverte, 1994.

EDELMAN G.M. - *Biologie de la conscience,* Éditions O. Jacob, 1992.
Édition originale en langue anglaise « *Bright air, Brilliant Fire: On the Matter of Mind* », Basic Books, 1992.

EDWARDS A. - *Venn Diagrams for Many Sets,* Newscientist, n° 1646, pp. 51-56, 1989.

GARFIELD J. - *How Students Learn Statistics,* International Statistics Review, 63, 1, pp. 25-34, 1995.

GENINET A. - *La gestion mentale en mathématiques, application de la sixième à la seconde,*

Éditions Pédagogie RETZ, 1993.

GRANDIN T. - *Penser en images,* Éditions O. Jacob, 1997.

JACQUARD A. - *L'équation du nénuphar, les plaisirs de la science,* Calmann-Lévy, 1998.

KAHANE J.-P. - *Mathématiques et formation,* La Pensée, 302, pp. 89-98, 1995.

LASZLO P. - *La découverte scientifique,* Que Sais-je ? n° 3473, 1999.

LE GUEN M. - *Statistique, imagerie et sciences cognitives,*
Bulletin de méthodologie sociologique, n° 49, pp. 90-100, 1995.

LE GUEN M. - *L'analyse exploratoire des données est au cerveau droit ce que l'analyse confirmatoire est au cerveau gauche : les deux doivent communiquer pour traiter l'information,*
CNRS-Matisse, Université Sorbonne-Panthéon, Document de travail n° F99-05,1999.

MARASINGHE M.G., MEEKER W.Q., COOK D. & SHIN T. - *Using Graphics and Simulation to Teach Statistical Concepts,*
The American Statistician, vol 50, n° 4, pp. 342-351, 1996.

LUCAS M. - *La communication graphique,* Le Courrier du CNRS, n° 80, p. 102, 1993.

MOORE D.S., COBB G. W., GARFIELD J. & MEEKER W.Q. - *Statistics Education Fin de Siècle,*
The American Statistician, Vol 49, n° 3, pp. 250-260, 1995.

ROSE S. - *La mémoire, des molécules à l'esprit,* Seuil, 1992.

SALANSKIS J.-M. - *L'interdit de l'image en mathématiques,* Alliage, n° 15, pp. 29-34, 1993.

SPIEGEL R.M. - *Probabilités et statistique, cours et problèmes,* Série Schaum, 3e tirage, 1975.

STEWART I. - *Les dentelures de l'esprit : quand les diagrammes de Venn deviennent fractals,*
Pour la Science, n° 138, 1989.

STEWART I. - *Dieu joue-t-il aux dés ? Les mathématiques du chaos,* Champs Flammarion, 1994.

STUART M. - *Changing the Teaching of Statistics,* The Statistician, 44, n° 1, pp. 45-54, 1995.

TROCMÉ-FABRE H. - *J'apprends donc je suis,* Éditions Organisations, 1997.

TROCMÉ-FABRE H. - *Réinventer le métier d'apprendre,* Éditions Organisations, 1999.

VARELA F. - *Connaître les sciences cognitives, tendances et perspectives,* Seuil, 1988.
Réédité en 1997 sous le titre « Invitation aux sciences cognitives », Point Seuil.

WILLIAMS L.V. - *Deux cerveaux pour apprendre, le gauche et le droit,*
traduit par TROCMÉ-FABRE H., Éditions Organisations, 1997.

ZEKI S. - *La construction des images par le cerveau,* La Recherche, n° 222, juin 1990, pp. 712-721, 1990.

Voir dans le Cerveau, La Recherche, n° 289 (numéro spécial), 1996.

Autres supports

AMADO, logiciel d'Analyse graphique d'une MAtrice de DOnnées,
distribué par CISIA (1, avenue Herbillon - 94160 Saint -Mandé).

JMP et SAS/Insight de SAS Institute, logiciels de statistique et visualisation.

EDELMANN C. - *Un violon dans la tête,* festival des films scientifiques, Palaiseau 1992.

TROCMÉ-FABRE H. - Vidéogramme en 7 cassettes « *Né pour apprendre* »,
coproduction Université de La Rochelle & École normale supérieure de Fontenay-Saint-Cloud :

Né pour découvrir (**CYRULNIK B.**, éthio-psychiatre) ;
Né pour reconnaître les lois de la vie (**NICOLESCU B.**, physicien théoricien) ;
Né pour organiser (**VARELA F.**, biologiste) ;
Né pour créer du sens (**VARELA F.**).
Né pour choisir (**JACQUARD A.**, généticien et mathématicien, **DE PERETTI A.**, psychosociologue) ;
Né pour innover (**VINCENT J.-D.**, neuro-endocrinologue) ;
Né pour échanger (**SCHWARTZ B.**, insertion professionnelle des jeunes).


NÉ
POUR
APPRENDRE
Vidéogramme en 7 cassettes
Une coproduction
Université de La Rochelle
&
Ecole Normale Supérieure Fontenay /St-Cloud
Auteur : Hélène Trocmé-Fabre Réalisateur : Daniel Garabédian

Qualité et statistique

En statistique, le succès du terme qualité ne se dément pas depuis quelques années. Publications, colloques, discussions de couloir, nous avons souvent ce mot à la bouche, sans savoir vraiment ce qu'il recouvre. Comme c'est souvent le cas lorsque les concepts sont riches, celui-ci se révèle en pratique polysémique, réfractaire à une approche simpliste, voire fourre-tout et chacun y met ce qu'il veut. Tout dépend en effet de quel point de vue on se place : utilisateur, répondant, responsable de l'enquête, responsable de la production, politique, avant ou après l'enquête, pour tel ou tel usage. Il existe, comme nous le verrons, des définitions très élaborées de la qualité, mais elles se déclinent différemment selon les rôles des uns et des autres dans le processus de production statistique : chacun agit à son niveau sur la qualité finale des statistiques, et n'en possède en général qu'une vision tronquée. Cela n'empêche pas de se doter de principes de qualité pour chaque étape du traitement statistique, cohérents avec une grille d'évaluation générale.

Si l'on arrive à s'entendre sur une définition, même floue, se pose ensuite la question de mesurer cette qualité, c'est-à-dire de déterminer une batterie d'indicateurs, effectivement calculables et censés représenter de près ou de loin le concept envisagé, puis de les appliquer à des cas réels.

Ce n'est pas tout : une fois les indicateurs calculés, que va-t-on en faire ? Les publier, les comparer, analyser leur évolution d'une période sur l'autre. Au minimum pour des besoins internes, ou parce que l'affichage de ces indicateurs présente un intérêt en soi. Mais surtout, cela peut permettre de détecter des points faibles, des voies d'amélioration et conduire à améliorer la qualité, en agissant sur les instruments adéquats (moyens, méthodologie, outils, organisation du travail...).

Le but du présent papier est de mettre en lumière les différents aspects de la notion de qualité en statistique, en insistant plus particulièrement sur les statistiques économiques. On essaiera au passage de faire le point sur l'état d'avancement de l'Insee en la matière et sur les travaux en cours au niveau européen. Il ne s'agira nullement de faire un cours sur le concept de qualité, mais plutôt de débroussailler un sujet confus, en essayant d'extraire quelques idées-forces.

La qualité en général

Deux notions fondamentales doivent être distinguées en matière de qualité : *la maîtrise de la qualité*[1] et *l'assurance de la qualité*. La première répond plus à un objectif interne d'amélioration des produits, et d'accroissement de productivité et de rentabilité. La seconde, plus récente, est partie de l'idée selon laquelle des garanties de qualité de produit pouvaient être exigées par les clients.

L'ISO, *International Organization for Standardization*, s'est emparée de ces concepts en 1979, aboutissant en 1986 à une définition normalisée de la qualité et en 1987 à des référentiels pour la certification de systèmes qualité d'entreprise.

Ainsi, la norme ISO 8402 de 1986 définit la qualité comme étant « l'ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confèrent l'aptitude à satisfaire des besoins exprimés et implicites ». Cette définition a l'avantage de la généralité, puisqu'elle n'apparaît pas comme spécifique à un secteur d'activité particulier. On notera en particulier qu'il existe une grande différence entre les concepts de *qualité* et *d'excellence* (souvent confondus dans le langage courant). En effet, la qualité visée est celle qui est *nécessaire* et *suffisante au besoin*[2] : la sur-qualité, comme la sous-qualité, est une non-qualité.

La norme ISO 9001 fournit un « modèle pour l'assurance de la qualité en conception, développement, production, installation et prestations associées ». Il s'agit ici d'une vision très large de la qualité, puisque toutes les composantes du travail de l'entreprise sont prises en compte. Cette certification, que l'AFNOR caractérise comme un référentiel d'organisation[3], se révèle donc en pratique la plus difficile à obtenir et requiert de la part de l'entreprise un

1. *Maîtrise de la qualité* est la traduction de l'anglais *quality control*. Ceci a conduit à provoquer de nombreux contresens en français, car l'expression « contrôle de qualité » a un sens tout différent, avec notamment l'idée de vérification du produit a posteriori (alors que la maîtrise de la qualité s'effectue au sein du processus de production).
2. On ne peut comparer la qualité d'une 2 CV et d'une Rolls-Royce, car les besoins sont différents.
3. On se référera utilement au site de l'AFNOR (www.afnor.fr), à la rubrique Normalisation.

travail préparatoire très important qui peut s'étaler sur des années.

Moins exigeante, la norme ISO 9002 exclut les aspects conception et développement. Il ne s'agit donc ici que de la qualité du processus de production, et en particulier de l'aptitude de l'entreprise à le maîtriser. Bien entendu, la qualité du processus contribue à la qualité du produit.

Enfin, la norme ISO 9003 est définie comme un « modèle pour l'assurance de la qualité en contrôle et essais finals ». En d'autres termes, on se limite ici à un aspect de la qualité du produit, ce que l'on peut en percevoir en l'analysant isolément, par des contrôles, sans se soucier aucunement de la manière dont il a été obtenu : il n'est nullement évident que le processus de production soit lui-même de qualité.

En toute rigueur, cette norme ISO 9003 est insuffisante pour le produit « statistique ». On voit bien qu'il est possible, au travers des caractéristiques et des fonctionnalités, de définir une qualité de produit pour une chaise, une imprimante, un yaourt : on peut imaginer des contrôles, des tests, qui s'appliqueront au produit lui-même et qui ne nécessiteront jamais de récupérer des informations auprès de la chaîne de production. En revanche, il est souvent difficile d'évaluer la qualité des statistiques à la seule vue de celles-ci. En effet, si l'on ne dispose pas d'éléments de comparaison, d'indicateurs supplémentaires (sur le sondage ou la non-réponse), qu'est-ce qui nous permet de dire que des statistiques sont fiables ou non ? Peu de choses *sur la base du produit lui-même.*

La qualité des statistiques, ou le point de vue de l'utilisateur

Il s'agit bien ici d'appréhender de façon générale la qualité des statistiques, en tant que produit. Simplement, lorsqu'on étudie la qualité d'un produit, on se place évidemment du point de vue de celui qui va l'utiliser, et qui aura au sujet de ce produit un certain nombre de desiderata à exprimer.

