STATISTIQUE

POUR HISTORIENS

Alain Guerreau 2004

Ce texte est le canevas du cours intitulé "méthodes pratiques de statistique et de cartographie" professé devant les élèves de première année de l'École des Chartes durant l'année 2003-2004. Il ne s'agit pas d'un cours de statistique, mais d'un cours d'histoire, fait par un historien pour des historiens. La visée en est essentiellement pratique : apporter aux étudiants une vue générale des possibilités riches et variées que les techniques statistiques ouvrent à toute recherche historique.

Il s'agit ici d'une toute première tentative ; ce canevas doit être amélioré, remanié, complété : je serai reconnaissant à tous les lecteurs qui voudront bien me faire part de leurs remarques, critiques, suggestions.

Aucune analyse statistique n'est possible sans une connaissance minimale des principes de la statistique (qui, pour le moment, font défaut à la plupart des historiens) ; mais, dans un cours d'initiation, il m'a semblé qu'il fallait traiter ces préalables assez rapidement, de manière à aborder suffisamment les applications proprement historiques ; au demeurant, les manuels qui exposent ces principes généraux sont nombreux, on s'y référera autant que nécessaire. Après un bref exposé des traitements propres aux données chronologiques, une présentation un peu plus détaillée est proposée des données spatiales. Après un chapitre consacré aux principes de base de la construction des graphiques, les quatre derniers chapitres sont dédiés à divers aspects de la sémantique historique qui est, à mon sens, le domaine jusquà présent le moins parcouru et sans doute le plus prometteur.

Dans ce cours, contrairement à ce que l'on trouve dans la plupart des manuels, on a laissé de côté tout ce qui ressemble à des démonstrations mathématiques, en privilégiant une approche essentiellement conceptuelle. J'ai insisté sur trois aspects :
* un aspect historique, en précisant le plus souvent possible quand et dans quelles conditions se sont développées les méthodes auxquelles il est fait allusion ;
* un aspect critique, c'est-à-dire des indications sur ce qu'il faut faire et ce qu'il faut éviter ; sur ce dernier point, il s'agit autant des pièges classiques (qu'il vaut mieux signaler énergiquement) que de certaines procédures, voire certains logiciels, qui ont pignon sur rue et dont l'usage me semble beaucoup plus néfaste que profitable ;
* un aspect prospectif, en signalant quelles sont, de mon point de vue, les principales lacunes de l'arsenal actuel, et les perspectives de développement de nouvelles procédures qui semblent prioritaires.

Je tiens à remercier publiquement et très chaleureusement plusieurs collègues qui, en diverses occasions, m'ont apporté une aide substantielle, en particulier Marc Barbut, Philippe Cibois, Claude Grasland, Ezio Ornato, Henry Rouanet.

SOMMAIRE GÉNÉRAL

Chapitre 1

LES NOTIONS CLÉS

Statistique, statistiques: ces termes désignent plusieurs objets différents, quoique plus ou moins liés. Cette ambiguïté apparente ne résulte pas d'un malencontreux hasard, mais d'une **évolution historique** qu'il est indispensable de connaître, au moins dans ses grandes lignes, pour peu que l'on ne veuille pas **partir sur des bases incertaines**, qui peuvent s'avérer dangereuses ou pour le moins très limitantes si l'on souhaite utiliser efficacement les procédures existantes et, le cas échéant, les modifier ou en créer de nouvelles, mieux adaptées aux matériaux historiques.
Nous procéderons donc en trois temps : a) un résumé des origines et de l'évolution de ces activités intellectuelles (et sociales) ; b) une rapide présentation des logiciels et matériels actuellement disponibles ; c) une analyse des notions de base, auxquelles les manuels ne font presque jamais référence (mais qui pourtant ne vont pas du tout de soi) et qu'il faut toujours avoir à l'esprit pour éviter de s'égarer au premier tournant (il est recommandé de relire périodiquement ces quelques pages).

SOMMAIRE

1.1. BREFS RAPPELS HISTORIQUES

1.1.1 origines et premiers développements

Il est sans objet de remonter au-delà du milieu du 17e siècle. Deux éléments fondamentaux apparurent alors : d'un côté la virgule et les tables (trigonométriques et de logarithmes), de l'autre les premiers éléments du calcul des probabilités.

La **virgule** (le point décimal des anglo-saxons), qui nous paraît un objet quasi naturel, est une invention du 17e siècle. Jusque là, il fallait se débrouiller avec des fractions quelconques, système lourd, lent, générateur d'erreurs continuelles. Son introduction ne fut pas instantanée, et les « quatre opérations de base » demeurèrent pendant encore deux siècles au moins l'outil d'une minorité spécialement entraînée. (Notons au passage qu'il fallut attendre nettement plus d'un siècle pour que fût inventé le « système métrique **décimal** », qui tirait parti concrètement de ce système de calcul et le rendait en pratique particulièrement efficace. Sur tous ces points, chapitre 12).

Le **calcul des probabilités**, quant à lui, est né d'une réflexion sur les jeux de hasard et les paris, due à Pascal et à Fermat. Immédiatement après, Huygens rédigea le premier traité de calcul des probabilités (*De ratiociniis in ludo aleae*, 1656). Le mot hasard vient lui même de l'arabe az-zahr, qui signifie le dé (notons encore que l'anglais *random* vient de l'ancien français *à randon*, issu lui-même de la racine germanique *rennen*, courir).

Les recherches, les discussions, les théorèmes s'enchaînèrent à grande vitesse dans la seconde moitié du 17e et tout le 18e, pour aboutir à un premier état développé et organisé dans l'œuvre du « prince des mathématiques », Carl Friedrich Gauss (1777-1855) qui établit la « loi » (= équation) qui porte son nom, et en laquelle beaucoup de statisticiens croient encore trouver le fondement de toute analyse statistique.

Il est crucial de ne pas perdre de vue ce point de départ : les énormes développements du calcul des probabilités reposent peu ou prou sur une réflexion à propos des « jeux de hasard », les dés, le pile-ou-face, les cartes, le loto (les fameuses « boules dans une urne ») ou la roulette : jusqu'à une date récente, les séries de « nombres au hasard » étaient empruntées aux publications du casino de Monte-Carle (ou à celles du loto national soviétique...). Or tous ces jeux, **entièrement artificiels**, sont fondés sur la mise en œuvre de l'équiprobabilité d'un nombre parfaitement connu et limité de possibilités. Deux caractères dont le moins que l'on puisse dire est qu'ils ne se rencontrent guère à l'état natif dans la nature et/ou la société.

1.1.2 calculs et société : évolution d'une technique liée à des usages sociaux limités

Ce fut aussi dans les années 1660 que vit le jour en Angleterre ce que l'on appelle l'« arithmétique politique », qui tentait de décrire la population et ses caractères numériques à partir de relevés partiels dont on extrapolait des nombres globaux.

Un siècle plus tard, ce furent des allemands qui les premiers forgèrent le terme de *Statistik*, pour dénommer une manière de décrire globalement les États (comme le nom l'indique). Dès le 18e, de fortes controverses se déchaînèrent, à propos de la question de l'application de ce que nous appellerions des « modèles mathématiques » aux réalités sociales. Ces débats se poursuivirent avec plus ou moins d'intensité jusqu'au milieu du 20e siècle.

Ce fut seulement au début des années 1830 que le terme *statistique* en vint, en français, à désigner une branche des mathématiques jusque là rangée sous « calcul des probabilités ». L'usage de calculs un tant soit peu élaborés, pour analyser les grands recueils de données numériques issues des travaux des « bureaux de statistiques » qui se mirent progressivement en place dans tous les pays industrialisés, ne se développa que très lentement. Le rapprochement du calcul des probabilités

et de la statistique descriptive, auquel certains auteurs contribuèrent dès la fin du 18e, n'entra dans une phase active que dans les deux dernières décennies du 19e. Entre 1880 et 1950 environ, les principaux développements des statistiques « pratiques » furent dus pour la plupart à des anglais (Galton, Pearson, Gosset, Yule, Fisher). Empirisme, pragmatisme ? Au milieu du 20e siècle, la cause paraissait enfin entendue : les « calculs statistiques » sont non seulement utiles mais même irremplaçables dans une série de domaines, où ils permettent de clarifier les connaissances mieux que tout autre moyen.

De cet état de la science, on trouve une présentation exceptionnellement claire et dense dans le Que-sais-je ? 281 d'André Vessereau (*La statistique*, Paris, 1947), continuellement réimprimé depuis, sans modifications autres que tout à fait mineures et accessoires [lecture de base, **strictement indispensable**]. (A l'autre bout, on peut, éventuellement, retenir le nom de Maurice George Kendall (1907-1983), qui publia *The Advanced Theory of Statistics*, en deux volumes en 1943-1946, qui furent constamment enrichis et réédités, avec l'aide d'Alan Stuart à partir de 1966 (dernière édition, 3 vol., 1977-1983) : volumes qui passèrent longtemps (ou peut-être *encore* pour certains ?) sinon pour la Bible de la statistique, au moins pour l'encyclopédie définitive).

En fait, ces statistiques pratiques correspondent à un **nombre relativement restreint de champs d'application**. Dans chacun de ces domaines, des méthodes précises ont été plus ou moins « routinisées », et donnent lieu à des manuels et à un enseignement appropriés (sans bien entendu que cela supprime les controverses, ou les « sous-champs » concurrentiels).

Le domaine auquel cet ensemble s'applique le plus facilement est celui des séries expérimentales (résultats numériques d'expériences plus ou moins nombreuses). Les diverses branches de la biologie appliquée utilisent ces statistiques intensivement : expérimentations bio-médicales (et psychologiques), expérimentations agronomiques. Dans l'industrie, des procédures statistiques sont employées dans la surveillance des chaînes de fabrication. Dans ces secteurs, la notion de **test** est fondamentale : l'expérience (ou le processus de fabrication en cours) peuvent-ils raisonnablement être dits correspondre à telle ou telle hypothèse ?

Un ensemble assez différent est celui de la **prévision**. Moyennant la connaissance d'un nombre suffisant d'états passés, peut-on prévoir l'évolution la plus probable ? En pratique, cela « marche » convenablement à court terme (le court terme des économistes n'est pas celui des historiens : c'est « quelques jours » ou « quelques semaines », mais moins de trois mois). En dépit de cette faible durée, il y a beaucoup de « ratés ». Les enquêtes d'opinion, malgré toutes les mises en garde (essentiellement rhétoriques) des instituts de sondage, sont du même type. Ce sont des secteurs où les « statisticiens » peuvent gagner beaucoup d'argent.

Un secteur en plein développement est celui du « data mining » : comment, dans une masse d'informations de plus en plus gigantesque, et en renouvellement apparemment constant, faire, quasi automatiquement, la part de ce qui est réellement nouveau ? De gros moyens informatiques sont nécessaires, les calculs étant, apparemment plus que dans d'autres domaines, éloignés de toute préoccupation théorique, ce qui d'ailleurs provoque un certain mépris de la part des "statisticiens purs".

1.1.3 des techniques à l'écart des préoccupations des historiens

Si l'on considère globalement ces applications, on constate sans peine qu'elles sont orientées sur la *prévision* et l'*aide à la décision*. Elles sont efficaces dans les cadres expérimentaux (on travaille « toutes choses égales d'ailleurs » et les facteurs de variation sont sous contrôle, c'est précisément le caractère spécifique des expériences) ; les prévisions macro ou microéconomiques comportent au contraire une marge d'erreur qui s'accroît très rapidement quand s'allonge la durée pour laquelle on construit la prévision. *L'explication courante, qui consiste à affirmer que cette*

erreur proviendrait de la nature sociale des phénomènes, est ***inconsistante*** : l'atmosphère est une réalité purement physique, et les prévisions météorologiques sont encore plus incertaines, même à trois jours. Il y aurait plutôt lieu de se demander s'il ne serait pas temps de **s'inquiéter du degré de réalisme des hypothèses de base sous-jacentes**. La plupart des économistes ne s'en soucient guère, à l'instar de Keynes, à qui l'on prête cette répartie qui mérite d'être retenue : « pourquoi voulez-vous que je m'intéresse aux prévisions à long terme ? Dans le long terme, nous serons tous morts ». Dans la "littérature économique", le "long terme", c'est un ou deux ans. En cherchant à "utiliser" les méthodes de l'économétrie (comme d'aucuns l'ont fait), l'historien se fourvoie immanquablement.

1.1.4 révolution technologique et invention de nouvelles procédures (1945-1980)

Le terme *mécanographie* date semble-t-il de 1911. Les fiches perforées se répandirent lentement dans l'entre-deux-guerres, tandis que se perfectionnaient les machines électro-mécaniques permettant de les manipuler (toute une quincaillerie qui a connu son apogée dans les années 60 et qui a totalement disparu ; espérons que quelques musées ont conservé des ensembles cohérents). La mécanographie facilitait les tris et les comptages, mais au prix d'un labeur complexe et d'un outillage onéreux. Les tout premiers emplois de l'électronique pour effectuer des calculs datent de la fin des années 40.

Le terme *informatique* apparut en 1962. Les machines d'IBM et de quelques autres firmes envahirent la planète dans le courant des années 60 et 70. L'interface homme-machine changea complètement de nature, on inventa les premiers langages de programmation (le Fortran date de 1956). Outre les grandes entreprises, les centres de recherche importants s'équipèrent dès la fin des années 50.

Quelques esprits originaux en profitèrent pour inventer de nouvelles méthodes fondées sur des procédures de calcul (ce qu'en jargon on appelle des "algorithmes") pas forcément très complexes, mais nécessitant une grosse capacité de calcul. En 1960, un ingénieur chez Renault, Pierre Bézier (1910-1999) inventa une méthode révolutionnaire de calcul et de dessin des lignes et surfaces courbes (les « **bezier curves** » que l'on trouve dans tous les logiciels de dessin). En 1964, un jeune mathématicien, Jean-Paul Benzécri (1932-), découvrit un mode complètement nouveau de calcul des distances entre les lignes et les colonnes d'un tableau numérique, d'où résulta l'« **analyse des correspondances** », qui est apparu dès le début des années 70 comme la procédure la plus générale et la plus efficace d'analyse factorielle.

Un autre mathématicien français, Benoît Mandelbrot (1924-), publia durant les années 50 et 60 une série d'articles soulignant le caractère particulier de la « loi de Gauss » et montrant, dans le prolongement des travaux de Paul Lévy (1886-1971), que les lois de « Pareto-Lévy », qui ont des propriétés bien différentes, ont au contraire un caractère très général et s'appliquent à une grande quantité de phénomènes aussi bien naturels que sociaux, qu'il désigna par un néologisme qui ne se répand que lentement, "fractales" [une large partie de la matière de ces articles fut reprise en 1975 dans *Les objets fractals : forme, hasard et dimension* – 4e édition, Paris, 1995 : **lecture indispensable**].

On pourrait aussi mentionner ici les travaux fondamentaux d'Edgar Frank Codd (1923-), ingénieur chez IBM, qui, entre 1969 et 1971, établit les règles fondamentales qui doivent présider à la constitution des « bases de données relationnelles » ; après quelques années de discussions intenses, on considéra (vers 1981) qu'il était impossible d'aller au-delà des « **5 formes normales** » qui définissent les critères d'optimalité d'une base de données.

La liste pourrait être allongée. Durant la période qui s'étend des années 50 à 1980 environ, une série d'innovations radicales éclata ainsi, en ordre dispersé, dans le monde des « statistiques appliquées ». Ces innovations ont en commun d'être assez directement liées aux nouvelles possibilités concrètes de calculs de masse fournies par les ordinateurs. En revanche, elles connurent

des sorts divers. Si les courbes de Bézier et les formes normales de Codd s'imposèrent universellement (elles comblaient un vide), l'analyse des correspondances ne se répandit que dans certains pays (vive résistance aux États-Unis), et la remise en cause de l'universalité putative de la « loi de Gauss » fut pour l'essentiel reléguée aux rang des curiosités mathématiques ; l'édifice gaussien résista par le silence (sinon le mépris) et, au début du 21e siècle, on ne trouve qu'exceptionnellement mention des lois de Pareto-Lévy dans les manuels de statistique, même les plus récents et les plus développés (c'est une lacune extrêmement gênante et préoccupante).

1.1.5 bouleversements accélérés du contexte matériel : un autre environnement, de nouveaux rythmes

En une quinzaine d'années (1981-1996 environ), l'environnement concret, technique et social, des calculs a connu un bouleversement sans aucun précédent dans l'histoire. Ce bouleversement tient en trois mots : universalisation des PC. La micro-électronique se développa dans les années 70 (premières calculettes vers 1972, premières calculatrices de poche programmables vers 1975, premiers micro-ordinateurs vers 1978). IBM, qui dominait de manière écrasante le marché mondial de l'informatique, considéra d'abord ces objets comme des accessoires ludiques. En 1981, par un revirement à 180 degrés, cette firme, après un accord bâclé avec un fabricant de micro-processeurs (Intel) et un producteur de petits logiciels (Microsoft), mit sur le marché *son* micro-ordinateur, IBM Personal Computer. On connaît la suite : le succès vertigineux de ce produit, l'irruption des « clones », la course effrénée à la puissance, le déclin d'IBM et la marche impériale de Microsoft. La politique tarifaire du seul concurrent un peu sérieux, Apple, le fit chuter. Au milieu des années 90, Microsoft avait repris à son compte toutes les innovations d'Apple et imposait son OS (operating system : ouindoze) sur plus de 90% des micro-ordinateurs en service sur la planète. En face de ouindoze 98, puis XP (et des logiciels de bureautique à peu près imposés), aucune entreprise de logiciels ne put résister ailleurs que dans de petits créneaux très spécialisés, à tel point que la justice américaine elle-même commença à s'offusquer de cet abus manifeste de position dominante.

L'on a assisté dans les trente dernières années du 20e siècle à un bouleversement du « **système technique** », au sens où l'a défini Bertrand Gille [« Prolégomènes à une histoire des techniques » in B. Gille (éd.), *Histoire des techniques*, Paris, 1978, pp. 1-118 : lecture **obligatoire**]. Les supermarchés des pays industrialisés vendent comme des grille-pain ou des cocottes-minutes des machines cent fois plus puissantes que les « supercalculateurs » des années 60. Les interfaces graphiques, à peu près inconnues jusque dans les années 70, sont devenues la chose la plus courante, d'usage quotidien pour une large partie de la population. Une machine qui ne fait pas plusieurs millions d'opérations par milliseconde est considérée comme paléolithique. Les effets pratiques d'une telle irruption sont encore largement à venir (e.g. dans les domaines de l'enseignement et de la conservation du patrimoine...).

On peut cependant d'ores et déjà saisir une conséquence considérable, qui ne paraît pas avoir été encore bien théorisée, alors même qu'elle le mérite, et d'urgence. L'augmentation rapide de la puissance des machines, en même temps que leur intrusion dans des domaines de plus en plus variés, ont conduit à la production de logiciels eux aussi de plus en plus puissants. Certains logiciels, courants à un moment donné, ont disparu, remplacés par d'autres. Les "versions" se sont succédées, et se succèdent encore, à un rythme difficile à suivre : **l'utilisateur est contraint à un réapprentissage permanent**. A peine s'est-on familiarisé avec les commandes d'un logiciel que celui-ci doit être remplacé. Bien entendu, tous les producteurs de logiciels commerciaux ont un avantage majeur à cette course-poursuite, qu'ils encouragent de toutes les manières, notamment en rendant de plus en plus difficile l'utilisation de fichiers produits par une version récente dans une version plus ancienne : celle-ci devient de facto, plus ou moins rapidement, inutilisable. L'apparition

et la généralisation de **nouveaux périphériques** (imprimantes laser, scanners, modems, lecteurs-graveurs de CD puis de DVD, appareils photo numériques, etc.) ont des effets tout aussi déstabilisants.

Le système génère, de par son organisation même, une ***instabilité structurelle***. Il faut donc se faire à cette idée : tout ce que l'on sait, ayant trait de près ou de loin à la micro-informatique (ou plutôt à l'informatique en général : l'écart est de plus en plus indiscernable), à un moment donné (disons l'année x), est condamné à une obsolescence rapide, c'est-à-dire n'est plus opérationnel à l'année x+3 ou x+4, et complètement anachronique à l'année x+5. Il faut donc partir, consciemment et clairement, du principe que la pratique de tout ce qui touche l'informatique doit nécessairement comporter une part de mise à jour quasi permanente. C'est un point que tous ceux qui ont fait leurs études jusque dans les années 80 peuvent avoir du mal à prendre en compte, mais le mouvement ne paraît pas réversible. Il est d'autant plus important de réfléchir à ce point (pour en tirer les conséquences adéquates) que l'ensemble des connaissances que recouvre l'expression "tout ce qui touche l'informatique" ne paraît pas devoir se restreindre, c'est le moins que l'on puisse dire.

1.1.6 un environnement qui offre aux historiens des outils de travail sans précédent

Les innovations des années 60-70, l'irruption de la micro-informatique et sa puissance sans précédent ne pouvaient pas ne pas avoir des effets profonds sur les pratiques statistiques. Si l'on peut à bon droit parler de bouleversement du système technique, c'est -notamment- en considérant précisément la situation et l'environnement immédiat desdites pratiques statistiques.

Durant trois siècles, du milieu du 17e au milieu du 20e, tout ce que l'on range (rétrospectivement) dans la catégories "statistiques" subissait **une contrainte extrêmement forte, celle de la difficulté des calculs**. Tout un attirail fut mis au point pour tenter d'alléger cette contrainte (en commençant par la fameuse machine de Pascal, en passant par les tables, les règles à calcul, les papiers fonctionnels, les abaques, etc.). En dépit de cet outillage, la contrainte demeurait pesante à un point qu'il nous est devenu difficile d'imaginer. Et cela avait un retentissement drastique sur toutes les recherches théoriques : aucune formule n'avait d'intérêt si elle conduisait à des calculs impraticables ; dès lors, l'objectif fondamental et central de toutes les réflexions visait la simplification des calculs. La question du réalisme des hypothèses était nécessairement seconde, une "bonne loi" était une loi qui renvoyait à des calculs exécutables. La domination de la « loi de Gauss » s'explique dans une très large mesure par cette contrainte. Les propriétés mathématiques fort remarquables de cette loi avaient en effet cet avantage incommensurable de faciliter les calculs et dès lors tout l'art du statisticien consistait, si nécessaire, à bricoler les données pour permettre de leur appliquer les procédures liées à ladite loi (abusivement baptisée « loi normale » en 1894 seulement, c'est-à-dire justement au moment du grand essor des statistiques appliquées).

Cette contrainte a disparu, mais **ses effets sont encore omniprésents** : un ensemble compact d'axiomes, de théorèmes, de modes de calculs, étroitement liés à des structures sociales douées de forte inertie (institutions académiques, ou ce qu'il faut bien appeler la "corporation des économistes") demeure parfaitement en place. On continue par exemple de publier des "tables statistiques" qui ne servent plus à rien ni à personne. On peut s'attendre à ce que cette situation perdure encore dix ou vingt ans (peut-être davantage). Il faut le savoir et s'adapter, c'est-à-dire effectuer sans état d'âme le tri entre ce qui demeure tout à fait valide et ce qui est biaisé, et surtout **s'orienter résolument vers la mise au point de nouvelles procédures** (indépendamment de toute considération de la masse de calculs nécessaires) fondées sur des **hypothèses plus réalistes** : l'extinction de la contrainte des calculs doit permettre de rendre le primat à une réflexion (complètement renouvelée) sur les bases mêmes des analyses statistiques.

Avant de procéder à un indispensable examen des grandes notions fondamentales, qui sont sous-jacentes à toute analyse statistique, nous allons brièvement survoler la situation concrète actuelle.

1.2. MATÉRIELS ET LOGICIELS

Cette partie devra être réécrite chaque année. Le fleuve ne remonte pas vers sa source, l'attente d'une éventuelle "stabilisation" est une manière de nier la réalité. Il est dangereux (en particulier pour un historien) de tenter de décrire la conjoncture présente, mais il est peut-être moins absurde de commencer par là que de lire une page d'horoscope. Il faut du matériel et des logiciels, lesquels choisir ?

1.2.1 éléments de conjoncture.

Deux éléments qui caractérisent la situation actuelle ne paraissent pas du tout en voie de ralentissement : la course à la puissance des matériels et le développement d'internet.

Bien entendu, les constructeurs de matériels et les vendeurs de logiciels ont partie liée dans le sens de l'**augmentation de la puissance**, car il s'agit là d'une des sources principales de leur fortune. En regardant de plus près, on s'aperçoit que les moteurs les mieux identifiables de cette course sont les producteurs de jeux vidéo et l'industrie cinématographique. Les procédures graphiques animées exigent des capacités de mémoire et des vitesses de micro-processeur qui sont encore loin d'être atteintes. On le voit d'ailleurs bien *dans un domaine qui concerne directement les conservateurs et futurs conservateurs, celui de la photographie numérique.* Les performances de la photographie numérique ne sont pas encore tout à fait équivalentes à celles de la photographie dite "argentique" (ne serait-ce que les pellicules 24x36, pour ne rien dire des "grands formats"). On imagine mal ce qui bloquerait le développement des capacités de la photographie numérique tant qu'elle n'aura pas rejoint celles de l'"argentique". Notons d'ailleurs que, d'ores et déjà, les appareils photo numériques offrent des possibilités pratiques que l'on ne trouve sur aucun appareil classique. Les fabricants de circuits électroniques et de disques durs ne manifestent aucune intention de freiner leur course.

L'**internet** en est encore à ses balbutiements. La simple comparaison avec les pays voisins montre instantanément le retard de la France. La plupart des chambres d'étudiants des cités universitaires en Allemagne fédérale sont munies d'une connexion directe sur internet. Le "haut débit" sur les lignes téléphoniques est cher et peu efficace. Or toute l'imagerie numérique, encore elle, avec ses fichiers gigantesques, a besoin d'internet. Malgré cette incertitude technique actuelle, le sens du mouvement ne fait aucun doute. Surtout, et cela est crucial pour nous, l'universalisation d'internet a d'ores et déjà complètement bouleversé les pratiques des utilisateurs de logiciels un tant soit peu avertis, et a fortiori des informaticiens. Des équipes de programmeurs travaillant sur un même projet peuvent être constituées de personnes résidant sur les cinq continents, et collaborant de manière plus simple que s'ils étaient dans le même bâtiment à deux étages de distance. Une part rapidement croissante de l'écriture et de la diffusion des logiciels s'opère à l'aide d'internet. Le mouvement a commencé lentement au cours des années 90, des résultats concrets de première importance sont visibles depuis trois ou quatre ans, et l'on vit en ce moment une période de croissance exponentielle.

Le bouleversement provient de l'***implantation irréversible des logiciels libres***, "open-source", conçus et diffusés en respectant les principes énoncés dans la GPL (General Public License), qui est destinée à garantir la possibilité de partager et de modifier les logiciels libres et de s'assurer que ces logiciels sont effectivement accessibles à tout utilisateur (le texte de la version 2 de cette GPL est disponible sur de nombreux sites, il date de 1991, voir

www.gnu.org/licenses/gpl.html). Le succès universel d'objets créés hors de toute finalité lucrative semble paradoxal dans un monde dominé par la recherche du profit capitaliste. L'explication, car il y en a une, tient d'abord à la conjonction des possibilités offertes par internet et d'un certain nombre de caractères intrinsèques de l'activité de programmation. La programmation est une opération d'ordre intellectuel, qui réclame du temps et des efforts mentaux importants, mais des investissements (financiers) minimes : un individu isolé peut tout à fait écrire en quelques mois un gros logiciel, qui marche. Mais ce premier jet n'est jamais parfait, à la fois parce qu'il y a toujours des défauts (le cas non prévu qui donne des résultats faux ou bloque tout le système), et parce que les utilisateurs potentiels s'aperçoivent plus ou moins rapidement que des fonctions qui seraient bien utiles manquent. Autrement dit, la phase la plus compliquée et la plus longue est celle de la maintenance et du développement.

Internet a permis de contourner l'obstacle constitué par la nécessité (jusqu'au début des années 90) de confier la "distribution" à un réseau commercial, en général surtout capable d'absorber la plus grande partie des profits. Mais c'est ici que l'on doit se souvenir d'un des caractères les plus frappants (quoique rarement mentionnés) de la programmation : son aspect de jeu, ou de sport. Cela vaut très largement les échecs, le bridge ou même les mots croisés. La création spontanée de très vastes "communautés" rassemblées autour de tel ou tel "projet logiciel libre" renvoie sans aucun doute à ce caractère. Lorsqu'une nouvelle "version" apparaît, des milliers, des dizaines de milliers de personnes de par le monde installent cette version et "cherchent la bogue", c'est la première phase du jeu ; dès que des bogues ou des insuffisances ont été repérées, les mêmes se jettent sur les sources, à la recherche de l'erreur et de la manière de la corriger, c'est la seconde phase. Il n'y a aucun droit d'entrée, n'importe qui peut jouer ; les gains sont purement symboliques, mais énormes : le sentiment d'appartenir à une communauté hors du commun, le sentiment de contribuer au progrès des techniques sinon au bien-être de l'humanité. Gains que peu d'autres activités peuvent procurer au même degré aussi commodément.

Mais deux circonstances viennent encore conforter cette structure. D'abord le fait que ce jeu n'est pas "politiquement neutre". Il est immédiat de constater que toute contribution pratique à un logiciel sous la GPL peut être perçue ou vécue comme un pied-de-nez à Microsoft, sinon à la domination US. Les grandes initiatives sont venues de Scandinavie, d'Allemagne et d'Autriche, d'Australie et Nouvelle-Zélande ; les contributeurs viennent largement des "pays de l'Est", de l'Asie du Sud-Est et des Indes ; les universitaires américains sont aussi très nombreux et toute une série de fondations universitaires américaines apportent des soutiens très importants. Mais surtout -second point-, il est remarquable de constater à quel point **ce système est efficace, en termes d'informatique pure** ! Tout simplement parce que la quantité de "testeurs", de "débogueurs" et de "développeurs" qui participent à l'évolution de ces logiciels est telle qu'aucune entreprise d'informatique ne peut, même de très loin, disposer d'un réservoir de matière grise équivalent. C'est pourquoi tout le monde s'accorde à reconnaître que les "grands logiciels libres" sont plus efficaces et plus stables que la plupart de leurs équivalents "propriétaires".

Une question clé, à propos de laquelle les pronostics sont plus que hasardeux, concerne l'avenir des OS (operating system) ; en pratique, Ouindoze ou Linux. Microsoft, qui dispose d'une position de domination écrasante, résiste par tous les moyens, en utilisant toutes les procédures de verrouillage disponibles et à inventer, et en passant toutes sortes d'accords occultes avec des constructeurs (ordinateurs portables, périphériques divers), pour essayer de faire en sorte que ces matériels soient inutilisables hors de Ouindoze. Accords aussi avec de nombreux producteurs de logiciels spécialisés (tous ceux qui sont destinés à des activités techniques très particulières, et dont la programmation requiert donc des compétences techniques élevées en plus des compétences proprement informatiques). Une très large partie de la "presse informatique", qui vit de ses

annonceurs bien plus que de ses lecteurs, fait le silence sur tout ce qui n'est pas commercial.

Mais de puissants facteurs jouent en sens inverse, notamment la généralisation des politiques de "réduction des déficits publics". Les serveurs officiels de la RFA offrent aux écoles publiques un "Schullinux", qui va permettre d'installer dans toutes les écoles des suites logicielles complètes à un côut à peu près nul. Lorsque tous les élèves se seront habitués à Linux.... Là encore, la France est en retard (malgré des circulaires de L. Jospin, guère suivies d'effets), mais la pression budgétaire augmente. Enfin, certains constructeurs, sentant le vent tourner, proposent désormais des pilotes LINUX pour tous leurs matériels (e.g. EPSON, publicité non payée). Aujourd'hui même, sans que l'on s'en aperçoive, la très grande majorité des serveurs de par le monde fonctionnent avec le tandem Linux-Apache, c'est-à-dire entièrement sous GPL. Bref, *s'agissant des institutions de conservation européennes publiques (aussi bien que des structures d'enseignement et de recherche)*, il semble que l'on ne s'avance guère en estimant que *dans les dix ans qui viennent Linux sera installé sur la quasi-totalité des machines*. Il vaudrait sans doute mieux y penser dès à présent et s'y préparer avec calme et méthode. (On comprend assez mal, par exemple, que les bibliothèques publiques françaises dépensent autant d'argent pour utiliser plusieurs dizaines de logiciels verouillés différents, qui rendent à peu près les mêmes services, alors qu'existe un logiciel open source, qui pourrait être adapté aux divers besoins...)

1.2.2 propositions concrètes.

S'agissant de matériels, se rappeler seulement ceci : lors de l'acquisition d'un périphérique, toujours se préoccuper *avant l'achat* de la disponibilité d'un pilote Linux. Pour les machines elles-mêmes, se préoccuper surtout de la capacité de l'unité centrale (512M minimum) et de la taille du disque dur (pas de limite supérieure, 60 gigas paraissent aujourd'hui un minimum). La vitesse du processeur (argument de vente courant) ne présente aucun intérêt pour la recherche.