Comment procéder ? À partir d'une description de tout ce qui constitue le produit « statistique », il est tout à fait possible d'en caractériser la qualité : en analysant a posteriori le produit (ensemble de tableaux, de graphiques…), mais aussi en essayant d'extraire du processus de production statistique d'autres informations utiles sur ce produit. Il existe ainsi des *définitions* de la qualité, même si le terme peut être jugé excessif : on approche plutôt le concept de qualité en le décomposant en plusieurs catégories.

Plusieurs instituts nationaux de statistique (INS) ont ainsi proposé des définitions, qui se révèlent assez proches les unes des autres. Nous allons ici expliciter celle qui a été proposée par Eurostat, car elle constitue une excellente synthèse, et désormais un standard reconnu. Cette définition a été discutée dans le cadre du groupe de travail Évaluation de la qualité des statistiques, et validée par les États membres.

Pour représenter la qualité, Eurostat distingue six composantes : pertinence, précision, actualité et ponctualité, accessibilité et clarté, comparabilité, cohérence.

Les six composantes de la qualité des statistiques selon Eurostat

Pertinence. Les statistiques produites doivent répondre aux besoins des utilisateurs.

Il est intéressant de constater qu'il s'agit, de loin, de la composante la moins détaillée… alors que, pour nombre d'utilisateurs, c'est la plus importante. Le document d'Eurostat ne consacre que quelques lignes à la pertinence. Cela peut s'expliquer : Eurostat n'est pas n'importe quel utilisateur des statistiques. Ce n'est pas un client au sens strict, directement intéressé par les résultats en termes d'actions à mener, mais plutôt un intermédiaire, qui va recueillir les statistiques des divers États membres pour fabriquer des statistiques européennes qui, elles, seront diffusées à de véritables clients. Cela dit, dans d'autres documents sur la qualité, on constate que les instituts statistiques ont beaucoup de mal à proposer des indicateurs de pertinence, ou des méthodes pour l'améliorer.

Précision. Les différents types d'erreur doivent être estimés.

De toutes les composantes, c'est bien évidemment celle qui est la plus commentée, qui fait l'objet du plus grand nombre d'indicateurs. Pour la plupart des statisticiens, c'est évidemment là qu'est l'essentiel. Mais étrangement, ce n'est pas un élément crucial pour les utilisateurs, qui ont souvent tendance à faire confiance aux statisticiens à ce sujet. Attention, la précision n'est pas un but en soi (l'accroître indéfiniment serait absurde), ce n'est pas le critère de qualité. Ce qui importe en premier lieu, c'est de connaître cette précision.

Actualité et ponctualité. Les utilisateurs demandent des données récentes mises à jour fréquemment et avec ponctualité.

Selon les enquêtes menées par Eurostat auprès des utilisateurs, la fraîcheur de la publication est un critère essentiel de qualité.

Accessibilité et clarté. Les données doivent être accessibles et lisibles aux utilisateurs.

En poussant plus loin sur ce sujet, il faudrait aussi évoquer la qualité de la publication (bonne organisation du document, commentaires, illustrations, etc.) mais aussi la facilité d'accès à la publication elle-même, qui peut souvent constituer un obstacle.

Comparabilité. Les statistiques doivent pouvoir être comparées dans le temps et dans l'espace.

Cohérence. Les statistiques provenant de sources différentes doivent être cohérentes (mêmes définitions, mêmes nomenclatures, mêmes méthodologies...).

Toujours selon Eurostat, la demande des utilisateurs pour des statistiques cohérentes entre elles est très forte. C'est même le facteur le plus important avec la fraîcheur des données.

Le point de vue du méthodologue sur la précision

Pour le statisticien, et plus particulièrement le méthodologue, précision est naturellement synonyme de qualité (on a souvent tendance à omettre les autres critères). C'est la plus importante des six composantes que distingue Eurostat. Mesurer la précision constitue une tâche très délicate, car les sources d'erreur en statistique sont multiples. On en distingue généralement deux types, celles qui sont dues au sondage et les autres.

L'erreur induite par le sondage, lorsque celui-ci est aléatoire, se décompose classiquement en biais[1] et variance. Le calcul de la variance peut être effectué aujourd'hui avec plusieurs logiciels : POULPE, développé à l'Insee, en est un très bon exemple. Lorsque le sondage est non aléatoire (technique des taux de couverture, par exemple), l'estimation de l'erreur est plus difficile à obtenir.

Les erreurs qui ne sont pas dues au sondage sont très diversifiées. Elles peuvent survenir au départ, dans la base de sondage : sur-couverture (prise en compte, à tort, d'unités hors-champ, comme les unités mortes), sous-couverture (oubli de certaines unités, par exemple les naissances), mauvaise classification… Tout cela peut donner lieu à des indicateurs, obtenus par estimation, comme par exemple le taux de faux actifs dans les répertoires d'entreprises. À vrai dire, la question de la qualité des répertoires est un problème en soi. C'est même un thème à part du groupe de travail d'Eurostat sur la qualité. Lors de la dernière table ronde sur les répertoires d'entreprises, tenue à Paris en septembre 1999, la question des indicateurs de qualité des répertoires a d'ailleurs fait l'objet de nombreux articles aux conclusions convergentes.

Tout au long du processus de production, les erreurs peuvent naître des traitements, notamment la saisie, le codage, ou le contrôle. Souvent mesurables (double saisie, double codage), les erreurs liées aux traitements n'en demeurent pas moins difficiles à intégrer dans une mesure globale d'erreur.

L'erreur due à la non-réponse (partielle ou totale, comme les éclipses) peut avoir un impact très important sur les résultats. L'effet de la non-réponse est à ce titre systématiquement étudié par la littérature statistique. Comme dans le cas du sondage, on distingue biais de non-réponse et variance due à la non-réponse. La mesure de l'erreur va dépendre de la manière dont la non-réponse est traitée (imputation ou repondération, et bien entendu les différentes méthodes possibles dans chaque cas).

On ne saurait omettre l'erreur de mesure, qui est faite au moment de la collecte des données, même si le contrôle peut permettre de la corriger partiellement. Elle peut provenir de multiples facteurs : influence de l'enquêteur, lisibilité de la question, non-concordance de la donnée demandée avec le système d'information du répondant…

Reste un dernier type d'erreur à mentionner : l'erreur due au modèle. Lorsque des techniques particulières d'estimation sont requises (correction pour variations saisonnières, pour écart entre année comptable et année civile…), l'erreur induite est en effet directement liée au modèle mathématique utilisé pour l'estimation.

Techniquement, un certain nombre de ces erreurs sont calculables à partir des données disponibles, mais il est très difficile de bâtir un indicateur qui les prenne toutes en compte. Et en pratique, comme sondage et non-réponse sont le plus souvent les composantes majeures de l'erreur, les mesures sur lesquelles on se fonde pour évaluer l'erreur sont en général des mesure de variance qui intègrent ces deux sources d'imprécision[2]

Même déterminées ainsi, les mesures de variance, prises telles quelles en tant qu'indicateurs de qualité, doivent mener le statisticien à réfléchir : il n'est pas évident qu'il s'agisse là d'un indicateur universel, conduisant à définir des normes de fiabilité. Michel Volle le rappelle, en écrivant [3] qu'une statistique ne doit pas être « précise », mais « exacte », au sens où elle doit pouvoir alimenter un raisonnement exact de l'utilisateur des données. Peu importe la précision des données, si le raisonnement que l'on en tire est juste, ce qui implique que les décisions qui en découlent sont fondées. Cette vision des choses confère à la qualité des statistiques un caractère non absolu, entièrement lié aux usages, dont on peut regretter qu'il soit insuffisamment connu par la communauté statistique.

1. Rappelons au passage que le biais peut apparaître même si le plan de sondage est irréprochable, en particulier dans le cas où l'on estime des ratios.

2. Il ne s'agit pas là d'un problème que l'on sait facilement résoudre. Il n'existe pas réellement, pour le calcul de la variance due à la fois au sondage et à la non-réponse, de méthodes générales standard reconnues par la communauté statistique, loin de là. Cf. pour cela l'excellent Model Quality Reports in Business Statistics *(volume 1 :* Theory and Methods for Quality Evaluation*), corédigé par des statisticiens et universitaires suédois et britanniques dans le cadre d'un projet financé par Eurostat.*

3. Cf. Système de pilotage de l'entreprise*, sur le site Web www.volle.com.*

Le **coût** (pour l'office statistique, mais aussi pour le répondant) n'est en revanche plus considéré comme une composante de la qualité des statistiques[4] mais comme une contrainte. Ce choix est discutable. Il est logique que le coût de la production statistique soit en dehors de la mesure de la qualité des statistiques : il se peut que l'on veuille minimiser les coûts à qualité donnée ou améliorer la qualité à coût constant, et l'on voit bien que dans ce cas, le fait que la qualité intègre le coût n'a pas de sens. En revanche, le fardeau que constitue la réponse pour le répondant, que l'on appelle également la charge d'enquête, devrait être pris en compte : finalement, c'est un effet de bord négatif induit par le processus de production. La qualité doit donc prendre en compte simultanément ces aspects, en trouvant un juste milieu entre intérêt pour l'utilisateur et coût de la réponse pour l'enquêté. Les choses se compliquent en statistique d'entreprise, où il arrive souvent que les utilisateurs soient également les enquêtés.

La définition proposée par Eurostat a l'avantage de fournir un cadre très complet et très général pour appréhender la qualité des statistiques[5] : tous les aspects de cette qualité, tels que les acteurs de la production statistique la perçoivent isolément, peuvent se raccrocher à la liste proposée. Cela va imposer, petit à petit, des normes aux statisticiens européens, les amener à réfléchir sur leurs pratiques, à mettre celles-ci en commun, tout en fournissant des principes, des méthodes, voire un langage, fort utiles dans notre métier.

La qualité de l'interaction : le point de vue du répondant... ou de l'enquêteur

La collecte de données n'est pas une opération neutre pour celui qui

4. Dans les versions initiales du document d'Eurostat, le coût était une 7e composante

5. Il est à noter qu'une grande partie des indicateurs proposés par Eurostat s'applique potentiellement aux statistiques fondées sur les sources administratives.

répond : cela prend du temps et de l'énergie, nécessite parfois de mobiliser plusieurs personnes, plusieurs sources d'information. Les travaux menés dans le cadre de l'opération SPE, Statistique publique et entreprises, ont permis de constater que les entreprises enquêtées se sentaient traitées comme des « machines à répondre », le statisticien considérant leur réponse comme un dû. Ce comportement jugé excessivement régalien a de nombreuses conséquences négatives, sur la propension à répondre, sur la qualité des réponses individuelles, ainsi que sur l'image de l'institut. La volonté désormais affichée d'améliorer les contacts avec les entreprises n'est pas spécifique à la France. Ainsi, le bureau statistique australien (ABS) a créé une charte des statistiques d'entreprises [6], qui définit un cadre pour les relations entre l'ABS et les entreprises qui lui fournissent de l'information.

Donc, la manière dont l'enquêté perçoit l'enquête est un aspect essentiel de l'ensemble du processus de production statistique car les données individuelles constituent notre matériau de base. Elles sont issues d'une **interaction**, parfois complexe, entre le service statistique et l'unité enquêtée. La qualité de cette interaction, telle qu'elle est perçue par le répondant, doit donc être envisagée en soi : si l'on se réfère au cadre initial, on touche ici aux deux composantes pertinence et précision (qualité de la donnée individuelle collectée, mais aussi taux de réponse), ainsi qu'à la contrainte de coût (coût pour l'enquêté, aussi bien objectif que subjectif).

Lors d'une interaction, l'objectif du service enquêteur est d'obtenir le meilleur niveau de réponse possible (au sens minimisation des risques de non-réponse et d'erreur de réponse), en essayant de limiter la charge de travail de l'unité et le tout en donnant une bonne image de la statistique publique. Ce dernier point ne doit pas être négligé, car l'image laissée au répondant conditionne peu ou prou les réponses aux enquêtes ultérieures, et, par là même, la qualité à long terme. Dès lors, réfléchir à la qualité de l'interaction entre le service statistique et l'enquêté, c'est définir tout ce qui la compose, pour être en mesure de l'améliorer sur la base de cet objectif.

On peut décrire l'interaction à partir de quatre éléments : une date, un support (ou un motif de discussion), un individu représentant l'unité enquêtée, un autre représentant le service enquêteur.

La qualité de l'interaction sera en grande partie celle de ses supports, même si le reste doit aussi être analysé. Il existe de nombreux supports : questionnaires, lettres d'envoi, retours d'information, informations sur les publications... Mais c'est bien sûr au questionnaire, en tant que support de collecte, que l'on va plus particulièrement s'attacher.

Que ce soit pour l'enquêteur ou l'enquêté, la qualité du questionnaire est toujours liée à la possibilité d'y répondre dans des délais raisonnables[7]. Cela nous ramène aux critères classiques : possibilité même de mobiliser l'information, lisibilité, temps nécessaire pour répondre. Et typiquement, ce sont les tests de questionnaire, les relectures par des comités d'utilisateurs, qui vont permettre de se prononcer sur ces sujets.

Il ne faut pas oublier les interactions orales : entretien avec l'enquêteur, discussion au téléphone avec le gestionnaire d'enquête (soit parce que ce dernier appelle, soit parce que l'entreprise demande des précisions). En effet, elles jouent un rôle crucial dans l'image que le répondant se fait de l'enquête, puisqu'il se trouve alors en contact non pas avec une entité abstraite, mais avec un être humain qui « représente » la statistique publique. L'expérience montre que des considérations très peu mesurables entrent en ligne de compte : par exemple, la présentation de l'utilité et du contexte de l'enquête, le ton employé au téléphone[8] pourront transformer un récalcitrant en un répondant... ou l'inverse.

Les autres supports écrits transmis aux enquêtés et qui ne constituent pas des supports de collecte au sens strict (lettres d'envoi, de rappel, retours d'information après l'enquête) font partie intégrante de la qualité de l'interaction telle qu'elle est perçue par le répondant.

Outre le support papier, l'échange oral, la collecte assistée par ordinateur, mentionnons les échanges de données informatisés (EDI), qui peuvent permettre, dans certains cas, de faciliter grandement la tâche de réponse. Les EDI, encore sous-utilisés aujourd'hui, sont appelés à se développer dans les années à venir ; à long terme, on peut espérer qu'ils conduisent à une réduction des coûts aussi bien pour l'enquêté que pour le service enquêteur.

Les autres termes de l'interaction, évoqués plus haut, jouent également sur la qualité de celle-ci. Ainsi, les statisticiens d'entreprise savent bien que la date de la collecte d'information et l'interlocuteur choisi dans l'entreprise ont un impact énorme sur la fiabilité des réponses et le taux

6. Effective depuis juillet 1998.
7. Pour cela, *l'homologie des formes cognitives*, matérielles ou non, est essentielle mais délicate à maîtriser : au fond, il faudrait assurer la correspondance entre les schémas mentaux de l'enquêté et de l'enquêteur, en ayant éventuellement recours à des outils susceptibles de fabriquer le produit « réponse à l'enquête ». Sur ces sujets, cf. L. Thévenot, *Les investissements de forme*, Conventions économiques, Paris, CEE/PUF, 1986.
8. Les enquêtés, et en particulier les entreprises, apprécient de moins en moins d'être considérés comme des « subordonnés ».

de réponse. Les études menées dans différents INS le confirment : un bon « ciblage », dans le temps et dans l'espace, est essentiel, non seulement pour avoir des réponses rapides, mais aussi pour établir une relation de qualité entre un service statistique et l'unité qu'il interroge.

On peut essayer de retourner le problème en se plaçant du point de vue du service enquêteur : la qualité de l'interaction avec l'unité serait alors liée à la capacité à obtenir des données pertinentes et fiables dans des délais brefs. Mais cela revient au même. En effet, si, du point de vue de l'enquêté, une bonne relation s'établit (rapports de confiance, bonne compréhension réciproque des tenants et aboutissants), les effets ne pourront être que positifs pour le statisticien en termes de taux de réponse, de délais, de pertinence et de précision des données fournies... pour tous les aspects de l'interaction entre service enquêteur et enquêté, il est possible de définir des indicateurs de qualité, parfois avant la réalisation de l'enquête (à travers des tests), mais plus souvent après : taux de réponse (totale ou par question), temps moyen de réponse[9], parfois indices de satisfaction, etc.

La qualité de l'enquête, a priori : le point de vue de la coordination statistique

Même si l'on va retrouver plusieurs caractéristiques, on utilise là un prisme un peu différent pour analyser la qualité.

La problématique est cette fois la suivante : soit une enquête, candidate pour faire partie des enquêtes du système statistique public.

L'on se pose deux questions :

- les objectifs visés, en termes de statistiques futures publiées, justifient-ils que l'on dépense les deniers publics pour la réaliser ?
- l'appareillage statistique mis en place est-il bien adapté pour atteindre ces objectifs ?

Le point important, ici, c'est que l'enquête n'a pas encore été réalisée et qu'il faut donc se prononcer sur sa qualité avec peu d'éléments.

La première question nous renvoie d'abord à la notion de pertinence de l'enquête[10]. Présente-t-elle un intérêt public, pour quels utilisateurs et pourquoi faire ? On ne peut se prononcer à ce sujet qu'à travers des réunions rassemblant statisticiens et organisations professionnelles touchées de près ou de loin par le thème de l'enquête. En France, ce sont les **formations du CNIS**, Conseil national de l'information statistique, qui jouent ce rôle, en délivrant un **avis d'opportunité**. Lorsqu'elles ont lieu, les réunions de **comités d'utilisateurs** des résultats de l'enquête, à caractère non formel, permettent aussi de se prononcer sur l'intérêt de l'enquête[11]. De tels comités jouent un rôle important avant le lancement des opérations, car ils donnent leur avis sur le questionnaire, et le font évoluer pour une meilleure utilité future, en éliminant les questions inutiles ou redondantes. Ce sont des réunions techniques, où il s'agit en pratique d'étudier dans le détail l'ensemble du questionnaire, les formations du CNIS n'ayant pas vocation à remplir cette fonction.

La pertinence n'est pas seule en cause pour juger de l'intérêt de l'enquête. En effet, sur un plan politique, on n'analyse pas isolément une enquête : elle fait partie d'un ensemble, d'un système d'enquêtes et de sources administratives, dans lequel on va devoir la situer, à la fois du point de vue de son contenu (vérifions que l'on n'a pas la même information ailleurs) et des unités interrogées (évitons d'enquêter toujours les mêmes). Ce qui est recherché, c'est la qualité de l'ensemble du système statistique. Cette fonction de coordination incombe elle aussi aux formations du CNIS.

La deuxième question se situe en aval et présuppose une réponse positive à la première question. Le contexte est ici clarifié : l'enquête est jugée opportune, le contenu et les objectifs sont validés. Il s'agit maintenant de vérifier que le plan de sondage répond aux objectifs visés, que l'on n'interroge pas trop d'unités, que toutes les précautions possibles seront prises au tirage de l'échantillon de l'enquête (pour mieux répartir la charge statistique pesant sur les unités), que le questionnaire est bien adapté (lisibilité, existence de l'information, coût de réponse acceptable). Sur ce deuxième aspect, on retrouve évidemment les critères de qualité indiqués plus haut, avec, en particulier, la nécessité de tester le questionnaire et de le donner à relire aux futurs utilisateurs des résultats de l'enquête. On retrouve également la volonté de donner une bonne image du système statistique public, la démarche de communication auprès des enquêtés relevant d'un choix politique.

9. Aux Pays-Bas et en Australie, entre autres, les questionnaires envoyés aux entreprises se terminent systématiquement par une question sur le temps passé à répondre.
10. Il faut bien distinguer la pertinence *de l'enquête* et la pertinence *des statistiques*. Dans le premier cas, on se demande, de façon générale, si les thèmes abordés par l'enquête présentent un intérêt : le terme d'*opportunité* est plus approprié. Dans le deuxième, l'enquête a été réalisée (on la suppose pertinente). Il existe alors une infinité de tableaux ou graphiques possibles, il s'agit de rechercher ceux qui ont une utilité.
11. L'existence de tels comités est parfois liée aux besoins de financements extérieurs (enquêtes régionales par exemple). Mais de plus en plus, ces groupes sont constitués sans idée de financement, mais pour faire connaître le travail et promouvoir l'utilisation des résultats pour des études théoriques ou débouchant sur des décisions.

En France, le travail d'expertise des instruments techniques de l'enquête est du ressort du **Comité du label**, qui délivre (ou ne délivre pas) un **avis de conformité**. Le Comité apprécie la qualité de l'enquête avant sa réalisation, à travers une documentation de synthèse rassemblée par le producteur[12] : comptes rendus de comités d'utilisateurs, comptes rendus de tests de questionnaires, imprécision induite par le sondage, justification du champ de l'enquête, de la taille de l'échantillon, du calendrier, etc. La France n'est pas la seule à procéder ainsi : les statisticiens australiens possèdent la *Statistical Clearing House*, qui effectue le même travail et analyse en outre la manière dont les données seront contrôlées et redressées. Dans tous les cas, de multiples informations sont ainsi demandées au producteur, dans une optique d'évaluation a priori de la qualité de l'enquête.

La qualité de la réalisation de l'enquête : le point de vue de la production

Malgré les progrès considérables de l'informatique, réaliser une enquête demeure une opération lourde, coûteuse, présentant de multiples difficultés. Pour travailler dans de bonnes conditions, le responsable de l'enquête doit disposer d'une série d'indicateurs afin de savoir à tout moment où en est le travail, mais aussi, près de la fin, quel sera le degré de fiabilité des résultats obtenus. Ainsi, le gestionnaire de la production intervient non seulement en amont de la réalisation de l'enquête (questionnaire, plan de sondage), mais également pendant (tableau de bord de la gestion) et après (analyse de la fiabilité des résultats).

Le **référentiel**[13], commun à toutes les enquêtes d'entreprises et premier en jeu dans l'enchaînement des opérations, constitue un élément fondamental ; ce sera la base dans laquelle on effectuera ensuite les tirages d'échantillons. La qualité est ici tout simplement liée à la fiabilité et

à la fraîcheur des variables caractérisant les unités. Ainsi, dans un répertoire d'entreprises, la proportion de NPAI (entreprises non présentes à l'adresse indiquée) est un critère important, qui conditionne l'efficacité de la collecte[14] ; c'est donc la validité de la variable «adresse de collecte» qui est en cause. Le taux de « faux actifs » (unités mortes considérées à tort comme actives) apparaît également comme un indicateur crucial employé très largement dans les INS. Par ailleurs, la fiabilité de la variable « activité principale » est à de nombreux égards essentielle, d'où l'intérêt d'estimer un taux d'unités mal classées.