Si les éléments de conjoncture décrits précédemment ne sont pas entièrement controuvés, il apparaît que le plus raisonnable (automne 2003) est de privilégier les logiciels dont il existe une version Ouindoze et une version Linux, fonctionnant de la même manière dans les deux environnements, sous GPL dans les deux cas (ce qui veut dire surtout, en français courant, téléchargeables et gratuits, reproductibles et modifiables en parfaite légalité). Tous ceux qui auront pris l'habitude d'utiliser ces outils sous Ouindoze n'auront rien à changer à leurs habitudes de travail lorsque, par choix ou par obligation, ils "migreront". (Nota : "migration" est le dernier concept "tendance"). (Nota 2 : la taille importante de ces logiciels amène à déconseiller très fortement de tenter un téléchargement par ligne téléphonique et modem ; soit trouver un moyen de se brancher directement sur le réseau, soit recopier ces fichiers à partir d'un exemplaire gravé sur CD).

Signalons d'abord, bien que ce ne soit pas directement l'objet du présent cours, que l'on trouve des logiciels de bureautique de très haute qualité répondant à ces critères : OpenOffice et GIMP. OpenOffice (www.OpenOffice.org, actuellement, 10/2003, version 1.1) est une suite comportant un traitement de texte (OpenOfficeWriter, avec lequel a été composé le présent document), un tableur, un logiciel de présentation et un logiciel de dessin. Les deux premiers sont totalement compatibles avec les formats *.doc et *.xls, les aides copieuses et bien faites. On ne voit pas ce que l'on pourrait reprocher à ces logiciels (sinon peut-être d'avoir poussé le zèle jusqu'à introduire divers "automatismes" -sur un modèle connu- qui sont plus désagréables qu'utiles). GIMP (www.gimp.org, version stable 1.2.3) est un logiciel de traitement des images bitmap qui comporte toutes les fonctionnalités de ses homologues "haut de gamme" commerciaux (fort onéreux). Pour les aficionados de la programmation, signalons qu'OpenOffice et GIMP incluent des langages de programmation puissants, permettant notamment de produire les macros les plus variées.

Une analyse statistique consiste à effectuer des séries de procédures sur des données. Celles-ci doivent donc d'abord être "entrées". S'il s'agit seulement d'analyser une série de 50 nombres, la

plupart des logiciels statistiques permettront de les entrer directement. Mais, dans la plupart des cas, les données sont nombreuses et variées. L'usage d'un tableur est approprié si l'on n'a qu'un type d'individus (individus statistiques = objets à analyser), et que les caractères qui décrivent ces individus ne peuvent prendre qu'une valeur (tiennent dans une seule colonne). Si ces deux conditions ne sont pas remplies, il faut passer à un logiciel de base de données (SGBDR -système de gestion de base de données relationnelle, en anglais RDBMS), la lettre importante étant le R, pour "relationnel". Un tel logiciel "gère" un ensemble de tables liées les unes aux autres, mobilisables simultanément de manière transparente, de plus en plus couramment au travers d'un langage à peu près normalisé, dénommé "langage de requête structuré" (structured query language, SQL). Ces considérations conduisent directement au choix de MySQL (www.mysql.com, charger la version 5.0-max). D'abord conçu pour Linux, ce logiciel puissant a été "porté" sous Ouindoze. *Depuis cette année*, il existe une interface graphique conviviale, qui permet d'utiliser les diverses tables comme les "feuilles" d'un tableur. A ceci près que la définition des types de champ est tout de même un peu plus large, puisque l'on peut avoir des champs-textes ou des champs-images (i.e. intégrer directement des images dans la base de données ; en fait tout fichier binaire peut être inclu et lu ensuite avec le logiciel adéquat). La grande majorité des services de conservation fonctionnent déjà autour d'une base de données relationnelle, les autres devront suivre rapidement. **Il paraît plus que souhaitable que les conservateurs en connaissent non seulement l'existence, mais aussi le fonctionnement**.

Pour en venir enfin au logiciel statistique, ***le choix se porte sans hésitation sur R*** (www.r-project.org, version 1.9 depuis avril 2004). Développé à partir de 1996, ce logiciel a rapidement atteint une grande maturité, et une notorité internationale de premier plan. Cette rapidité s'explique aisément : les créateurs bénéficiaient d'une longue expérience de langages de programmations divers et de logiciels mathématiques et statistiques eux aussi assez nombreux. Les chercheurs qui ont débuté la programmation dans les années 70 commencent à avoir des idées assez précises sur ce que sont les avantages et les inconvénients potentiels d'un langage. Les contraintes (de place, de rapidité) qui ont longtemps orienté les travaux, s'estompent. Il devient possible de dresser la liste de toutes les qualités que l'on aimerait trouver dans un langage... et de le créer !! C'est à peu près ce qu'ont fait Ross Ihaka et Robert Gentleman en 1995, d'où est résulté R, qu'ils définissent comme un "environnement statistique". Nous verrons en pratique que l'on peut utiliser R pour manipuler les données et leurs appliquer des algorithmes statistiques ("langage déclaratif"), mais que l'on peut très facilement écrire n'importe quel algorithme nouveau selon les méthodes de programmation traditionnelles ("langage procédural") et convertir ces "bouts de programme" en fonctions elles-mêmes intégrées dans le langage. A cela s'ajoute une grande variété de procédures graphiques, et des modalités d'entrée-sortie adaptées à (presque) tous les formats, notamment MySQL, que l'on peut utiliser directement à partir de R. De nombreux statisticiens ont écrit des "éléments de programmes" ("packages") en R et les ont mis à la disposition du public (on les trouve sur le site de R). La liste de discussion "r-help" (à laquelle vous pouvez vous inscrire en quelques secondes) témoigne d'une activité internationale foisonnante. En 2002 sont apparus des manuels en français, en italien, en espagnol, en allemand, ce qui constitue un indice clair de la pénétration rapide de R dans toutes les zones, notamment au niveau de l'enseignement.

La domination des grands logiciels commerciaux (SAS et SPSS au premier chef) pourrait s'effriter plus rapidement qu'on ne le pense. Comme on le verra un peu plus tard, le grand logiciel lyonnais d'analyses factorielles (ADE-4) existe maintenant sous forme d'un "package" R, et tous les outils sont disponibles pour brancher R sur le principal SIG (système d'information géographique) distribué sous la GPL (Grass).

Tous ces logiciels sont, à bien des égards, surpuissants par rapport aux besoins d'une

recherche historique de taille habituelle. Mais rien n'oblige à se servir de toutes les fonctions disponibles. Selon le vieil adage : "abondance de biens ne nuit".

1.2.3 instabilité structurale

Comme on l'a indiqué plus haut, nous sommes entrés dans un nouveau "système technique". Une **transformation continuelle rapide est devenue un élément de la structure**. Ce qui implique que **toutes les connaissances, tant en matière de matériels que de logiciels, sont à la fois strictement indispensables et hautement transitoires**. Il faut se faire le plus vite possible à l'idée que, durant plusieurs décennies d'activité professionnelle, **chacun devra consacrer une part non négligeable de son temps à une mise à jour permanente et continue de ses connaissances**. Il vaut mieux s'y préparer explicitement...

1.3. QUELQUES NOTIONS FONDAMENTALES

Nous allons opérer ici un premier débroussaillage à propos de notions sur lesquelles nous serons amenés à revenir à de multiples reprises. Ce sont des notions qu'il ne faut jamais perdre de vue mais au contraire mobiliser à chaque instant pour éviter de tomber dans les innombrables pièges que recèlent les procédures de manipulation des nombres (même les plus simples en apparence) : ordre de grandeur, indicateur, biais, différence (opposition) entre imprécision et approximation, multiplicité (assez générale) des "solutions optimales", exploration, formalisation.

1.3.1 ordre de grandeur

La notion d'***ordre de grandeur*** est souvent rendue par le terme d'"échelle" ; on dit, par exemple, si l'on étudie le budget d'une famille ou celui d'un État, que l'on "ne travaille pas à la même échelle". Pourtant, formellement, on peut, dans les deux cas, partir d'un tableau d'entrées-sorties à peu près identique, aux unités près, et appliquer les mêmes procédures de calcul. Le statisticien prend rarement en compte **le détail : "aux unités près"**. Pour l'historien, ce point est au contraire déterminant, et l'on ne saurait dresser la liste, bien trop longue, de tous les ouvrages irrecevables simplement parce que ce "détail" a été traité par prétérition. Les géographes savent depuis longtemps que l'on n'analyse pas de la même manière une vallée du Jura et une vallée des Andes, même si, dans les deux cas, il y a un talweg, des versants et des crêtes. L'historien a toujours affaire à une société humaine, et la taille et l'extension du groupe considéré sont des caractères de base de l'objet, qui conditionnent radicalement toute l'interprétation que l'on peut donner des phénomènes observés.

La notion corrélative de ***limite*** n'a pas du tout le même sens en mathématiques et en histoire (ou en sociologie). En mathématiques, les "limites" sont des **outils abstraits**, dont d'ailleurs il est fait un large usage dans les calculs : on manipule couramment l'infiniment grand et l'infiment petit, et cela est bien utile. En histoire, il existe des **limites réelles** simples, dont on ne peut pas sortir sans divaguer : du berger isolé sur son alpage à l'humanité considérée dans son entier. Il n'y a rien en-deçà, rien au-delà. L'unité minimale est insécable (comme l'indique bien le nom lui-même, *individu*) et représente inévitablement, comme les Grecs l'avaient déjà noté, "la mesure de toute chose". Et les relations entre un individu et un groupe, ou entre deux ou n groupes (relations dont la composition forme à proprement parler la structure sociale, c'est-à-dire l'objet propre du travail de tout historien) dépendent foncièrement de la taille des groupes considérés.

Dès que l'on a compris que ***le sens de tout objet historique renvoie sans aucune exception à une structure sociale déterminée***, on saisit aussitôt que l'ordre de grandeur, ou la taille, de la structure pertinente par rapport à l'objet considéré doivent donner lieu à une réflexion préalable approfondie. Bien entendu, l'ordre de grandeur n'est qu'un élément parmi d'autres, au sein du groupe des propriétés intrinsèques d'un ensemble social ; mais c'est un élément crucial, car la nature même des relations sociales varie radicalement selon "l'échelle" considérée. Rien d'aussi absurde que de

traiter les relations entre deux groupes comme celles entre deux individus, ou entre deux villages et deux États.

Ces remarques générales entraînent divers corollaires, que l'on ne pourra que survoler. Signalons d'abord une confusion courante : il ne faut pas confondre *ordre de grandeur* et *effectif*. Chacun sait qu'un sondage bien conçu sur 1000 Français donne des informations plus intéressantes qu'une enquête exhaustive sur toute la population de Paris intra muros, par exemple ; or l'effectif dans le second cas est au moins 2000 fois plus important ; il n'empêche que l'ordre de grandeur correspondant au sondage est très largement supérieur.

Les corollaires concernent d'abord les relations entre les diverses échelles. B. Mandelbrot a attiré l'attention sur la notion d'« homothétie interne » ("selfsimilarity"), terme qui évoque la propriété particulière de nombreux phénomènes de se présenter à peu près sous la même *forme* à plusieurs échelles successives. On a évoqué plus haut la question des budgets ; on peut penser aussi à la structure spatiale auréolaire (centre, zones intermédiaires, périphérie), schéma que l'on peut observer au niveau du village, de zones d'étendues variées, de pays et même de continents. Certains "styles" ont fait également un assez grand usage de cette relation, par exemple l'art gothique, qui employait avec prédilection une même forme (arc brisé avec remplage, spécialement quadrilobé) aussi bien dans les miniatures, les petits objets (monnaies, sceaux), la statuaire et l'architecture. La présence (ou l'absence) de ce phénomène dans diverses sociétés est un phénomène qui mérite toute l'attention. Mais son interprétation est délicate et l'on ne saurait guère imaginer une solution générale passe-partout.

En revanche, on ne saurait trop insister sur l'intérêt qu'il peut y avoir à examiner avec soin *les relations entre les formes d'organisation sociale aux divers "niveaux" où l'on peut les observer.* Ces relations sont le plus souvent complexes et surtout extrêmement variables. Une notion élémentaire comme celle d'emboîtement (et/ou de hiérarchie) est le plus souvent insuffisante et donne lieu aux affirmations les plus gratuites et dévastatrices (quand on imagine que toute détermination circule "du bas vers le haut", ou "du haut vers le bas"). L'indétermination affirmée (du genre : il y a une explication à chaque échelle, toutes les échelles se valent) n'est pas moins délétère, car il ne s'agit de rien d'autre que d'un camouflage malhabile du refus métaphysique de toute recherche d'une cohérence dans les phénomènes sociaux. Bref, il s'agit là d'un domaine très peu balisé, où de nombreuses recherches sont à la fois possibles et nécessaires.

Un autre problème que l'on peut ranger dans cette même catégorie est celui de *la relation entre sens et fréquence* (sur lequel on reviendra bien plus longuement par la suite, chapitres 9-10-11). En termes banals, quels rapports entre la règle et l'exception, entre le courant et le rare, voire entre la structure et l'anomalie ? D'un point de vue plus spécifique, on peut aussi se demander si les mots qui portent le plus de sens sont les mots les plus fréquents ou les plus rares. La réponse classique consiste à dire qu'il s'agit d'un paralogisme, celui de l'œuf et de la poule : ce n'est pas telle ou telle fréquence qui fait sens, mais la combinaison de plusieurs fréquences. Une telle réponse doit bien entendue être présente à l'esprit de tout chercheur, mais ne règle pas toutes les difficultés. L'histoire de l'art accorde systématiquement une place privilégiée, si ce n'est exclusive, aux "grandes œuvres" et le succès des biographies montre que l'on n'a pas réglé la question des "grands hommes". Dans la plupart des sociétés historiques, la classe dominante représente bien moins de 1% de la population, alors même que son rôle peut être déterminant. Ces quelques cas de figure peuvent, dans une certaine mesure, être ramenés à la question de l'œuf et de la poule. Mais cela ne résoud pas du tout la question de l'anomalie. Car, à considérer seulement les fréquences, il existe une quasi-similitude entre le fascicule de poèmes, unique et génial, d'un poète maudit (mais reconnu après quelques temps comme un des plus grands artistes de son époque) et le fascicule, exactement

analogue, tiré à 100 exemplaires, que personne n'a jamais lu et dont quelques très rares spécimens dorment encore dans les exceptionnelles bibliothèques où l'on n'a pas encore décidé de renouveler régulièrement les stocks.

[**Nota bene** : il suffit d'un historien en mal de notoriété pour exhumer ledit fascicule, lui attribuer des vertus hors du commun par l'effet d'une rhétorique, elle, tout à fait ordinaire, et l'on assiste à la "résurrection" d'un "auteur méconnu". Pour peu que l'historien en question soit entouré d'un petit cénacle d'adulateurs exaltés, la mayonnaise prend, les publications se multiplient et l'auteur méconnu est hissé irrémédiablement sur un socle de bronze (exegi monumentum... : pour ceux qui ont des lettres). Il est manifeste qu'une telle opération n'est possible que parce que le "bon public", et les historiens en particulier, ne réfléchissent que très rarement en termes explicites de fréquence ; l'ordre de grandeur (i.e. l'importance concrète) d'un phénomène historique peut être carrément inversé sans que personne ne s'émeuve.]

1.3.2 indicateur

Le terme ***indicateur*** a des sens multiples. En matière de statistiques, il s'agit encore d'une notion cardinale. En partie complémentaire de la précédente ; nous avons en effet signalé un peu plus haut que l'objet d'une analyse historique est toujours une structure sociale, ou un objet très directement dérivé. Or **une structure est un objet abstrait qui ne s'observe jamais directement**. Les relations, qui sont les éléments constitutifs d'une structure, se mesurent à l'aide d'indicateurs. Il en va d'ailleurs de même dans toutes les sciences : la température d'un liquide ne se "voit" pas, mais se mesure par divers montages qui, par l'effet de cette température, déclenchent des variations visibles qui en sont un indicateur. Avez-vous déjà vu de l'électricité, ou des électrons ? L'idée que tous les "data" accumulés (et accumulables) sont des indicateurs possède cette vertu éminente de susciter une réflexion constante sur la nature des objets étudiés, et la relation entre les structures sociales et certaines réalités substantielles qui les déterminent, qui les entourent ou qu'elles produisent.

L'histoire n'a pas pour objet l'espèce animale homo sapiens. L'erreur catastrophique de la « démographie historique » a été de croire et de *faire croire* qu'un décompte sophistiqué des naissances, mariages et décès pouvait suffire à constituer une discipline historique autonome. Mais, par la suite, on a jeté le bébé avec l'eau du bain : les paramètres biologiques ainsi mesurés (encore que le mariage...), s'ils contraignent à chaque moment une structure sociale, en sont bien davantage encore un produit ; et il n'est pas indispensable de disposer d'une capacité d'abstraction supérieure pour faire l'hypothèse que **tout produit d'une société peut être un bon indicateur de certains modes de fonctionnement de cette société**. Lesquels ? C'est là que la pente devient raide. Mais il n'y a pas d'autre issue.

Les "démographes" continuent impertubablement à considérer "la population" comme une substance en soi et pour soi, à propos de laquelle ils calculent des coefficients de plus en plus artificiels, qui ne nous apprennent rien sur quelque société que ce soit. Les historiens qui, partant en sens inverse, considéreront les séries démographiques comme des indicateurs et parviendront à découvrir les ensembles intriqués de relations sociales sous-jacents, obtiendront sans aucun doute des résultats novateurs. On peut toutefois supposer qu'il faudra décomposer ces séries, tenir soigneusement compte des groupes et sous-groupes, et se défaire une fois pour toutes du tristement célèbre "schéma d'urne" qui, ici encore, pousse à considérer les individus comme des boules indifférenciées dans un bocal opaque. On peut tenir pour assuré a priori que les ensembles sous-jacents seront différents d'une société à une autre, raison de plus pour tenir l'idée de "démographie historique" générale comme un piège redoutable.

On pourrait faire des remarques analogues sur "l'histoire des prix", qui a pareillement sombré ; les erreurs commises furent encore plus flagrantes, car dans ce cas des historiens ont

substantifié des grandeurs qui, en elles-mêmes, ne représentent déjà que des rapports. En soi, un prix isolé n'a aucun sens, puisqu'il s'agit d'une équivalence, et que les unités (grandeurs monétaires) n'existent que comme des équivalents ; l'histoire des prix ne peut être qu'une histoire de l'évolution (modifications permanentes) de rapports d'équivalence. Équivalences entre quoi et quoi, et *dans quelles limites* ? Quant aux incessantes modifications de ces équivalences, elles renvoient de manière aveuglante à des variations de rapports sociaux de toutes natures : il est inepte d'imaginer que l'on puisse faire une "histoire des prix" qui ne soit pas une histoire de toute la société considérée, dans la mesure (hypothèse minimale) où les prix sont un indicateur des fluctuations des rapports d'échange, lesquels constituent un des rapports sociaux les plus fondamentaux, selon des modalités *propres* à chaque société. Là encore, une "histoire des prix des origines à nos jours" est une impasse lamentable.

L'idée de considérer tous les "data" historiques comme des indicateurs **heurte le sens commun**. L'idée qu'une carte des églises construites à telle époque ou qu'une chronologie des maisons fortes de telle région soient de simples indicateurs d'une structure et d'une évolution sociale n'est pas encore très répandue. Mais on ne doit pas se désespérer : une thèse qui vient de paraître, fondée sur une analyse systématique des maisons fortes de Côte-d'Or, sur le terrain et dans les archives, se définit elle-même comme « une approche quantitative de la société féodale ». Si la notion d'indicateur n'est pas explicite dans cet ouvrage, elle y est de facto constamment présente et c'est grâce à elle que ledit travail apporte autant de nouveautés.

Accessoirement, on doit dissiper dès à présent deux équivoques. Ne pas confondre *indicateur* avec *indice*. Un indice est un nombre abstrait qui exprime l'évolution d'une grandeur quelconque, en général en considérant un point donné (au début ou à la fin) auquel est affecté par construction la valeur 100. Autrement dit, si, au moment tx une grandeur est à l'indice 130, cela signifie simplement qu'elle a augmenté de 30% entre t0 et tx. Le seul intérêt de cette procédure simplette est de pouvoir assez facilement comparer l'évolution relative de plusieurs grandeurs. En fait, c'est un système dangereux (on y reviendra à propos des séries chronologiques) dans la mesure où toutes les grandeurs sont exprimées par construction à partir de leur valeur au même moment t0, qui peut avoir des sens très variés selon les grandeurs ; si bien que les pourcentages que l'on croit pouvoir comparer peuvent avoir des significations toutes différentes.

Une autre ambiguïté frappe le terme de "data" que l'on a employé un peu plus haut, ou son équivalent français "données". C'est une grande banalité de rappeler que les "données" ne sont jamais données mais toujours construites (que ce soit par l'auteur du document ancien ou par l'historien contemporain qui compile). Mais cette mise en garde rituelle est le plus souvent oubliée dès que l'on a fini de l'énoncer, et surtout le raisonnement s'arrête pile, juste au moment où il faudrait au contraire poursuivre : la procédure de compilation, où qu'elle se situe, vise à fournir des renseignements sur un objet déterminé, et **le résultat de la compilation ne peut en aucune manière être confondu avec l'objet sur lequel il doit renseigner**. Or les statisticiens parlent en permanence de traiter des data ou d'analyser des données, supposant en fait inconsciemment une conformité parfaite entre l'objet étudié et les data disponibles, conformité qui tend vers une quasi identité. C'est une faute de raisonnement qui peut, dans certaines conditions (notamment expérimentales) n'avoir que peu de conséquences, mais peut au contraire être dirimante en sociologie ou a fortiori en histoire. ***L'objet de l'analyse est toujours "sous" ou "derrière" les data***.

1.3.3 biais

Cette dernière remarque nous conduit logiquement à la notion de ***biais***. Si les data ne sont que des indicateurs, il existe eo ipso un écart entre l'objet étudié et les tous les renseignements collectés, sous quelque forme que ce soit. On ne voit pas très bien comment ces ensembles de

renseignements pourraient inclure une information à la fois parfaitement fidèle et exhaustive sur l'objet étudié, dès lors surtout qu'il s'agit d'une société du passé. La seule hypothèse raisonnable consiste à supposer que les "data" fournissent **une information à la fois insuffisante et infidèle**. Ce que l'on peut exprimer de manière lapidaire : *il y a toujours un biais dans les données*.

Cette affirmation, au premier abord un tantinet déroutante pour le profane, doit en principe laisser le statisticien de marbre. Car le statisticien professionnel (qui travaille sur des données contemporaines et le cas échéant sur une société qu'il connaît bien) passe une grande partie de son temps à "redresser" des biais, c'est-à-dire à les identifier et à faire en sorte que l'interprétation finale n'en soit pas affectée. Le bon statisticien est celui qui "sent" les biais et parvient (par habitude et/ou par intuition) à les évaluer correctement. Et l'on comprend bien dès lors pourquoi une telle compétence, pas donnée à tout le monde, appliquée aux sondages d'opinion, aux études de marché, aux prévisions micro-économiques à court terme, puisse être si bien rémunérée. Diverses procédures statistiques peuvent permettre de repérer des irrégularités, des incongruités, des décalages inattendus que l'on peut ensuite, éventuellement, analyser comme des biais. Et cela dans le domaine historique comme ailleurs. Le handicap apparent de l'historien est qu'il ne peut pas reprendre les expériences ni refaire un sondage complémentaire en modifiant le questionnaire. Les sources sont ce qu'elles sont, on ne peut pas en inventer. On ne doit cependant pas exagérer cette différence, car bien souvent l'expérience médicale ou le sondage ne peuvent pas non plus être refaits et le statisticien doit se débrouiller avec les data dont il dispose, sans supplément. On peut même dire cum grano salis que l'historien est dans une situation plus confortable, dans la mesure où une erreur d'interprétation de sa part aura rarement des conséquences concrètes désastreuses (ce qui ne justifie pas l'irresponsabilité !).

Les biais les plus nocifs sont naturellement **ceux que l'on n'aperçoit pas** et qui peuvent soit bloquer complètement une analyse, soit entraîner une interprétation fausse. Un exemple classique est celui de l'**hétérogénéité** invisible à première vue : on examine un ensemble d'objets, ou d'individus, et l'on tente de mettre en évidence les structures de cet ensemble. Si l'ensemble en question est constitué en fait de deux sous-population d'effectifs équivalents et de structures bien distinctes, il sera à peu près impossible de découvrir ces structures ; l'analyse sera bloquée tant que l'on n'aura pas réussi à faire le tri. En face d'une telle constatation, la réaction irréfléchie consiste à conclure que la statistique historique propose un programme irréalisable, exigeant de grands efforts pour un résultat plus qu'incertain ; un peu d'attention suffit pour montrer tout au contraire la nécessité de l'inclusion de la statistique dans tout effort de recherche historique sérieux ; car **si l'on a compris que toutes les données sont plus ou moins biaisées, on ne voit pas au nom de quel principe l'on s'autoriserait à renoncer aux seules procédures disponibles** pour traquer, sinon pour éliminer lesdits biais. Sauf à admettre l'identité de l'histoire et du roman, ou de la dentelle...

1.3.4 imprécision et approximation

Supposer méthodiquement un biais n'indique pas l'importance que l'on attribue à ce biais. Or la situation change du tout au tout selon que l'on est en face d'un biais minuscule ou d'un biais massif. Ce qui nous amène à évoquer la question de ***la différence entre imprécision et approximation***. On arrive ici au cœur de la statistique. La connaissance commune est imprécise, et le bon sens croit que l'on peut distinguer entre ce que l'on sait et ce que l'on ne sait pas. Cette logique binaire est simple à manipuler dans la plupart des cas, et suffit le plus souvent dans la vie courante. Le passage à la connaissance scientifique implique d'y renoncer, pour passer à la notion basique d'approximation, c'est-à-dire d'**incertitude contrôlée et mesurée**. La finalité de toute science étant de réduire l'amplitude des approximations. Le réflexe professionnel qui définit l'homme de science est qu'à chaque instant il se pose ces questions : jusqu'à quel point suis-je sûr de ceci ou de cela ? dans quelles limites telle ou telle affirmation est-elle valide ? quel est le degré de

précision des informations qui permettent telle ou telle déduction ? (c'est au demeurant pour cette raison qu'un scientifique est le contraire d'un expert, qui est, lui, l'homme qui sait et répond par oui ou par non). Deux notions dépendent directement de cette opposition, celle de *fourchette* et celle de *seuil*.

La notion de **fourchette** a été popularisée par la diffusion des sondages pré-électoraux. Mais le public ignore de quoi il s'agit. Les "fourchettes" sont un élément parmi beaucoup d'autres du calcul des probabilités et de la statistique théorique ; elles n'en sont naturellement pas indépendantes. Leur utilisation implique diverses hypothèses, en particulier sur la forme des distributions des erreurs possibles (nous examinerons dans le prochain chapitre la question des "distributions"). Le cas des pourcentages de voix dans des élections est un cas élémentaire pour lequel les hypothèses classiques ont été largement validées. Mais il n'en va pas toujours ainsi, et il faut savoir que, contrairement à ce que croient beaucoup de gens, **les incertitudes ne se compensent pas, elles s'additionnent**. Si bien que l'on se trouve aisément dans des situations où une évaluation grossière du degré d'incertitude aboutit à des résultats décourageants... Encore une fois, il vaut mieux le savoir que l'ignorer.

L'approximation est liée de plusieurs manières à la notion de **seuil**, par exemple au travers du problème des limites de validité, mais aussi de la question du choix du taux de probabilité qui délimite le certain et l'incertain. Il faut savoir également que **la sémantique repose dans une large mesure sur l'emploi des seuils et sur la manière de les déterminer**, et c'est d'ailleurs au travers de cette notion que s'opère le plus nettement le rapprochement entre statistique et sémantique. L'opposition du jour et de la nuit repose sur des considèrations astronomiques qui ne sont guère discutables ; mais les deux passages ne sont pas instantanés et il faut donc, dans certains cas, fixer les critères permettant de préciser cet instant ; l'opposition du blanc et du noir n'est pas plus contestable, mais le passage de l'un à l'autre s'opère tout à fait continûment, il n'existe aucun seuil "naturel". La « charte de couleurs » de Kodak (KODAK Gray Scale) utilisée par les photographes professionnels comporte 20 cases, une pour le noir, une pour le blanc et 18 pour les gris. On pourrait sans peine en distinguer bien davantage, mais les ingénieurs ont estimé qu'un tel découpage était suffisant pour un repérage précis. L'efficacité de tel ou tel découpage, plus ou moins fin, avec des intervalles équivalents ou non, dépend de l'usage que l'on doit en faire.

Dans la société, ce genre de difficulté se rencontre à chaque instant, lorsqu'il s'agit de répartir entre les jeunes et les vieux, les pauvres et les riches, mais aussi les doués et les incapables, les charmeurs et les revêches, les progressistes et les réactionnaires, etc. La nature des échelles varie considérablement, depuis l'échelle numérique unique (âge), l'échelle numérique plus ou moins complexe (richesse), l'évaluation pifométrique multicritère (charme ?), jusqu'à l'indication vaguement arbitraire sur une échelle dont ne sait pas si elle est linéaire ou même si elle ne doit pas être décomposée ("échiquier politique"). Il est apparent que les découpages issus de ces échelles 1. sont **hétérogènes** (ne sont pas de même nature), 2. sont **arbitraires** (on peut en imaginer plusieurs pour chaque échelle). Mais on doit immédiatement ajouter que **ces découpages constituent le fondement de l'organisation de la société**, et même que **c'est précisément ce caractère "arbitraire" qui distingue la société humaine de toutes les autres sociétés animales. Tout processus social met en jeu des classements**, qu'il s'agisse seulement d'utiliser des classements établis dans des opérations routinières, ou d'opérer le classement de tel ou tel, tels ou tels individus, ou encore de discuter voire de modifier des critères de classement.

Et c'est pour cela que l'on a dit plus haut qu'il existe **une parenté qui confine à l'identité entre structure sociale et sens**. Les processus cognitifs élémentaires et le langage ordinaire reposent d'abord sur des classements, très souvent fort simples, par pure opposition (découpage binaire : fort/faible, éloigné/rapproché, gentil/méchant, etc.). Et les sociétés se distinguent entre elles, profondément, par la manière dont elles organisent leurs critères de classement et la façon

dont elles les mettent en œuvre (d'où, soulignons le au passage, le rôle clé que devrait jouer la sémantique à la base de toute recherche historique).

L'historien qui se lance dans une analyse statistique doit à la fois déterminer les modes de classement (seuils) utilisés par la société qu'il étudie, et élaborer les siens propres, en les organisant de telle sorte qu'il parvienne à mettre au jour la logique de la société qu'il étudie, logique que les classements "indigènes" mettaient en œuvre **en même temps qu'il la camouflaient** (tout système social de représentation du monde doit permettre à la société de fonctionner tout en masquant le mieux possible ses tensions et contradictions profondes).

Si l'on dispose de mesures précises et continues, on peut avoir intérêt à les utiliser directement, il existe une foule de méthodes statistiques pour cela. Mais même dans ce cas, la question des seuils ressurgit, ne serait-ce que pour déterminer les limites de la population (objets) considérée : on a rappelé un peu plus haut que l'hétérogénéité entraîne les pires difficultés pour toute analyse statistique. Le dénombrement et a fortiori la mesure sont inconcevables si l'on n'a pas défini les objets à dénombrer, et définition implique strictement classement. C'est une question difficile, que l'on tentera d'éclairer autant que possible au cours des prochains chapitres, mais il est indispensable d'avoir dès l'abord une idée du caractère central et récurrent de cette difficulté, et de ses implications décisives.

1.3.5 seuils

La question du caractère "arbitraire" des seuils est, comme on vient de le voir, à la fois centrale et difficile. En gros, trois cas de figure :

a) il existe des seuils naturels, des frontières de fait, comme jour/nuit, homme/femme, etc.

b) chaque société, à tel moment, fixe des limites qui prennent dès lors un caractère factuel (e.g. âge de la majorité légale, plafond de la Sécurité Sociale, etc.).

c) le sens commun aussi bien que la démarche scientifique découpent en permanence dans des continuums, tout simplement pour dénommer, procédure initiale sans laquelle aucune pensée ni aucun processus social ne sont possibles.

Dans les faits, ces trois cas sont en général imbriqués les uns dans les autres. Si la rotation de la terre détermine sans choix possible des unités "réelles", le jour et l'année, des grandeurs comme l'heure d'un côté, ou la semaine et le mois sont d'origine purement "sociale" ; ces grandeurs sont clairement définies (dans chaque société, avec les innombrables variantes liées notamment aux techniques de la mesure du temps) ; mais que faire de "l'époque", ou de "la génération" ? Ou, pire encore, des incontournables "dates-charnières" ? On est renvoyé à la question des degrés de "réalité" et d'"utilité" des seuils. Et, ici même, les techniques statistiques, intelligemment employées, peuvent rendre des services décisifs. Notamment à partir de l'idée de la ***possible multiplicité des solutions optimales***.

Expression savante pour parler de ce que tout le monde sait : il existe souvent (mais pas toujours) deux itinéraires "équivalents" entre le point A et le point B. Et cette "équivalence" n'est possible que s'il y a accord préalable sur les critères (dits "d'optimalité"). Or ces critères, dans le cadre d'une analyse scientifique, sont peu variables : cohérence logique, économie de moyens, conformité maximale avec les données. Toutefois, **la hiérarchie (l'ordre de mise en œuvre) de ces critères est à peu près indéterminable**. En informatique, la situation est courante : deux programmes organisés différemment et impeccables l'un et l'autre peuvent permettre d'obtenir le même résultat. Une réflexion très formalisée a porté dans les années 70 sur la question de l'architecture des "bases de données relationnelles", et l'on a démontré que dans de nombreux cas deux architectures différentes pouvaient être dites "optimales" (i.e. non améliorables), et bien entendu permettre d'exploiter les données de la base de la même manière.

En statistique, la situation est différente, quoiqu'en disent certains statisticiens. Car, en face

de nombreux problèmes, il existe **un éventail de méthodes disponibles, cohérentes et économiques, mais qui n'aboutissent pas toujours au même résultat !** Cette constatation est de prime abord un peu troublante, ce qui conduit certains auteurs de manuels à proposer, sans trop de justification, des "raisons" de choisir telle ou telle méthode dans tel ou tel cas. D'autres, plus nombreux, font semblant de n'avoir rien vu, exposent "leur" méthode et ne font aucune allusion aux autres. Un peu de réflexion montre au contraire que le trouble n'est qu'apparent, et que cette variété est au contraire une force de la statistique (bien comprise). La facilité actuelle de mise en œuvre des procédures les plus variées doit conduire à **employer systématiquement diverses méthodes pour traiter un même problème**, précisément pour voir dans quelle mesure les "résultats" convergent ou divergent. Si les résultats convergent, c'est (probablement) que l'on a affaire à une structure forte et claire, on peut passer à l'étappe suivante. La divergence est, d'une certaine manière, plus instructive, car elle oblige à se demander où sont les causes de cette divergence qui peut n'être qu'apparente, c'est-à-dire résulter d'un ou plusieurs choix opérés explicitement ou implicitement dans le cours des procédures.

Et c'est là que l'on peut le mieux **voir l'effet des découpages, et du caractère, réel ou fallacieux, de tel ou tel seuil**. Si plusieurs découpages d'un même caractère aboutissent au même résultat, cela signifie probablement que l'on a affaire à un vrai continuum ; si au contraire des découpages différents aboutissent à des résultats statistiques différents, cela constitue un indice appréciable du fait que l'on est en présence d'un ou plusieurs seuils "effectifs". D'une manière générale, on doit partir du principe qu'au cours de toute analyse statistique, il faut faire varier les catégories et employer successivement sinon toutes les méthodes disponibles, du moins plusieurs, aussi bien d'ailleurs celles réputées "bonnes" que celles réputées "inappropriées".

On comprend sans peine que, pour procéder ainsi, un "seuil de compétence" minimal est requis. *Des calculs simplets sont souvent pires que pas de calcul du tout.* C'est ce qui fut une des causes de la débâcle de "l'histoire quantitative". Dès lors, comme on l'a rappelé ci-dessus, que l'historien travaille toujours avec des *indicateurs biaisés*, une statistique quelque peu **élaborée** peut apporter une aide déterminante voire irremplaçable, tandis que quelques sommes, moyennes et pourcentages donneront aux biais, sans que l'on sache ni pourquoi ni comment, l'allure d'une pseudo-objectivité.

1.3.6 exploration

Une notion qui synthétise en partie ce qui vient d'être dit est celle d'***exploration***, en ce sens que ce terme indique schématiquement la nature et l'orientation du travail statistique de l'historien. Comme on l'a dit plus haut, les statistiques appliquées visent communément la prévision, la surveillance, l'aide à la décision. Rien de tout cela dans le labeur de l'historien. Celui-ci a pour fonction, à partir des *sources*, de reconstituer autant que faire se peut le fonctionnement et l'évolution des sociétés du passé. Les procédures statistiques interviennent donc comme outil d'aide au progrès des connaissances abstraites, au sein du processus de va-et-vient constant entre les connaissances acquises, les lacunes qu'elles comportent et les questions qui en résultent, et les sources, qui constituent la matière première propre à l'historien. On l'a rappelé : le stock des sources accessibles varie peu, sinon du côté de l'archéologie (en principe, il ne devrait pas être nécessaire de rappeler aux chartistes le rôle déterminant de la confection des inventaires et catalogues, que les universitaires ignorent en général).

Beaucoup se sont imaginé, au 19e siècle, que des sources bien ordonnées équivalaient à peu près à une histoire bien constituée. **Erreur** considérable qui a abouti dans l'impasse de ce que l'on a appelé depuis (par abus de langage) l'histoire "positiviste". Si les sources doivent demeurer (ou peut-être même revenir) au centre des préoccupations de l'historien, elles constituent pour ainsi dire la matière première et nullement le produit fini. On a évoqué ci-dessus les notions clés d'indicateur

et de biais. Si l'on se place à un degré un peu plus élevé d'abstraction, on peut se hasarder à parler de *configuration de l'information*. Une approche traditionnelle consisterait à dire que les sources contiennent l'information à l'état brut, tandis que les historiens livrent l'information à l'état élaboré. Mais, à y regarder de plus près, et en examinant empiriquement les procédés concrets de la recherche historique, on se rend compte qu'il s'agit de bien plus que de cela. Les sources, surtout les sources écrites, ont tout autant pour effet de déguiser la réalité que de la mettre en évidence (si, comme un collègue l'a écrit, les sources matérielles, objets et bâtiments, ne mentent pas, c'est simplement qu'elles sont muettes !) Dès lors, **le travail de l'historien est bien moins de trier que de reconstituer les fondements des processus d'énonciation** qui ont abouti à l'élaboration des sources telles qu'elles se présentent. L'information décisive, celle qui est la plus pertinente, est marginale, camouflée, sous-jacente ; il faut la reconstituer à partir d'éléments dérivés et toujours partiels : les structures ne sont jamais visibles, alors même que ce sont elles qui constituaient la société en tant que société et formaient l'ossature du sens de tous les documents. Autrement dit, pour parler en termes d'information, le travail de l'historien consiste à transformer une **information *potentielle*** en **information *disponible***. Dans cette perspective, les méthodes statistiques apportent une aide déterminante au repérage, à l'organisation et à la structuration de l'information. C'est ce que l'on peut résumer en disant que les statistiques historiques ont pour finalité l'*exploration*.

Cette conclusion permet de lever une autre équivoque courante, qui empoisonne depuis des décennies les relations entre histoire et statistique : celle qui consiste à prétendre que la statistique permettrait, là où elle intervient en histoire, d'apporter des "preuves" que l'histoire seule serait incapable de fournir. Cette affirmation est grossièrement fausse pour au moins deux raisons : 1. aucun domaine de l'histoire n'échappe a priori à l'utilisation des méthodes statistiques ; 2. les méthodes statistiques peuvent fournir une aide importante, voire décisive, pour reconstituer des structures, mais les raisonnements sont toujours des raisonnements historiques, les nombres permettent de cerner des relations, mais jamais d'en saisir la nature. (Au demeurant et au surplus, **il est permis de douter de la pertinence de la notion de "preuve" en matière scientifique en général**, historique en particulier).

Cette même erreur se retrouve dans l'opposition inepte entre "qualitatif" et "quantitatif" ; *tout se classe et dès lors tout se dénombre* ; mais aucun empilement de dénombrements ne restitue une structure. Fréquence, intensité, poids, sont des caractères essentiels des relations sociales, de toute relation sociale (voir plus haut : « ordre de grandeur »), mais jamais le caractère unique. La statistique peut et doit intervenir partout en histoire, précisément pourrait-on dire parce que le syntagme « histoire quantitative » ne renvoie qu'à un fantasme inconsistant.

1.3.7 formalisation

Le principe de base "tout se classe" entraîne à réfléchir sur la notion de **formalisation**. Le sens commun le plus ordinaire est fondé sur l'emploi permanent de qualificatifs pris (naïvement) pour élémentaires et intangibles : petit/grand, jeune/vieux,pauvre/riche, devant/derrière, rond/carré, etc. L'analyse historique sérieuse implique :

a) de se rendre suffisamment compte que *ce "sens commun" est toujours **daté*** (donc variable et instable), assez pauvre sinon indigent, souvent au bord de l'incohérence. La trop fameuse "familiarité avec les sources", souvent valorisée, n'est en général qu'une variante de sens commun prétentieuse et autosatisfaite, qui n'a pour effet que de légitimer le traditionalisme et l'irréflexion.

b) de réfléchir énergiquement aux propriétés des objets examinés, susceptibles d'être définies, c'est-à-dire à propos desquelles une grille de lecture pourra être ***construite***. Ici se conjuguent toutes les difficultés évoquées plus haut à propos d'*approximation*, de *seuil*, d'*exploration* ; à quoi il faudrait peut-être encore adjoindre des réflexions sur des notions comme celle de *pertinence* et de *cohérence*. Mais ces considérations abstraites (préalable nécessaire) trouvent leur point de chute et

leur efficacité au moment concret de **la formalisation, c'est-à-dire la construction et l'explicitation du système de caractères et de catégories** au moyen duquel on passe des objets considérés ("sources") à une description méthodique et ordonnée, qui seule rend possible des comparaisons claires (c'est-à-dire précisément des procédures statistiques, tant il est vrai que la statistique n'est que l'art des comparaisons raisonnées).

Cette formalisation est un moment décisif de l'analyse historique sérieuse : on en retire toujours des bénéfices importants, et c'est peut-être un des principaux mérites de la statistique que d'y contraindre. Mais l'expérience montre que cette formalisation se heurte, dans le domaine historique, à des obstacles récurrents, qui donnent aux data historiques quelques caractères propres, qui obligent le chercheur à adapter, parfois assez vigoureusement, les procédures statistiques standard.

POUR CONCLURE : caractères propres des objets et de la statistique historiques

La plupart des manuels de statistique évoquent, en introduction et très brièvement, les caractères requis des "données" sur lesquelles on peut appliquer une analyse statistique. Le plus souvent, il est dit que l'information doit être **homogène** et **exhaustive**. Or, tout au contraire, l'information spécifiquement historique est plutôt **hétérogène**, **lacunaire**, **déséquilibrée**.

On reviendra, le moment venu, sur les effets de la durée : des données étalées sur une durée, même parfois assez courte, sont ab ovo hétérogènes : toute structure sociale est par nature en mouvement, et ainsi des éléments qui se rapportent à des états successifs de cette structure renvoient à un cadre transformé, donc à des significations différentes. C'est une des difficultés majeures du métier d'historien d'apprécier dans quelle mesure ou jusqu'à quel point un mot, ou tout autre objet, peut être considéré comme ayant la même signification malgré des évolutions de la structure englobante. Strictement parlant, seule la contemporanéité exacte apporte quelques garanties (limitées) ; en pratique, "il faut voir" ! En tout cas, ne jamais oublier que la "longue durée" est un piège, et que les séries chronologiques historiques sont par nature hétérogènes (l'économétrie part de l'hypothèse inverse, fausse, et c'est une des causes majeures de ses échecs constants).

Les données "à trous" sont le lot ordinaire de l'historien. Si les trous sont d'ampleur mesurée, on peut les ignorer ("interpoler"), il existe des méthodes appropriées ; mais il est encore plus habituel de n'avoir que des bribes dispersées. Dans cette situation, les méthodes statistiques sont encore plus nécessaires : elles seules apportent le cadre méthodologique apte à évaluer le degré d'incertitude qui entre dans les reconstitutions que l'on peut tenter à partir de ces bribes.

La pratique historienne est perpétuellement confrontée au "trop ou pas assez". Les sources étant ce qu'elles sont, il y a surabondance de matériaux sur certains points et silence sur d'autres. Ici encore, les modes de raisonnement statistiques aident à ne pas confondre l'ampleur des sources avec l'importance des éléments structurels, moyennant des procédures appropriées. Notons au passage que les "petits effectifs" ne sont nullement rédhibitoires, au contraire : la statistique historique n'a **pas besoin** de "grands effectifs", les procédures de formalisation sont précisément une aide irremplaçable lorsqu'on travaille sur des populations de taille très réduite.

Au total, il apparaît que ces caractères (hétérogénéité, lacunes, déséquilibres) sont en réalité une raison déterminante de faire appel aux raisonnements statistiques. La difficulté (actuelle) provient du fait que la plupart des travaux statistiques ont été appliqués à des données présentant le plus souvent des caractères inverses : un effort substantiel d'adaptation et d'expérimentation est indispensable. C'est pourquoi il est légitime de parler de "statistique historique" (comme on parle de

"biostatistique"). Si le champ est balisé, si des expériences concluantes ont été réalisées, il n'en demeure pas moins qu'une mise sur pied organisée, systématique et cohérente est encore à venir : l'aventure est devant nous...

Ces quelques réflexions auront, espérons-le, permis au lecteur de comprendre, au moins intuitivement,

1. pourquoi il est déraisonnable de se jeter tête baissée dans des manuels généraux de statistique qui exposent les théorèmes du calcul des probabilités ou des procédures de calcul sans jamais se demander à quoi cela sert ni a fortiori quelle est la relation entre ces méthodes et les objets traités ;
2. qu'une réflexion préalable, et continue, sur les objets mêmes de la recherche historique est indispensable si l'on veut introduire, et utiliser avec profit, une nouvelle technique au sein de cette recherche. D'une telle réflexion, on peut attendre un double bénéfice : une meilleure compréhension de la nature intrinsèque de la recherche historique (d'où résulte presque à coup sûr une capacité plus élevée à orienter et à conduire sa recherche) ; la possibilité de maîtriser un attirail technique profus, dont le développement a été largement orienté par des problèmes et des objets qui ne sont pas ceux de l'historien, et qu'il importe donc de remodeler progressivement. Vulgo dictu : place aux jeunes, on embauche !
3. que l'on ne peut plus faire semblant d'ignorer cette possibilité, même si cela oblige à repenser les fondements mêmes du raisonnement historique.

Chapitre 2

DISTRIBUTIONS UNIVARIÉES

C'est sur ce point que les manuels sont les plus diserts, pour ne pas dire prolixes. On s'y référera pour trouver tout ce dont on peut avoir besoin pour analyser les "lois" (= équations) discrètes et continues. Le manuel de Jacques Calot est sans doute le plus précis et le plus accessible, c'est un bon outil de travail. Toutefois, l'extrême faiblesse des outils d'analyse des distributions parétiennes (= non gaussiennes) demeure un handicap très gênant. Le présent chapitre n'a pas du tout pour objet de résumer ces manuels, mais de fournir un ensemble de notions générales susceptibles d'orienter une stratégie de recherche et d'éviter de tomber dans les pièges les plus grossiers et les plus courants. Qui se doute qu'une simple "moyenne" peut entraîner les pires erreurs d'interprétation ??

SOMMAIRE

2.1. REMARQUES GÉNÉRALES PRÉALABLES

2.1.1 la terminologie et ses pièges

On emploie en statistique le terme distribution avec le sens de " répartion, arrangement ". Il s'agit d'un terme technique, qui implique que le (ou les) caractère que l'on cherche à examiner soit ordonné ou pour le moins regroupé. Une distribution concerne une population, constituée d'un ensemble d'individus; chaque individu est muni d'un caractère.

On doit bien prendre garde que la terminologie de la statistique, comme de toute autre science ou technique, est composée en partie de termes dits "savants" (c'est-à-dire plus ou moins artificiels et sans usage dans le langage courant, e.g. covariance, orthogonalité, paramètre, bijectif...), mais surtout de vocables ordinaires, qui prennent, dans le cadre de cette science, un sens bien défini, souvent **complètement différent de celui de l'usage courant** ; le cas est particulièrement fréquent en statistique, et induit inévitablement chez le néophyte des risques permanents de contresens ; outre les termes que l'on vient d'utiliser, on peut citer confiance, espérance, événement, indépendance, loi, hypothèse, normal...). **NB** il sera indispensable de se familiariser avec les équivalents anglais (voire allemands), la majeure partie de la littérature disponible aujourd'hui étant rédigée dans ces langues.

2.1.2 un binôme essentiel : "observé / théorique"

On distingue couramment " distribution observée " (ou " distribution empirique ") et " distribution théorique " ; la première ne pose pas de difficulté apparente : il s'agit des données dont on dispose. La seconde au contraire n'est pas du tout intuitive, puisqu'elle renvoie au calcul des probabilités et à la théorie qui la fonde. Elle implique la notion de " modèle " ou de " loi ".

On doit souligner dès le départ que toute pratique statistique doit combiner les deux, c'est-à-dire l'examen des distributions empiriques et la recherche des " distributions théoriques " correspondantes, ou pour le moins une réflexion approfondie sur le type possible de ces distributions. Comme on va essayer de le montrer, une erreur dans cette démarche (absence de réflexion sur la distribution théorique, ou emploi d'une distribution théorique erronée) peut conduire (et conduit souvent) à des **erreurs graves d'interprétation des données empiriques** (voir notamment le chapitre 11).

2.1.3 perspective de la statistique historique

Le présent chapitre est consacré à la question en apparence la plus simple, c'est-à-dire à l'examen d'une distribution observée à un seul caractère. Cette ***exploration,*** si elle est conduite convenablement, c'est-à-dire si elle aboutit à l'identification de la distribution théorique correspondante la plus plausible, apporte toujours des informations inédites sur la population munie du caractère considéré.

L'avantage de la situation technique actuelle est de fournir des outils très commodes. On peut et on doit utiliser dans cette première phase **la gamme la plus large possible de méthodes numériques et graphiques** permettant d'inspecter la distribution empirique sous toutes ses coutures, de manière à en obtenir une vue détaillée, laissant échapper le minimum d'aspects pertinents. Si cette phase est contournée, ou si elle aboutit à une conclusion erronée, la suite des opérations peut être radicalement compromise : la littérature regorge d'exemples.

2.1.4 les principaux types de caractères (variables)

On utilise l'expression " distribution univariée " pour parler d'une " variable à un seul caractère ". Chaque objet (= " individu " au sens statistique) est " muni " d'un seul caractère (ou du

moins on ne le considère que sous cet angle). On distingue ordinairement quatre principaux types de caractères : *catégoriel*, *ordonné*, *numérique discret*, *numérique continu*.

a) catégoriel signifie simplement que chaque individu est affecté d'une **modalité** du caractère considéré. Ex. : la couleur, le type d'objet, l'appartenance à un sous-ensemble (topographique, socio-professionnel, politique, etc). Dans un fichier, la variable se présente soit sous forme explicite (“ bleu ”, “ vert ”, etc) soit sous forme de code, alphanumérique ou purement numérique.
[Dans le cadre du logiciel R, il faut faire en sorte que la variable soit définie comme “ facteur ” (commande `as.factor(x)`)].
Il faut **strictement éviter de parler de “ variable qualitative ”**, ne serait-ce que parce que les modalités peuvent correspondre à des classes de taille (ou de fréquence, ou de n'importe quel caractère numérique). Les variables catégorielles peuvent être éventuellement ordonnées (ordre de taille, de préférence, période, etc).

b) ordonné renvoie à une variable qui n'est spécifiée que par son **rang** dans l'ensemble (avec peu d'ex-aequo, sinon il s'agit d'une variable catégorielle ordonnée).
NB Toute variable numérique peut être convertie en variable ordonnée, au prix d'une certaine perte d'information. C'est une procédure utilisée en particulier dans le cadre de la recherche de “ corrélation ” entre deux variables. Mais on peut aussi convertir une variable numérique univariée en variable bivariée, en considérant pour chaque individu une valeur numérique et le rang correspondant. C'est une procédure dont il est fait un usage courant dans le cadre des statistiques “ non-gaussiennes ”.

c) numérique discret correspond en général aux “ entiers naturels ” (éventuellement négatifs) ; autrement dit, les valeurs se succèdent par sauts, sans intermédiaires possibles. Exemples classiques : le nombre d’enfants d’une famille, le nombre d’occurrences d’un terme dans un texte. Le plus souvent, un caractère numérique discret résulte d’un **comptage** ou dénombrement.

d) numérique continu peut correspondre à n’importe quel nombre réel (concrètement, autant de décimales que la machine en supporte). En théorie, entre deux valeurs si proches qu’elles soient, on peut toujours en intercaler une troisième. En général, un caractère numérique continu résulte d’une mesure ; soit d’une **mesure au sens étroit** (ex. poids, longueurs, etc : si l’on agrège deux individus, les deux mesures s’agrègent par simple addition), ou de **mesure au sens large** (ex. température : si l’on agrège deux ensembles, il faut trouver une valeur intermédiaire, au moyen d’une procédure de pondération adéquate).
NB Il faut *se méfier de tous les indices*, dont on doit user de manière différente selon les procédures ; si par exemple on considère le pourcentage en voix de deux candidats dans deux zones, on peut additionner les pourcentages des deux candidats dans la même zone, mais pas les pourcentages du même candidat dans deux zones différentes (dans ce cas, on doit faire une moyenne pondérée en fonction du nombre de voix exprimées dans les deux zones). Comme on le verra par la suite, les procédures de calcul ne se préoccupent pas de telles considérations, et *additionnent ou moyennent des nombres sans rechigner*. Ce qui peut entraîner des résultats surprenants, soit que le résultat paraisse absurde, soit au contraire que le résultat soit tout à fait intéressant alors que l’on a utilisé une procédure qui paraissait a priori illégitime. Inversement, de nombreuses procédures numériques refusent de traiter des valeurs négatives ou même nulles (cas classique d’une transformation d’une valeur absolue en valeur logarithmique, il existe des parades !).

Dans la réalité du travail de l'historien, les choses se présentent le plus souvent de manière bien moins nette. On verra en particulier (à l'expérience...) que la distinction (indispensable au plan

abstrait) entre "continu" et "discontinu" (discret) est *souvent problématique*, sinon même exaspérante.

2.1.5 trois questions préalables à toute exploration

a) les divers types de données renvoient à des " **degrés de précision** " variables. Ici, rien ne peut remplacer une " bonne connaissance des données " ; s'agissant par exemple de mesures de longueur, il vaut mieux savoir exactement comment les longueurs ont été mesurées pour avoir une idée du nombre de chiffres significatifs utilisables. La réflexion sur la précision relève de ce qui a été dit dans le premier chapitre sur les ordres de grandeur. On n'a en général aucune raison de tronquer les chiffres, même si les nombres disponibles paraissent comporter une précision fictive. De même, *tous les calculs doivent être effectués avec la précision maximale* (raison pour laquelle R utilise systématiquement la " double précision "). En revanche, *on doit s'interroger très attentivement sur la précision des résultats obtenus*. On n'est en aucun cas autorisé à donner des résultats comportant plus de chiffres significatifs que les données (indiquer par exemple 35,58%, alors que les données ne comportent que deux chiffres significatifs : erreur fréquentissime). En règle générale, il faut essayer d'évaluer la précision des données, et fournir des résultats compatibles avec cette précision de départ (c'est-à-dire souvent d'une précision moindre). Dans le domaine historique, il s'agit d'une question primordiale, souvent un tantinet déstabilisante, mais qu'il est nécessaire de prendre en considération méthodiquement tout au long de l'exploration.

b) similairement, il faut se demander **ce que signifie l'effectif** de la population que l'on étudie. **Population totale, population tronquée, échantillon, mélange de populations ?** Dans la majorité des cas, l'historien n'en sait rien a priori. Même dans des cas en apparence limpides, comme un texte dont on analyse le vocabulaire : on s'imagine avoir devant soi "tous les mots" dudit texte. Mais est-on toujours sûr qu'un paragraphe, voire toute une partie, ne sont pas de la plume d'un autre auteur ? Il est de loin préférable de partir de l'hypothèse que, peut-être, certains individus n'appartiennent pas à la population que l'on étudie, ou qu'au contraire certains individus qui en font partie n'ont pas été pris en considération. Bien des anomalies résultent de ce fait.

c) en général, on se contente de distinguer catégoriel et numérique. Certains logiciels distinguent une grande quantité de " types de variables ", soit pour gagner de la place, soit pour faciliter certains calculs (par exemple sur les variables " date "). Un logiciel statistique comme R distingue bien moins, et se contente de " `numeric` " et " `factor` ". En pratique, et selon les besoins, il est le plus souvent (pas toujours) possible de **transformer un type de variable dans un autre type**. Contrairement à ce que suggèrent bien des manuels (qui laissent croire à une certaine " substantialité " des types de données), on a en général intérêt à procéder à diverses transformations, **pour voir**. Certaines transformations "facilitent la lecture", mais il arrive souvent que la recherche d'une transformation adéquate soit le moyen privilégié pour découvrir le processus sous-jacent qui a généré les données.

Pour de multiples raisons, on peut être amené à découper un caractère numérique en classes (transformation d'une variable numérique en variable catégorielle). C'est souvent une quasi-nécessité dans le cadre des analyses multivariées (analyse simultanée de plusieurs caractères simples se rapportant à une même population). Un tel découpage peut s'inspirer de diverses règles (qui sont plutôt des recommandations), et parmi celles-ci figure la règle qui veut que l'on place les bornes des classes en fonction de la forme de la distribution. C'est manifeste dans certains cas, moins dans d'autres : encore faut-il avoir bien vu la forme de la distribution.

2.1.6 finalités de l'exploration d'une distribution observée

Avant de passer en revue les principales procédures d'examen des distributions univariées, il convient encore de clarifier quelque peu les finalités d'une telle opération. On peut distinguer des

finalités techniques (i.e. étapes dans une analyse globale d'un ensemble de données, ensemble dans lequel figurent des données univariées en plus ou moins grand nombre, et de natures différentes) et des finalités intrinsèques, c'est-à-dire constituant elles-mêmes des résultats ou, en d'autres termes, des éléments d'information sur les données auxquelles s'appliquent ces caractères univariés.

a) finalités techniques

La finalité classique, celle sur laquelle insistent les manuels qui évoquent cette question, consistait à *résumer* une variable. Soit un caractère numérique connu pour une population de 1000 individus ; si l'on peut " résumer " l'information contenue dans ces 1000 nombres en trois nombres en ne perdant que 1 ou 2% de l'information, on simplifie les procédures suivantes, et en particulier toutes les tâches de comparaisons entre distributions à comparer. En termes techniques, on dit " identifier la loi de la distribution et ses paramètres ". C'est le grand art des " statistiques classiques ". En fait, comme on l'a déjà souligné, la notion de résumé et de simplification, qui pouvait avoir une grande importance à l'époque des calculs manuels, a perdu la plus grande partie de son intérêt.

Inversement, il faut souligner l'importance de cette étape dans la perspective du découpage en classes d'une distribution numérique (procédure courante). Un choix raisonnable des bornes entre classes implique de les placer à des endroits significatifs de la courbe de densité.

b) finalités intrinsèques

La statistique est **l'art des comparaisons raisonnées**. A propos d'une distribution univariée, le sens commun distingue assez aisément trois objectifs :

* définir la place d'un individu dans la population observée ;
* comparer le même caractère dans deux populations ;
* estimer la place de la population observée dans une population " parente ", plus ou moins reconstruite à partir de la population observée (e.g. évaluer la distribution du vocabulaire d'un auteur à partir de la distribution du vocabulaire d'un texte connu).

Les procédures sont bien entendu différentes, mais toutes reposent sur une connaissance aussi exacte que possible de la distribution observée, en particulier de sa forme, qui est un élement déterminant dans la recherche d'une " distribution théorique ", sans la connaissance de laquelle on aboutit sans s'en apercevoir aux interprétations les plus fantasistes. Ce qui amène à dépasser les questions du sens commun, pour voir apparaître l'**enjeu majeur de cette exploration**, à deux faces :

1. établir la meilleure approximation possible du processus qui a engendré les données,
2. par différence, faire ressortir les écarts, les anomalies, les déséquilibres.

Si l'on y parvient, une information inédite apparaît, bien souvent déterminante par rapport à la recherche que l'on conduit et dans le cadre de laquelle se déroule cette analyse statistique.

2.2. CONDUITE CONCRÈTE DES OPÉRATIONS D'EXPLORATION

On suppose que les données sont contenues dans un " vecteur " x, de type `numeric` ou `factor`. Dans tous les cas, on doit commencer par la fonction élémentaire `summary(x)`. C'est une instruction souple et puissante, qui identifie le type de l'objet, et donne des informations de base, en fonction du type.

2.2.1 variables catégorielles

Dans ce cas, la fonction `summary(x)` renvoie les effectifs des modalités du caractère, classées par ordre alphanumérique. On obtient une vue graphique en tapant seulement `plot(x)` : graphique en barres, une barre pour chaque modalité. Les effectifs sont en ordonnées. On peut aussi taper `pie(table(x))`. On obtient un graphique " en camembert ", bien moins lisible que le

précédent, mais qui donne une vue complémentaire. Peut être utile pour une présentation, dans la mesure où lecteurs ou spectateurs sont habitués à ce genre de graphique (intrinsèquement très difficile à lire, souvent même trompeur, *à éviter strictement dans le cadre d'un travail sérieux*).

On ne doit pas se priver d'une observation de l'ordre des modalités en fonction de leur effectif. On peut ordonner les modalités en ordre ascendant ou descendant : `barplot(sort(table(x)))` ou `barplot(rev(sort(table(x))))`. Les instructions sont un peu plus complexes, dans la mesure où l'on doit utiliser **la fonction de tri**, `sort()`, qui, dans le cas de variables catégorielles, exige que les individus soient d'abord regroupés par modalité, fonction `table()`. Cette instruction `barplot()` utilise par défaut un dégradé de couleurs qui facilite la lecture et permet d'impressionner sans peine le spectateur. [On peut aussi essayer `pie(sort(table(x))`, mais cela n'apporte rien de plus].

On peut paramétrer tous ces graphiques en choisissant d'autres valeurs que les valeurs par défaut. Il peut être utile d'indiquer les valeurs extrêmes des ordonnées, si l'on veut pouvoir comparer plusieurs distributions `(ylim=c(a, b))` ; on peut aussi écrire les noms des modalités verticalement pour éviter les recouvrements `(las=3)`. Voir les tableaux du manuel d'E. Paradis.

Si les modalités sont nombreuses (en gros plus d'une dizaine), les effectifs constituent une variable numérique, qui doit être analysée comme telle. Sinon, on doit **réfléchir sur les rapports entre les effectifs des diverses modalités**, et se tenir en alerte maximale
1. si une modalité écrase toutes les autres (plus de la moitié),
2. si, dans l'autre sens, une ou plusieurs modalités ne renvoient qu'à des effectifs marginaux (moins de 1 ou 2%).

2.2.2 variables numériques : valeurs de position et courbe de densité

On doit aussi commencer par `summary(x)`, que l'on doit compléter par `length(x)`, ce qui permet d'obtenir l'effectif de la variable observée. La fonction `summary()`, dans le cas d'une variable numérique, renvoie 6 nombres : le minimum, le premier quartile, la médiane, la moyenne, le troisième quartile et le maximum.

Si l'on excepte la moyenne, ce sont toutes **des valeurs de position**, qui ne résultent d'aucun calcul, mais seulement d'un tri. La **médiane** est la valeur qui correspond à la position " du milieu " (autant d'observations inférieures que d'observations supérieures) et le **premier et le troisième quartile** sont l'équivalent de la médiane, appliqué à chaque moitié. Il faut se concentrer d'abord sur les extrêmes, qui donnent une idée de l'ordre de grandeur de la variable et de la **largeur de l'intervalle concerné**.

Avant de passer à des calculs, on doit absolument procéder à un **examen graphique détaillé**. C'est là un point qui résulte de l'existence de logiciels graphiques puissants, donc relativement récents ; pour cette raison, cette règle ne figure généralement pas dans les manuels courants.

En faisant `plot(x)`, on obtient un graphique simple de points : pour chaque valeur, x est le rang (dans l'ordre d'apparition, pas le rang issu d'un tri) et y la valeur elle-même. En observant l'axe des y, on obtient une première image de la répartion des valeurs.

La notion de densité, relativement complexe d'un point de vue mathématique strict, renvoie cependant à une intuition commune : si l'on place les points *sur une échelle régulière*, quelle est la zone (ou les zones) où les points ont tendance à se concentrer ? Autrement dit, quelles sont les valeurs les plus courantes dans la distribution, et les plus rares ? L'objectif (essentiel pour toute l'analyse) consiste à repérer **la forme de cette distribution**. Deux fonctions de base permettent

d'avoir deux vues instantanées : `hist(x)` et `plot(density(x))`.

La première dessine un *histogramme* avec des *classes* de même largeur. Le logiciel trie les valeurs par ordre croissant, coupe l'intervalle qui sépare les extrêmes en un nombre " raisonnable " de segments, compte le nombre de points dans chaque segment et affiche le résultat sous forme de colonnes verticales contiguës (jadis appelé " graphique en tuyaux d'orgue "). Les calculs sont simples, le résultat un peu grossier, voire trompeur : ce ne peut être qu'une première approche.