Les premiers indicateurs importants pour la production se résument à quelques dates-clés (dates d'envoi et de réception des questionnaires) et à des comptages liés à la **gestion de la collecte**. Cela commence par le nombre de questionnaires envoyés (car ils le sont souvent en plusieurs étapes). On détermine également la situation de chaque enquêté par rapport à la collecte, en termes d'aptitude à répondre, ce qui permet de produire en premier lieu le **taux de réponse**, mais aussi quelques comptages simples par catégorie de réponse. Plusieurs situations peuvent en effet se présenter. L'unité peut accepter de répondre, elle peut refuser explicitement. Elle peut aussi ne pas être présente à l'adresse indiquée : cela joue un rôle important en statistique d'entreprise, où la collecte s'effectue le plus souvent par courrier ; dans les statistiques auprès des ménages, effectuées le plus souvent par déplacement d'enquêteur ou par téléphone, on distingue les « impossibles à joindre », les ALD (absents de longue durée). L'unité peut aussi se révéler inapte à répondre. Enfin, les enquêtes auprès des entreprises exigeant en général deux ou trois courriers, on gérera aussi la répartition des non-réponses (qui ne sont, le plus souvent, ni des refus ni des NPAI) selon ces différentes vagues de courrier.

On élabore aussi des indicateurs élémentaires sur la situation de l'unité quant au champ de l'enquête, qui induisent un changement par rapport à la base de sondage. L'unité peut avoir changé en quelque sorte de « statut » pour cause de restructuration (c'est une situation que l'on retrouve aussi dans les statistiques sociales, avec les éclatements ou fusions de logements), auquel cas la délimitation de l'unité statistique a changé. En statistiques économiques, la mise hors champ peut être liée à n'importe quelle variable de définition de champ directement collectable : état d'activité (l'entreprise peut être morte ou en sommeil), activité principale (cas très fréquent), effectif, voire catégorie juridique.

Après la collecte, au fur et à mesure, ont lieu la **saisie** et le **codage**, pour lesquels plusieurs mesures de qualité sont envisageables, même si elles ne sont en général pas effectuées : taux d'erreur à la saisie (ce qui requiert une double saisie sur un sous-échantillon), taux d'erreur de codage, manuel ou automatique (un double codage semble là aussi inévitable). Le taux de rejet par le codage automatique (libellés impossibles à coder automatiquement) peut en revanche être immédiatement obtenu.

Dans le reste du processus de production, il faut mettre en exergue un aspect reconnu comme particulière-

12. Cf. dossier-type du Comité du label.
13. Pour les statistiques économiques, le répertoire des unités économiques, en l'occurrence SIRENE.
14. À telle enseigne que le taux de NPAI est un critère de base pour les entreprises de marketing direct, par exemple.

ment coûteux : le **contrôle** et le **redressement** des données (ce que les Anglo-Saxons appellent *Data Editing and Imputation*). Dans les enquêtes, ces opérations peuvent donner lieu à de nombreuses métadonnées : le fait que la variable ait été vérifiée ou non, qu'elle ait été modifiée ou confirmée en l'état par le gestionnaire, le mode de redressement utilisé s'il y a lieu... L'existence même de telles informations est une condition sine qua non au calcul d'indicateurs de qualité fort utiles : par exemple, le pourcentage d'unités jugées erronées par le contrôle (pour chaque variable), les parts respectives des unités vérifiées (resp. confirmées, modifiées) manuellement, l'écart entre données brutes et données modifiées. Cela permet en particulier de déterminer si l'on ne « survérifie » pas les questionnaires. D'autres indicateurs s'appliqueront aux redressements par imputation : taux d'unités redressées, poids relatif (sur les agrégats), poids des unités extrapolées pour non-réponse totale, répartition par type de redressement... On détermine ainsi quelle est la part des agrégats qui est « inventée ».

Enfin, même si on les utilise peu à ce stade, les calculs de **variance** (mais aussi, lorsque c'est possible, de biais) ont un intérêt évident pour la gestion de la production. On l'a vu, le calcul se révèle délicat dans la pratique, car l'erreur totale résulte principalement de deux facteurs (sondage et non-réponse). Mais en tant qu'indicateur pour orienter la production (vérifications par les gestionnaires, suivi des non-répondants), on peut se permettre d'avoir recours à des évaluations grossières[15].

Que tirer de tout cela ? On a vu que pendant l'enquête, les données de gestion procurent au responsable de la production un tableau de bord pour avoir à tout moment des idées précises sur l'état d'avancement du processus. Cela constitue en même temps une aide pour savoir où les équipes de gestionnaires doivent porter leurs efforts. Ainsi, si l'objectif est d'avoir des agrégats par secteur d'activité, on peut déterminer régulièrement les coefficients de variation par secteur, et faire porter l'effort (suivi des unités non-répondantes, vérifications manuelles) sur les secteurs où la précision est la plus faible.

Les indicateurs mesurés (part des unités hors - champ, taux de réponse, poids des unités redressées, coefficients de variation[16]) évoluent au fur et à mesure de l'avancement de la production et permettent de construire des tableaux de bord. Lorsque l'enquête est achevée, ils servent d'indicateurs de qualité des statistiques, et ils ont l'avantage d'être comparables d'une enquête à l'autre.

Mais soulignons que l'élaboration de tels indicateurs a un coût élevé et qu'il est possible de gagner du temps en normalisant, en ayant des méthodes communes. Comme le dit J.-C. Labat dans son rapport sur les indicateurs de qualité dans les statistiques démographiques et sociales, *l'utilisation d'outils généraux* (CAPI, SICORE, CALMAR), outre ses avantages en matière de qualité, d'homogénéité et d'efficacité, facilite la production d'indicateurs normalisés et assure une meilleure comparabilité de ces indicateurs.

Une question de compromis...

Les indicateurs relatifs au contrôle permettent de mieux gérer le compromis entre le coût[1] de la vérification des questionnaires et le coût d'une mauvaise fiabilité des résultats. Certes, plus on effectue de contrôles, meilleure est la qualité finale des résultats ; à l'inverse, il est largement inutile sur un plan statistique de vérifier beaucoup de questionnaires. Dès lors, un contrôle-redressement de bonne qualité est un processus qui trouve un juste milieu : avoir une bonne qualité finale des données, tout en limitant la quantité de vérifications manuelles. C'est une différence notable entre qualité des statistiques *et* qualité du processus de production *tel que le perçoit le responsable de l'enquête : pour celui-ci, un taux de contrôle manuel faible sera plutôt un gage de qualité (entre autres, cela améliore les délais de diffusion), alors que cette information n'est pas nécessairement utile pour juger de la fiabilité des statistiques obtenues.*

1. Nous avons indiqué plus haut que le coût ne devait pas faire partie de la qualité des statistiques et ce du point de vue du client de celles-ci. Il ne faut pas y voir une contradiction. En effet, du point de vue du responsable d'une enquête*, le coût est un critère de qualité évident du* processus de production *qu'il a à gérer. La qualité n'est pas une sorte de pur concept :* la qualité de quoi ? perçue par qui ?

***CAPI**, collecte assistée par informatique : l'enquêteur n'a plus de questionnaire papier, mais un questionnaire électronique et un ordinateur portable, ainsi qu'un logiciel permettant la saisie de l'information et certains contrôles.*

***SICORE**, système informatique de codage des réponses aux enquêtes, est un logiciel de codage automatique (libellés de communes, de professions...).*

***CALMAR**, calage sur marges : Ensemble de macros SAS permettant d'effectuer des repondérations (i.e. de calculer de nouveaux poids, à partir des poids de sondage) en prenant en compte des variables auxiliaires.*

La qualité totale : le point de vue du manager

Initialement, nous avons vu que *maîtrise de la qualité* et a*ssurance de la qualité* étaient les deux concepts de base. En statistique, le rapport sur la qualité, tel qu'Eurostat le préconise, relève d'une démarche d'*assurance qualité* : il s'agit bien de se doter d'un cadre normalisé en vue de donner des garanties à des utilisateurs.

15. En particulier : faire les calculs avec « l'échantillon des répondants », sans autre forme de procès. Ce qui revient à considérer que les unités répondantes sont un sous-échantillon de l'échantillon initial.

16. Toutes ces mesures sont évidemment à effectuer par catégorie de population (région, secteur...) et non sur la population totale.

L'expérience suédoise

En Europe, ce sont les statisticiens suédois qui se sont orientés vers la qualité totale. Les raisons en sont fort simples. Une grande réforme de l'INS suédois a eu lieu en 1993, avec une concurrence accrue, un budget moindre, et une volonté de privatisation partielle des activités. Il s'agit là d'un contexte habituel pour l'adoption d'une démarche TQM : en statistique ou ailleurs, c'est souvent la pression externe qui conduit à entrer dans une logique de qualité totale. Après la réforme, Statistics Sweden est devenue pour moitié une agence gouvernementale, pour moitié une sorte de firme statistique. Selon ses dires, elle est aujourd'hui plus orientée vers les utilisateurs, et ses processus de production sont beaucoup plus efficaces.

Les objectifs affichés étaient les suivants : répondre aux besoins des utilisateurs, améliorer la qualité des processus et des produits, développer un système pour les statistiques officielles, mettre en place de bonnes relations avec les répondants, impliquer l'ensemble du personnel et utiliser au mieux les compétences. Soulignons que l'approche TQM a été choisie dans un pays où le contexte n'est pas le même qu'en France : la question de la concurrence y est importante, alors que la notion de service public se révèle très peu présente.

Statistics Sweden a maintenant, semble-t-il, bien intégré les changements radicaux dans la manière de travailler provoqués par le TQM. Cela dit, une grande implication de tout le personnel, et en premier lieu des directeurs, a été nécessaire. Plus de 110 projets TQM ont vu le jour, ayant mobilisé 500 personnes sur 1300. Il est d'ailleurs apparu rapidement qu'il était inutile de sélectionner des projets pour en choisir certains qui seraient TQM : à partir du moment où l'Institut franchissait le pas, intégrait totalement la démarche, le retour en arrière n'était plus possible et distinguer les projets TQM des autres n'avait pas grand sens.

Mais, de l'aveu même des responsables de cette opération, la rénovation n'a pas été vécue sans douleur. Au départ, au sein de l'Institut, 60 % étaient pour, 20 % contre, 20 % sans opinion. De nombreuses barrières psychologiques sont apparues, avec des réactions du type : on sait ce que les utilisateurs veulent, nous avons toujours travaillé comme ça, TQM, c'est juste un mot de 3 lettres*. Le démarrage a été difficile, avec une forte résistance au changement et une notable absence d'enthousiasme. Est apparu également à ce moment-là un problème de ressources et de manque de formation à ce changement au niveau de la hiérarchie intermédiaire.*

Comment ont-ils procédé, techniquement, pour entrer dans la logique TQM ? Pour chaque étape du processus, ont été sélectionnées plusieurs variables-clés. Par exemple, pour la réduction du taux de non-réponse, on trouve parmi les variables-clés : le taux de non-réponse dans une certain nombre de catégories (par région, par secteur…), le taux de non-réponse selon le mode de collecte, le nombre moyen de rappels et son évolution au cours du temps, le coût de la collecte pour les 10 % derniers répondants, le taux de non-réponse partielle pour chaque variable...