La seconde fonction `(plot(density(x)))` est nettement plus efficace ; elle fait appel à des calculs longs, raison pour laquelle son développement est récent (son emploi n'a guère commencé avant le milieu des années 80, les manuels courants ne la mentionnent quasiment pas). Le principe est simple : le logiciel trie les valeurs puis se place successivement dans un grand nombre de points équidistants tout au long de l'échelle (couramment 512) ; à chacun de ces points, il " estime " la densité, c'est-à-dire décompte le nombre de valeurs de la variable situés à proximité de part et d'autre, en les affectant d'une pondération en fonction de leur éloignement. Le résultat dépend en partie de la formule de pondération (R en propose plusieurs), mais surtout de la largeur de la " fenêtre " considérée : si la fenêtre est large, l'estimation varie sans à-coup, si au contraire la fenêtre se rétrécit, le résultat est une ligne de plus en plus zig-zagante. Cette méthode se nomme, en anglais, *kernel density estimator* (abrégé en général KDE), en français on emploie **estimateur de densité** *ou estimateur de densité du noyau*. Le paramètre crucial est la **largeur de fenêtre**, en anglais **bandwidth** (dans R, paramètre `bw=`). R calcule un `bw` par défaut, qui est affiché sous le graphique. Le résultat est le plus souvent tout à fait lisible.

La bonne méthode consiste à afficher une série de graphiques en faisant varier le `bw` dans un sens et dans l'autre. Avec R, une telle procédure est extrêmement simple et rapide. Après un nombre d'essais suffisant, on trouve en général une valeur (parfois deux ou trois) qui permet d'obtenir un graphique d'où les aspérités locales ont disparu, mais qui donne une idée claire de la forme de la distribution ; si l'on n'est pas à court de papier et d'encre, on peut imprimer deux ou trois graphiques. L'utilisation de l'estimateur de densité introduit en fait **la notion clé de lissage** (smoothing).

La supériorité flagrante de cette méthode provient précisément de ce dernier point : on obtient un graphique qui comporte les éléments que l'on veut voir parce qu'ils ont un sens (**la forme** de la courbe), mais pas d'**éléments adventices** plus ou moins artificiels et inutiles qui brouillent la perception (se reporter aux principes énoncés dès les années 60 par Jacques Bertin, chapitre 8), comme les tuyaux d'orgue, qui produisent un " effet d'escalier " tout à fait artificiel.

La très grande souplesse de R permet de réaliser la plupart des combinaisons que l'on peut imaginer. Une petite fonction simple, baptisée `hisden()`, permet d'afficher sur le même graphique les 5 paramètres de position, un **his**togramme avec classes d'effectifs égaux et une courbe de **den**sité ; on peut choisir le nombre de classes et surtout faire varier la largeur de bande, ce qui permet d'obtenir très rapidement un lissage convenable, donnant une idée acceptable de la forme de la distribution.

2.2.3 formes de la distribution

En utilisant directement les fonctions de R ou en construisant une petite fonction ad hoc comme `hisden()`, on peut obtenir une vue graphique assez claire de la forme de la distribution. Les cas de figure principaux ne sont pas très nombreux. **Deux questions** seulement se posent : **combien de sommets ? où sont-ils situés ?** Les sommets de la courbe de densité sont appelés en statistique "valeurs modales" ; un sommet = distribution unimodale, deux sommets = distribution bimodale. Il est rare de trouver trois sommets clairement distincts. On peut rencontrer, surtout avec des effectifs modestes, des distributions d'allure cahotique ; dans ce cas, il est difficile de conclure, il faut reprendre les données directement.

Les **distributions bimodales**, surtout lorsque les deux modes sont séparés par un creux

profond, correspondent le plus souvent au mélange de deux populations. Il faut alors essayer de comprendre selon quel caractère les deux populations sont effectivement distinctes, trier, et reprendre les analyses avec deux fichiers séparés. En pratique, il arrive souvent que la "bimodalité" ne soit *pas excessivement nette*, et apparaisse ou disparaisse selon la largeur de bande choisie. Il faut regarder de près, revenir aux données et, par simple prudence, adopter plutôt l'hypothèse de la bimodalité, et l'explorer attentivement.

Le cas courant est celui de la **distribution unimodale**. Trois dispositions se présentent :
* **le mode est à peu près au milieu**,
* **le mode est assez nettement dans la partie gauche** (ou dans la partie droite) de la distribution,
* **le mode est complètement repoussé à une extrémité**.

Le mode à l'extrémité droite est excessivement rare, le mode décalé à droite est souvent (pas toujours) transformable en mode à gauche, par complémentarité (ex. un taux modal de 90% d'alphabétisation peut aussi se lire comme un taux de 10% d'analphabétisme, etc).

En pratique, la question décisive est de déterminer s'il existe ou non une valeur centrale. Dans les deux premiers cas évoqués ci-dessus, la densité maximale se situe à l'écart des bords : l'exploration va devoir cerner le mieux possible cette valeur centrale, et la signification de sa position par rapport à la distribution globale. Au contraire, dans le troisième cas, la densité maximale est "collée" contre un bord (cas standart des distributions lexicales : les hapax sont la classe la plus fréquente dans tous les textes ; mais c'est aussi le cas de la distribution des lieux habités : les hameaux sont plus nombreux que les villages, qui sont plus nombreux que les bourgs, qui sont plus nombreux que les villes, etc.). Dans ce cas, *la notion de "valeur centrale" est dénuée de toute signification*, il faut employer d'autres méthodes (point sur lequel la grande majorité des manuels sont désespérément muets).

A. DISTRIBUTIONS AVEC VALEUR CENTRALE

Dans le cas d'une distribution centrée à peu près symétrique, les valeurs centrales classiques, moyenne et médiane, sont très voisines, et la moyenne n'est pas un artefact, elle est effectivement représentative du centre de la distribution et apporte donc une information fiable sur la distribution dans son ensemble. Faire très attention au fait que quelques valeurs tout à fait isolées à droite ou à gauche peuvent suffire pour créer l'impression de dissymétrie : le graphique se déporte par construction (et *modifier aléatoirement la moyenne*, hypersensible aux valeurs extrêmes). Surtout examiner la relation du sommet (le mode) avec les trois valeurs de position les plus importantes, les deux quartiles et la médiane. **Si le mode est à peu près au milieu de l'écart inter-quartile**, on peut considérer sérieusement qu'il s'agit d'une distribution à peu près symétrique.

La **dissymétrie** est extrêmement fréquente dans les données historiques : le mode n'est ni au milieu ni sur un bord, mais dans l'entre-deux. Ici, il faut redoubler de prudence, et essayer d'évaluer au plus juste les distances entre mode, médiane et moyenne. Si la médiane s'écarte "sensiblement" de la moyenne, alors il faut faire très attention : *il y a peu de chance que la moyenne arithmétique brute soit une valeur centrale fiable,* ***ni d'ailleurs le mode***.

Dans le cas de dissymétrie, si le mode est dans "l'entre-deux", **il faut impérativement tenter de comprendre pourquoi**. La procédure la plus simple consiste à effectuer diverses **transformations des données** et à observer les résultats.

Les transformations

Les deux transformations les plus simples et courantes consistent à **prendre le logarithme** et à **extraire la racine carrée** : deux manières différentes de modifier l'échelle sur laquelle sont portées les données, ce qui modifie eo ipso l'allure de la distribution et la courbe de densité (il s'agit

d'une variante d'anamorphose). Pour cela, il suffit de remplacer, dans toutes les instructions, `(x)` par `log(x)` ou par `sqrt(x)`.

On doit au minimum essayer `summary(log(x))` et `plot(density(log(x))`. Ne pas le faire peut conduire à des erreurs rédhibitoires. Dans de nombreux cas, on s'aperçoit 1. que l'écart entre moyenne et médiane se réduit considérablement, 2. que le mode, précédemment dans "l'entre-deux", vient se placer au milieu, et que la nouvelle distribution peut être dite symétrique. Si tel est le cas, on a identifié un caractère essentiel de la distribution. Sinon, essayer `sqrt(x)`. Il arrive assez fréquemment que la transformation logarithmique, appliquée à une distribution avec mode à gauche, donne une distribution avec mode à droite. Dans ce cas, il y a lieu d'essayer ce que l'on appelle la **transformation de Box-Cox**, qui est une transformation que l'on pourrait dire "partiellement logarithmique" (la fonction, très simple, fait intervenir un coefficient compris entre 0 et 1, que l'on détermine par essais successifs).

Ici, il importe de comprendre **à quoi correspondent concrètement ces transformations**. Contrairement à ce qu'annoncent beacoup de manuels, il ne s'agit pas d'artifices de calcul ou de dessin. Lorsque les valeurs "brutes" donnent directement une distribution centrée symétrique, cela signifie que **les écarts par rapport à la valeur centrale** (mode, médiane et moyenne à peu près confondus) sont **de type additif** : la distribution observée s'explique par des oscillations (positives et négatives) autour de cette valeur centrale ; quelques écarts importants, un certain nombre d'écarts "moyens", un grand nombre d'écarts faibles. Si la distribution est centrée après passage aux logarithmes, cela signifie que les écarts sont **de type multiplicatif** ; au lieu que les écarts consistent en additions et soustractions, ils correspondent ici à des multiplications et divisions, dans quelques cas par un coefficient important, dans davantage de cas par un coefficient "moyen" et le plus souvent par un coefficient "faible". Dans les données historiques, c'est une situation des plus communes, l'expérience montre surabondamment que **la majorité des échelles "sociales" sont de type multiplicatif**. Pour le dire de manière approximative, on peut dire que le passage d'une classe à la suivante ne s'opère pas par addition d'une certaine quantité, mais par multiplication par un coefficient. Cependant, il arrive que ce coefficient ne soit pas constant, tendant en général à diminuer au fur et à mesure que l'on se déplace vers des valeurs plus élevées. Si le coefficient est constant, les logarithmes conviennent, si le coefficient diminue, il faut utiliser la transformation de Box-Cox.

L'extraction d'une racine peut s'expliquer soit directement, soit par analogie, avec l'oscillation centrée "brute" du côté d'un carré (racine carrée) ou d'un cube (racine cubique). Si l'on a des surfaces assimilables à des cercles, et que le rayon oscille additivement, la distribution des surfaces sera nettement dissymétrique, la racine carrée permettra de retrouver une distribution centrée symétrique.

Méthodes simples pour déterminer la valeur centrale

Les procédures graphiques sont manifestement, à l'heure actuelle, les outils les plus commodes.

***** Dans les manuels français, même anciens, on parlait de la **droite de Henri** ; les ouvrages anglais (et R) utilisent l'expression de **quantile-quantile plot** (QQ-plot). Très schématiquement, une des échelles (celle des x dans R) est graduée en écart-types, tandis que sur l'autre échelle on représente les données observées. Il s'agit d'un autre type d'anamorphose qui a cette propriété remarquable de **représenter par une droite toute distribution de forme gaussienne classique**. Avec R, l'opération est particulièrement simple, puisqu'il suffit de faire `qqnorm(x)`. Si la distribution est centrée symétrique et à peu près gaussienne, les points représentant la distribution x observée doivent se trouver alignés. Si au contraire la courbe est nettement incurvée, au-dessus ou au-dessous d'une ligne droite, alors la distribution s'écarte plus ou moins nettement d'une

distribution gaussienne.

La méthode consiste à essayer successivement diverses transformations, en faisant varier les coefficients, jusqu'à obtenir un ensemble de points à peu près alignés. L'essentiel est de parvenir à un alignement des 2/3 médians de la courbe, il arrive fréquemment que les queues s'écartent : si elles tendent vers l'horizontale, la distribution est moins étalée qu'une distribution gaussienne standart, si elles tendent vers la verticale, les valeurs extrêmes sont plus nombreuses que ne le prévoit la formule classique. Dans le cas où les queues s'écartent l'une et l'autre du même côté de la droite, cela tend à montrer une certaine dissymétrie de la courbe, c'est plus ennuyeux.

***** Une autre méthode consiste à surimposer à un graphique de densités ordinaire le dessin de la courbe normale "théorique" correspondant aux paramètres choisis. Une petite fonction ad hoc (baptisée, elle, `ajnorm()`) permet d'obtenir directement ce résultat. On peut modifier ad libitum la moyenne et l'écart-type, ce qui permet un **ajustement** assez fin, et un repérage relativement simple des écarts entre distribution observée et distribution théorique.

L'évaluation de la dispersion

Pour simplifier, on peut dire que la recherche d'une valeur centrale significative revient à déterminer le type de transformation, et éventuellement les paramètres, qui permettent d'obtenir une courbe de densité centrée et symétrique. Dans ce cas, il apparaît la plupart du temps (pas toujours) que la médiane et la moyenne sont très voisines. Ce qui conduit à énoncer fortement une vieille règle : **la médiane est presque toujours une valeur centrale de premier intérêt**, dès lors que la distribution est centrée (conditio sine qua non, trop souvent négligée).

Cela établi, il reste à examiner la dispersion, qu'elle soit de nature additive ou multiplicative. Dans tous les cas, il est facile et efficace de déterminer l'**écart inter-quartile** (intervalle compris entre le premier et le troisième quartile) qui contient par construction la moitié centrale des observations (en anglais, inter-quartile range, `IQR()`). La loi de Gauss et, à sa suite, toute la statistique classique, font jouer un rôle essentiel à la **variance** (moyenne des carrés des écarts à la moyenne), et à sa racine carrée, l'**écart-type** (en anglais, standart deviation, `sd()`). [La variance possède divers avantages, et en particulier d'être extrêmement facile à calculer (elle est égale à la moyenne des valeurs-carrées moins le carré de la moyenne ; c'est ce qu'on appelle le "théorème de König" ; il suffit donc, pour la calculer, de connaître la somme des x et la somme des x^2). La variance est, par construction, très (trop !!) sensible aux valeurs extrêmes.]

Il faut principalement retenir que, dans le cas d'une distribution gaussienne, 95% des observations sont situées dans une fourchette ±1.96 écarts-types de part et d'autre de la moyenne. Cette fourchette ne laisse en principe que 2.5% des observations à gauche et à droite. Cette relation offre un moyen simple et rapide de tester l'allongement des deux queues de courbe.

Les ajustements graphiques évoqués plus haut et ce petit test permettent de se faire une idée non seulement de l'amplitude de la dispersion, mais aussi de sa forme. Si l'on n'observe nulle part de différence marquée entre courbe observée et courbe théorique, on a une distribution gaussienne, rien à signaler. Il est assez rare que l'on note un déficit symétrique à gauche et à droite. Dans ce cas, on peut soupçonner une distribution tronquée. Plus fréquemment, un excédent important à gauche et à droite. Dans ce cas, on sort du "modèle gaussien", et l'on peut faire l'hypothèse d'une "loi stable de Lévy" : il s'agit de *variations aléatoires plus complexes que de simples oscillations pendulaires.* Un excédent dissymétrique suscite l'hypothèse d'une **anomalie**. Il faut revenir aux données, c'est à peu près le seul moyen de comprendre quelque chose.

D'une manière tout à fait générale, on ne saurait assez répéter que *les ajustements, qui permettent de préciser la forme de la distribution et ses paramètres de base, doivent tout autant servir à repérer* ***tous les écarts*** *entre "courbe observée" et "courbe théorique"*. Ces écarts, invisibles à l'œil nu, constituent souvent une information importante, dont l'analyse peut mettre sur la voie de phénomènes significatifs.

B. DISTRIBUTIONS SANS VALEUR CENTRALE

Lorsque la densité est maximale à l'une des extrémités de la distribution, la notion de valeur centrale devient non seulement encombrante, mais même dangereuse. Appliquer à une distribution de ce type les procédures classiques aboutit à des erreurs grossières. Des erreurs de ce genre sont malheureusement monnaie courante, ce qui s'explique en large partie par le silence de la plupart des manuels. Des statisticiens comme Benoît Mandelbrot ou Marc Barbut ont rappelé opportunément que le premier à avoir examiné de près ces distributions fut l'italien Vilfredo Pareto. D'où le nom d'"**univers parétien**" qu'ils emploient, pour bien le distinguer de l'"univers gaussien". A la suite de B. Mandelbrot, on utilise également le terme de "**fractales**", ou "objets fractals".

Ces distributions ont le plus souvent une forme hyperbolique. La littérature sur ce domaine demeure relativement confidentielle, les méthodes d'analyse et d'ajustement sont encore peu nombreuses et peu pratiquées. On peut cependant s'attendre à des progrès sensibles au cours des prochaines années.

Pour simplifier, on peut dire que, s'gissant de distributions hyperboliques, l'hypothèse la plus simple est celle d'un processus hiérarchique sous-jacent. Il existe une forte analogie entre une distribution hyperbolique et un arbre hiérarchique. Tous les nœuds sont autant de points où se produit un processus aléatoire élémentaire. La combinaison de nombreux processus élémentaires **engendre des aléas complexes**, doués de propriétés sensiblement différentes de celles des processus élémentaires. B. Mandelbrot a suggéré de parler dans ce cas de "hasard sauvage", par opposition au "hasard bénin" dont s'occupent les statistiques classiques. Dans la nature comme dans la société le hasard sauvage est de loin le plus fréquent. Soulignons immédiatement que les populations concernées ont des **propriétés déconcertantes**, notamment d'être (le plus souvent) indénombrables au plan mathématique, ce qui rend inutilisables les procédures classiques fondées sur la notion basique de population fermée (toute la Wahrscheinlichkeitslehre classique et tous les axiomes de la théorie des probabilités).

On considérera ici deux procédures graphiques assez aisées à mettre en œuvre : les représentations "rang-taille" et le graphe des moyennes et médianes conditionnelles.

rang-taille

Tous les graphiques destinés à examiner la densité utilisent une échelle des x (abscisses) régulière, en ce sens que les valeurs possibles de x sont représentées sur une échelle “naturelle” : les intervalles sur l’échelle sont proportionnels aux variations numériques (qu’il s’agisse des valeurs absolues, des logarithmes ou d'une transformation de Box-Cox).

Une autre manière de visualiser les données de manière ordonnée consiste à les ranger par ordre de taille, à les placer côte-à-côte, en indiquant pour chacune la valeur sur l’axe des y (ordonnées). Dans ces conditions, les valeurs portées sur l’axe des x sont tout simplement le rang de chaque individu. On obtient ansi un graphique dit “rang-taille”. On fait généralement en sorte de commencer par la plus grande valeur (= rang 1), et ainsi de suite. Par construction, le sommet de la courbe est contre l’axe des y, et la courbe s’abaisse continûment (éventuellement par palliers, s’il y a des individus ayant même valeur).

Cette fonction n’est pas implémentée directement dans R, mais s’écrit facilement. Dans

certains cas, on obtient une courbe hyperbolique très creusée. Dans ce cas là notamment, il est utile d'utiliser *deux échelles logarithmiques*. Il arrive que cette transformation permette d'obtenir une courbe à peu près linéaire. Dans ce cas, on a identifié une "**loi rang-taille**", qui constitue l'un des cas les plus simples des distributions parétiennes. Ce cas se rencontre dans les analyses de fréquences de vocabulaire. Si l'on admet que la distribution observée est bien linéaire sur un graphique de ce type, le seul paramètre qui caractérise la distribution est alors sa **pente**, qui constitue un indice simple de la dispersion de la distribution. Le cas élémentaire est la pente valant -1, ce qui indique que le produit rang x taille est constant.

moyenne et médiane conditionnelles

Dans une série de textes remarquables, Marc Barbut a bien mis en évidence une propriété très intéressante des distributions parétiennes : **le rapport entre une valeur et la moyenne (ou la médiane) de l'ensemble des valeurs plus élevées (ce qu'on appelle moyenne ou médiane conditionnelles) est constant**.

Cette propriété comporte deux avantages :

1. elle permet un test graphique d'une grande simplicité,
2. elle permet de déterminer avec une bonne précision les paramètres de la loi de Pareto observée.

Il n'existe pas de fonction de R permettant d'obtenir directement ce résultat, il faut donc écrire une petite fonction ad hoc.

Marc Barbut a montré qu'il n'y a (en général) pas lieu de tenir compte des 10 ou 20 premiers rangs (en haut à droite du graphique). On examine donc les autres points : s'ils sont à peu près alignés, il s'agit d'une distribution parétienne. Dans certains cas, les points sont alignés sur deux droites distinctes : on peut alors faire l'hypothèse de deux populations (parétiennes l'une et l'autre) distinctes. En repérant le point d'inflexion, on peut découper le vecteur x et reprendre deux analyses, toujours selon le même principe, **pour voir**. La plupart des distributions de type gaussien fournissent au contraire des lignes nettement incurvées.

Cette question des ajustements aux distributions parétiennes, ou de type parétien, sera reprise plus en détail dans le cadre de l'étude des statistiques lexicales (chapitres 9-10-11).

ÉLÉMENTS DE CONCLUSION

Une familiarité minimale avec les principes et méthodes de l'exploration des ensembles univariés constitue un préalable crucial. Il s'agit des **fondements**. Or on rencontre à ce niveau une série de pièges variés et redoutables. Déterminer la forme d'une distribution et les anomalies qu'elle peut comporter apporte des éléments d'information qu'une lecture naïve des chiffres ne permet jamais d'acquérir. Une série de valeurs numériques peut résulter de processus complètement différents, des méthodes d'investigation existent, les ignorer est une faute.

Les manuels courants offrent une présentation satisfaisante de la plupart des "lois statistiques" de l'« univers gaussien » Il faut savoir dès le départ que les distributions non-gaussiennes sont très fréquentes dans l'étude des sociétés, et l'histoire en particulier.

Chapitre 3

DISTRIBUTIONS BIVARIÉES

La notion de distribution bivariée paraît intuitivement simple : tandis que l'on parle de distribution univariée lorsque chaque individu est muni d'un caractère, on parlera de distribution bivariée lorsque les individus ont **deux caractères**, la spécificité de l'analyse visant dans ce cas à éclairer la relation existant entre les deux séries (*que l'on suppose avoir été au préalable soumises, l'une et l'autre, aux procédures d'examen des distributions univariées*).

Dans la pratique de la recherche, les choses sont moins nettes : selon le type de caractère et davantage encore en fonction de la nature intrinsèque des données et de leur signification, on est amené à transformer les données originelles de diverses manières et à leur appliquer une gamme de traitements variés qu'il n'est pas facile, ni d'ailleurs très utile, de classer dans des catégories prédéterminées parfaitement tranchées. Ce qui se comprend aisément dès lors que l'on réfléchit un peu à la notion de **relation**. Ce terme recouvre en fait des situations, et donc des objectifs de recherche, extrêmement divers, raison pour laquelle on ne saurait trop répéter que *l'analyse statistique empirique requiert d'abord une connaissance approfondie des données* et ensuite seulement des compétences techniques, d'ailleurs plutôt fondées sur une expérience pratique que sur des connaissances mathématiques dont le minimum indispensable se réduit à assez peu de choses.

Il importe de clarifier d'abord la situation en présentant de manière organisée les principaux cas de figure qui peuvent se présenter, en fonction de la nature des séries et du type de relation entre elles. L'exposé lui-même sera organisé en fonction des méthodes d'exploration, en commençant par les méthodes graphiques les plus simples, en introduisant par étapes certains calculs, pour aboutir à la notion générale et fondamentale de **distance**, qui sera la pierre angulaire de l'étude des distributions multivariées (prochain chapitre).

SOMMAIRE

3.1. LES PRINCIPAUX CAS DE FIGURE

3.1.1 données appariées et non appariées

Deux distributions peuvent être liées de diverses manières sans être strictement appariées, et il y a donc lieu de mettre en œuvre dans ce cas des procédures appropriées. Dans le cadre de la démarche classique, une situation ordinaire consiste à comparer une distribution « observée » avec la distribution « théorique » censée lui correspondre (à la suite de ce que l'on appelle un *ajustement*). La question qui se pose est alors de savoir si cet ajustement est acceptable ou non, ou mieux, de mesurer l'écart entre la distribution observée et la distribution théorique. Un cas analogue consiste à comparer la distribution d'une population globale et celle d'un de ses sous-ensembles (question liée la théorie de l'échantillonnage, qui remplit des bibliothèques entières). De façon naturelle, on est amené à comparer plusieurs sous-ensembles : on passe ainsi imperceptiblement aux distributions multivariées ; mais certaines procédures de comparaison graphique élémentaires s'appliquent de la même manière à deux, trois ou n distributions, il ne s'agit pas alors de multivarié stricto sensu. Notons d'ailleurs que *ce problème peut se présenter d'abord sous forme bivariée stricte*, lorsqu'un caractère est de type catégoriel et l'autre de type numérique. Ce genre de comparaisons, pour lequel une procédure graphique simple est particulièrement efficace, peut donner lieu également à des calculs plus ou moins complexes sur lesquels nous ne nous étendrons pas, que l'on appelle globalement des **tests**. Ceux-ci portent en général sur tel ou tel paramètre de la distribution, spécialement les paramètres de valeur centrale et de dispersion (deux distributions ont-elles même moyenne ? même écart-type ? etc). Ces comparaisons de distributions relevant de populations distinctes ne sont possibles en général que si le caractère comparé est le même. C'est un point dont les logiciels ne s'inquiètent pas, ils font les graphiques ou les calculs sur les deux séries de chiffres qu'on leur indique, c'est au chercheur de savoir si la comparaison a un sens, et lequel.

Les statistiques bivariées stricto sensu portent sur des données appariées : les deux distributions considérées comportent alors le même nombre d'éléments, et chaque élément de l'une correspond à un élément de l'autre (« bijection »). Dans ce cas, il peut encore s'agir du même caractère, par exemple à des dates différentes, mais le plus souvent, on met en relation des caractères différents, des exemples courants étant surface / prix, ou âge / poids. D'autres procédures graphiques sont alors recommandées, susceptibles de montrer la forme de la relation, voire son intensité. Si une relation apparaît, il est possible de rechercher une forme « théorique » et de procéder également à des *ajustements*, portant cette fois non plus sur les distributions, mais sur la relation qui peut les lier.

3.1.2 nature des données en relation

On peut se borner à rappeler ici brièvement ce que l'on a dit dans le chapitre précédent: il existe quatre formes principales, *catégoriel*, *ordonné*, *discret*, *continu*. Dans la pratique courante, du point de vue des méthodes à utiliser, **on distingue essentiellement catégoriel et numérique**. Ce qui aboutit à trois cas possibles de croisement : catégoriel / catégoriel, catégoriel / numérique, numérique / numérique. Il convient ici de faire trois remarques : 1. s'agissant de distributions numériques, on traitera uniquement de distributions ayant une valeur centrale ; *les distributions sans valeur centrale (de type parétien) nécessitent une réflexion et des procédures particulières* ; si l'on n'y prend pas garde, le logiciel fera sans rechigner les graphiques et les calculs demandés, mais les résultats seront presque toujours sources d'erreurs d'interprétation grossières. 2. le cadre des analyses bivariées est le cadre privilégié de l'utilisation des valeurs ordonnées (**rangs**) : on peut toujours transformer une distribution numérique en distribution des rangs ; cette procédure, qui n'a guère d'intérêt dans le cadre de l'étude des distributions univariées à valeur centrale, constitue au

contraire une transformation assez efficace (impliquant un minimum de présupposés) dans le cadre bivarié ; les statistiques classiques s'y sont beaucoup intéressées, et l'on en trouve les résultats dans le chapitre habituellement intitulé « tests non-paramétriques » (expression vaguement ésotérique simplement destinée à indiquer que l'on ne raisonne plus sur des caractères à valeur numérique, mais seulement sur les rangs). 3. d'une façon plus générale, on peut toujours passer du numérique aux rangs, et des rangs au catégoriel, ou directement **du numérique au catégoriel** : il suffit de *découper des classes* dans une distribution numérique ou ordonnée. Si ce découpage est bien fait, la perte d'information est faible, et en contrepartie s'ouvrent des possibilités de traitement d'une tout autre ampleur.

3.2. MÉTHODES GRAPHIQUES ÉLÉMENTAIRES DE COMPARAISON DE DISTRIBUTIONS NUMÉRIQUES

3.2.1 juxtaposition ou superposition de graphes de densité

Comme on l'a vu dans le chapitre précédent, *le graphe de densité, obtenu à l'aide de la méthode dite du kernel* (KDE), est de loin le moyen le plus efficace pour se faire une idée à la fois globale et précise de la forme d'une distribution. Il est donc logique de construire deux graphes de ce type pour comparer deux distributions. Mais cela n'est pas automatique et requiert des précautions. On a vu en effet que la construction de ce graphe nécessite le choix d'une « largeur de bande » (bandwidth) appropriée ; le logiciel détermine une valeur plausible, mais la pratique montre que l'on obtient le plus souvent des résultat plus nets en modifiant quelque peu, dans un sens ou dans un autre, cette valeur par défaut. Dans le cas d'une comparaison, *il faut appliquer le même paramètre aux deux distributions*, sans quoi l'on comparerait des graphiques non comparables... La superposition est probablement le moyen visuellement le plus efficace de comparer de manière fine deux distributions, en utilisant par exemple deux couleurs bien repérables (type bleu et rouge). Il faut procéder à divers essais, régler les paramètres, les limites des x et des y. La juxtaposition verticale est en principe plus efficace, dans la mesure où elle permet de repérer plus commodément les analogies et les écarts partie par partie (queue à gauche, mode, queue à droite, dispersion).

Cette procédure a toutefois deux limites. D'un côté, elle ne permet pas de comparer plus de trois ou quatre distributions ; d'un autre, il faut souligner qu'elle est très intéressante pour comparer des formes, mais ne permet pas de repérer convenablement des valeurs de position ou de dispersion précises (sinon le mode, mais ni la médiane ni la moyenne). Cette procédure demeure cependant irremplaçable dans le cas de distributions plus ou moins irrégulières, notamment les distributions bimodales, dont les caractères spécifiques disparaissent avec d'autres méthodes.

On doit compléter cette analyse graphique en comparant les valeurs numériques des principales positions (`summary(x), summary(y)`).

3.2.2 le boxplot

Le **Box-and-Whiskers-Plot** (en français « boîte à moustaches ») est une forme de graphique d'une remarquable efficacité, imaginée dans les années 70 par le statisticien John W. Tukey. De manière courante, on dit simplement « faire un boxplot ». La plupart des logiciels statistiques actuels ayant un minimum de capacités graphiques font des boxplots. La fonction R `boxplot()` comporte un grand nombre de paramètres qui permettent d'obtenir des résultats parfaitement contrôlés, et si besoin est très élégants.

Le graphique est fondé principalement sur *une visualisation des principales valeurs de position*. La « boîte » centrale (un rectangle) est limitée par le premier et le troisième quartile ;

autrement dit, **cette boîte enferme la moitié centrale de la distribution**, laissant à l'extérieur un quart à gauche et un quart à droite. Un segment est tracé à l'intérieur du rectangle pour situer la médiane (que l'on peut rendre plus visible par des « encoches » (notch)). De chaque côté de la boîte sont tracées des « moustaches », qui peuvent s'étirer, le cas échéant, jusqu'à une distance représentant une fois et demi l'écart interquartile ; à cet endroit, le trait horizontal est barré par un petit segment vertical (1,5 est la valeur usuelle, utilisée par défaut ; une option permet de choisir toute autre valeur, si on l'estime pertinent). Si des individus se situent à l'extérieur de ces limites, ils sont représentés par des points (ou de petits cercles), et correspondent plus ou moins à ce que l'on appelle couramment des « outliers » (« valeurs extrêmes » ou « valeurs aberrantes »). Si les distributions correspondent à des effectifs différents, une option disponible avec R permet de dessiner des rectangles de largeur plus ou moins proportionnelle à l'effectif concerné. La comparaison peut porter sur deux à plusieurs dizaines de distributions.

Il est facile de voir que ce graphique fait jouer un rôle décisif à l'écart interquartile. Dès lors, la représentation est d'autant plus efficace que les distributions sont mieux centrées et symétriques. Si tel n'est pas le cas s'agissant des données brutes, il faut essayer les diverses transformations possibles évoquées dans le chapitre précédent (logarithme, transformation de Box-Cox, racine carrée, notamment).

3.3. DISTRIBUTIONS NUMÉRIQUES STRICTEMENT APPARIÉES

3.3.1 le nuage de points

Lorsqu'à chaque individu correspondent deux valeurs, on considère l'une comme x et l'autre comme y, et il suffit de faire `plot(x,y)`. Dans la grande majorité des cas, on aperçoit instantanément la forme du nuage. Il faut cependant se méfier par principe de cette première impression visuelle, qui peut facilement être biaisée. En effet, l'œil identifie plus ou moins clairement une forme en fonction de la densité relative des points dans les diverses parties de la surface rectangulaire du graphique. Il suffit que, dans telle ou telle partie de ce rectangle, des points se trouvent superposés, ou quasi-superposés, pour qu'ils passent en fait inaperçus ; or l'expérience montre que cette situation est très fréquente, et presque inévitable. Il faut donc trouver un moyen de tenir compte de manière équivalente de tous les points, quelle que soit leur position relative. Ce moyen existe, c'est l'évaluateur de densité, en abrégé **kde2D** (kernel density estimator in 2D). On retrouve donc en 2 dimensions l'équivalent de ce que l'on a utilisé sur un axe.