Puis ont été mis en place, de façon systématique, des outils simples, voire très élémentaires, permettant de bien comprendre et maîtriser ce qui se passe : diagrammes cause-effet, diagrammes de flux permettant de visualiser le processus, mesures permettant de déterminer les variables critiques, tableaux de classement des problèmes par difficulté décroissante, enquêtes de satisfaction auprès des utilisateurs, larges et fréquents brainstormings (de type cercles de qualité).

Chacune des phases de la production a fait ainsi l'objet d'un projet TQM. En voici quelques exemples : réduction de la non-réponse, vérification des données, conception de questionnaire, estimation de variance, traitement de la non-réponse. L'implication des gestionnaires d'enquête dans ces projets a été beaucoup plus grande que par le passé, notamment avec les brainstormings où l'on cherche à recueillir le maximum d'idées sans être critique a priori. Ce que ces projets ont produit a été ensuite analysé par des examinateurs, formés à ce métier, qui ont attribué à chaque projet un certain nombre de notes, selon un protocole fixé à l'avance, en l'occurrence les Swedish Quality Awards*.*

La *maîtrise de la qualité* des statistiques devrait, quant à elle, s'observer plutôt au niveau de la production : il s'agit, pour bien maîtriser le processus, de répertorier, documenter, vérifier et faire évoluer chacune des opérations qui le composent. Vaste programme.

La **qualité totale** constitue l'étape suivante, voire ultime. Les « qualiticiens » ont constaté qu'assurance de la qualité et maîtrise de la qualité n'étaient que des mécanismes, et que le problème était de les actionner correctement, ce qui est la fonction du management. Il est apparu que l'implication de la hiérarchie de l'entreprise, depuis la tête, permettait d'insuffler une démarche qualité à l'ensemble de l'organisation, en visant à mobiliser tout le personnel. Cela justifie l'expression de « qualité totale », ou encore de *Total Quality Management (TQM).* Qu'en est-il en statistique ? Clairement, l'approche TQM a ceci de particulier qu'elle ne peut s'appliquer à une seule enquête, puisque c'est une démarche globale, qui implique l'ensemble de l'INS et qui a des conséquences importantes sur les travaux des agents. Mais ne tournons pas autour du pot : le TQM, c'est d'abord une optique de réduction des coûts, de meilleure maîtrise de l'ensemble, de flexibilité accrue permettant de faire face rapidement à des demandes et de s'imposer dans un contexte concurrentiel. Dans l'INS, le manager[17] est le point nodal des contradictions, confronté à quatre impératifs difficiles à concilier : le triptyque habituel gestion du personnel, contraintes budgétaires, satisfaction des utilisateurs, mais aussi la limitation de la charge d'enquêtes pesant sur les enquêtés.

La gestion de la qualité totale constitue donc, pour une organisation, une opération très lourde, une véritable révolution culturelle, qui serait vouée à l'échec sans une vraie adhésion de la haute hiérarchie. Elle serait tout aussi difficile à mettre en place sans l'adhésion des personnels. Car si le sigle TQM provoque souvent l'ironie, semble sonner creux, les conséquences de son application sont très concrètes : dès lors que la volonté est réelle, l'organisation dans son ensemble est fortement modifiée. Ses

17. Le terme est volontairement vague, et s'applique plus à la situation de services statistiques privatisés.

conséquences sociologiques sont importantes, la démarche qualité étant à double tranchant (cf. Cochoy & al.) : en associant l'ensemble des personnels et en leur fournissant la reconnaissance écrite de leur travail, elle valorise les individus, mais, à l'inverse, elle offre des possibilités accrues de contrôle de ce travail individuel.

L'amélioration de la qualité

Les mesures de qualité selon des critères reconnus ont un intérêt en soi, entre autres pour donner aux utilisateurs, aux « clients », des garanties de qualité du produit. Mais bien évidemment, le fait de posséder indicateurs et documentations normalisés du processus de production statistique permet aussi d'accroître la qualité et de visualiser cet accroissement.

De fait, l'amélioration de la qualité fait partie intégrante de la réflexion générale sur la qualité, mais elle n'est applicable que si l'on s'est déjà préoccupé d'effectuer des mesures standardisées. En effet, on ne peut évaluer l'écart de qualité entre deux situations A et B (par exemple une enquête en 1999 contre la même en 1998) que si l'on possède des indicateurs sur A et B, comparables entre eux.

Si l'on se réfère à l'expérience des INS suédois et britannique, le simple fait de « décortiquer » l'ensemble du processus de production tâche par tâche[18] (préalable indispensable) a permis de faire émerger certains problèmes qui étaient ignorés jusque-là (vérifications redondantes, erreurs dans les consignes de saisie), et certains ont pu être rapidement résolus. La première étape de l'amélioration de la qualité consiste donc souvent à dénicher les principales sources d'inefficacité et d'erreur.

Mais pour faire évoluer les processus vers une plus grande qualité, se limiter aux aspects techniques ne suffit pas : le facteur humain est absolument essentiel. Le fait d'associer les personnels à la documentation, à la standardisation de leurs procédures de travail, les conduit à prendre conscience des différents aspects de la qualité, à mieux appréhender l'impact de ce qu'ils réalisent. Le statisticiens britanniques et suédois ont observé une réelle influence des mesures de qualité sur le comportement des gestionnaires d'enquête, ce qui a conduit tout naturellement à des améliorations substantielles.

Sur un plan méthodologique, une façon courante de procéder consiste à rechercher quelles sont les « meilleures pratiques » (*Current Best Practices*), afin de les utiliser éventuellement en lieu et place des techniques courantes. Pour chaque étape importante (par exemple : suivi des non-répondants, méthodes d'imputation, macrocontrôles), l'INS édite un manuel de *Current Best Practices*, qui sert ensuite de référence pour le traitement d'enquêtes. Un tel manuel peut tout à fait évoluer : l'innovation dans les méthodes est un aspect très important de l'amélioration de la qualité.

En termes d'organisation, la veille technologique sur les méthodes ne doit pas être du ressort du service chargé de la qualité. Dans l'INS, il faut en effet distinguer l'unité « méthodologie » (qui élabore des méthodes nouvelles, se tient au courant de ce qui se crée dans les autres INS, etc.), et l'unité « qualité », qui applique les méthodes proposées par l'unité méthodologie, et surtout met en oeuvre et organise l'ensemble de la démarche qualité.

Très concrètement, l'amélioration de la qualité passe donc par :
- la définition de principes et d'objectifs clairs et quantifiés pour chacune des opérations élémentaires de la production statistique ;
- une analyse sans concession du fonctionnement réellement à l'oeuvre pour chaque étape ;
- une description des techniques existantes, pour chaque étape, que l'on pourra utilement confronter avec ce qui est fait réellement.

De façon plus générale, et si l'on vise une progression globale et à long terme, l'amélioration de la qualité nécessite un travail de **coordination,** de **normalisation** et de **docu-**

Petite remarque sur l'amélioration continue des programmes de traitement statistique

On a souvent tendance à penser qu'améliorer la qualité en jouant sur quelques instruments bien ciblés, par exemple la modification des chaînes de production et notamment des programmes informatiques, est au fond une opération assez simple : c'est une erreur. Si l'on souhaite faire évoluer très régulièrement les programmes à des fins de qualité, ceux-ci doivent être pensés autrement : en particulier, ils doivent avoir un degré de généralité suffisant pour être adaptables (programmation déclarative, paramétrage).

De même, on a coutume de croire que le processus de traitement de l'information, qui vise à obtenir comme « produit » une base de données individuelles propre, a pour seul output cette base de données ; c'est là encore une erreur profonde, car il y a bien deux outputs : une base de données et un ensemble d'informations sur le processus, qui permettront par la suite de l'améliorer. Ainsi, le processus de production s'autogénère en partie.

Les outils permettant la modification de la chaîne de traitement doivent donc être envisagés dès le départ : d'une certaine manière, cette chaîne doit être dotée d'un « œil », d'une capacité d'auto-observation, de la faculté de mesurer ses propres performances ; mieux encore, cette chaîne de traitement doit être également aisément transformable. En d'autres termes, cela revient à dire qu'il faut doter l'outil de la plasticité nécessaire à ses futures évolutions.

18. Les outils d'administration de processus, dits aussi de workflow, tendent à se généraliser. Ils permettent de représenter graphiquement le processus, de mieux le visualiser et le gérer.

mentation dans les techniques utilisées[19]. Sans cela, il est difficile de parler de qualité : c'est en se référant à des standards reconnus et documentés, reliés à la qualité finale du produit statistique, que l'on peut se poser la question de la qualité d'un travail qui y concourt. Sans norme commune ni possibilité d'un jugement extérieur, chacun peut se définir isolément ses propres critères de qualité (puis s'autodécerner des lauriers).

Où en est-on ?

Nous n'en sommes plus aux grands mots, aux idées creuses : aussi bien en France qu'en Europe, il existe une véritable effervescence autour de la qualité, qui se traduit par des avancées visibles, concrètes.

Au niveau européen, le groupe de travail d'Eurostat sur l'évaluation de la qualité des statistiques, déjà cité, a permis de définir un modèle de rapport qualité incluant des indicateurs quantitatifs et des fiches de documentation sur les méthodes utilisées (exemple : traitement du biais de non-réponse). Il s'agissait, soulignons-le, d'une optique d'assurance de la qualité. Ce rapport a déjà été testé dans plusieurs pays. Certains des indicateurs qui y figurent, jugés particulièrement importants, sont désormais demandés systématiquement par Eurostat aux États membres (coefficients de variation dus au sondage et à la non-réponse, taux de réponse, erreurs sur la base de sondage, éléments relatifs à la comparabilité temporelle) dans le cadre des règlements européens sur les statistiques d'entreprise. Il existe également un LEG (Leadership Group) européen sur la qualité, plus récent, mené par les statisticiens suédois, et qui est orienté sur les questions de qualité totale[20]. La dernière table ronde sur les répertoires d'entreprises a fait la part belle à plusieurs articles de fond sur la qualité des répertoires. Le colloque annuel sur le thème du *Data Editing*, qui s'est déroulé en juin 1999 à Rome, a fait l'objet, pour la première fois, de nombreuses contributions sur l'impact des traitements (contrôle, imputation) sur la qualité des statistiques. Si l'on analyse la situation pays par pays, on observe de nombreuses évolutions autour de ce thème, notamment en Suède, mais aussi au Royaume-Uni, aux Pays-Bas et en Italie par exemple.

En France, il existe de nombreux outils généraux et documentés, ce qui, on l'a vu, facilite considérablement l'approche qualité. Outre CAPI, SICORE et CALMAR, déjà cités, on peut également mentionner OCEAN pour le tirage d'échantillons d'entreprises, ou encore CITRUS pour l'information sur les restructurations d'entreprises. Enfin et surtout, l'échantillon-maître, pour les statistiques démographiques et sociales, et SIRENE, pour les statistiques d'entreprise, jouent un rôle essentiel de normalisation.

Un certain nombre d'indicateurs de qualité ont également été définis, ou sont en cours de définition. Ainsi, le dossier-type du comité du label (configuration entreprises), effectif depuis 1997, permet de fixer un cadre d'analyse des dossiers d'enquête pour évaluer la « qualité a priori ». Le groupe de travail sur la qualité de SIRENE a conduit à la rédaction d'un rapport proposant des principes généraux et des indicateurs pour la qualité du répertoire. Une « charte qualité » est en cours de rédaction, à destination des organisations professionnelles qui gèrent des enquêtes de branche dans l'industrie. Tout récemment, un modèle de rapport qualité pour les enquêtes auprès des entreprises a été soumis aux services enquêteurs : l'objectif est que ce rapport, après modifications diverses, serve ensuite de cadre pour les bilans qualité des enquêtes auprès des entreprises.

Sur un plan méthodologique, il faut d'abord souligner le travail de fond qui a été fait sur le calcul de la variance grâce au logiciel POULPE (qui est en soi un instrument de qualité), et qui a permis de développer l'utilisation des calculs de précision dans les enquêtes auprès des ménages. En statistique d'entreprise, les travaux récents portent sur l'estimation du taux de faux actifs[21], les indicateurs de qualité de production dans les EAE, enquêtes annuelles d'entreprise[22], ou la méthodologie de comparaison entre sources (appliquée à la comparaison entre EAE et données fiscales).

Cela dit, la possibilité pratique d'évaluer la qualité n'est pas seulement liée aux outils, mais aux données de gestion et **métadonnées** conservées dans les bases de données d'enquête. Là encore, nous disposons d'un outil adapté avec le DDS, dictionnaire de données statistiques. La mise en commun se joue aussi au niveau des données. Dans ce cas, soit la coordination est organisée par les statisticiens (tronc commun d'enquêtes ménages, tronc commun des enquêtes annuelles d'entreprise), soit leur est imposée de fait (cas des sources administratives : état civil, déclarations fiscales annuelles, déclarations de TVA, déclarations annuelles de données sociales).

Maîtriser un processus tendu vers l'utilisateur

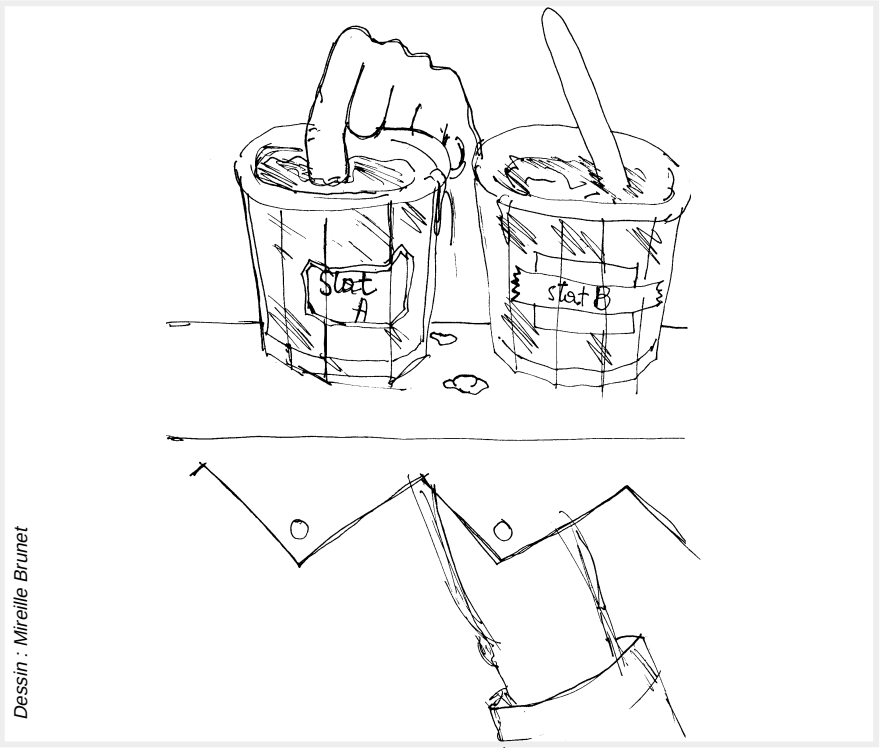
Différents acteurs, on l'a vu, interviennent dans le processus de production statistique, animés par des objectifs de qualité qui leur sont propres, mais qui doivent concourir à la qualité finale des statistiques.

Au total, on constate donc que la qualité, c'est en premier lieu la maî-

19. Il ne s'agit pas de normaliser les besoins, ou les types de statistiques produites, en raison des importantes différences qui existent entre statistiques régionale et nationale, et de l'hétérogénéité des besoins régionaux.
20. Ce groupe se réunira à Paris en mars 2000.
21. Cf. l'article qu'avait rédigé H. Mariotte en prévision de la dernière réunion de la table ronde sur les répertoires d'entreprises.
22. Cf. V. Lamour, *Indicateurs de qualité pour les enquêtes annuelles d'entreprise*, rapport de DEA de statistique, septembre 1998.

trise du processus : savoir ce que l'on fait et pourquoi, savoir de quelle succession d'opérations est issu le « produit » statistique (traçabilité), être conscient des faiblesses, et bien sûr savoir évaluer. **Sans mesure, il n'y a pas de qualité**. Une fois le processus maîtrisé, une fois qu'une batterie d'indicateurs standardisés est mise en place, il est beaucoup plus facile de s'interroger sur l'amélioration de cette qualité, sur les actions à mettre en œuvre, mais aussi sur la possibilité de comparer entre eux des produits statistiques. À ce sujet, on pourra regretter qu'il n'y ait pas réellement à l'Insee de norme commune, ni du point de vue des concepts, ni même du point de vue des outils (ont pourtant été développés de nombreux instruments à vocation générale...), sur ce que doit être, fondamentalement, le processus de production statistique.

Deuxième point, à ne jamais perdre de vue : le processus de production statistique n'a de sens que s'il est tendu vers **l'utilisateur** (effectif ou potentiel), qui veut avoir des données utiles, lisibles, fiables, fraîches, cohérentes. En effet, l'intérêt du chiffre ne se justifie, pour l'essentiel, que par les actions qu'il conduira à mener. Or cet aspect se trouve souvent en dehors de l'univers du statisticien, bien que l'on insiste sur la pertinence en tant que critère de qualité. Le fait est que notre tendance naturelle ne nous pousse pas à nous intéresser aux usages. Il y a là un paradoxe, même si l'on ne peut certainement pas limiter le travail du statisticien à la satisfaction des besoins d'un ensemble prédéfini d'utilisateurs[23].



Dessin : Mireille Brunet

Pour l'instant, nos habitudes de travail excluent une approche normalisée et généralisée de la qualité. Mais les mentalités évoluent : il est tout à fait possible d'engager dès aujourd'hui un travail de fond sur la qualité de nos statistiques. Ce travail ne portera toutefois ses fruits que si la volonté commune, à tous niveaux, est réelle, et appuyée sur une démarche institutionnalisée. Un préalable indispensable est une coordination forte des méthodes. Il faut aussi adopter des principes communs, notamment la conservation des métadonnées, et mettre en place des indicateurs de qualité standardisés et comparables. Enfin, une diffusion systématique de ces résultats et des fiches de documentation méthodologique, permettant la comparaison, sera également nécessaire.

Pascal RIVIÈRE
Insee, chef de la division "harmonisation d'enquêtes auprès des entreprises"

23. Contrairement à une autre « machine à mouliner des chiffres », la comptabilité d'entreprise.

Bibliographie

AFNOR - *Gérer et assurer la qualité, qualité et efficacité des organisations*,
Paris, Afnor, 1996.

P. Arondel, **R. Depoutot** - *Overview of Quality Issues when Dealing with Socio-Economic Products in an International Environment*,
Actes du congrès de l'ASU, Rennes, 1998.

CNIS - *Dossier-type du Comité du label (configuration entreprises),*
Note CNIS n° 108/D130, mars 1997.

F. Cochoy, **J.-P. Garel**, **G. de Terssac** - *Comment l'écrit travaille l'organisation : le cas des normes ISO 9000,*
Revue française de sociologie,XXXIX-4, pp. 673-699, 1998.

L. Cruchant - *La qualité*,
Que sais-je ? juillet 1998.

P. Domergue - *L'approche de la qualité à l'Insee*,
Note Insee n° 964/D201, mai 1999.

P. Engström, L. Granquist - *Improving Quality by Modern Editing,UN/ECE Work Session on Statistical Data Editing,*
Rome,Working Paper N° 23, juin 1999.

Eurostat - *Quality in Structural Business Statistics,*
N° D3/96/03, 1996.

Eurostat - *Proposal for a Quality Report on Structural Business Statistics,*
N° D3/Quality/96/04 final, 1996.

Eurostat - *3e réunion du CPS, LEG sur la qualité*
CPS 1999/32/11/FR, mars 1999.

G. Griffiths, S. Linacre - *Quality Assurance for Business Surveys*
Business Survey Methods, Wiley, pp. 673-690, 1995.

T. Holt & T. Jones - *Quality Work and Conflicting Quality Objectives*
84e conférence des DGINS, Stockholm, 28-29 mai 1998.

J.-C. Labat - *Les indicateurs de qualité des opérations statistiques de la direction des statistiques démographiques et sociales,*
Rapport de mission, 1998.

L. Lyberg & al. - *Survey Measurement and Process Quality*,
Wiley 1997.

A. Manzari, G. Della Rocca - *A Generalized System Based on a Simulation Approach to Test the Quality of Editing and Imputation Procedures,*
UN/ECE Work Session on Statistical Data Editing, Rome, Working Paper N° 13, juin 1999.

P. Rivière - *Proposition de structure-type de bilan qualité pour les enquêtes auprès des entreprises*,
Note Insee n° 228/E210, octobre 1999.

M. Volle - *Système de pilotage de l'entreprise*
(site Internet www.volle.com).

M. Volle - *Rapport général sur l'évolution à moyen terme de l'appareil statistique français*,
présenté au CNIS en 1989.

Arguments

Intégrer l'économie spatiale dans les études régionales

La demande d'informations économiques et sociales à un niveau géographique fin a toujours été importante, et la décentralisation des années quatre-vingt n'avait fait qu'accentuer une tendance déjà à la hausse.

Curieusement, la globalisation des économies, la mondialisation, semblent susciter un nouveau surcroît d'intérêt pour la connaissance du « local ». Le paradoxe n'est qu'apparent. Les possibilités accrues de mobilité du capital liées à la mondialisation éveillent en effet de nouvelles interrogations sur le degré de dépendance des territoires vis-à-vis de formes d'organisation économique qui se sont largement internationalisées. L'évolution rapide des technologies, l'apparition de nouvelles formes de coopération entre les entreprises, ont également fait émerger de nouvelles questions, portant sur l'ancrage territorial de l'innovation[1] et de la coordination inter-entreprises. *La globalisation, au contraire de l'internationalisation ou de la multinationalisation, ne se réduit pas à un simple phénomène de répartition des activités, mais elle exprime la multiplicité des territoires d'où l'innovation peut émerger selon des processus et des combinaisons variés* [Denis Maillat, 1995].

Dans ce contexte, les chargés d'études socioéconomiques des directions régionales de l'Insee apparaissent particulièrement fondés à prendre en considération les apports de l'économie spatiale.