Le processus est un peu plus complexe. Dans un premier temps, le programme définit une **grille** (grid) de points plus ou moins fine (la plupart des programmes ont une valeur par défaut, qui est souvent 50 ou 100, tant pour les x que pour les y, mais on peut choisir une autre valeur ; en pratique, les résultats sont rarement différents) ; dans un second temps, le programme calcule, pour chaque point de la grille, le nombre de points (observés) situés « à proximité » dudit point, en effectuant en général une pondération en fonction de la distance. Bien entendu, on retrouve ici le même problème que sur un axe : il faut choisir une « **largeur de bande** » (bandwidth), qui permettra un lissage plus ou moins marqué. Chaque point de la grille se trouve donc muni d'une valeur de densité, et finalement, dans un troisième et dernier temps, le programme dessine ce que l'on peut appeler des **courbes d'isodensité**, c'est-à-dire l'équivalent de courbes de niveau sur une carte. Bien que chaque programme soit muni de valeurs par défaut, il est en fait indispensable de préciser soi même ce que l'on veut. Le plus efficace consiste à ne tracer qu'une ligne, au plus deux, en essayant de trouver (par habitude ou par essais successifs) les valeurs du ou des paramètres permettant d'obtenir la courbe qui définit le mieux la forme du nuage. On y arrive en pratique très rapidement. Et l'expérience montre surabondamment qu'**une telle ligne est à la fois beaucoup plus**

lisible et beaucoup plus fiable que le nuage brut.

Cette méthode demande une quantité de calculs énorme, et c'est pourquoi son développement est tout à fait récent et que son emploi est encore loin d'être généralisé. C'est la méthode d'aide à la visualisation des nuages qui implique le moins de présupposés : elle ne demande aucune hypothèse préalable sur la forme du nuage, et fait jouer à tous les points un rôle identique. Elle est particulièrement efficace lorsque le nuage a une forme irrégulière et/ou se trouve fragmenté en deux ou plusieurs parties distinctes. On peut poser comme règle qu'**il faut l'essayer sur tous les nuages de points que l'on tente d'analyser**.

3.3.2 analyse de la forme du nuage, transformations

Comme on l'a expliqué précédemment, les transformations ne se ramènent pas à des « artifices graphiques », mais correspondent au fait que les écarts par rapport à une valeur centrale sont très fréquemment d'une nature autre qu'additive. Cette question clé dans le cadre de l'examen des distributions univariées devient un passage obligé lorsqu'il s'agit de cerner la forme de la relation entre deux distributions.

Une méthode ancienne, simple pour ne pas dire simpliste, consiste à **ne retenir que les rangs des individus dans les deux distributions**. Avec R, c'est immédiat, il suffit de remplacer `x` par `rank(x)` et `y` par `rank(y)`. D'un point de vue technique, *cette transformation revient en fait à passer d'une distribution quelconque à une distribution uniforme*. C'est un peu brutal, mais cela ne coûte rien d'essayer, et en général c'est instructif. Les points se retrouvent distribués de manière parfaitement uniforme tant sur l'axe des x que sur celui des y, le centre du graphique est occupé par la médiane des x et la médiane des y. Si les x et les y sont liés par une relation simple, elle apparaît le plus souvent. Inversement, si l'on n'aperçoit aucun regroupement, il y a bien des chances que les deux distributions soient indépendantes.

La bonne méthode consiste à essayer de trouver, pour les x et pour les y, la transformation la plus pertinente. En général, il vaut mieux procéder sur chaque distribution indépendamment de l'autre, notamment avec la droite de Henri (fonction `qqnorm()`). Mais cela n'interdit nullement de faire à nouveau divers essais sur le nuage de points. L'objectif étant à peu près le même pour une distribution bivariée que pour une distribution univariée : parvenir, par une procédure ayant un sens, à faire en sorte que la médiane soit à peu près au centre du graphique, et que, par conséquent, les points soient répartis de manière « équilibrée » sur les deux axes.

Ce résultat étant supposé atteint, les diverses situations observables peuvent en gros se ramener à trois cas principaux :

a) le nuage prend *une forme ovoïde à peu près régulière, parallèlement à l'axe des x ou à celui des y* ; en modifiant plus moins l'échelle des x ou celle des y, on obtient sans peine un nuage de forme grossièrement circulaire : il y a indépendance des deux distributions, le caractère a et le caractère b ne sont liés par aucune liaison fonctionnelle.

b) le nuage prend *une forme ovoïde ou plus ou moins fuselée oblique (voire filiforme, mais le cas est rare)* ; ce qui signifie qu'à une augmentation de la valeur de x correspond tendanciellement, de manière plus ou moins nette, une augmentation - ou une diminution - de y ; les deux caractères (le cas échéant après transformation) sont liés par une fonction linéaire (du type bien connu $y = ax + b$).

c) le nuage prend une autre forme assez nettement définissable, qui peut être par exemple une forme en C ou en U, ou bien se fragmente manifestement en deux ou plusieurs nuages distincts ; la liaison non-linéaire est un phénomène que l'on rencontre assez souvent, et qui doit s'expliquer au cas par cas ; un exemple classique est la liaison âge / revenu, le revenu augmente d'abord avec les années, mais, passé un certain seuil, diminue. Le fractionnement du nuage peut avoir des causes variées ; il s'agit assez souvent d'un indice intéressant d'hétérogénéité, ce qui signifie que, au moins sous certains aspects, la population considérée se subdivise en sous-ensemble qui se comportent de

manière bien distincte par rapport à telle ou telle variable. Le cas n'est pas rare où l'on se trouve devant une population en évolution, certains éléments ayant déjà subi cette évolution, d'autres non.

3.3.3 la régression

Par ce terme technique traditionnel un peu bizarre, on désigne toute procédure qui, à toute valeur d'une des deux distributions, par exemple les x, fait correspondre une seule valeur de y, censée représenter de manière approchée, mais convenable, les diverses valeurs observées, plus ou moins dispersées, de y. Concrètement, *on remplace le nuage de points par une ligne*. Les méthodes de calcul sont très nombreuses, mais se classent en deux types principaux : a) les régressions empiriques, dites aussi locales, qui sont fondées sur le principe de la fenêtre mobile ; b) les régressions linéaires, qui supposent une relation linéaire entre les deux variables et cherchent à calculer les coefficients a et b de la fonction y = ax + b, de manière à obtenir le « meilleur ajustement » possible. Les premières aboutissent à une ligne plus ou moins sinueuse, qui n'a guère d'autre intérêt que de permettre de calculer des interpolations plus ou moins plausibles ; il est rare que l'historien puisse tirer parti de ce genre de manœuvre. L'ajustement linéaire (que l'on peut toujours calculer, malheureusement...) n'a de sens que si l'observation directe du nuage, et en particulier la ou les lignes d'isodensité, font nettement apparaître la linéarité de la relation entre les deux variables. Dans ce cas, mais dans ce cas seulement, et si l'on a des raisons sérieuses de considérer qu'il peut s'agir d'une relation fonctionnelle réelle, on peut calculer un ajustement, dont *l'intérêt principal réside bien moins dans la ligne elle-même que dans les écarts par rapport à cette droite « théorique »*. Les x étant donnés ainsi que la fonction, on peut calculer immédiatement les y « théoriques » et donc, par simple soustraction, les écarts « observé - théorique », que l'on peut classer par importance décroissante, au dessus et en dessous de la ligne. On dispose ainsi d'un moyen simple de repérer méthodiquement les individus qui s'écartent le plus de la tendance générale. Ce repérage effectué, on peut faire l'hypothèse qu'un ou plusieurs facteurs sont la cause de ces déviations : là, il faut « se débrouiller »...

3.4. LE CROISEMENT DE VARIABLES CATÉGORIELLES : LE GRAPHE DE BERTIN

3.4.1 tableau de contingence simple

Lorsque les deux variables sont des variables catégorielles, la première chose à faire est un **tri croisé**. Dans R, la fonction `table(x,y)` fournit directement le tableau qui en résulte, qui comporte autant de lignes que la variable x comporte de modalités, et autant de colonnes que la variable y comporte de modalités. Chaque case indique l'effectif qui correspond à la fois à une modalité de x et à une modalité de y. Les totaux en lignes et en colonnes (**marges** ou **distributions marginales**) fournissent les effectifs globaux des modalités de chacun des deux caractères.

Ce genre de tableau, qui ne comprend que *des entiers positifs ou nuls* est généralement appelé **tableau de contingence**. Il s'agit là des données brutes, simplement présentées sous une forme condensée. Dès que le tableau dépasse deux colonnes et trois lignes, il est impossible d'en rendre compte par simple lecture. D'ailleurs même un tableau de 2 x 2 comporte bien plus de pièges que ce que l'on pourrait croire naïvement.

3.4.2 les notions clés d' « **indépendance** » et d' « **écarts à l'indépendance** »

La principale difficulté de lecture de ces tableaux provient du fait que les effectifs des modalités (de chaque caractère) sont rarement égaux. Dit autrement : les pourcentages des effectifs

globaux des modalités sont différents ; telle modalité représente 20% de l'effectif total, telle autre 12%, telle autre 32%, etc. Partant de cette considération, l'idée de base consiste à supposer que, *si les deux caractères étaient indépendants, les pourcentages dans toutes les colonnes seraient tous les mêmes que ceux de la marge de droite, et que les pourcentages dans les lignes seraient tous les mêmes que ceux de la marge du bas*. Et qu'en définitive chaque case serait ainsi la combinaison (multiplication en l'occurrence) du pourcentage global de la ligne et du pourcentage global de la colonne. Le total de toutes les cases du tableau étant donc, par construction, de 100%.

Connaissant ainsi le pourcentage de chaque case par rapport à l'effectif total, et cet effectif total, on peut calculer instantanément la valeur de l'effectif « théorique » de chaque case dans l'hypothèse où toutes les colonnes seraient proportionnelles entre elles, de même que les lignes, situation que l'on convient de désigner du terme d'**indépendance**. Chaque case se trouvant ainsi munie d'une valeur « théorique » dite d'« indépendance », *on calcule immédiatement par soustraction l'«écart à l'indépendance »*. A la différence de l'effectif brut, l'indépendance est le plus souvent une valeur avec décimales, et l'écart est, quant à lui, un réel algébrique (positif ou négatif). Ce tableau dérivé, dit des écarts à l'indépendance, est sensiblement plus facile à interpréter, dans la mesure où il fait apparaître des déficits et des excédents. On comprend facilement que la signification (approximative) de cet écart dépend de la valeur de l'indépendance : le même écart en valeur absolue risque fort de n'avoir pas la même signification s'il concerne deux cases dans lesquelles l'indépendance est très différente ; on est donc tenté de considérer plutôt le rapport que la valeur absolue. Diverses solutions ont été proposées, et peuvent être mises en œuvre ; pour des raisons qui tiennent à la théorie des probabilités classique, *on utilise le plus souvent le carré de l'écart divisé par la valeur théorique* (ce que l'on appelle le « khi-deux », nous y reviendrons).

3.4.3 représentation graphique des écarts à l'indépendance

Etant de la sorte parvenu à constituer un tableau, pondéré, des écarts à l'indépendance, on doit tenter de trouver un procédé capable de donner une vue globale compréhensible de ces écarts. Diverses possibilités existent ; la pratique tend à montrer que le graphique le plus lisible est un graphique dont l'idée a été exposée (sans doute) pour la première fois en 1977 par Jacques Bertin (ce qui n'empêche pas les anglo-saxons, qui ne lisent guère le français, de désigner couramment ledit graphique du nom de « Cohen-Friendly », alors même que la première publication de Cohen remonte seulement à 1980). En France, c'est le sociologue Philippe Cibois qui en a fait un usage systématique et a écrit un programme permettant de produire aisément le graphique. Dans R, il s'agit de la fonction graphique `assocplot(x, y)`.

Les calculs, comme on vient de le montrer, traitent de manière absolument équivalente lignes et colonnes : la permutation des x et des y ne change en rien les résultats. Le graphique, lui, privilégie les lignes ; à chaque ligne du tableau, on fait correspondre un segment horizontal, figurant l'indépendance, et l'on dessine verticalement des rectangles proportionnels aux écarts, au dessus de la ligne pour les excédents, en dessous pour les défcits ; le résultat est encore plus lisible si l'on emploie deux trames ou mieux encore deux couleurs différentes, classiquement du rouge et du bleu. La position et l'importance des excédents (ou des déficits) apparaissent clairement. Bien entendu, cette construction n'étant pas parfaitement symétrique, on a tout intérêt à construire les deux graphes correspondants. L'expérience montre que les deux lectures ne diffèrent pas, et qu'en général une seule suffit. En fait, avec des moyens différents, on parvient à un résultat assez analogue à celui du nuage de points pour les valeurs numériques : et la lecture de la forme des nuages est très rarement modifiée par une permutation des x et des y.

3.4.4 alignement sur une diagonale des écarts de même signe

Même dans le cas où les modalités sont naturellement ordonnées (échelle de tailles, de

préférences, succession chronologique), il est presque toujours utile et instructif de *regrouper au mieux sur l'une des deux diagonales* (le choix n'a guère d'importance) *les écarts d'un même signe, en général les écarts positifs*, en permutant les lignes entre elles et les colonnes entre elles. Dans les années 60 et même 70, on procédait « à la main », ce qui était long et fastidieux. Il existe divers algorithmes permettant d'obtenir par calcul un résultat satisfaisant, l'un des plus simples étant celui dit des « *moyennes réciproques* », qui procède par itérations (on calcule la position du « centre de gravité » des colonnes, que l'on reclasse en fonction de ces positions ; on procède de même pour les lignes ; on revient aux colonnes, et ainsi de suite ; l'on s'arrête dès que le tri ne produit plus de permutation, ce qui est le plus souvent presque immédiat).

Le graphique ainsi réordonné permet de voir quelles sont les modalités de x qui « **attirent** » telle ou telle modalité de y, et réciproquement (le traitement est complètement symétrique). La pratique montre qu'il s'agit là d'une procédure à la fois **extrêmement simple** et **redoutablement efficace**.

3.4.5 généralité et limites du graphe des écarts (graphe de Bertin)

L'extrême généralité de cette procédure tient tout simplement au fait que la quasi-totalité des caractères observables peuvent être subdivisés en modalités ; pour les valeurs numériques ou les rangs, il suffit de découper la distribution d'origine ; tout caractère plus ou moins définissable, du simple fait que l'on puisse en parler, peut être considéré sous l'angle de ses variations, quelle qu'en soit la nature. *La notion de modalité frôle l'universalité*. Dès lors que ces modalités s'appliquent à une série d'individus, un tri croisé est possible, et le graphe s'ensuit... Si l'on peut faire en sorte que les modalités aient des effectifs du même ordre de grandeur, on tend, d'une certaine manière, vers ce que l'on pourrait appeler une distribution uniforme non ordonnée. Cette situation permet d'observer les liaisons entre deux caractères, **indépendamment de la forme de la liaison**. C'est ce qui fait la souplesse, et finalement la **supériorité** de cette méthode sur la plupart des autres.

Mais *on ne peut en aucun cas faire l'économie d'une réflexion approfondie sur la nature des modalités et la signification du découpage qui les définit*. Une procédure qui peut être un peu lourde, mais souvent éclairante, consiste à modifier le découpage et les regroupements, pour essayer de percevoir les conséquences de ces modifications sur la place et l'importance des écarts correspondants.

Au surplus (et ceci n'est pas toujours aisément perceptible), la définition de l'« indépendance » n'est pas exempte d'un certain arbitraire, puisque l'on admet que la répartition des modalités dans l'ensemble de la population (en termes techniques, les « distributions marginales ») constitue en quelque sorte le cadre automatique de référence. Lorsque l'on observe un tableau d'écarts, construit à partir de cette hypothèse, il arrive fréquemment que l'on s'aperçoive que cette méthode fait disparaître des variations locales, dont la connaissance directe des données permet au contraire de penser qu'elles ne sont pas sans signification. C'est parfois le cas quand une modalité a un effectif nettement supérieur à celui des autres (ce cas est facile à déceler), mais aussi lorsque deux modalités sont en très forte conjonction, si bien que les effectifs d'une ligne et d'une colonne sont presque tous regroupés dans une seule case. En pratique, lorsque l'on subodore un déséquilibre ou un biais, et si les effectifs ne sont pas trop faibles, une procédure efficace consiste à *découper de diverses manières la population globale, de manière à s'assurer de la stabilité ou au contraire de l'instabilité des écarts*. Ce qui renvoie à une observation très générale : **la notion d'homogénéité d'une population est toujours relative, voire très relative**, les relations entre caractères sont rarement identiques d'un bout à l'autre d'une population. L'analyse statistique vise précisément à déceler des inflexions, éventuellement des lignes de fracture, souvent invisibles en lecture directe.

3.5. LA NOTION GÉNÉRALE DE « DISTANCE »

3.5.1 comparer les comparaisons : position du problème

Jusqu'ici, on s'est inquiété de savoir dans quelle mesure le caractère A était « lié » au caractère B, et selon quelle relation (forme du nuage). On dispose en général de bien plus de deux caractères, et l'on procède identiquement pour la relation entre le caractère A et le caractère C, et ainsi de suite. Se pose inévitablement alors la question : la « liaison » entre A et B est-elle plus ou moins forte que la liaison entre A et C ? Il faut au moins pouvoir classer les liaisons par ordre (rang), ou mieux encore les « mesurer ». A cette question ancienne a été apportée une grande variété de solutions, nous ne pourrons donner ici que quelques indications générales. D'ailleurs, dans la pratique, on peut se contenter de connaître trois ou quatre coefficients usuels et, le cas échéant, construire soi-même un coefficient en fonction de tel ou tel problème spécifique.

3.5.2 les coefficients les plus courants

Les trois coefficients les plus fréquents sont la distance euclidienne, le coefficient de corrélation linéaire et la distance du khi-deux.

a) distance euclidienne. Il s'agit de la généralisation du théorème de Pythagore : dans un triangle rectangle, le carré de l'hypothénuse est égal à la somme des carrés des deux autres côtés. Sur un graphique cartésien, la distance entre deux points est égale à la racine carrée de la somme des carrés de la différence des abscisses et de la différence des ordonnées. Ce qui vaut pour deux points se généralise à n points : on prend la racine carrée de la somme des carrés des écarts entre toutes les paires. Le calcul est simplissime.

b) coefficient de corrélation linéaire. On a exposé brièvement, un peu plus haut, la notion de régression linéaire : étant donné les x, définir une droite qui « résume » le mieux les y. Pour diverses raisons, au premier chef la facilité des calculs, on détermine la droite qui minimise la somme des carrés des écarts entre points observées et valeurs théoriques ; on obtient ainsi ce que l'on appelle la « droite des moindres carrés ». On procède pour les y par rapport aux x, puis pour les x par rapport aux y ; sauf dans le cas où les points sont strictement alignés, les deux droites sont différentes. Si le nuage est à peu près circulaire, les deux droites sont à peu près perpendiculaires, plus le nuage est oblong, et plus l'angle qu'elles forment diminue. Un calcul relativement rapide permet d'obtenir un coefficient proportionnel à cet angle, qui varie entre 0 (droites perpendiculaires) et 1 (droites confondues) ; ce coefficient reçoit un signe selon la pente générale des droites : coefficient positif si une augmentation des x correspond à une augmentation des y, coefficient négatif si à une augmentation des x correspond une diminution des y. La limite et l'inconvénient de ce coefficient tiennent à sa nature *linéaire* : il mesure assez convenablement l'intensité d'une relation linéaire, mais risque fort d'induire en erreur s'il existe une relation *non-linéaire*.

c) le khi-deux (ou chi-carré, squared chi). S'applique à un tableau comportant au moins deux lignes et deux colonnes, et porte sur des effectifs. Selon la procédure exposée plus haut, on calcule pour chaque case une valeur « théorique », un écart, et le rapport du carré de l'écart à la valeur théorique (tous ces nombres sont par construction positifs), et l'on fait la somme de toutes ces valeurs ; on obtient ainsi un khi-deux total pour le tableau entier. Le khi-deux est élevé dès que quelques cases présentent des écarts importants, quelle que soit leur position relative. La liaison mesurée par le khi-deux est indépendante de la forme de la relation.

3.5.3 l'interprétation des coefficients : le point de vue probabiliste

Ces coefficients ne sont directement comparables que s'ils portent sur des effectifs identiques, des valeurs du même ordre de grandeur et, dans le troisième cas, des tableaux ayant le

même nombre de cases. Ce n'est pas une situation rare, mais ce n'est tout de même pas la plus fréquente. Lorsque les situations diffèrent, on ramène les coefficients calculés à une valeur « en probabilité », qui tient compte du nombre de paires (coefficient de corrélation linéaire) ou du nombre de cases (khi-deux). A partir de considérations mathématiques et/ou de simulations, on peut déterminer quelle est la probalité (entre 0 et 1) pour que, dans telle condition, le coefficient dépasse, ou non, une valeur fixée d'être « le produit du hasard » (on utilise en général le « seuil de 95% »). Le logiciel, si on lui pose la bonne question, répond le plus souvent : `p = .xxx`. Ce qui veut dire qu"avec une marge d'erreur de 5%, on peut considérer que le coefficient a 50% de chances de résulter du hasard (absence de liaison) ou au contraire 0.00003%, ce qui indique une liaison forte. C'est une petite « cuisine » que les logiciels actuels rendent très simple, et qui fournit des indications assez grossières, mais faciles à utiliser avec un minimum d'habitude.

Ce point de vue probabiliste est fondamentalement lié aux structures du hasard « gaussien », ce qui signifie qu'il est loin d'être universel. Il ne faut jamais perdre de vue cette limite drastique. Il est surtout important de se rappeler qu'entre deux caractères il est presque toujours possible de calculer plusieurs distances différentes, sans que l'on puisse proclamer (ce que font pourtant la plupart des manuels) que telle ou telle distance est « la meilleure ». Ici comme bien souvent, la méthode la plus rationnelle consiste à essayer plusieurs possibilités et à comparer les résultats.

Chapitre 4

DISTRIBUTIONS MULTIVARIÉES

On arrive ici au cas le plus complexe *en apparence*, qui est aussi le plus courant : le tableau de données sous n'importe quelle forme. Des colonnes et des lignes, remplies d'indications chiffrées, ou codées, ou pour le moins transformables en codes (modalités). Depuis les années 60, de manière globalement corrélée à l'augmentation des capacités de calcul, des méthodes sont nées et se sont développées, qui rendent l'analyse d'un tableau presque aussi aisée que celle d'une série unique. Aujourd'hui, des bibliothèques de programmes open source sont disponibles, qui devraient permettre de généraliser l'emploi de telles méthodes, particulièrement appropriées à l'exploration de données historiques. Les statisticiens français ont joué dans ce domaine un rôle majeur, ce pour quoi les anglo-saxons paraissent encore répugner à l'emploi de ces programmes. Citons la bibliothèque toulousaine « *multidim* » et surtout la bibliothèque lyonnaise « *ade4* » (qui est accompagnée d'une documentation surabondante et très précieuse) ; nous avons fortement utilisé cette dernière, tout en cherchant, par l'écriture de fonctions dérivées, à en rendre l'usage plus commode. Soulignons aussi ce que nous devons aux travaux du sociologue Philippe Cibois, qui fut l'un des tout premiers, dans les années 80, à diffuser les sources de ses programmes gratuits d'analyse factorielle, et à qui l'on doit une procédure originale et efficace, la méthode « *TRI-DEUX* », que nous avons implémentée sous R, et dont la présentation sera faite dans le présent chapitre.

SOMMAIRE

4.1. PRINCIPES DE BASE : REPRÉSENTATION D'UN TABLEAU QUELCONQUE

4.1.1 le nuage des points-colonnes

Les nombreuses méthodes que l'on range sous l'étiquette globale, un peu déconcertante au premier abord, d'**analyses factorielles**, reposent toutes sur l'*utilisation méthodique de la notion de distance*, introduite dans le chapitre précédent. On choisit l'une ou l'autre des diverses distances disponibles, et l'*on calcule la distance entre toutes les paires de colonnes*. On obtient un tableau carré symétrique (la distance A-B étant la même que la distance B-A). Le plus facile est de raisonner par analogie avec les tableaux (triangulaires) de distances entre villes que l'on trouve sur certaines cartes. Ces distances étant à peu près linéaires et mesurées sur un plan, il suffit de dessiner tous les triangles possibles pour reconstituer, à très peu près, la forme générale du réseau des villes concernées. C'est en gros ce que les analyses factorielles tentent de faire : ***représenter chaque colonne par un point, et disposer l'ensemble des points-colonnes en tenant compte le mieux possible de toutes les distances***. Deux colonnes presque identiques (distance faible) devront se trouver proches sur le graphique, et au contraire deux colonnes d'allure inverse (distance forte) se trouveront aux deux extrémités du graphique.

Le principe général est donc particulièrement élémentaire. L'exécution l'est nettement moins, parce que les distances calculées ne sont pas représentables toutes ensemble, convenablement, dans un espace à deux dimensions, ou même à trois. Les points sont en fait dispersés dans un espace à **n** dimensions, **n** étant égal au nombre de colonnes concernées. Conceptuellement, cela ne présente aucune difficulté (c'est de l'algèbre linéaire tout ce qu'il y a de plus ordinaire), mais ce que l'on souhaite doit tenir sur un plan, c'est-à-dire un espace à deux dimensions seulement. On peut raisonner à nouveau par analogie. Dans le chapitre précédent, l'on a évoqué la procédure dite de **régression** : résumer un nuage de points plus ou moins ovoïde (deux dimensions) par une simple droite (une dimension) ; par calcul, on a donc essayé de trouver *le moyen de réduire le nombre de dimensions des données en abandonnant un minimum d'information* et en « résumant » le mieux possible lesdites données. L'analyse factorielle généralise ce principe, en procédant par étape, et en décomposant ainsi la difficulté. Le programme commence par rechercher la droite qui représente le mieux tout le nuage (qui correspond à ce que l'on définit comme « la direction du plus grand étirement du nuage »), et calcule la position de tous les points sur cette droite. Cette répartition (la meilleure possible) est désignée comme « le premier axe » ; le programme « retire » alors en quelque sorte des données l'information traduite sur cet axe, et cherche la droite qui va le mieux résumer l'information restante : deuxième axe ; et ainsi de suite. Finalement, il y a autant d'axes que de colonnes, mais l'information est concentrée sur les premiers, en ordre décroissant. Chaque axe est en somme une correction par rapport à l'approximation formée par l'ensemble des axes précédents. Dans la plupart des cas, la structure globale du tableau est tout à fait convenablement résumée par la position des points sur les trois premiers axes. On produit donc en général trois graphiques, correspondant aux axes 1-2, 1-3 et 2-3. C'est l'équivalement approximatif d'une représentation dans un espace à trois dimensions. L'expérience montre surabondamment que c'est tout à fait suffisant et efficace.

En simplifiant grossièrement : une analyse factorielle, partant d'un nuage de points-colonnes dans un espace à n dimensions, projette ces points dans un espace à 2 ou 3 dimensions, de telle manière que cette projection assure la meilleure représentation graphique possible des distances calculées entre toutes les colonnes.

4.1.2 propriétés des axes, premiers principes de lecture et d'interprétation

Le lecteur qui ne sait pas ce qu'est une analyse factorielle est surtout déconcerté par l'*absence d'échelle sur les axes* ; comme on l'a indiqué précédemment, les « distances » sont en fait des coefficients de ressemblance, elles n'ont de signification que relative : la distance A-B est-elle plus grande ou plus petite que la distance A-C ? Chaque type d'analyse factorielle utilise un mode de calcul de cette distance, le même pour toutes les paires, si bien que les distances, dans une analyse donnée, sont comparables ; ces distances sont donc traduites numériquement, mais il n'y a aucune unité significative.

Dès lors, on conçoit aisément le mode de lecture : l'analyse d'un graphique suppose que l'on sache ce que représente chaque point-colonne, l'examen du graphique permet de repérer plus ou moins rapidement (cela dépend en grande partie du nombre de points affichés...) *les colonnes qui se ressemblent et celles au contraire qui diffèrent le plus fortement.*

On identifie assez rapidement, par exemple sur l'axe 1, les points les plus à droite et les points les plus à gauche (l'orientation est tout à fait aléatoire), et l'on peut dire, e.g. : « le premier axe oppose surtout les premières périodes aux dernières », ou bien « le premier axe oppose surtout les zones riches et les zones pauvres », etc. On procède de même pour la suite : « le second axe oppose surtout les jeunes et les vieux », ou « le second axe oppose surtout les actifs et les inactifs ». On interprète encore le troisième axe, bien plus rarement le quatrième. **La « signification » de chaque axe ressort ainsi principalement des « grandes oppositions » que l'on parvient à détecter**.

La seule difficulté réelle provient de la relation entre les axes. Il faut absolument savoir que toutes les analyses factorielles sont construites sur le même principe : les axes sont tous « orthogonaux » entre eux, ce qui veut dire simplement que *la corrélation linéaire entre les valeurs des points sur l'axe a et la valeur des points sur l'axe b (quels que soient a et b) vaut par construction zéro*. On ne trouvera pas deux axes sur lesquels les points soient répartis de la même manière. Mais *l'absence stricte de la moindre corrélation linéaire ne signifie nullement l'absence de corrélation en général* car, comme on l'a vu précédemment, il arrive assez couramment que deux variables soient liées par une relation non-linéaire. C'est pourquoi il est relativement fréquent que l'interprétation mette en jeu simultanément deux axes et que l'on soit ainsi amené à **élucider la signification d'un « plan factoriel », très souvent le plan 1-2, plus rarement 2-3 ou 1-3** (les chiffres renvoyant au rang des axes). Une forme globale du nuage en U ou en C a , en général, une signification importante qui dépasse nettement la somme des significations propres des deux axes considérés. Au-delà des oppositions plus ou moins claires qui définissent chaque axe, *la présence éventuelle d'un nuage (2D) ayant une forme définissable traduit presque à tout coup une structure que l'on doit essayer d'identifier*. On ne doit d'autre part jamais oublier le principe de construction : un axe est une correction de l'approximation donnée par les axes précédents, il faut donc procéder par ordre ; il est absurde de prétendre interpréter l'axe 3 ou 4 si l'on n'a pas une idée suffisante de ce que l'on trouve sur les axes 1 et 2.

4.1.3 les points « supplémentaires »

L'analyse factorielle consiste en un calcul qui affecte à chaque point une position sur une suite d'axes, en ordre d'importance décroissant. Supposons une ou plusieurs autres colonnes (correspondant au même nombre de lignes) : rien n'empêche de calculer leurs « distances » par rapport aux colonnes « actives » et, à partir de là, de *calculer leur position sur les axes déjà existants*. On appelle cette opération « projeter des points supplémentaires ». Les bibliothèques d'analyse factorielle effectuent toutes cette opération, à la fois simple et particulièrement efficace. La répartition des points sur les axes, donc en fait la forme et la signification des axes, sont déterminées exclusivement par les points « actifs ». *L'usage des points « supplémentaires » peut se faire dans deux perspectives apparemment inverses, au fond complémentaires* :

a) soit *on veut savoir où se situent un ou plusieurs éléments autres* ; si, par exemple les colonnes

actives se rapportent à une année déterminée, on peut disposer des mêmes colonnes à une autre année ; en les projetant comme éléments supplémentaires, et en examinant les paires corespondant à la même entité, on pourra analyser le ou les sens d'évolution. Un cas tout à fait classique en sociologie consiste à faire une analyse factorielle sur des variables socio-économiques (âge, sexe, profession, résidence, revenu) et à projeter en variables supplémentaires des indicateurs culturels, ou d'opinion. Les variables actives constituent alors en quelque sorte une « carte » sur laquelle on projette les indicateurs culturels ou d'opinion : on peut ainsi voir, plus ou moins nettement, comment se distribuent les indicateurs culturels en fonction des structures sociales de base.

b) mais il peut aussi se faire que l'*on songe surtout à éclaircir la signification des axes produits par les variables actives*. En examinant la position d'un certain nombre de points supplémentaires, on enrichit et précise l'interprétation de la position relative des variables actives. On peut reprendre l'exemple évoqué ci-dessus, mais dans l'autre sens : on fait une analyse factorielle avec des indicateurs culturels ou d'opinion, et l'on projette en variables supplémentaires les indicateurs sociaux de base, les points supplémentaires permettant alors d'examiner la répartition sociale des variables culturelles.

On voit sans peine que les deux procédures inverses qui viennent d'être évoquées sont complémentaires, et qu'il est fort recommandé de les exécuter l'une et l'autre afin de comparer les résultats : si les deux structures sont à peu près homologues, les deux analyses fourniront des résultats voisins (cas courant), sinon il faudra se demander en quoi consiste la différence. Il n'existe aucune règle pour définir a priori ce qui doit être actif et ce qui doit être supplémentaire. Tout dépend de la signification des variables ; une stratégie simple consiste à commencer par tout mettre en variables actives, puis à procéder à divers regroupements actifs / illustratifs, de manière à cerner ce qui est stable, ce qui varie, et comment. Il s'agit là d'un instrument d'une très remarquable efficacité. La seule chose que l'on puisse considérer comme une règle est qu'*il ne faut jamais se contenter d'un seul essai*, **on ne fait pas *une* analyse factorielle, il faut toujours en faire *une série***.

Accessoirement, notons dès à présent que les variables supplémentaires sont également souvent *employées pour des raisons dites « techniques »* : si une ou plusieurs colonnes présentent par exemple des effectifs disproportionnés (une colonne qui « écrase » toutes les autres, ou au contraire des colonnes à effectifs infimes), on place ces colonnes en éléments supplémentaires, ce qui permet une analyse équilibrée, mais donne cependant le moyen d'observer la position des colonnes en question. C'est un « truc » très commode.