L'espace à sa juste place

Depuis quelques années déjà, l'économie spatiale est en ébullition, ainsi qu'en témoignent les nombreux ouvrages qui lui sont consacrés. Et si concepts et théories ne sont peut-être pas bien établis, c'est plutôt d'un trop-plein qu'il s'agit. En effet, les angles d'attaque sont nombreux et variés. Ainsi, certains économistes spatiaux s'intéressent à des combinaisons particulières de facteurs de production issues de l'histoire locale, d'autres étudient les effets de l'apprentissage[2] et de l'innovation sur le local, ou inversement les effets du local sur l'apprentissage et l'innovation, d'autres encore, se situant dans une perspective de création de ressources locales spécifiques, considèrent la capacité des acteurs locaux à résoudre leurs problèmes productifs...

Parallèlement, on assiste à l'émergence de préoccupations communes entre économie spatiale, économie industrielle, économie publique locale, et économie du commerce international. Encore plus révélateur, les convergences entre économie spatiale et géographie économique sont devenues tellement évidentes que les deux sciences paraissent aujourd'hui quasiment confondues.

Économie et géographie

Historiquement, géographes et économistes sont partis de points de vue bien différents pour ne pas dire opposés.

Les géographes ont d'abord nommé les lieux et dressé des cartes, puis montré l'influence des lieux sur les comportements humains (ainsi, la présence d'un fleuve ou d'une montagne, ou l'existence de ressources naturelles, impliquent des modes de vie différents). Dans les années cinquante, leur démarche va se transformer, privilégiant désormais l'étude des interactions entre les lieux, jusqu'alors considérés de façon séparée.

L'économie a suivi un chemin inverse. Ses investigations ont longtemps porté sur *un monde merveilleux sans dimensions spatiales* (Isard), *un monde ponctiforme* (Ponsard). Puis l'espace y a été introduit, d'abord sous la forme de la distance, ensuite sous celle des relations entre agents économiques localisés.

Aussi est-il maintenant difficile (et d'ailleurs est-ce bien nécessaire ?) de distinguer entre géographie économique et économie spatiale.

Espace et territoire

L'économie spatiale plonge ses racines dans les théories de la localisation : dans un comportement de recherche d'un profit maximum, les entreprises sont censées fonder leurs décisions de localisation sur des considérations de coût et sur l'importance du marché local.

Ces théories se sont progressivement complexifiées, dans plusieurs directions : prise en compte de varia-

1. Conception et mise au point de nouveaux produits ou de nouveaux procédés.
2. L'apprentissage résulte de la pratique et des interactions entre les agents, c'est de lui que naît le savoir-faire.

bles de plus en plus nombreuses, passage d'une modélisation des choix individuels de localisation à une approche de niveau collectif, évolution vers une analyse dynamique.

Mais au delà de ces questions de localisation, les économistes spatiaux s'intéressent aussi et surtout à la façon dont s'organisent les relations économiques et sociales sur le territoire. En effet, ce qui est en cause, dans l'espace économique, ce n'est pas seulement la fixation ponctuelle des objets économiques et sociaux sur l'espace physique, mais aussi et surtout les rapports qu'y entretiennent ces objets, fluctuants dans le temps et selon les lieux.

Les pratiques spatiales font ainsi de chaque lieu un espace original, avec ses propres structures historiques, sociales, culturelles et économiques. Dans cette approche, s'interroger sur les modes d'organisation de l'espace économique revient donc à s'interroger sur les spécificités des différents territoires, en rapport avec l'évolution des formes dominantes d'organisation économique et technologique au plan national.

Bien sûr, les territoires sur lesquels va se pencher l'économiste spatial ne doivent être ni trop petits ni trop grands : étudier un trop petit territoire ne mènerait qu'à des évidences, étudier un trop grand territoire conduirait à noyer les phénomènes localisés dans une moyenne sans grand intérêt. Pour situer les idées, les concepts de l'analyse spatiale sont particulièrement adaptés à l'analyse du fonctionnement des agglomérations urbaines, dans lequel la notion de proximité joue un rôle-clé.

La proximité

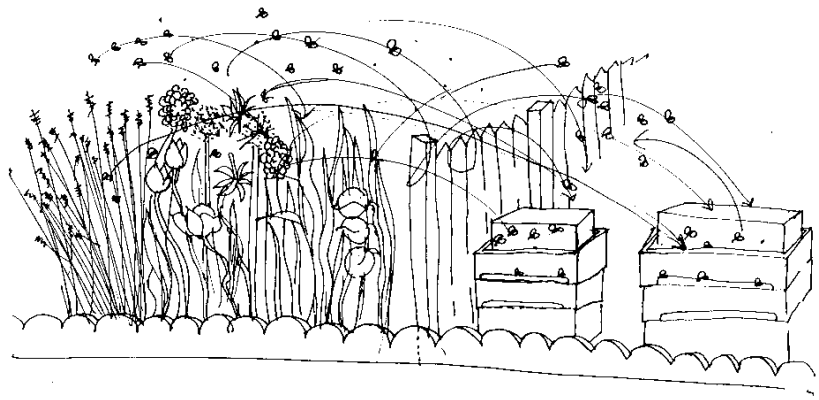
Marshall, considéré comme le chef de file de l'école néoclassique, avait observé au début de ce siècle que des entreprises voisines pouvaient tirer de leur seule proximité (indépendamment de tout échange à caractère marchand) des bénéfices mutuels, favorisant leur développement et donc celui du territoire sur lequel elles étaient implantées. Le concept d'externalité de proximité était né. À la même époque, Weber élaborait une théorie aux termes de laquelle le choix de la localisation d'une unité de production devait également s'appuyer, au delà des habituelles considérations sur les coûts de transport, sur les économies réalisables en matière de salaires (le coût de la main-d'oeuvre varie selon les lieux), ainsi que sur les économies nettes que pourrait procurer un voisinage adapté.

Cinquante ans plus tard, Isard distinguera les économies de localisation, externes à l'entreprise mais liées à la nature de ses activités, et les économies d'urbanisation, externes à l'entreprise et non liées à ses activités, tenant en particulier à la concentration de la population et aux infrastructures disponibles.

Aujourd'hui, proximité, externalités de proximité et économies d'agglomération sont au coeur de la réflexion des économistes spatiaux.

L'apiculteur et l'horticulteur

Marshall, pour illustrer son propos sur les externalités de proximité, avait cité l'exemple de l'apiculteur et de l'horticulteur[1]. Installés côte à côte, tant le premier que le second verront leur production augmenter, indépendamment de toute modification de leurs facteurs de production respectifs.



Dessin : Mireille Brunet

1. Selon les dernières informations, la démonstration ne serait pas de Marshall mais de Meade (un disciple de Keynes).

Des territoires et de leur dynamique

Au delà des débats sur les délocalisations et relocalisations des activités économiques, l'évolution des rapports entreprise/espace se caractérise par la transition d'une géographie statique des coûts à une géographie dynamique de l'organisation : les décisions de localisation des entreprises ne sont plus seulement fondées sur des considérations de coût, elles s'inscrivent également dans des stratégies d'organisation, de coopération et d'anticipation, c'est-à-dire de création de ressources nouvelles et d'innovation. Dans cette optique, la marche des entreprises va étroitement dépendre de leur ancrage territorial. Symétriquement, le développement des territoires sera fonction des formes de coopération construites localement.

Prolongeant les travaux pionniers de Marshall sur l'organisation industrielle, le concept de district industriel s'est imposé comme l'un des pôles de la réflexion théorique sur la dynamique territoriale. Par la suite, les

travaux sur les milieux innovateurs, et ceux, plus récents, sur les systèmes localisés de production et d'innovation, ont permis de faire émerger l'importance du territoire dans la dynamique industrielle, en particulier dans le développement de relations de coopération entre les entreprises. Ainsi, la dynamique industrielle va s'affirmer en se « territorialisant », et l'organisation territoriale des activités productives contribuer à la progression de la capacité de création de ressources nouvelles. Le cas de la *Silicon Valley* illustre parfaitement ces processus d'articulation, et de construction dans le temps, entre des formes d'organisation industrielle et des modes de relations avec l'environnement local.

Cette territorialisation de la dynamique économique reste cependant fragile. En cas de changement technologique, d'évolution des formes d'organisation du travail ou des entreprises, certains territoires vont bifurquer de manière positive, d'autres au contraire rentrer dans une spirale récessive. Il appartient justement à l'économiste spatial de comprendre et expliquer ces différences de trajectoire.

Études régionales et économie spatiale

Dans les directions régionales de l'Insee, les chargés d'études qui manient, traitent et commentent l'information socioéconomique disponible sont souvent amenés à s'interroger sur sa signification et sa cohérence. Ils constatent aussi que les interrogations soulevées par leurs propres travaux sont parfois difficiles à lever en l'absence de sources adéquates. Bref, ils ressentent la nécessité de bien comprendre, pour à leur tour pouvoir bien expliquer.

Dans cette attente, sans doute pourraient-ils utilement mettre à profit les enseignements des économistes spatiaux : *S'efforcer de parvenir à des résultats aussi robustes que possible est la première des préoccupations. Cela implique un choix de méthodes appropriées, qui renvoie lui-même à la conjonction d'une bonne connaissance des progrès théoriques les plus récents et d'une solide expérience dans la mise en oeuvre des méthodes éprouvées...* (Les études à l'Insee : selon quelles modalités ? - note n° 27/G001 du 13 janvier 1995 de la direction des études et synthèses économiques).

Une autre très bonne raison de s'intéresser à l'économie spatiale est que nos partenaires peuvent parfois se situer clairement dans ce champ. En Haute-Normandie par exemple, le débat en matière de développement régional tourne autour de l'alternative suivante : la région doit-elle se mettre en position de capter une croissance qui vient de l'Île-de-France, ou bien doit-elle construire sa propre capacité de

Territoires et théories
Quelques premières notions

La théorie des pôles de croissance. Les grandes firmes exercent des effets d'entraînement (ou au contraire de stoppage) sur les autres entreprises, elles « polarisent » la région[1], qui va donc s'organiser autour d'elles. Le raisonnement peut être étendu aux régions elles-mêmes, dont certaines seraient motrices à l'égard d'autres régions, en raison de leur importance économique et du déséquilibre des échanges avec ces autres régions.

La théorie des districts industriels. L'organisation industrielle du territoire est fondée sur un mélange de concurrence/émulation/coopération entre petites et moyennes entreprises spécialisées autour d'un même métier.

La théorie de la base. La prospérité de la région est fondée sur ses activités d'exportation (appelées activités basiques) vers les autres régions.

Les trois grandes théories de la division spatiale du travail :

- Les branches nouvelles à technologie élevée apparaissent dans les régions où l'environnement technologique est déjà élevé, les activités plus anciennes sont reléguées dans des régions à faible niveau technologique général.

- Au sein d'une même branche, les fonctions de direction/conception, de fabrication qualifiée et de fabrication peu qualifiée se déploient dans des régions différentes. Ce déploiement différencié est conditionné par les gisements en main-d'oeuvre, l'adaptabilité au travail et le coût de reproduction de la force de travail dans chaque système local d'emploi.

- Le travail se localise en fonction du besoin en informations qu'il requiert : le travail qualifié va ainsi se déployer dans les grandes régions urbaines, où le réseau informationnel est étendu, les travaux peu qualifiés et standardisés vont se localiser dans des régions plus rurales.

La théorie des milieux innovateurs. Le milieu innovateur se caractérise par une forte dynamique d'apprentissage, source d'innovation et donc de développement.