4.1.4 les « contributions »

La plupart des ouvrages de statistique mathématique qui sont consacrés en totalité ou en partie aux analyses factorielles comportent un ou plusieurs développements relatifs à ce qui est habituellement dénommé « décomposition de l'inertie du nuage ». Par-delà les démonstrations et la terminologie, on doit retenir un point pratique très intéressant : des calculs adéquats permettent de disposer d'un moyen lisible d'évaluer la « qualité de la représentation » des points. Concrètement, on dispose de deux tableaux, l'un qui se lit en lignes et l'autre en colonnes ; les deux tableaux sont organisés de la même manière : une ligne pour chaque point, et une colonne pour chaque axe.

Un des deux tableaux se lit en lignes : pour chaque point, on voit (en pourcentages ou plus souvent en pour 1000) comment ses caractères propres sont répartis sur les axes successifs ; on peut voir en particulier si les valeurs les plus élevées se trouvent dans l'une ou l'autre des 2 ou 3 premières colonnes (= 2 ou 3 premiers axes) ou si on les rencontre plus ou moins loin sur la droite ; dans ce dernier cas, cela signifie que le point est mal représenté sur les 2 ou 3 premiers axes, il est donc prudent de ne pas trop se préoccuper de la position dudit point sur ces premiers axes, le point présente un « profil » singulier (on peut tenter de se rendre compte pourquoi, ce n'est pas toujours

facile).

L'autre tableau se lit en colonnes. Dans chaque colonne, c'est-à-dire pour chaque axe, les valeurs les plus élevées sont celles des points qui « contribuent » le plus à la formation de l'axe, c'est donc l'examen attentif de la signification de ces points qui doit permettre le mieux de se faire une idée de la signification globale de l'axe. Inversement, les valeurs faibles (qui, grosso modo, correspondent aussi à des points qui sont positionnés en général vers le milieu de l'axe), n'ont que peu de rapport avec cet axe.

Au total, c'est donc en combinant ces deux lectures que l'on peut le mieux vérifier comment se répartit et s'organise sur les différents axes l'information d'ensemble contenue dans le tableau de données de départ. C'est un complément quasi indispensable à la lecture des graphiques : comme on l'a indiqué pour commencer, *les axes et les projections sur les axes sont des approximations ; or on dispose avec ces tableaux d'***un indicateur précis sur ces approximations,** il est facile de comprendre qu'une vérification de la qualité de ces approximations est nécessaire avant toute interprétation.

4.2. AFFICHAGE SIMULTANÉ DES LIGNES ET DES COLONNES, LES PRINCIPAUX TYPES D'ANALYSES FACTORIELLES

4.2.1 le principe du biplot

Jusqu'ici, l'on a parlé de points-colonnes. On peut appliquer aux points-lignes à peu près tout ce qui a été dit des points-colonnes. Tableau de distances, décomposition, projections sur des axes successifs. La seule chose à retenir est que le nombre d'axes sera au plus égal à la plus petite des deux dimensions du tableau ; en général, il y a moins de colonnes que de lignes, mais le fait de changer le sens (transformer les colonnes en lignes, les lignes en colonnes, ce qu'en algèbre on appelle « transposer une matrice », résultat qu'avec R on obtient sur toute matrice par la simple instruction `t(X)`) ne modifie pas le nombre maximal d'axes.

On peut donc décomposer *les points-colonnes et les points-lignes* de la même manière. Et donc *les projeter ensemble sur le même graphique*, que la littérature anglo-saxonne appelle pour cette raison un **biplot**.

On peut (et on doit) se livrer sur les deux nuages à la même démarche d'analyse et d'interprétation esquissée ci-dessus. Mais comment mettre en rapport les deux ensembles ? Concrètement, cela dépend des distances utilisées, et tout spécialement du choix : utilise-t-on la même distance pour les lignes et les colonnes ou deux distances différentes ? Cela dépend du type d'analyse factorielle et c'est pourquoi il convient à présent d'examiner, à grands traits, les caractères originaux de ces types principaux.

4.2.2 l'analyse en composantes principales (ACP, alias *principal components analysis*)

La mise au point et la première présentation de cette méthode remontent apparemment à un article de l'économiste américain Harold Hotelling (1895-1973) de 1933. Bien que cet auteur ait proposé une méthode de calcul nettement plus simple que la technique de calcul matriciel héritée du 19e siècle, l'exécution demeurait excessivement lourde. Ce fut seulement avec l'apparition des ordinateurs dans les années 50 et surtout 60 que cette méthode se répandit. Elle reste prépondérante dans le domaine anglo-saxon.

En simplifiant à l'extrême, on peut dire que l'**analyse en composantes principales utilise le coefficient de corrélation linéaire entre les colonnes et la distance euclidienne usuelle entre les**

lignes. On peut donc mettre des nombres de n'importe quelle nature (notamment des nombres négatifs en cas de mesures), mais les colonnes et les lignes ne sont pas traitées de la même manière. L'ACP sur données brutes considère chaque paire (de colonnes et de lignes) indépendamment des autres, ce qui peut être un avantage. Mais *le coefficient de corrélation linéaire a deux limites* : il ne mesure que les relations linéaires, et pas les autres ; il est très (trop) sensible aux valeurs extrêmes, qui peuvent donc biaiser les résultats.

Pour atténuer ces défauts, on a imaginé un grand nombre de procédures, qui consistent le plus souvent en une *transformation des données brutes*. Le **centrage**, appliqué en général aux colonnes, s'apparente à un simple changement d'origine : on soustrait de toutes les valeurs de la colonne la moyenne de ladite colonne ; ainsi *la somme algébrique de la colonne devient nulle* : toutes les colonnes ont alors **même somme**. La **normalisation** est une opération un peu plus complexe : on divise chaque valeur par la racine carrée de la somme des carrés des valeurs. Résultat : *la somme des carrés des nouvelles valeurs vaut 1* dans tous les cas. On l'applique aussi aux colonnes. On obtient des colonnes ayant des **dispersions voisines**. Dans la bibliothèque ade4, la fonction d'analyse en composantes principales (`dudi.pca()`), qui comporte diverses options, effectue par défaut le centrage et la normalisation des colonnes. Il n'y a pas plus de règle ici qu'ailleurs : il faut essayer les diverses possibilités et interpréter les différences de résultat.

L'ACP a pour premier objectif de simplifier la représentation des individus en fonction d'axes calculés à partir de l'ensemble des colonnes (axes qui sont, pour cette raison, définis comme « composantes principales »). Logiquement, *l'interprétation devrait donc commencer par l'examen de la position des points-lignes*. La position des points-colonnes, d'une certaine manière, est déduite de la position des points-lignes, considérés globalement. *L'interprétation de la position des points-colonnes n'a donc de sens que par rapport à l'ensemble du nuage de points-lignes*.

Notons enfin que le type de distance entre colonnes entraîne fréquemment ce que l'on appelle un « **effet de taille** ». Si toutes les variables sont plus ou moins corrélées (cas qui se produit aisément si toutes les lignes représentent des individus de taille variable, mais aussi dans d'autres configurations), les variables vont toutes se trouver ensemble d'un côté du premier axe qui, du coup, n'a pas d'autre signification que de classer les individus en fonction de leur taille globale. Cela peut être utile, mais cela peut aussi ne présenter aucun intérêt. **Dans ce cas, l'information pertinente est sur le plan factoriel 2-3**. Il arrive enfin que les variables se regroupent en deux (éventuellement trois) « paquets » assez groupés : cela signifie alors que toutes les mesures se ramènent pour l'essentiel à deux (ou trois) grandeurs, pas davantage ; l'ACP peut alors servir à simplifier des ensembles de mesures plus ou moins redondantes (comme on l'a indiqué ci-dessus, c'était l'objectif de départ, mais il est souvent perdu de vue, car l'ACP peut en effet servir à bien plus que cela).

4.2.3 l'analyse des correspondances (AFC)

C'est un statisticien français, Jean-Paul Benzécri, qui a mis au point en 1964 une procédure de calcul (et des programmes) qui **traite de manière identique les lignes et les colonnes** ; avec l'analyse factorielle des correspondances (AFC), la transposition de la matrice d'origine n'entraîne aucune modification des résultats. La distance utilisée (tant pour les colonnes que pour les lignes) s'apparente au khi-deux évoqué dans le chapitre précédent. *Toutes les cases sont pondérées par les sommes des lignes et des colonnes (fréquences marginales), et ainsi les lignes ne sont traitées que comme des écarts par rapport au « profil moyen » identifié à la ligne-somme et les colonnes, comme des écarts par rapport au « profil moyen » identifié à la colonne-somme*.

Cette procédure présente des avantages considérables. Avec une AFC, on peut directement *interpréter la position relative des points-lignes et des points-colonnes* : un point-ligne et un point colonnes seront proches si, dans le tableau, la case correspondant à leur croisement présente un fort « écart à l'indépendance » positif. On observe empiriquement que le fait de travailler avec des écarts

plutôt qu'avec des valeurs brutes aboutit, d'une manière générale, à une répartition plus équilibrée des points sur le graphique et donc à une **meilleure lisibilité**.

La seule limite à l'usage de l'AFC est l'absolue nécessité de *n'avoir dans les cases du tableau aucune valeur négative* (la fonction `dudi.coa()` renvoie un message d'erreur et s'interrompt). On peut trouver un peu étrange a priori de traiter en AFC un tableau hétérogène (les colonnes comportant par exemple des effectifs, des mesures, des pourcentages, etc), puisque le programme attribue un rôle décisif à la somme en ligne. Les puristes peuvent être choqués, mais la pratique montre que cela marche en général parfaitement, alors pourquoi se priver d'une méthode qui marche ?

En réalité, la difficulté principale est tout autre. D'un côté, le rôle décisif des distributions marginales donne par construction une importance excessive à une colonne ou à une ligne qui « écrase » les autres par son effectif, puisqu'alors elle définit presque à elle seule l'effectif marginal. Le point en question se retrouve au centre du graphique et les autres points se définissent par rapport à lui. Ce n'est pas en général ce que l'on cherche : soit l'on pondère cette colonne (ligne), soit on la traite en élément supplémentaire. D'un autre côté, le fait d'analyser des écarts aboutit à donner une importance considérable aux profils présentant une forte différence par rapport au profil moyen, et ce quel que soit l'effectif du profil considéré ; de telle sorte qu'une colonne (ou une ligne) d'effectif très réduit, mais complètement discordante par rapport à ce profil moyen jouera automatiquement un rôle important dans la définition des axes. Ainsi, dans une population que l'on pourrait définir comme « mollement structurée », *il suffit d'un tout petit groupe présentant plusieurs caractéristiques atypiques pour que ce groupe définisse presque à tout coup le premier axe de l'AFC*. En face d'une situation de ce genre, c'est le chercheur seul qui peut savoir ce qu'il convient de faire, en fonction de la signification des lignes ou des colonnes « bizarres » : il peut s'agir d'un groupe dominant (en général à effectifs minimes, mais très structuré) ou au contraire d'une population marginale, ou de caractères annexes sans importance ; dans cette seconde hypothèse, la solution simple consiste à faire passer en éléments supplémentaires le groupe de lignes ou de colonnes concerné (qu'il n'est de toute manière pas inutile de repérer et d'analyser, comme toutes les anomalies).

4.2.4 autres formes d'analyses factorielles

Il est important de signaler - sans plus - qu'il existe, en dehors de l'ACP et de l'AFC (et de leurs multiples variantes), plusieurs autres procédures, qui peuvent s'avérer plus appropriées dans certaines circonstances.

Mentionnons seulement, à titre d'exemple, l'*analyse en coordonnées principales* (1966), qui permet de traiter n'importe quel tableau de distances. Il existe, selon les situations et les objets, des dizaines de distances, qui ont été proposées à tel ou tel moment, et d'ailleurs il peut être nécessaire d'en créer d'autres, en fonction de tel ou tel problème spécifique. Dans ce cas, l'analyse en coordonnées principales permet de décomposer le tableau de distances, et de représenter les points (dans ce cas ni vraiment lignes ni vraiment colonnes) sur une suite d'axes ; cela peut rendre d'intéressants services. La littérature concernant les analyses factorielles est abondante.

4.2.5 premières remarques générales sur l'emploi de l'ACP et de l'AFC

a) Il importe d'avoir une idée suffisante des calculs effectués par ces deux procédures de manière à pouvoir comprendre, au moins approximativement, les origines des différences de résultats, quand on en observe. Car **s'il est une règle qui paraît vraiment devoir être retenue, c'est bien celle de la confrontation systématique de ces deux types principaux**. C'est la manière la plus efficace de s'assurer de la stabilité générale du résultat, et des particularités éclairées par les écarts.

b) Lorsque l'on dispose d'un peu de temps, que l'on souhaite contrôler plus nettement les regroupements de points et de colonnes, la stratégie consiste à **transformer de diverses manières le tableau des données de départ**. Tout dépend de la nature de ces données. On peut remplacer les effectifs par des pourcentages ou des coefficients (par exemple une densité de population à la place d'un effectif), mais on peut aussi procéder aux diverses transformations suggérées par les analyses des distributions univariées (racine carrée, logarithme, transformation de Box-Cox, rangs). Il est classique de procéder au centrage et à la normalisation des colonnes avant une ACP, mais on peut faire l'un sans faire l'autre, ou laisser les données brutes. Philippe Cibois propose astucieusement de procéder à une ACP sur les écarts bruts à l'indépendance.

c) La manipulation la plus intéressante demeure malgré tout l'**essai de diverses répartitions entre variables actives et variables illustratives**. Les combinaisons théoriques sont en nombre énorme, il faut procéder par essais successifs, et comparer attentivement au fur et à mesure. C'est une opération qui peut être longue, mais c'est en pratique le seul moyen de vérifier l'homogénéité d'une population et / ou d'un ensemble de variables, et de repérer des sous-ensembles.

d) Il reste que toute interprétation ou toute explication ne peut reposer que sur les données et, au-delà, sur les structures dont elles sont l'écho plus ou moins déformé. *L'interprétation ne porte pas, contrairement à ce que semblent parfois croire certains statisticiens, sur les graphiques factoriels ou les tableaux de contributions, mais sur les données.* Si l'on repère des proximités, des oppositions, des disjonctions, et que l'on puisse en tirer des hypothèses sur les données, alors, il est impératif de revenir à ces données elles-mêmes, et d'en extraire des tableaux synthétiques, des listes, des effectifs ou des fréquences tangibles qui permettent de montrer explicitement et clairement *à quoi correspondent concrètement les configurations mises en lumière par les analyses factorielles.* Sans cela, les meilleures statistiques du monde demeurent en l'air, dénuées de la moindre portée. **Le « retour aux données » est un impératif catégorique**.

4.3. L'ANALYSE FACTORIELLE MULTIPLE

4.3.1 position du problème

Si une série d'individus donne lieu à plusieurs mesures, on obtient un tableau de mesures, que l'on peut soumettre à une analyse factorielle. Si une série d'individus est munie de deux caractères catégoriels (modalités), on peut, en croisant les deux groupes de modalités, obtenir un tableau croisé d'effectifs, que l'on appelle ordinairement **tableau de contingence**, et qui peut aussi être soumis à analyse factorielle. Notons toutefois que, dans ce cas contrairement au précédent, *les individus ne seront pas représentés*. Reste le cas le plus fréquent : une série d'individus munis d'un ensemble de caractères. que l'on souhaite croiser systématiquement. On s'est aperçu [semble-t-il par étapes, depuis les années 40, je n'ai pas réussi à identifier les auteurs ayant joué un rôle décisif, s'il y en eut...] qu'une solution facile à mettre en œuvre était de constituer un **tableau en codage logique** (dit aussi codage disjonctif)

4.3.2 le codage disjonctif

On garde une ligne par individu. *Pour chaque caractère, on constitue autant de colonnes qu'il existe de modalités différentes.* Si l'on considère le caractère A à 5 modalités, et l'individu x muni de la 2e modalité du caractère A, on place un 1 dans la deuxième colonne et des 0 dans les quatre autres colonnes ; ce qui donne

0 1 0 0 0

On procède de la même manière pour tous les caractères et tous les individus. On obtient au total *un*

tableau rempli de 1 et de 0 (surtout des 0), *ayant autant de lignes que d'individus et autant de colonnes que le total des modalités de tous les caractères*. Le total de chaque colonnes indique l'effectif des individus correspondant à la modalité définie par cette colonne. (**NB** le programme effectue lui-même toutes ces opérations à partir d'un tableau qui indique, pour chaque case, le code numérique correspondant).

Si cette règle est strictement respectée (les statisticiens paraissent y tenir beaucoup), on parle de codage disjonctif complet (ou codage disjonctif strict). Il vaut mieux en effet essayer de s'y tenir autant que possible, cela donne des résultats plus clairs. Mais les données historiques ne sont pas toujours aussi maléables qu'on le souhaiterait. Ce qui entraîne deux divergences par rapport à cette règle :

☞ soit le caractère est inconnu pour cet individu, c'est le cas « donnée manquante » ; le plus simple est alors de suivre l'exemple des sociologues confrontés aux non-réponses dans les données d'enquête ; leur habitude, simple et logique, est d'attribuer aux non-réponses le code numérique 0 (que l'on traite en élément supplémentaire) ;

☞ soit au contraire un individu présente deux, voire trois modalités en même temps (cas courants lorsque le caractère définit des attributs d'objets) ; dans ce cas, on met autant de 1 que nécessaire (ce que la plupart des logiciels ne prévoient pas, raison, parmi d'autres, qui a obligé à écrire des fonctions pour analyse factorielle multiple adaptées aux données historiques).

En pratique, on s'écarte encore plus de la règle théorique *en considérant le plus souvent comme des colonnes « supplémentaires » d'une part les colonnes correspondant au code 0* (non-réponse, donnée manquante), *d'autre part les colonnes à effectifs très réduits* (le seuil se déterminant par habitude et par essais successifs). Cette procédure vise simplement à écarter de l'analyse les colonnes ayant peu de signification et susceptibles de provoquer des perturbations. Bien entendu, on peut projeter sur le graphique ces colonnes supplémentaires et les réintroduire en colonnnes actives si cela paraît avoir du sens.

4.3.3 le codage : avantages, précautions à prendre

L'avantage considérable est que l'**on peut de cette manière croiser autant de caractères que l'on veut, et de n'importe quelle nature** ; car, comme on l'a vu précédemment, il est toujours possible de transformer en variable catégorielle une variable numérique ou ordonnée. Il s'agit donc d'une méthode de présentation des données en tableau numérique qui, sans être absolument universelle, présente une très grande généralité et permet donc de traiter simultanément des caractères extrêmement variés. Cette méthode, par sa souplesse, est particulièrement bien adaptée aux données historiques ; son emploi est, jusqu'ici, demeuré confidentiel parmi les historiens, le potentiel apparaît considérable !

Cette souplesse même est la source de la difficulté principale que rencontre l'utilisateur : on peut presque tout coder, mais comment ? Les manuels sont, sur ce point, d'un mutisme remarquable. Dans certains cas, classiques, on n'a pas de choix, par exemple « homme / femme », mais cette situation est, au total, peu fréquente. *Le principe est qu'il faut un codage assez fin pour transcrire l'information avec le moins de perte possible*, mais qu'*un nombre de modalités excessif produit en général des artefacts qui perturbent l'analyse* de façon plus ou moins gênante. Bien entendu, la marge de manœuvre est plus grande si l'effectif global de la population est plus élevé. Au dessus d'un millier d'individus, on peut distinguer sept ou huit modalités sans trop de risque ; inversement, si la population n'est que de quelques dizaines, on se contentera de trois ou quatre modalités. Dans une configuration comme dans l'autre, il faudra autant que possible tenter de parvenir à *des modalités ayant des effectifs du même ordre de grandeur* ; si les données ne s'y prêtent vraiment pas, les colonnes (modalités) à effectifs minimes devront être traitées en éléments supplémentaires.

Soulignons d'ailleurs que le principe des modalités multiples permet d'utiliser simultanément un code principal correspondant à des effectifs équilibrés et suffisants, et un *code annexe*, correspondant à des variantes à effectifs faibles, que l'on traitera en colonnes supplémentaires, ce qui permettra de les représenter sur le graphique sans que leur présence perturbe l'équilibre général de l'analyse.

Dans le cas des variables numériques continues, le découpage porte le nom technique de « **discrétisation** » et diverses procédures ont été imaginées pour « optimiser » ce découpage. Dans la majorité des cas, les critères de cette optimisation sont discutables. Le mieux est de commencer par *inspecter la distribution* à l'aide des procédures d'examen des distributions univariées. Dans certains cas (peu fréquents), des seuils ou des césures apparaissent nettement. Si la distribution peut se ramener plus ou moins facilement à une distribution centrée symétrique, le plus simple consiste à découper en 5 ou 7 parties égales (quantiles) ; si les queues sont très étalées, on peut diminuer les effectifs des modalités extrêmes, selon des proportions du type 5 - 10 - 10 - 10 - 5.

Les variables catégorielles demandent surtout une connaissance précise des données et de la signification des modalités.

Dans tous les cas, on a plutôt intérêt à partir d'un découpage relativement fin, quitte à opérer ensuite des regroupements. C'est là un problème technique auquel il vaut mieux réfléchir dès le moment où l'on entre les données dans un logiciel : **il faut prévoir dès le départ des possibilités pratiques de recodage aussi simples et rapides que possible**.

4.3.4 vers une réflexion générale sur la formalisation

Mais il s'agit d'un problème qui n'est pas seulement technique. On peut partir de l'observation des géographes qui ont bien montré que les mêmes chiffres et la même carte pouvaient donner lieu à plusieurs interprétations complètement différentes selon la discrétisation choisie. S'il est indispensable de ne pas perdre de vue le problème de l'équilibre numérique de la répartition des modalités, on doit malgré tout accorder la priorité à la sémantique. La nécessité de coder a justement cette grande vertu d'obliger à se poser la question de la signification des coupures que l'on utilise ; les coupures produisent une certaine information, des erreurs de codage peuvent créer des artefacts ou au contraire faire disparaître des indications importantes. Il est nécessaire de réfléchir constamment à la relation entre les objets que l'on tente de décrire et les indices qui amènent à placer des césures, lesquelles peuvent modifier sensiblement l'interprétation. Les médiévistes peuvent reconnaître là deux difficultés bien connues, et jamais résolues de manière univoque : combien de « campagnes de construction » dans un bâtiment, combien de « mains » dans un manuscrit ?

4.3.5 les diverses formes possibles de visualisation des résultats

L'affichage classique consiste à *faire figurer sur un graphique les points-colonnes et les points-lignes*. On a tout intérêt à **afficher différemment les colonnes et les lignes, et à distinguer nettement points actifs et points supplémentaires**. Si l'on affiche les noms, il est en général possible, avec les logiciels actuels, d'utiliser des caractères romains, italiques, gras, soulignés, et de varier les couleurs. Si l'on se contente de points, il faut de même essayer de choisir quatre symboles clairement reconnaissables.

Une difficulté qui est apparue dès les origines des analyses factorielles, et n'a pas trouvé à ce jour de solution entièrement satisfaisante, est celle des *points superposés ou quasi superposés*. Les programmes des années 70 et 80, dans ce cas, affichaient un seul point et fournissaient, sous le graphique, une liste des points superposés. Les sorties graphiques actuelles autorisent sans restriction les superpositions : le résultat est un gribouillis illisible ou bien l'affichage des points se trouvant par hasard en fin de liste, les autres disparaissant purement et simplement. La difficulté est

atténuée en agrandissant la taille de l'image ou en diminuant la taille des caractères, mais la question n'est pas pour autant résolue sur le fond. Je propose *un algorithme qui, partant du centre du graphique, décale progressivement vers l'extérieur les points dont les étiquettes se recouvrent plus ou moins*. Les étiquettes sont alors toutes lisibles ; si les points ne sont pas excessivement nombreux (en gros, moins d'une centaine) et si l'on utilise une taille de caractère modeste, les décalages restent limités et n'ont que des effets locaux, donc sans importance. Lorsque les étiquettes sont vraiment nombreuses, ou fortement groupées dans des zones restreintes du graphique, les déplacements entraînent nécessairement une déformation plus ou moins prononcée de l'équilibre général du nuage. Pour éviter des interprétations fallacieuses, on peut imprimer deux fois le graphique, avec recouvrements et sans recouvrements : le premier traduisant exactement la forme du nuage, le second permettant la lecture des étiquettes.

Pour les analyses factorielles multiples, c'est-à-dire celles qui traitent simultanément plusieurs caractères, il faut pouvoir contrôler l'affichage des points-modalités *en ne conservant que les caractères que l'on souhaite*. Lorsque les modalités sont ordonnées, il peut être très suggestif de *joindre les points-modalités selon leur ordre* « naturel » (ce que l'on appelle souvent un « **trajet** » ou une « **trajectoire** »). Notons au passage que les résultats de l'analyse factorielle (de n'importe quel type) seront identiques quel que soit l'ordre des modalités (numérotation du codage). Cet ordre n'a d'importance que par rapport au problème technique d'affichage des trajets en cas de modalités ordonnées.

Une possibilité particulièrement suggestive consiste à afficher **les nuages de points-individus correspondant aux diverses modalités d'un caractère**. Chaque modalité, au lieu d'être simplement traduite par un point-colonne, est représentée par l'ensemble des individus qui sont munis de cette modalité. On peut ainsi observer directement dans quelle mesure l'analyse factorielle sépare clairement les modalités ou les laisse au contraire largement enchevêtrées. Il arrive souvent que l'on note qu'une ou deux modalités se détachent nettement, tandis que les autres s'étalent et se recouvrent. C'est une indication précieuse, qui permet d'une part d'affiner les hypothèses sur la signification plus ou moins distinctive de telle ou telle modalité, et, éventuellement, de procéder à des remaniements du codage, notamment par regroupements.

Cette procédure d'affichage des nuages de points-individus définis par une modalité *peut aussi s'appliquer, si cela semble intéressant, à des modalités de divers caractères*. Si, en particulier, deux points-modalités (points-colonnes appartenant à deux caractères) sont proches sur le graphique général, on peut se demander si les deux nuages de points-individus correspondants ont, ou non, même forme et même étendue. Dans ce cas (comme d'ailleurs dans le précédent) **l'usage d'un estimateur de densité et le tracé d'un contour permettent une visualisation commode et fiable** (d'autres méthodes existent, comme l'ellipse, l'« oursin » (alias « étoile ») ou l'enveloppe convexe, mais ces méthodes sont peu recommandables, dans la mesure où elles attribuent a priori un rôle majeur au « point moyen » - les deux premières - ou au contraire aux points extrêmes, la troisième). L'estimateur de densité repose sur des hypothèses de départ bien plus faibles et il est, notamment, *irremplaçable* dans le cas de nuages segmentés.

Tout programme bien conçu offre la possibilité de sélectionner, avant tout affichage de nuage, une *sous-population*. On peut, par exemple, observer les nuages des modalités d'un caractère dans une partie seulement de la population d'origine, puis dans une autre partie, ce qui permet de constater, éventuellement, que les nuages sont bien distincts dans un cas, en grande partie enchevêtrés dans l'autre. C'est un des moyens que l'on peut utiliser lorsque l'on soupçonne une hétérogénéité gênante au sein d'une population.

4.4. LA MÉTHODE TRI-DEUX *DE Philippe CIBOIS*

Confronté aux problèmes spécifiques et épineux du dépouillement d'enquêtes, le sociologue Philippe Cibois a, depuis la fin des années 70, conçu et mis en œuvre un ensemble de procédures originales qui, s'appuyant sur l'analyse factorielle multiple (telle qu'elle vient d'être décrite à grands traits), ont pour objectif de *faciliter la lecture des résultats, en la rendant à la fois plus* ***analytique*** *et bien plus* ***interactive***. Ces procédures sont tout à fait adaptées aux données historiques, j'ai donc tenté de les implémenter sous R, en essayant de-ci, de-là, d'ajouter quelques options supplémentaires.

4.4.1 fichiers analytiques

La première orientation consiste à produire plusieurs fichiers (lisibles sous n'importe quel éditeur ou traitement de texte) fournissant des indications systématiques sur les caractères et les modalités.

Le **tri à plat** donne, pour chaque caractère, *l'effectif de chaque modalité et le pourcentage* du total qu'elle représente.

Les **profils** sont destinés à caractériser chaque modalité par la *liste des modalités qui lui sont les plus proches*. La procédure repose sur une idée simple : étant donné le tableau général des distances entre toutes les modalités, on considère successivement chaque modalité et ses distances à toutes les autres ; on obtient ainsi un vecteur que l'on trie par ordre croissant et l'on ne garde que les 15 ou 20 valeurs les plus faibles : on dispose ainsi de la liste triée des modalités « les plus proches ». Comme on l'a indiqué dans le chapitre précédent, il existe une multitude de manières de calculer la distance entre deux vecteurs de même longueur, et les méthodes sont encore plus nombreuses lorsque l'on considère des vecteurs seulement constitués de 0 et de 1, ce qui est le cas ici. Il me paraît raisonnable de *calculer deux distances assez différentes*, et de placer les résultats exactement l'un sous l'autre : on peut ainsi se rendre compte par simple lecture des analogies et des différences entre les deux listes. L'idée (un peu approximative) étant que les modalités qui apparaissent aux 10 ou 15 premières places dans les deux listes, résultant de calculs différents, peuvent être considérées comme effectivement assez voisines de la modalité considérée. Par construction, les distances varient de 0 (identité) à 1 (distance maximale) ; on peut, au moins en première lecture, prendre en compte toutes les distances inférieures à 0.8, voire 0.9. [j'ai choisi la *distance de Jaccard*, qui ne considère que deux colonnes et ne tient compte que des lignes comportant au moins un 1 ; et le *phi de Pearson* qui combine plus intimement toutes les lignes].

Les **tris croisés** ont déjà été examinés à propos des distributions bivariées. Ici, il s'agit de *fournir pour toutes les paires de caractères les deux tableaux croisés, celui des effectifs bruts et celui des écarts à l'indépendance*, suivis de la valeur du coefficient khi-deux et de la probabilité qui lui correspond. A la réflexion, il m'a semblé plus utile de donner ces tableaux après réorganisation par *report vers la diagonale des écarts positifs* (par la procédure des moyennes réciproques). La lecture est facilitée, en ce sens qu'il suffit d'une observation des écarts sur la diagonale du tableau des écarts pour se faire une idée concrète de l'importance (ou de l'insignifiance) de la relation entre les deux caractères considérés. Bien entendu, le graphique de Bertin-Cibois offre un complément très utile au tableau numérique, mais l'interprétation de ce graphique n'a pas de sens en dehors d'une considération attentive des effectifs concernés (*on ne saurait trop recommander la bipartition de l'écran, une moitié pour les tableaux d'effectifs, l'autre moitié pour le graphique*).

4.4.2 le graphe TRI-DEUX : principe

On prend le tableau général de toutes les distances entre paires de modalités ; on le classe

par ordre ascendant ; sur un plan factoriel (en général le plan 1-2), on affiche la paire correspondant à la plus petite distance, en indiquant le nom des deux points et en traçant un segment pour les relier ; puis, après un <retour charriot>, on procède de même pour la distance classée au deuxième rang, et ainsi de suite. De cette manière, le graphique s'enrichit peu à peu d'étiquettes et de segments les reliant. Si la population est suffisamment structurée par les caractères considérés, *on voit peu à peu se former des groupes de modalités en divers endroits du graphe*, deux, trois, quatre, parfois davantage. On peut ainsi avancer aussi longtemps que ces groupes demeurent à peu près distincts. Vient toujours un moment où des segments finissent par relier des groupes jusque là séparés. Le programme, s'il est interactif (c'est un de ses caractères principaux), prévoit la possibilité de repartir en sens inverse, c'est-à-dire d'effacer un par un les segments qui ont été dessinés. On peut ainsi avancer et reculer autant qu'on le souhaite, pour parvenir au graphique qui inclut le plus possible de modalités tout en laissant apparaître des ensembles distincts.

4.4.3 le graphe TRI-DEUX : stratégie

Notons d'abord que *tout ce que l'on vient de dire des modalités s'applique aussi, dans le domaine historique, aux individus*. S'il est vrai que, dans une enquête sociologique ordinaire, l'identité des individus, si on la connaît, n'a en général aucune signification, la situation est toute différente en histoire : les objets que l'on définit comme individus statistiques sont le plus souvent connus et l'un des intérêts de l'analyse est justement de compléter l'information ponctuelle dont on dispose par une mise en lumière des relations.

Si l'on obtient du premier coup trois ou quatre « paquets » de modalités (ou d'individus) bien groupées, c'est sans doute que la population est assez fortement structurée et qu'en plus on a choisi les bonnes options. Il reste de toute manière à vérifier concrètement (i.e. numériquement) la réalité de ces structures. Mais *il arrive fréquemment que la situation soit beaucoup moins nette.* Il faut alors envisager successivement, et combiner, une série de possibilités :

☞ *utiliser d'autres distances* ; la fonction ade4 utilisée (`dist.binary()`), offre le choix de dix distances différentes ;

☞ *étudier les trois combinaisons possibles* : actives seules, supplémentaires seules, actives et supplémentaires ensemble ;

☞ vérifier si *un autre plan factoriel* (1-3 ou 2-3) ne donne pas des résultats intéressants ;

☞ *partir d'une autre analyse factorielle*, soit en changeant le type (AFC, ACP, éventuellement NSC), soit surtout en modifiant la répartition entre caractères et modalités actifs et supplémentaires ; on peut par exemple passer en supplémentaires des modalités qui contribuent très peu aux deux premiers axes, mais il y a une infinité de possibilités. Une situation tout à fait courante et, reconnaissons-le, pas toujours aisée à identifier, est constituée par le mélange de deux populations structurées différemment : une (souvent plusieurs) modalité liée à une autre au sein d'une partie de la population ne l'est pas du tout dans une autre partie : les calculs sur le mélange ne laissent apparaître aucune relation. Découvrir selon quel critère les deux sous-ensembles se distinguent est parfois une affaire de longue patience...

Avec un minimum d'habitude, on peut aisément exécuter une douzaine d'essais différents en moins d'une heure, voire davantage. **Des donnés historiques réelles et qui n'ont pas été regroupées arbitrairement comportent toujours des éléments de structure**. Ce que d'aucuns appellent « l'infinie diversité du réel » n'est à aucun égard un capharnaüm, un élément de société ne s'apparente jamais à un tas, les individus ne sont jamais des boules dans une urne opaque. Les objets sociaux ont tous un sens, et un sens n'est rien d'autre qu'un ensemble de relations. Pour le chercheur, la difficulté consiste à identifier les relations et leur organisation. Il arrive que cela soit ardu, mais

les graphiques sur lesquels on ne voit rien ne montrent nullement l'absence de structure, ils traduisent seulement le fait que l'on n'a pas sélectionné les bonnes variables et trié convenablement la population étudiée.

Considérations finales

Notons d'abord que cette présentation de l'étude des distributions multivariées diffère nettement de ce que l'on trouve dans la majorité des manuels qui abordent cette question (qui sont loin de représenter la grande masse). *Les manuels et logiciels accordent en général une large place à la « régression multiple », et des développements substantiels aux « classifications automatiques »*.

La régression simple, qui touche deux variables, est une procédure de calcul qui introduit à la notion de distance. Son intérêt pratique, limité, est de permettre de repérer rapidement des écarts à une tendance simple. En général, on effectue une régression des y sur les x, puis des x sur les y ; autrement dit, on traite de manière analogue et symétrique les deux variables, les calculs n'impliquent en fait aucune hypothèse sur le sens de la relation. La régression multiple généralisant ce principe à plusieurs variables, la situation change en fait du tout au tout, car on choisit une variable comme « variable expliquée », toutes les autres étant regroupées comme « variables explicatives » ; en pratique, *cela peut avoir un sens (quoique contestable) dans le cadre des travaux des prévisionnistes* ; dans le domaine historique, une telle manière de procéder n'a de sens que dans des situations exceptionnelles, et lorsque l'on utilise certains types de régression (type « stepwise ») qui permettent, tant bien que mal, de distinguer les variables potentiellement « explicatives » de celles qui ne le seraient pas. C'est un exercice périlleux, qui au surplus implique une série d'hypothèses complémentaires, en particulier sur la nature mathématique des relations en jeu. Seule l'« économétrie rétrospective » fait usage de ces méthodes, avec des résultats indigents et le plus souvent hautement contestables. **A éviter**.

Les « classifications automatiques » partent, comme les analyses factorielles, de la notion clé de distance. Un tableau de distances entre objets étant donné, une multitude de procédures se proposent de regrouper les objets « au mieux ». L'idée peut paraître séduisante. Mais *deux obstacles* dirimants se présentent. D'abord, *on ne sait pas pourquoi tels ou tels objets se trouvés regroupés* ; bien entendu, diverses méthodes ont été suggérées pour tenter d'identifier les caractères propres de chaque groupe ; mais ces méthodes sont la plupart du temps lourdes et peu lisibles, et l'on constate en général que les critères ainsi mis en lumière ne sont pas cohérents : tel caractère dans un groupe se retrouve aussi dans un autre, et certains caractères au contraire disparaissent. Mais surtout, il est frappant de constater **la fondamentale instabilité des résultats**. L'expérience montre que toute variation dans le choix des distances et dans le choix des algorithmes de regroupement produit des groupes sensiblement différents, jamais les mêmes (et un minimum de raisonnement permet de montrer aisément qu'*il ne peut pas en être autrement*). Dès lors, ces méthodes ne peuvent avoir d'utilité que dans des conditions pratiques où il s'agit de produire un *découpage quelconque*, pourvu qu'il soit à peu près cohérent ; l'exemple classique étant la ventilation d'une clientèle en cinq ou six groupes, de manière à développer des statégies de promotion mieux ciblées. Je ne connais aucune donnée historique méritant un tel traitement. **A proscrire strictement**, car générateur d'artefacts indébrouillables.

Les analyses factorielles sont très peu appropriées à l'« aide à la décision » et à la « prévision », qui sont, de loin, les emplois dominants de la statistique. C'est certainement ce qui explique que, plusieurs décennies après leur mise au point, elles soient encore absentes de la

majorité des manuels de statistique les plus répandus et n'apparaissent, dans les principaux logiciels, que sous forme de bibliothèques annexes. D'un autre côté, il faut reconnaître que les ouvrages (même didactiques) qui leur sont consacrés, rédigés par des statisticiens mathématiciens, sont à peu près hermétiques au commun des chercheurs en sciences sociales. *Il n'est paru qu'une poignée de volumes consacrés à la pratique des analyses factorielles, et ils sont pour ainsi dire tous épuisés et introuvables*.

Rappelons donc brièvement, avant de terminer, *trois points généraux essentiels*. **1.** les analyses factorielles **ne sont nullement réservées aux grands effectifs** et aux vastes tableaux de données, comme l'ont malencontreusement affirmé certains de leurs promoteurs ; quelques centaines d'individus, au plus quelques milliers, sont des effectifs bien suffisants ; il faut même dire que la clé de bien des problèmes réside dans la prise en compte séparée de deux sous-populations (ou davantage). De tout petits tableaux (5 colonnes et 8 lignes, par exemple) peuvent être traités utilement par les analyses factorielles. **2.** on ne fait jamais **une** analyse factorielle, mais toujours **une série**, et l'on doit systématiquement contrôler les résultats d'une analyse par plusieurs autres. **3.** les analyses factorielles sont un outil, pas une fin en soi : elles visent à clarifier la structure des données, toute analyse doit se terminer par une partie consacrée au **retour aux données**, c'est une règle à laquelle on ne peut tolérer aucune exception.

Il me paraît hors de doute que **les analyses factorielles sont aujourd'hui la boîte à outils statistique la plus utile et la plus efficace pour les historiens, quels que soient la période ou le domaine étudiés**. Aucune autre méthode ne peut leur être comparée s'agissant de leur puissance d'exploration et de structuration des données. En imposant couramment l'usage de données formalisées et en manipulant méthodiquement des ensembles de relations, elles apportent une aide irremplaçable. On ne s'avancera pas beaucoup en disant que leur rôle dans la recherche historique est appelé à prendre une grande ampleur.

Chapitre 5

DONNÉES CHRONOLOGIQUES

Le temps des sociétés se mesure et se repère avec le temps astronomique (stable à l'échelle historique) ; mais la mesure du changement social dans le temps est une affaire bien plus complexe : les erreurs de perspective sont une faute courante. Avant tout calcul et même toute formalisation, il faut réfléchir sur le sens des durées que l'on prend en compte. Il s'agit en fait d'une première étape de toute sémantique historique.

SOMMAIRE

5.1. SOURCES ET PROBLÈMES : PRINCIPALES PERSPECTIVES

Il est indispensable de prendre une vue assez générale des divers types que l'on rencontre, afin de pouvoir placer convenablement, au sein d'un ensemble assez riche et complexe de méthodes et de problèmes, les données que l'on doit traiter.

5.1.1 types de « données »

On pense d'abord, en général, à la « série simple », suite de valeurs numériques correspondant à des *moments régulièrement espacés dans le temps*. Cette simplicité est trompeuse, car on peut avoir à faire à des objets très différents : il peut s'agir de mesures ponctuelles de phénomènes continus ou quasi continus, comme la température journalière ou l'effectif de population dans une zone donnée ; mais il peut aussi s'agir de dénombrements correspondant à l'intervalle de temps choisi : nombre de décès, chiffre d'affaires. On retrouve *l'opposition déjà évoquée entre « discret » et « continu »*, mais **compliquée** par la manière de prendre en compte la durée : mesure ponctuelle, mesure issue d'une moyenne à partir des valeurs d'un phénomène continu, somme d'éléments de toutes natures.

Même dans les cas les plus classiques, il faut se demander précisément de quoi il s'agit. Que veut dire « prix de la coupe de froment sur le marché de X au jour Y », ou, encore pire, « prix de la coupe de froment sur le marché de X, année Y » ? Des pondérations interviennent, que l'on contrôle rarement, et qui peuvent cacher des structures de première importance. S'il s'agit des « décès enregistrés dans la paroisse de X, année Y », de qui parle-t-on ? la lecture la plus superficielle de n'importe quel registre paroissial montre que l'on a enterré dans telle paroisse des individus qui y sont décédés plus ou moins par hasard (vagabonds, enfants en nourrice, etc) et l'on ne connaît ni le lieu ni l'année de décès de paroissiens morts ailleurs, parfois dans une paroisse voisine, parfois très loin. Dans ces conditions, que peut signifier la mise en relation d'une « courbe des décès » avec la « courbe des naissances » censée lui correspondre ? La combinaison, ou juxtaposition, de telles séries démultiplie les difficultés, *surtout lorsque les séries ne sont pas de même nature.*

Un cas un peu plus délicat consiste dans la succession d'états présentés sous forme de distributions : pyramide des âges ou hiérarchie des fortunes connues, par exemple, tous les cinq ans. Le type le plus fréquent, et le plus embrouillé, est la suite, pas forcément régulière, d'objets dont on connaît certains paramètres, ou certains caractères.

5.1.2 espacement

Les manuels qui traitent de données chronologiques considèrent comme matériau de base *l'« intervalle élémentaire », pris sans discussion comme unité.* Pour l'historien, **le jour, l'année ou le quart de siècle ne sont pas interchangeables**, quelle que soit la période considérée. Le temps de l'historien est **par définition un temps social**, ce qui entraîne des contraintes fortes au plan même de l'analyse formelle. L'historien est en permanence confronté au problème des « lacunes » ; les séries sans trou ne se rencontrent guère qu'en histoire contemporaine, et encore. On comprend sans trop de peine qu'une lacune de quelques jours n'a pas la même signification qu'une lacune d'un siècle, même si, pour le statisticien, c'est la même chose. On passe insensiblement d'une série lacunaire à des bribes de données réparties de manière irrégulière. C'est le lot commun de l'historien, c'est pour lui une difficulté de base, sur laquelle les manuels sont parfaitement silencieux.

5.1.3 rythmes et durée

Les données historiques comportent des **fluctuations**, on pourrait presque dire toujours et

partout, tant il est vrai que c'est *un caractère de base des sociétés humaines d'être par nature instables*. On observe des *fluctuations à toutes les échelles temporelles*. Fluctuations journalières, hebdomadaires, saisonnières sont marquées et le plus souvent faciles à reconnaître. Les variations **interannuelles**, très prononcées dans les climats tempérés, ont des conséquences aujourd'hui quelque peu atténuées, mais jouaient un rôle majeur (et fort mal étudié) dans toutes les sociétés préindustrielles. Si les variations climatiques à long terme (e.g. glaciations) sont un phénomène parfaitement établi (mais encore *très mal expliqué*), les variations à moyen et court terme (quelques décennies ou quelques siècles) demeurent un sujet de recherche, plus ou moins mêlé à des considérations extrascientifiques. A cette échelle, les relations avec l'évolution des structures sociales demeurent un sujet de spéculations plutôt obscures.

Cela étant, tous les indicateurs dont on dispose ont cette propriété générale de présenter « des hauts et des bas », selon *le vieux paradigme « après la pluie, le beau temps »*. Avant les historiens, les économistes du 19e siècle, intéressés à des titres divers par la spéculation (financière !), ont dépensé des trésors d'énergie pour identifier des « cycles ». Quelques historiens, heureusement pas trop nombreux, ont tenté de « retrouver » des cycles dans les données historiques. C'est un mirage grossier, que Pierre Chaunu a heureusement qualifié de « cyclomanie ». Les fluctuations historiques sont dénuées de toute périodicité mathématique, ce sont des « **pseudo-cyles** », ce qui n'est pas du tout la même chose, et ne s'étudie pas du tout de la même manière. Or la plupart des manuels consacrés aux « time series » consacrent la majeure partie de leurs développements aux méthodes de recherche de la périodicité (conçue en termes mathématiques, c'est-à-dire à l'aide de toute la quincaillerie des fonctions périodiques). La plupart de ces méthodes impliquent explicitement *l'hypothèse dite de « stationnarité », qui est strictement incompatible avec tout raisonnement proprement historique*.

5.1.4 durée et sens

Restent les deux questions de fond, qui sont le substrat même de la réflexion historienne :
1. **peut-on distinguer des « fluctuations » d'« évolutions à long terme »**, et si oui, comment ?
2. **comment articuler des comparaisons portant sur quelques jours, quelques années, voire quelques siècles** ? En d'autres termes : existe-t-il des *seuils au-delà desquels les comparaisons perdent leur sens*, ou pour le moins en prennent un tout différent ? (i.e. quelle est la pertinence historique de la notion de « série longue » ou de « longue durée » ?) Concrètement : des paramètres ***en apparence*** aussi simples que « le prix du pain » ou le « taux de mortalité » sont-ils des notions « transhistoriques » qui auraient un sens intrinsèque, indépendant de la société considérée ? et que l'on pourrait donc « comparer » à plusieurs siècles de distance ? C'est une faute de ne pas se poser la question.

5.2. SÉRIES SIMPLES

5.2.1 examens et manipulations élémentaires

Un ensemble de données numériques (vecteur), ordonné dans le temps, n'en reste pas moins *une distribution univariée*, qui doit, comme telle, être soumise aux procédures adaptées aux séries univariées. C'est une règle de méthode simple, et pourtant rarement respectée. Toute exploration de série chronologique doit commencer de cette manière car c'est le seul moyen d'en prendre une vue globale précise, par rapport à laquelle on pourra articuler le facteur chronologique. C'est d'ailleurs ainsi que l'on pourra, par exemple, s'apercevoir de l'utilité de procéder à telle ou telle *transformation* susceptible de rendre les données plus cohérentes.

Les **logarithmes** sont un des outils fondamentaux de l'analyse des séries chronologiques : l'écart entre deux logarithmes dépend du rapport et non de la différence entre les valeurs absolues correspondantes. Les courbes construites avec les logarithmes traduisent en général les évolutions beaucoup mieux que les courbes en valeurs absolues : les pentes indiquent des taux d'évolution, quelle que soit la valeur considérée : un passage de 20 à 40 sera figuré de la même manière qu'un passage de 5000 à 10000 et ainsi de suite. Dans des cas plus rares, on pourra faire appel à la racine carrée ou à la transformation de Box-Cox.

Mais surtout, une manipulation propre aux séries chronologiques est l'utilisation des **différences**, principalement de la *différence première*. On entend simplement par là le remplacement des valeurs brutes par la différence (algébrique) entre une valeur et la valeur précédente. Notons d'ailleurs que l'on obtient ainsi une *distribution dérivée* qui gagne à être soumise aux procédures classiques. Il arrive par exemple que cette distribution des différences soit nettement dissymétrique, ce qui signifie que l'on observe un grand nombre de différences positives d'ampleur limitée et quelques différences négatives importantes (ou l'inverse), phénomène qu'il est presque impossible de voir directement par simple lecture de la courbe. Les *différences secondes* sont obtenues en différenciant les différences premières (et ainsi de suite), l'interprétation est rarement simple. On peut aussi choisir un autre intervalle que la simple succession, et calculer la différence par rapport à la valeur figurant deux rangs avant, ou trois, etc. Ce genre de procédure n'est utile que si la série est clairement périodique : il peut être intéressant d'utiliser le coefficient -7 pour une série journalière (ou -12 pour les mois).

5.2.2 décompositions

Un siècle de considérations obscures et contradictoires sur les soi-disant cycles n'ont nullement convaincu les économistes de la futilité d'une telle démarche, et les historiens leur ont emboîté le pas dès avant la seconde guerre mondiale. Dans les années 50 et 60, les noms de Juglar, Kondratieff, Simiand passaient pour renvoyer à des réalités cruciales. Pire : on a vu se routiniser à cette époque une procédure inventée, semble-t-il, à la fin des années 40, *la décomposition des séries en trois éléments (additifs) : la tendance (« trend »), les variations cycliques, et le résidu aléatoire, baptisé « bruit blanc »*. La seule règle (approximative) concerne la structure de ce dernier élément, qui doit, en principe, être en moyenne nul et de variance stable. Mais la tendance peut être définie ad libitum : on peut se contenter d'une fonction linéaire, mais on peut aussi, pour les besoins de la cause, utiliser n'importe quelle fonction polynomiale. Il est inutile d'entrer dans de grands détails : avec un peu d'habileté et d'habitude, ce petit bricolage permet presque à tout coup de trouver une « composante cyclique » présentable, et souvent plusieurs différentes. Ainsi, les résultats dépendent en très grande partie de la procédure utilisée, et *finalement on interprète l'effet d'un découpage tout à fait arbitraire*. Même si, chez les historiens, la fréquence d'emploi de cette manipulation a nettement décru, on continue de parler de « tendance », de « variations cycliques », de « chocs aléatoires ».

Indéniablement, ce genre de décomposition pifométrique rend des services aux prévisionnistes, qui peuvent ainsi extrapoler de manière plausible les données des 12 ou 30 derniers mois (méthode célèbre sous le nom de « Box-Jenkins », les deux statisticiens anglais qui ont mis au point dans les années 60 les algorithmes les plus efficaces). Il suffit de savoir en gros de quoi il s'agit. La cuisine des prévisionnistes est à l'exact opposé des perspectives de toute exploration sérieuse de données historiques. Il faut reprendre la question de manière analytique précise.

5.2.3 l'autocorrélation

En termes abstraits, il est tout à fait légitime, et même indispensable, de se demander dans quelle mesure la valeur de la série au moment Tn « dépend » des valeurs en T-1, T-2, etc. La

réponse est fonction avant tout de la nature du phénomène considéré et de la périodicité. Si l'on a une série journalière, de températures ou de prix, quelles que soient les oscillations, on sera nécessairement loin de l'« indépendance ». Si l'on dispose de moyennes annuelles, la question sera plus ouverte, a fortiori si l'on examine des moyennes décennales.

La procédure de base est simple : *on calcule le coefficient de corrélation linéaire entre la série brute et la même série décalée d'un cran, de deux, de trois, etc.* Un graphique en bâtonnets permet de visualiser facilement les résultats, et l'interprétation de premier degré est presque immédiate. S'il existe un cycle (semaine ou année), le diagramme le montre instantanément : on observe des pointes régulièrement disposées en fonction des multiples de la période (par exemple, 7, 14, 21, 28, etc.). Avec R, il suffit d'employer la fonction `acf()` (autocorrelation function).

Cette reproduction des pics à intervalles réguliers étant notée, il peut être intéressant de faire disparaître ce qui n'est qu'un artefact, en éliminant cet effet de reproduction ; par exemple, pour essayer de voir, au delà d'un cycle de sept jours, jusqu'à quelle distance l'autocorrélation reste significative. On utilise pour cela ce que l'on appelle l'**autocorrélation partielle** (R : `pacf()`).

5.2.4 tendance, taux moyen

Si, dans une zone donnée, la population a doublé en 50 ans, quelle que soit l'allure de détail de la courbe, on parlera de tendance sans trop d'arrière-pensée. Signalons ici **un piège grossier et très meurtrier** : des régiments d'historiens y sont tombés. Pour reprendre cet exemple : 100% d'augmentation en 50 ans, quel est le taux moyen annuel de croissance ? **Eviter à tout prix la division fatale !** Si l'on désigne ce taux par x, on part de 1 à l'année 0 ; on a ensuite 1(1+x), puis (1+x)(1+x), puis (1+x)(1+x)(1+x), etc (c'est que l'on appelait jadis la « *loi des intérêts composés* ») ; finalement, au bout de 50 ans, on a (1+x) puissance 50. Si donc (1+x)^50 = 2, x vaut 2 racine 50e (ce qui s'écrit 2^(1/50), soit à peu près 1,014, ou 1,4%). Faire ici un autre calcul est une faute inexcusable, qui bien souvent entraîne des conclusions farfelues. (Ceci est naturellement tout aussi vrai dans l'autre sens : on n'additionne pas des pourcentages annuels).

Cela n'est pas sans rapport avec la manière de construire les graphiques. Si l'échelle des ordonnées est une échelle arithmétique ordinaire, une courbe correspondant à un taux moyen de variation constant sera toujours courbe : concavité vers le haut si augmentation, vers le bas si diminution. Une courbe rectiligne traduit une variation constante en valeur absolue, c'est-à-dire un taux moyen variable (en diminution ou en augmentation...). Au contraire, un taux constant sera représenté par un segment droit *si l'échelle des ordonnées est une échelle logarithmique* : comme on l'a déjà indiqué, avec une échelle logarithmique, la pente correspond au taux de variation, et non à la différence en valeur absolue. Une règle simple consiste à essayer méthodiquement les deux modes de représentation.

Mais la tendance est rarement résumée convenablement par la comparaison brute entre la situation de départ et la situation d'arrivée. Il faut essayer de repérer une forme générale.

5.2.5 la fenêtre mobile

L'outil essentiel d'examen des séries chronologiques est constitué par la **fenêtre mobile**. Comme le nom le laisse supposer, on choisi un intervalle plus large que l'intervalle unité, que l'on fait en quelque sorte *glisser le long de la série*. Et, à chaque cran, on peut calculer divers paramètres, comme moyenne, moyenne pondérée, écart-type, médiane, écart à la médiane, écart inter-quartile, etc., c'est-à-dire *les principaux paramètres de valeur centrale et de dispersion.* On obtient ainsi plusieurs séries dérivées, que l'on peut immédiatement reporter sur un graphique. La facilité des calculs faisait jadis préférer la moyenne mobile. Celle-ci, éventuellement pondérée, représente un choix acceptable si la fenêtre est étroite, de deux à cinq éléments ; la médiane est plus fiable, même si, graphiquement, le résultat est souvent moins agréable, du fait de sauts par palliers.

Il ne faut surtout pas se priver d'examiner aussi (on peut le faire sur le même graphique) un ou plusieurs paramètres de dispersion. Rappelons encore une fois que, *dans le cas d'une distribution non symétrique, l'utilisation de la moyenne peut conduire aux erreurs les plus grossières*. Dans certains cas, on peut avoir à faire à des distributions sans valeur centrale.

Un des intérêts des outils numériques et graphiques actuels est que l'on peut sans difficulté aucune *faire varier autant qu'on le souhaite la largeur de la fenêtre mobile, et comparer les résultats*. Au fur et à mesure que l'on élargit la fenêtre, la courbe des valeurs centrales prend un aspect de plus en plus régulier. La courbe de tel ou tel paramètre de dispersion suit au contraire une évolution imprévisible, qui tient à la structure même des fluctuations. Il faut regarder de près.

On voit donc que la procédure de la fenêtre mobile peut, à partir d'une série simple unique, engendrer un grand nombre de courbes dérivées. Six est un ordre de grandeur acceptable : deux paramètres avec une fenêtre étroite, une fenêtre moyenne et une fenêtre large. Cette procédure fort simple reste sans doute le meilleur moyen de visualiser et d'analyser sans a priori caché l'évolution des caractères de la série sur toute sa longueur.

5.2.6 lissages

La moyenne mobile a longtemps été considérée comme le principal sinon le seul outil de « lissage ». Remplacer les valeurs brutes par une valeur centrale définie dans une fenêtre mobile a bien entendu pour effet de faire disparaître les « pics » résultant (c'est l'hypothèse implicite) de fluctuations brèves moins significatives. C'est une *hypothèse contestable et dangereuse*. L'exemple classique du prix du blé est assez parlant : les fluctuations interannuelles, et dans le cours de chaque année, avaient la plupart du temps une signification sociale bien supérieure aux « évolutions à long terme », pas forcément bien perceptibles, et qui pouvaient résulter tout autant d'un allongement (ou d'un raccourcissment), dans l'année, des semaines de cherté que d'une élévation globale du niveau moyen. Il est insensé de se contenter d'une considération mathématique d'une suite de chiffres d'où l'on tire *des paramètres que l'on croit pouvoir interpréter indépendamment de l'unité de temps considérée*. D'ailleurs, de nos jours encore, les agriculteurs savent pertinemment qu'un seul jour de gel en avril peut avoir des effets bien plus massifs qu'une « température moyenne » un peu basse durant plusieurs semaines (alors même que ce second phénomène sera beaucoup plus visible sur une courbe lissée).

Les moyens de calcul actuels rendent possible le remplacement des « courbes » à l'ancienne, constituées en fait d'une succession de segments formant des angles plus ou moins aigus, par des « **splines** », qui passent exactement par les mêmes points, mais les joignent par des segments sinueux, sans aucun angle [en fait, une succession de fonctions, habituellement du 3e degré, auxquelles on impose d'être tangentes à la même droite à chaque point de suture]. Les pointes aigües disparaissent ainsi par une simple modification de la procédure de dessin. C'est un avantage important, dont il ne faut pas se priver.

Dans certains cas, assez spécifiques, il peut être utile de rechercher un ajustement fonctionnel. C'est notamment le cas lorsque l'on passe d'un état à un autre par un flux (mouvement quelconque), qui démarre lentement, s'accélère et gonfle, puis, après un instant de tension maximale, décroît peu à peu jusqu'à stabilisation dans un nouvel état. On peut alors, le plus souvent, envisager un ajustement sur **une fonction logistique**. Cette procédure, comme tous les ajustements, permet de « résumer » convenablement le mouvement, et donc de permettre des comparaisons plus claires et plus simples de mouvements parallèles ou analogues.

Dans le cas habituel, la notion intuitive de « tendance » doit donner lieu à une réflexion approfondie ; le repérage, qui peut être décisif au plan de l'interprétation, des « *ruptures de tendances* » (alias « *dates charnières* ») est un exercice délicat, pour lequel n'existe aucun outil passe-partout ni aucune règle vraiment générale. C'est spécialement vrai lorsque les variations

brèves ont une amplitude égale ou supérieure aux écarts de la « tendance ».

Cette difficulté de base est d'ailleurs étroitement liée au problème général (et le plus souvent traité par prétérition) de la construction du graphique. La plupart des logiciels commencent par identifier la valeur minimale et la valeur maximale de la série, et déterminent les valeurs portées sur l'axe des ordonnées en fonction de ces deux nombres. De telle manière que la représentation des variations utilise toute la hauteur disponible de l'espace graphique. Ce choix par défaut accorde ainsi toujours la même valeur graphique à l'écart (max-min), quelle que soit la valeur de cet écart, et surtout quelle que soit l'ampleur de la variation relative, par rapport à la mesure ou à l'effectif considéré. Dans R (comme dans la plupart des logiciels qui dessinent des graphes), on peut choisir les valeurs extrêmes portées en ordonnées, et il est le plus souvent impératif de procéder à divers essais, sans quoi l'on risque de *confondre les Alpes et une motte de terre...*

5.3. DEUX SÉRIES

5.3.1 traitements préalables

Il faut, dans un premier temps, considérer deux séries chronologiques correspondant à la même période comme *deux séries appariées ordinaires*, et leur appliquer *les traitements prévus* dans ce cas : transformations appropriées et dessin de graphiques, sans oublier l'examen des nuages. Ce type de traitement doit permettre de déterminer s'il existe ou non une liaison instantanée entre les deux courbes.

5.3.2 analyses numériques

Un examen visuel des deux courbes portées sur le même graphique peut déjà donner une idée de leur relation. Bien entendu, en particulier si les valeurs absolues des deux séries ne sont pas exactement du même ordre de grandeur, **un graphique semi-logaritmique s'impose**.

La possibilité existe d'utiliser deux échelles arithmétiques différentes. Cela peut donner des graphiques suggestifs, mais il faut faire ici très attention à éviter les effets artificiels (en pratique, c'est un des moyens les plus couramment employés pour faire dire à des chiffres le contraire de ce qu'ils disent).

Il faut éventuellement procéder à plusieurs visualisations (en découpant la série en tronçons successifs) si le nombre de points est élevé. En second lieu, on peut visualiser les **corrélations croisées** (R : `ccf()`). Le programme affiche sous forme d'un diagramme en bâtonnets les corrélations linéaires de la série A avec la série B(n), B(n-1), B(n-2), etc. On peut ainsi, dans certains cas, mettre en évidence la liaison entre les deux séries avec un décalage de x intervalles. Dans ce cas, reprendre les deux séries en tenant compte de ce décalage et revenir aux procédures du paragraphe précédent.

Lorsque l'on étudie deux séries, disons approximativement moyennes ou longues, il peut arriver que les deux séries soient bien liées durant une ou plusieurs périodes, et très peu pendant d'autres. Il peut alors être efficace d'employer la fenêtre mobile, pour calculer cette fois un « **coefficient de corrélation linéaire mobile** », dont l'affichage permet de visualiser, le cas échéant, les périodes de liaison et celles de non-liaison. Comme précédemment, il est nécessaire d'essayer plusieurs largeurs de fenêtre. Il n'est pas interdit non plus d'afficher simultanément deux médianes mobiles et deux écarts interquartiles mobiles (par exemple en utilisant deux couleurs).

5.4. PLUSIEURS SÉRIES

5.4.1 précautions

Un graphique comportant plus de trois ou quatre courbes devient rapidement illisible, même en utilisant toutes les techniques graphiques disponibles (forme du trait, épaisseur, couleur), sauf naturellement dans des cas particuliers, comme le parallèlisme presque parfait, ou tout au moins des formes simples et non-sécantes.

5.4.2 la recherche de « profils » : analyse factorielle des correspondances

Des cas tout à fait habituels sont constitués par des séries de sous-ensembles (postes de recettes ou de dépenses), des séries de même nature dans un même cadre (prix sur le marché de x), ou des séries identiques dans des cadres différents (population de diverses localités). La question dans ce cas est de distinguer quelques groupes évoluant de manière plus ou moins analogue, ce que l'on appelle aussi « trier les profils ». **L'outil privilégié est ici l'analyse factorielle des correspondances** (AFC) : on observe presque toujours, sur le plan factoriel 1-2, le regroupement des éléments ayant même profil. « presque » : il arrive en effet qu'un élément « perturbateur » (parfois un petit groupe), ayant une évolution complètement erratique par rapport aux autres séries, « produise un axe », qui peut même est l'axe 1 ou 2, si cette perturbation est forte. Mais cette anomalie est facilement repérable : l'AFC affiche les points-dates en même temps que les points-matières ; ces points doivent se retrouver grossièrement dans leur ordre naturel sur le plan factoriel déterminant. Cet outil est d'un usage commode. Comme d'ordinaire avec les analyses factorielles, il est vivement recommandé de procéder à de nombreux essais, en prenant des parties du tableau ou en utilisant plusieurs types d'analyse factorielle.

5.4.3 série de distributions

On peut évoquer ici les suites de distributions. Dans certains cas, si les objets successifs sont très homogènes, un découpage en classes est possible, à partir duquel on revient au cas précédent, c'est-à-dire l'analyse factorielle des correspondances (puisque le découpage produit en définitive autant de séries que l'on a choisi de classes). Si la série n'est pas trop longue (disons moins d'une centaine d'éléments), **un alignement (horizontal) de boxplots (verticaux)** est simple et très efficace : chaque élément représente assez clairement les principaux paramètres de la distribution, et leur juxtaposition permet donc de saisir globalement l'évolution.

5.5. DONNÉES ÉPARSES, DONNÉES NON DATÉES

5.5.1 données manquantes

Si une série à peu près régulière et homogène comporte quelques lacunes, il faut essayer de la traiter comme une série complète. La plupart des logiciels ignorent le problème, qui est le lot commun des historiens (les logiciels ne sont conçus ni par ni pour des historiens). Sur un graphique, il vaut mieux laisser un blanc : tracer un segment au travers de la lacune a souvent pour effet d'attirer l'attention sur un point qui justement n'en mérite aucune. Si la lacune est tout à fait ponctuelle, elle n'est pas très gênante ; il faut toutefois s'assurer que la procédure de la fenêtre mobile est capable de tenir compte de l'absence d'un ou plusieurs éléments. « *Interpoler* », c'est-à-dire boucher le trou à partir des éléments situés de part et d'autre, est plus ou moins simple, selon l'allure de la courbe. En particulier, si les fluctuations instantanées sont faibles, on ne court guère de

risque. Si au contraire les variations à très court terme sont fortes par rapport aux évolutions plus longues, l'exercice devient beaucoup plus périlleux. En effet, on doit, en principe, tenir compte des divers paramètres de la distribution des valeurs environnantes ; une interpolation avec une valeur proche de la valeur centrale va laisser celle-ci intacte, mais faire baisser (indûment) le paramètre de dispersion ; inversement, en supposant un écart sensible, on va conserver la dispersion, mais on risque de modifier la valeur centrale... Les analyses factorielles multiples présentent l'immense avantage de prévoir les cases vides, sous la forme du code zéro (« non-réponse »).