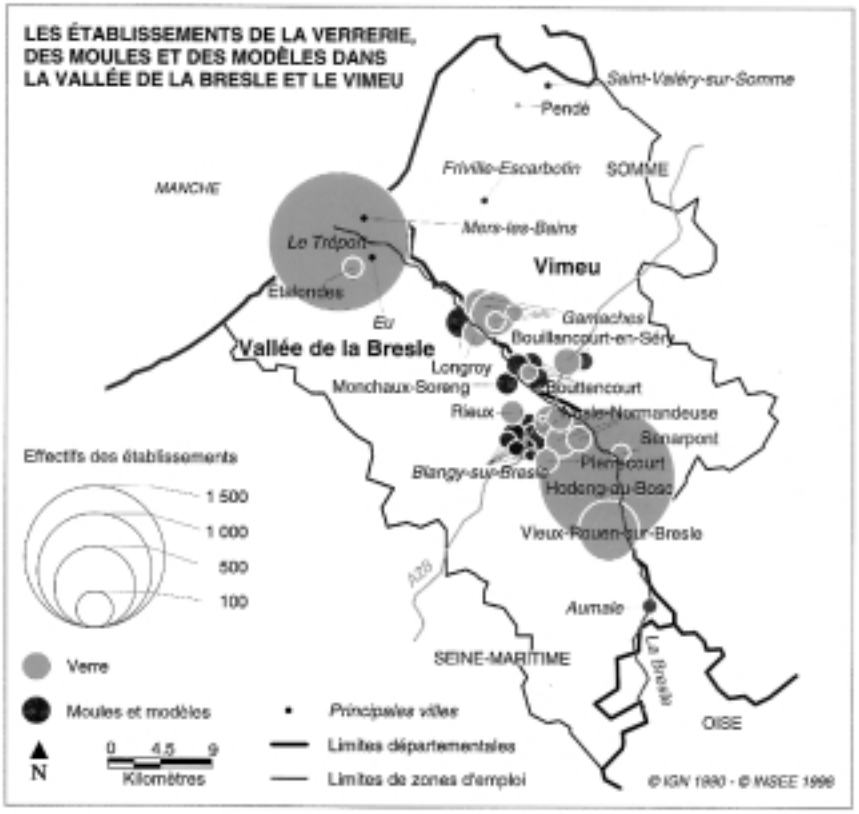
Les théories basées sur la notion de croissance endogène. Les modèles de croissance endogène distinguent de multiples facteurs de croissance : investissement en capital physique, public et humain, apprentissage (par la pratique), division du travail, recherche et innovation technologique. Dans les théories de l'économie spatiale basées sur ce même concept de croissance endogène, il est ainsi considéré que les régions à même d'enclencher une croissance forte et soutenue sont celles qui produisent le plus d'externalités, ou qui les gèrent le mieux.

1. Le mot région est ici employé dans l'acception spatiale du terme. L'économie spatiale appelle en effet ses découpages territoriaux des régions, à ne pas confondre avec les régions administratives. Mais celles-ci peuvent bien sûr se comprendre comme une somme de régions spatiales.

développement ? On est bien là dans une approche de type spatial. Bien sûr, les études et leurs conclusions pourront être différentes selon que l'on se situe dans l'une ou l'autre perspective. Dans la première, il sera sans doute prioritaire d'étudier l'économie de l'Île-de-France, dans l'optique de déceler des complémentarités haut-normandes, et de se pencher sur les facteurs de localisation pouvant inciter des entreprises franciliennes à venir s'installer en Haute-Normandie. Dans la seconde, il sera probablement utile d'étudier les complémentarités internes à la Haute-Normandie, les réseaux et coopérations qui s'y développent, dans une optique d'amélioration de l'intégration de l'économie régionale.

Plus généralement, les orientations actuelles du contrat de plan État-Région privilégient l'entrée territoriale. Dans le même ordre d'idées, la DATAR, Délégation à l'aménagement du territoire et à l'action régionale, s'intéresse depuis quelque temps aux systèmes localisés de production.

Le verre et son réseau de sous-traitants dans la vallée de la Bresle et le Vimeu
(Aval, Insee Haute-Normandie, 3e trimestre 1998, n° 82)



Quelles perspectives à l'Insee ?

Depuis quelques années déjà, la direction générale (département de l'action régionale) organise des stages d'analyse spatiale, consacrés aux phénomènes de localisation ou aux performances locales des entreprises.

Dans le domaine des dynamiques territoriales et systèmes localisés de production et d'innovation, une formation a été mise en place, avec le concours d'universitaires rouennais, dans l'interrégion Ouest de l'Insee (Haute-Normandie, Basse-Normandie, Bretagne et Pays de la Loire), mais beaucoup reste à faire. Un investissement dans cette voie pourrait conduire, à moyen terme, à d'importants progrès dans les études régionales, pour la plus grande satisfaction des utilisateurs de l'information économique et sociale.

La récente mise en place, à la direction régionale de Midi-Pyrénées, d'une mission à compétence nationale sur le développement des études économiques régionales, devrait permettre de progresser dans ce sens.

Frédéric LAINÉ
et **Serge TILLARD**
Insee

Frédéric Lainé fait partie de la division « études territoriales » à la direction générale, Serge Tillard est chargé d'études à la direction régionale de Haute-Normandie.

Bibliographie

L'aspect théorique

Collectif d'auteurs - *Encyclopédie d'économie spatiale,* Economica, 1995.

Collectif d'auteurs - *Économie industrielle et économie spatiale,* Economica, 1995.

DAVEZIES L. - *L'intégration contrariée de l'espace dans la théorie économique,* Courrier du CNRS, 1994.

DERYCKE P.H. & GILBERT G. - *Économie publique locale,* Economica, 1988.

POLESE M. - *Économie urbaine et régionale,* Université du Québec, 1994.

Quelques applications à l'Insee

BÉNARD R. & JAYET H. - *Les préférences de localisation des entreprises,* Insee Nord-Pas-de-Calais, Les dossiers de Profils, n° 51, octobre 1998.

GUILLEMET M., HÉMEZ C., LAINÉ F. & TILLARD S. - *Économie spatiale, initiation aux concepts,* document de travail H9801.

HECQUET V. & LAINÉ F. - *Structures productives locales et formes d'organisation économique : une analyse typologique,* document de travail E9811.

ILAL M. - *Les zonages agricoles,* in document de travail E9601.

JAYET H. (sous la direction de) - *L'espace économique français,* 1988.

LAINÉ F. & RIEU C. - *Le tissu productif régional : diversité et concentration,* Insee Première, n° 630, janvier 1999.

LAINÉ F. & RIEU C. - *La diversité industrielle des territoires,* Insee Première, n° 650, juin 1999.

LAURENT L. - *Le fonctionnement économique des bassins d'emploi : réhabilitation de la théorie de la base,* in document de travail H9506.

LAURENT L., LE JEANNIC T. & TERRIER C. - *De nouvelles frontières pour comprendre l'espace : concevoir des zonages, analyser le territoire,* in Cahier d'Aval, Insee Haute-Normandie, n° 38, 1996.

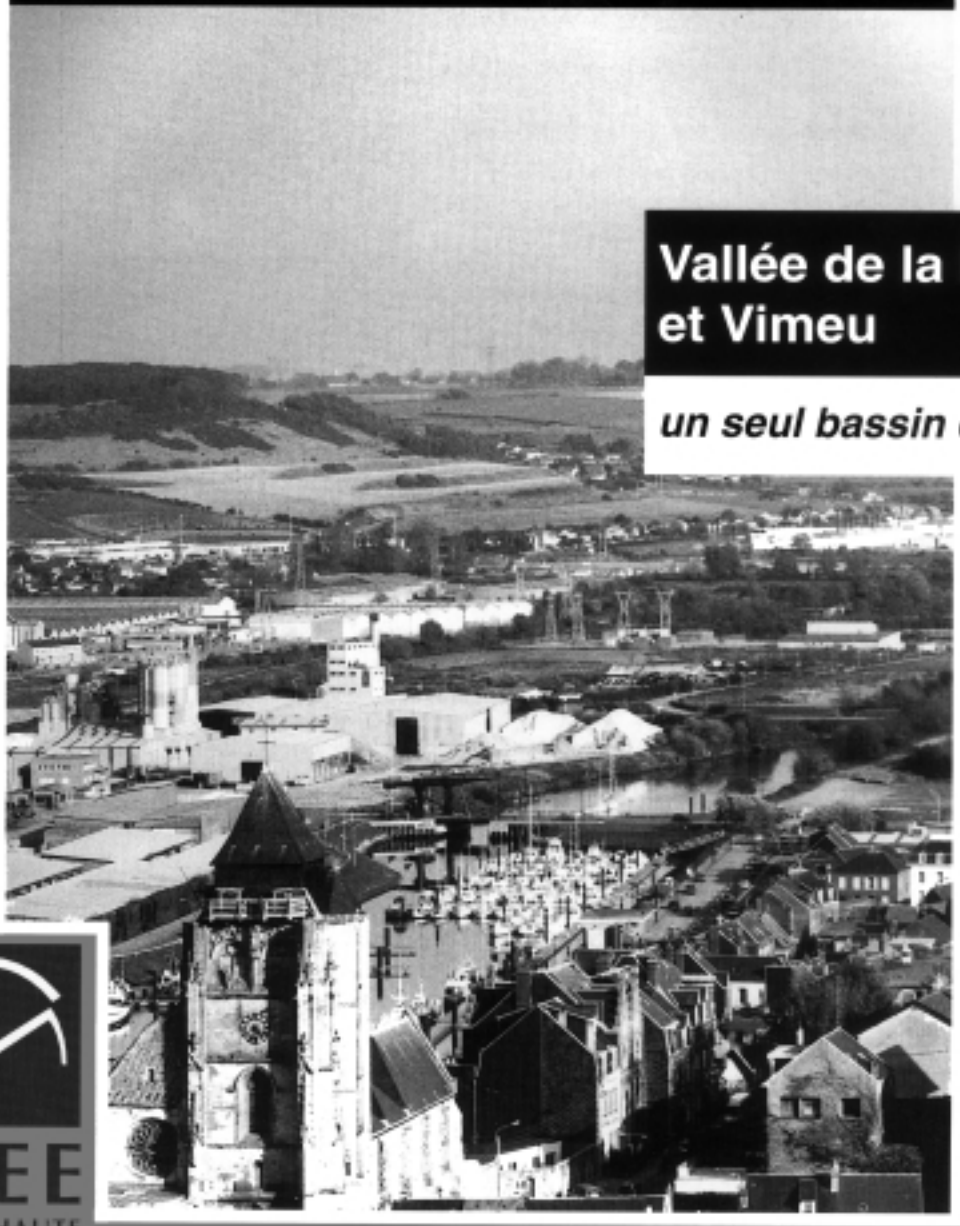
ROUALDÈS D. - *La restructuration des grands établissements industriels,* Insee Première, n° 512, 1997.

Atlas des zones d'emploi, cédérom, 1998.

Tendances régionales 1998, Insee Synthèses n° 22, 1999.

AVAL

REVUE STATISTIQUE ET ÉCONOMIQUE DE HAUTE NORMANDIE



Vallée de la Bresle et Vimeu

un seul bassin d'emploi

INSEE
HAUTE
NORMANDIE

PRIX : 30 F - 3e TRIMESTRE 1998 - N° 82