5.5.2 données éparses

Ici, pour ainsi dire, *tout est cas d'espèce*. Si par exemple on dispose de fragments de série séparés par de vastes blancs, on analysera chaque fragment comme une distribution indépendante, et l'on comparera les paramètres ainsi obtenus.

Si les fragments concernent plusieurs séries, on peut, moyennant quelques précautions, presque toujours présenter ces données sous la forme d'un tableau traitable par analyse factorielle. Dans ce cas, il faut procéder à de nombreux essais, notamment en faisant passer en *éléments supplémentaires* autant d'éléments que nécessaire pour éviter un déséquilibre au profit d'un fragment surreprésenté. Eviter cependant d'éliminer une partie de l'information disponible, en aucun cas se satisfaire d'une quelconque valeur centrale pour chaque fragment, car il est courant que l'évolution touche autant la dispersion que la valeur centrale (si celle-ci a vraiment un sens, ce qui, on ne le rappellera jamais assez, n'est pas toujours le cas).

5.5.3 données non datées : la sériation

Pour toutes les périodes anciennes, le problème des éléments non datés a souvent bien plus d'importance que celui des évolutions, qu'en fait on ne connaît pas a priori. Ce qui peut d'ailleurs aussi se produire à des époques bien plus récentes. La situation-type est bien connue : on dispose d'un certain nombre d'« objets », qui peuvent être des chartes dans un cartulaire, des lettres dans un recueil, des tombes dans un cimetière mérovingien, des églises romanes dans une contrée, etc. On peut décrire ces objets plus ou moins finement, mais on n'en connaît pas la date (chronologie absolue), ni même l'ordre chronologique (chronologie relative). En gros depuis les années 60, des chercheurs se sont aperçus que, **dans le cas d'ensembles homogènes, la mise en ordre des modalités des caractères des objets permet d'obtenir une « sériation », fournissant un ordre approximatif de succession tant des individus que des modalités des divers caractères**.

L'idée fondatrice est assez simple : on suppose que, pour chaque caractère, les modalités se succèdent, mais que tous les changements ne sont pas simultanés, ce qui permet de faire apparaître, au moyen de manipulations adéquates, une sorte de *« chaînage » des modalités* des divers caractères considérés. Les possibilités et les limites d'une telle procédure ont été peu à peu étudiées, tout simplement en observant le comportement des modalités et des individus ainsi décrits et manipulés, mais munis également d'une date précise.

La conclusion principale est que *la méthode de loin la plus puissante est l'analyse factorielle des correspondances*, qui ordonne simultanément, selon les mêmes critères, les modalités et les individus. Dans les cas favorables, on obtient ce que l'on appelle un « **scalogramme** », *c'est-à-dire un nuage de points régulier en forme de parabole*. Cette situation n'est pas rare.

Dans des cas plus nombreux, le scalogramme est perturbé par un ou deux caractères dont les modalités varient sans lien avec la chronologie. D'un autre côté, les analyses menées sur des objets datés montrent nettement que la sériation fait ressortir une chronologie approximative : l'ordre ainsi établi diffère quelque peu de l'ordre chronologique réel ; ce qui se comprend aisément : à chaque moment, certains objets sont « archaïques », tandis que d'autres sont « en avance sur leur temps ». Si l'on connaît les dates réelles, la sériation est particulièrement fructueuse de ce point de vue,

puisqu'elle permet ainsi de se rendre compte, à chaque moment, *quels sont les objets (individus) en retard ou en avance par rapport à l'évolution générale.* Dans l'état actuel de la recherche, on peut encore attendre de sérieux progrès d'études à venir, qui permettront de mieux préciser l'amplitude chronologique de ces écarts et leur fréquence. Il est hautement plausible que l'unité de référence dans ce domaine soit la longueur de la vie humaine et sa variabilité : dans un cimetière, on trouve côte à côte des individus dont l'âge au décès peut différer de cinquante ans ; si l'on observe les bijoux, on risque ainsi de trouver des parures dont les moments de fabrication sont répartis sur plusieurs décennies, alors même que les tombes sont contemporaines. Cet exemple se généralise aisément.

On ne doit toutefois pas se cacher que certaines situations sont complexes sinon inextricables. Il peut arriver que *les caractères sans évolution* pèsent d'un poids considérable, de telle sorte que, dans les résultats de l'analyse factorielle, le facteur indicateur de la succession ne sera pas le premier, ni même éventuellement le second. Surtout, on peut avoir à faire à des populations non homogènes, dont les éléments évoluent selon des successions différentes. En principe, de la patience et de l'imagination doivent permettre de débrouiller l'écheveau, c'est-à-dire de séparer les sous-populations mélangées, et finalement d'ordonner chacune d'entre elles. Cela peut demander beaucoup de temps... On peut aussi se trouver, notamment si l'on considère des séries longues (pluriséculaires), devant des caractères ayant valeur chronologique durant une partie de la période considérée, et n'en ayant pas ou plus durant une autre partie.

Les quelques essais concrets (dans le domaine médiéval) permettent de cultiver **des espoirs raisonnables**. En examinant les données sous toutes leurs coutures, il est à peu près sans exemple que l'on ne repère pas au moins un caractère dont on puisse penser, ou tout au moins supposer solidement, qu'il peut être considéré comme un marqueur chronologique fiable. En essayant une multitude de combinaisons actifs-supplémentaires (notamment au sein de la population, c'est-à-dire en analysant des sous-populations) on parvient à des reconstructions qui « collent » d'assez près au critère chronologique. Il est rare également que l'on ne trouve pas, de-ci de-là, quelques éléments, même ténus, de datation absolue.

Signalons, pour finir par une indication qui ouvre la porte au prochain chapitre, *une combinaison très astucieuse due à François Djindjian*, la **toposériation**. Les tombes mérovingiennes ont donné lieu à un ensemble de sériations plus ou moins précises. L'idée de F. Djindjian a consisté à combiner l'analyse abstraite des tombes considérées comme des « ensembles clos » d'objets typés, avec la prise en compte de leurs positions respectives, en partant de l'hypothèse que les cimetières de ce genre se sont développés par extension topographique progressive, et que par conséquent l'ordre reconstruit par sériation simple doit être corrigé par la prise en compte topographique du mode d'accroissement de la zone d'inhumation. Les résultats sont remarquables, et montrent que le chercheur doit toujours partir de l'hypothèse que des éléments de formalisation simples et pertinents n'ont sans doute pas encore été repérés, et qu'une réflexion empirique sur les données en apparence les plus pauvres peut apporter des résultats inattendus et novateurs.

Notes finales

Se rappeler avant tout que :

* toute considération de série chronologique doit commencer par une réflexion approfondie sur l'unité de temps employée ; de cette unité et de son sens dépendent toutes les conclusions ultérieures ;

* le raisonnement en proportion est ici plus indispensable que nulle part ailleurs, ce qui implique le recours systématique aux logarithmes ;

* la notion de « cycle » est à proscrire strictement ;

* les analyses factorielles peuvent rendre dans ce domaine des services considérables, et sont encore regrettablement sous-employées ;

* la notion même de « série », dans le domaine historique, doit être considérée avec une extrême prudence ; toutes les données historiques sont « dans le temps », ce qui implique le devoir d'examiner à fond ce facteur, même s'il se présente sous des formes irrégulières, difficiles à formaliser. Beaucoup de recherches dans cette perspective sont encore possibles et nécessaires.

Chapitre 6

DONNÉES SPATIALES: CARTOGRAPHIE

L'histoire d'une représentation à peu près convenable de la surface terrestre commença très lentement au 17e siècle. Il est anachronique et absurde de parler de cartes à propos des documents graphiques antérieurs. Des outils (objets matériels), d'une précision acceptable, de mesure des distances et des angles furent mis au point et diffusés peu à peu au cours du 18e siècle. La question de la forme exacte du "globe terrestre" ne fut abordée sérieusement que dans la seconde moitié du 19e. La France ne se dota d'une couverture cartographique soignée qu'à partir de 1870. Les techniques de calcul et de représentation furent complètement bouleversées par l'irruption des satellites et de l'électronique. Il n'est pas excessif de dire que toutes les méthodes théoriques et pratiques de la cartographie ont été renouvelées de fond en comble au cours des vingt dernières années.

Sans entrer dans des détails inutiles pour le profane, il est indispensable de connaître les grandes lignes de cette évolution, ne serait-ce que pour comprendre ce qui est imprimé aujourd'hui sur une simple carte au 1/25000, ou a fortiori pour savoir ce que représentent les indications des appareils GPS vendus dans les supermarchés. Il faut procéder succinctement à cette étude descriptive avant d'aborder la question des statistiques spatiales.

SOMMAIRE

6.1. LA REPRÉSENTATION PLANE DE LA SURFACE TERRESTRE

6.1.1 la géodésie : description géométrique de la terre

Que la terre soit une sphère est indiscutablement établi au moins depuis le 16e siècle. Mais en réalité, ce n'est pas tout à fait une sphère. Au 19e siècle, il est apparu de plus en plus clairement que cette sphère est légèrement "aplatie" aux pôles. Si l'on considère un méridien, le demi-grand axe situé dans le plan équatorial (6378km) est environ 22km plus long que le demi-grand axe allant d'un pôle à l'autre (6356km). Ce fut un officier britannique, *Alexander Ross Clarke* (1828-1914) qui établit le premier de manière convaincante à partir de 1860 que la meilleure représentation géométrique de la terre est un ellipsoïde de révolution. Durant plus d'un siècle, la France utilisa l'**ellipsoïde de "Clarke1880"**.

La première difficulté provient du fait que, dans le détail, il est impossible de définir un ellipsoïde incontesté qui convienne parfaitement en tout point de la terre. On compte actuellement plus de 200 définitions différentes de l'ellipsoïde, dont au moins trois ou quatre d'usage courant. Il n'y a que quelques centaines de mètres de différences sur les deux axes, mais ces différences sont bien trop importantes pour que l'on puisse les négliger si l'on souhaite travailler avec précision. A l'heure actuelle, la référence est le système **WGS84** (World Geodetic System 84). C'est ce système qui sert de référence aux satellites qui permettent de repérer n'importe quel point avec une précision de l'ordre du centimètre (outils du **GPS** - Global Positioning System, plus ou moins précis selon les appareils au sol, la précision centimètrique n'est obtenue que par des appareils assez encombrants d'une vingtaine de kg). Ces matériels ont rendu obsolètes tous les outils de visée angulaire et de mesure des distances qui étaient utilisés jusque dans les années 80 (du moins pour une utilisation cartographique).

6.1.2 les représentations planes

Ni une sphère ni un ellipsoïde ne sont "développables" sur une surface plane. Il faut donc définir un procédé mathématique de transformation, appelé **projection**, qui permette de traduire le mieux possible une surface courbe en surface plane. En pratique, les difficultés sont d'autant plus grandes que l'échelle est petite (c'est-à-dire que la zone considérée est plus vaste, ou que l'on se rapproche d'une planisphère). A moyenne ou grande échelle, les diverses projections donnent des résultats peu différents. Mais les référentiels sont bien distincts, et l'on ne passe pas sans calculs des uns aux autres. La France a adoptée à la fin du 19e siècle *la projection conique de Lambert*, i.e. un cône tangent au méridien de Paris (défini sur la base de l'ellipsoïde de Clarke 1880). Pour obtenir une bonne précision, il a été décidé d'utiliser quatre projections distinctes du nord au sud de la France continentale, désignées comme Lambert I, Lambert II, Lambert III et Lambert IV. Pour les représentations globales de la France, on utilise ce que l'on appelle le **Lambert II étendu**. Ce système de projection a l'*énorme avantage d'aboutir à un quadrillage kilométrique*, qui facilite de manière décisive tous les calculs de distances et de surfaces. Son emploi reste tout à fait général en France. C'est le système dans lequel sont repérés tous les points de la **NTF** (Nouvelle Triangulation de la France), élaborée à partir de la fin du 19e et précisée jusque dans les années 80. C'est le système employé par exemple pour définir la position d'une zone de fouille, d'un monument (église, château, etc.). C'est un système planimétrique (deux dimensions) établi principalement par visées angulaires. Il est matérialisé dans le paysage par plusieurs dizaines de milliers de bornes.

Après la seconde guerre mondiale, l'armée américaine, dans le cadre de l'OTAN, a constitué un système uniforme pour toute l'Europe, baptisé **ED50** (European Datum 1950), dans lequel furent intégrées par calculs toutes les triangulations nationales. Le référentiel est l'ellipsoïde de Hayford

1909, le méridien d'origine celui de Greenwich, et la projection, la projection Mercator transverse universelle (**UTM - Universal Transverse Mercator**), c'est-à-dire une projection par tranches longitudinales (entre deux méridiens) sur un cylindre perpendiculaire à l'axe de rotation de la terre.

Toute la géodésie géométrique a été entièrement revue sur la base des observations satellitaires. Le World Geodetic System 84 a été universellement adopté. Il repose sur un nouvel ellipsoïde, sur le méridien de Greenwich et la projection UTM (divisée en *60 fuseaux de 6 degrés de largeur : la France est à cheval sur les fuseaux 30, 31 et 32* (qui correspondent aux longitudes -6 à 0 degrés, 0 à 6 et 6 à 12). Dans les années 90, les autorités françaises ont donc décidé de redéfinir complètement le système de référence français, d'où ce que l'on appelle le système **RGF93** (Réseau Géodésique Français, qui remplace le système NTF ; c'est un système tridimensionnel, largement appuyé sur les observations satellitaires), système calé étroitement sur le système WGS84. On aboutit à plusieurs nouveaux quadrillages kilométriques ; le système **Lambert93**, projection conique calée sur l'ellipsoïde WGS84 (en gros une simple translation par rapport au Lambert II étendu), mais aussi l'UTM lié au WGS84, à présent imprimé en bleu sur les dernières cartes au 1/25000 (pour le repérage par GPS). Comme il s'agit de deux méthodes de projection différentes, les décalages sont sensibles, surtout à longue distance. La largeur de la France (Ouessant est à 5° W et Lauterbourg à 8° E, soit 13° d'écart) fait que le pays est projeté sur 3 fuseaux successifs : 30, 31 et 32, ce qui complique dangereusement un usage simultané de coordonnées cartésiennes issues, éventuellement, de trois projections distinctes. Pour un usage à l'échelle nationale, le système Lambert93 est plus commode (une majorité de documents administratifs sont encore en Lambert traditionnel ; de toute manière, les conservateurs vont devoir s'y adapter, comme à la chronologie d'ancien et de nouveau style !).

6.1.3 hauteurs et altitudes

Depuis la fin du 19e siècle, il est devenu clair que la terre ne correspond précisément à aucune forme géométrique régulière, on parle donc de "**géoïde**", les différences (en altitude) peuvent atteindre plus ou moins 100 mètres (par rapport à un ellipsoïde). La masse terrestre n'est pas répartie de manière parfaitement régulière ; la seule manière (jusqu'à aujourd'hui) d'évaluer ces irrégularités est de mesurer exactement, en chaque point, la valeur de la pesanteur. On obtient ce que l'on appelle une "*cote géopotentielle*". L'adoption d'une *surface équipotentielle de référence* permet alors, par soustraction, de traduire l'altitude par une mesure linéaire. La surface de référence est le géoïde, défini de telle manière que la surface moyenne des mers, supposée prolongée sous les terres émergées, en soit une approximation. En matière d'altimétrie, on aboutit ainsi à deux valeurs : l'altitude, déterminée par une mesure gravimétrique locale, et qui est calculée en référence au géoïde, et la hauteur, définie géométriquement par rapport à la surface d'un ellipsoïde. En France, les deux mesures ont été effectuées dans plusieurs milliers de points (dits "en collocation"), ce qui donne une grille empirique à partir de laquelle on peut effectuer, par interpolations, les conversions hauteur-altitude de n'importe quel point du territoire (je simplifie beaucoup). Ce que l'on trouve sur les cartes n'est donc pas l'altitude mais la hauteur !! (également fournie par le GPS, encore que celui-ci se réfère à son ellipsoïde propre, tandis que les cartes françaises sont toujours conçues en référence à l'ellipsoïde Clarke1880....).

6.1.4 conversions

En dépit des simplifications abusives des lignes qui précèdent, il apparaît que l'on se trouve, au début du 21e siècle, dans une situation où coexistent (sans doute encore pour longtemps) plusieurs systèmes de référence bien distincts. Les actuelles cartes au 1/25000, issues des travaux dans le cadre de la NTF, ne portent, dans leurs marges, pas moins de cinq graduations différentes. Tous les systèmes de gestion de données géographiques doivent impérativement inclure des outils

de conversion. L'IGN (www.ign.fr) met gratuitement à la disposition du public un logiciel free, intitulé **Circé 2000**, qui facilite grandement toutes les conversions concernant la France, et qui comporte également *un fichier d'aide* qui expose de manière assez précise tous les problèmes de la géodésie et les méthodes mathématiques de conversion : il faut en prendre connaissance. On trouve un grand nombre de pages internet consacrées à ce sujet, de qualité variable.

6.2. GESTION DES DONNÉES SPATIALES

6.2.1 évolutions

Les données de base aptes à permettre la création de documents plans de représentation de portions de la surface terrestre ont subi une évolution extraordinaire. Il ne semble pas que la chaîne d'arpenteur métallique soit antérieure au 16e siècle. Comme on l'a signalé plus haut, des outils de mesure et de visée précis, (métalliques, i.e. surtout laiton, cuivre, acier) ne se sont répandus qu'à partir du 18e siècle. Le terme « *vernier* » ne fit son apparition qu'en 1795 (la première description de l'outil date de 1631), ce qui montre la lenteur de la mise en œuvre des procédés purement mécaniques destinés à améliorer la précision. Les règles à échelles ont été inventées et diffusées par la famille viennoise *Kutsch* à partir de la fin du 18e. Cet outillage permettait des mesures visuelles, dont les résultats devaient être reportés sur des feuilles ou des cahiers, avant d'être traduits en dessins précis, opération que l'on appelle le report (nous n'aborderons ici que les reports en plan ; il existe depuis longtemps, notamment dans le domaine de l'architecture, tout une gamme de procédés visant à rendre compte des volumes, évolution qui s'est récemment accélérée avec les techniques informatiques dites 3D ; les résultats sont souvent frappants, donc attractifs, il y a lieu de les considérer avec *une infinie prudence* ; on ne saurait les aborder raisonnablement avant de maîtriser complètement les représentations planes). Le croquis et la feuille de relevés sont demeurés des instruments nécessaires jusque dans les années 80, et peuvent d'ailleurs parfaitement être encore employés pour des relevés simples (une série de précautions de bon sens doivent être prises, voir notamment Jean-Paul SAINT-AUBIN, *Le relevé et la représentation de l'architecture*, Paris, 1922, très documenté et clair).

Ce fut à l'occasion de la première guerre mondiale que se développa la pratique de la *photographie aérienne*, à partir des premiers avions et surtout de ballons ; l'objectif était naturellement l'observation militaire (on conserve des documents de Marc Bloch, officier de renseignement, commentant des clichés de ce genre). La première couverture photographique aérienne soignée et complète de l'Angleterre a été réalisée en 1940 par la Luftwaffe, les clichés ont été ensuite saisis par l'armée américaine et sont conservés à Washington. Ce sont d'ailleurs des avions américains qui ont réalisé la première couverture complète de la France en 1945 (clichés que l'on peut consulter et dont on peut acheter des reproductions à l'IGN). Dès 1901, l'allemand Pulfrich a imaginé le premier appareil permettant de restituer 3 dimensions à partir de 2 clichés du même objet pris à une certaine distance, selon le principe ensuite connu sous le nom de *stéréophotogrammétrie*. L'application en a été systématisée au domaine cartographique durant les 30 années qui ont suivi la seconde guerre mondiale. Le matériel et son usage étaient lourds, compliqués, onéreux.

La photographie aérienne continue d'être systématiquement utilisée, elle fournit des images d'une extrême précision, des filtres et des pellicules ad hoc permettent d'obtenir des informations qui débordent le spectre visuel (d'où tout l'attirail dit de la "photointerprétation"). Depuis les années 60, le point de prise de vue s'est sensiblement élevé avec les satellites. Les images produites par ces engins sont de plus en plus précises, leur cadence de plus en plus élevée. Toute une technologie ad hoc s'est développée sous le nom (pas idéalement choisi) de "*télédétection*". Le spectre des

émissions captables et captées s'est largement ouvert, la masse d'informations ainsi générées dépasse l'entendement.

La *télémétrie*, elle aussi développée surtout d'abord en vue d'applications militaires, s'est affinée et répandue. Par suite de progrès récents et rapides, on trouve dans le commerce des *appareils légers et faciles d'emploi* qui mesurent les distances avec une précision millimétrique sur plusieurs centaines de mètres (pour un prix abordable pour un simple particulier). Les théodolites électroniques mesurent les distances et les angles avec une précision supérieure à tous les outils antérieurs, et transmettent les résultats directement à un support électronique. Comme on l'a déjà indiqué, un réseau de satellites-balises permet un repérage centimétrique de n'importe quel point du globe terrestre.

Si les photographies aériennes sont encore recueillies sur des "plaques grand format", elles sont à présent numérisées, et les données des satellites parviennent bien entendu directement sous cette forme. Le papier et les méthodes d'archivage qui lui sont liées sont en passe d'être relégués à un rôle subsidiaire sinon anecdotique. Ce qui ne veut pas du tout dire que la situation soit simplifiée.

6.2.2 les grandes catégories de données

Il subsiste naturellement une considérable quantité de matériel topographique et cartographique traditionnel, qu'il importe de conserver soigneusement car il s'agit de documents historiques souvent irremplaçables, témoignant des évolutions au moins depuis le 18e siècle ; une comparaison minutieuse de la "carte de Cassini" avec la situation actuelle est toujours très riche d'enseignements, et devrait être un passage obligé de toute recherche d'histoire locale. Il est plus que probable que l'on numérise dans les années à venir la plupart des "cartes anciennes" (le processus est largement entamé), mais on ne saurait se contenter de produire des séries de "fichiers-images", il va falloir les intégrer dans des bases de données spécifiques (voir ci-après), et ce ne sera pas forcément facile.

On tend à distinguer à présent trois types principaux de données spatialisées :
a) les "images", ce que l'on appelle en termes techniques le **"mode raster"**. En fonction de la précision des données contenues dans l'"image", celle-ci est considérée comme une sorte de grille orthonormée, et à chaque point de la grille correspond une valeur numérisée.
b) les objets en **"mode vecteur"**. Il s'agit d'objets n'incluant que des indications purement géométriques, le plus souvent sur un plan : points, lignes (points reliés), polygones (ensemble de points fermé délimitant une surface), graphes (points reliés de manière plus ou moins complexe).
c) les objets dits **"site data"**, comportant des informations de n'importe quelle nature, situées dans l'espace, en général en des points définis, mais il peut aussi s'agir de surfaces, voire de lignes (e.g. des flux).

Chacun de ces types correspond dans la réalité actuelle à une multiplicité de "formats" (modes d'enregistrement) et de repères géodésiques....

6.2.3 les systèmes d'information géographique (SIG, alias GIS)

Il est apparu que l'informatique devait offrir un moyen approprié de "gérer" simultanément des données de types très divers, repérés originellement avec des référentiels géodésiques eux-mêmes fort variés. De manière à ce que tout type d'information concernant un point donné soit mobilisable aussi commodément que possible, sans oubli ni confusion. On a évoqué ci-dessus, très superficiellement, les problèmes de conversion entre référentiels différents, et la variété des données en entrée. Il faut donc procéder à ce que l'on appelle un *"géocodage" parfaitement homogène*. On conçoit donc assez aisément que les logiciels capables de venir à bout de ces difficultés soient inévitablement des outils très spécifiques. Il n'est sans doute pas indispensable de s'étendre

longuement sur le fait que tous les métiers de la conservation du patrimoine sont très directement concernés et le seront encore bien davantage dans les prochaines années. *Un conservateur du patrimoine devra obligatoirement savoir utiliser personnellement un ou plusieurs SIG*.

Dans les années 70 sont apparues de "grandes usines", complexes, extrêmement onéreuses, ne fonctionnant que sur ce que l'on appelait à l'époque les "grosses machines". La situation a fortement évolué, sans cependant qu'aient disparues des difficultés qui risquent de ne pas disparaître de sitôt. Dès les années 80, plusieurs SIG ont été conçus pour les microordinateurs, et naturellement ces SIG ont aujourd'hui des possibilités proportionnelles à la puissance incommensurablement accrue des matériels actuels. Toutefois : 1. les logiciels commerciaux disponibles demeurent fort onéreux (ils ne sont pas destinés au grand public !) et 2. d'un maniement difficile, en dépit de leurs commandes soi disant "intuitives".

Le seul SIG libre et open source de grande envergure est **GRASS** (*Geographical Resources Analysis Support System*) qui a été entrepris et développé d'abord par l'armée américaine au milieu des années 80, puis diffusé dans le domaine public et "porté" à l'heure actuelle par des collègues des universités de Trento et Hannover. Ce logiciel est en plein essor et paraît à l'heure actuelle le meilleur choix si l'on raisonne à moyen terme (a fortiori à long terme) car le principe de la GPL (General Public License) permet à la fois de suivre sans frais les développements (à l'inverse de ce qui se passe pour les logiciels commerciaux), de savoir exactement comment fonctionnent les divers programmes et, si l'on en ressent le besoin et que l'on ait un peu de patience, d'adapter telle ou telle fonction à un besoin particulier. (Une liaison directe peut être établie entre GRASS et R aux fins d'analyse statistique).

De toute manière, les problèmes traités par les SIG resteront forcément relativement compliqués, raison pour laquelle il est déraisonnable d'imaginer que l'on puisse les confier à des "petites mains" : le conservateur et le chercheur devront avoir une idée suffisante des diverses caractéristiques des données, et plus encore de leur signification, pour pouvoir les manipuler de manière utile. Cela ne signifie pas que le conservateur ou le chercheur devront tout faire eux-mêmes, mais qu'ils devront *maîtriser parfaitement l'outil et son fonctionnement* pour pouvoir, selon les possibilités, confier telle tâche bien définie à tel collaborateur.

6.3. REMARQUES SUR QUELQUES DONNÉES DISPONIBLES

6.3.1 données anciennes

Tous les documents cartographiques "traditionnels" posent des problèmes particuliers de conservation, du simple fait de leur format. Raison pour laquelle ils ont souvent été retirés des liasses où ils se trouvaient pour être conservés à part. Leur communication ne va pas non plus sans difficultés.

On ne doit pas perdre de vue que les données géographiques ne sont *pas seulement des plans ou des cartes*. D'abord, il est du plus haut intérêt, lorsque c'est possible, de conserver tous les documents de base (cahiers, relevés, etc.) qui comportent souvent des renseignements qui n'ont pas été reportés, ou qui permettent au contraire de préciser dans quelle mesure la carte est une interpolation à partir de renseignements épars. C'est particulièrement vrai pour les plans (notamment de bâtiments), qui, jusqu'à une date très récente (sinon même aujourd'hui encore) résultent de relevés très rapides et insuffisants. On peut citer ici aussi, par exemple, les dossiers résultant des opérations de délimitations des communes après 1790, qui occupent plusieurs mètres linéaires dans la plupart des départements : on y trouve quelques croquis, des plans rarement, mais surtout une masse impressionnante de paperasses relatant les controverses, les expertises, les arbitrages. Tous ces textes sont très directement liés à la représentation de l'espace et à son évaluation à la fin du 18e

et dans la première moitié du 19e, une analyse géographique portant sur cette période ne peut pas les laisser de côté. C'est un lieu commun, mais incontournable, de rappeler que le "cadastre napoléonien" comporte trois documents indissociables : l'état des sections (définition des parcelles), la matrice (liste des parcelles de chaque propriétaire) et les plans. C'est une erreur aussi commune que grossière de prendre l'un et d'oublier les deux autres (combien de toponymistes ont cru que **tous** les microtoponymes figuraient sur les plans...)

6.3.2 qu'est-ce qu'une carte ?

On ne peut pas non plus faire l'économie d'une réflexion, même très rapide, sur la notion de carte. *Il y a encore bien plus de différence entre une photographie aérienne et une carte qu'entre un manuscrit et une édition*. Dans les deux cas, on ne passe du premier état au second qu'au travers d'un processus long, méticuleux et complexe d'analyse et d'interprétation, interprétation qui se traduit par le choix des éléments que l'on juge seuls pertinents. Dans le cas de la carte, le tri est beaucoup plus drastique, mais, d'un autre côté, le cartographe ajoute une grande quantité d'informations que l'on serait fort en peine d'extraire d'une photographie aérienne, ne serait-ce que les toponymes. L'histoire de la cartographie n'est pas seulement une histoire de techniques, c'est aussi (peut-être surtout) l'histoire de ces règles d'interprétation et de traduction graphique (histoire très finement analysée dans les travaux du Père de Dainville). C'est pourquoi il est important de bien saisir que l'irruption des SIG et de l'électronique, si elle modifie les supports, ne fait en aucune manière disparaître les cartes, qui demeurent une traduction signifiante d'une masse de données informes non maîtrisable.

Malheureusement, pour les cartes comme pour les autres documents imprimés, des considérations juridiques retreignent les possibilités d'utilisation des documents récents, même lorsqu'ils sont le résultat de l'activité de services publics (il est amusant de noter que l'on est passé des restrictions liées au "secret militaire" aux restrictions liées aux intérêts mercantiles de la "puissance publique"). Il reste que, grosso modo, tous les documents imprimés antérieurs aux années 30 sont "tombés dans le domaine public" et que toute restriction à leur reproduction est difficile à justifier. Le scanner dans bien des cas, et l'appareil photo numérique dans presque tous, constituent des outils privilégiés de passage au numérique.

Quel que soit l'outil, quelques précautions élémentaires sont strictement indispensables. C'est ici le lieu de rappeler qu'**une carte stricto sensu** se définit par la présence de deux éléments :
a) une **échelle très précise**, dessinée - autant que possible **en longueur et en largeur**. On doit se défier de toute indication purement numérique car les supports sont instables. *Le carton et le papier se rétrécissent ou s'allongent*, sans que l'on puisse prévoir ni comment ni dans quel sens. Il arrive malheureusement souvent, sur des documents anciens, que l'échelle ne soit portée que dans un sens ; il faut alors opérer toutes les vérifications possibles pour déterminer l'échelle perpendiculaire, qui est le plus souvent différente de la première (le seul support à peu près stable est le verre...). Il faut savoir également que tous les moyens de reproduction, à peu près sans exception, déforment, de manière irrégulière et imprévisible. Avec une photocopieuse, on peut faire une série de tests, surtout si l'on emploie des taux d'aggrandissement ou de réduction. De même avec un scanner ou un appareil photo (papier ou numérique). Le plus simple est de prendre une feuille de papier et d'y dessiner un carré exact, par exemple de 20cm, et de mesurer ensuite les reproductions. On est toujours surpris. (il est généralement possible de rectifier une image numérique). Au total, le seul procédé solide consiste à établir *un cadre soigneusement gradué tout autour de la carte ou du plan*. A partir de là, toutes les déformations seront repérables et pourront être mesurées et compensées.
b) un **système de repérage précis par rapport à un référentiel géodésique**. Jadis, on recommandait de dessiner au moins un parallèle et un méridien. Si l'échelle est indiquée avec une grande précision, un point bien géoréférencé et une orientation sérieusement définie peuvent à la

rigueur suffire. *Quatre points clairement géoréférencés* aux quatre extrémité du graphique devraient être une règle.

Il est hors de question d'appeler "carte" un objet graphique qui n'inclut pas convenablement ces deux éléments. Ceci posé, il reste que l'on trouve de *bonnes cartes*, lisibles voire agréables, et de *mauvaises cartes*, embrouillées et pâteuses, d'où l'on ne tire presque rien. Cet écart résulte du respect, ou du non-respect, de quelques règles simples, et de portée très générale, qui énoncent les caractères d'un bon et d'un mauvais graphique. Nous y reviendrons (chapitre 8). Disons seulement, schématiquement, que le métier de cartographe, jusque dans les années 80, était un métier de dessinateur (encre de chine, tire-ligne, tracé à main levée, etc.) et qu'aujourd'hui tout est réalisé avec des logiciels ad hoc, logiciels de dessin plus ou moins adaptés à la spécificité des données cartographiques. Il ne s'agit plus d'habileté manuelle mais de compétence informatique. Pour autant, les règles distinguant une bonne carte d'une mauvaise n'ont pas varié.

6.3.3 supports actuels

Les objets disponibles sont en mode raster ou en mode vecteur. On trouve de plus en plus de "fonds de carte" dans divers formats sur internet. Avec de la patience et un peu de chance, on trouve presque n'importe quoi. Il est probable que ce mouvement va se développer. D'un autre côté, divers organismes, publics ou privés, vendent des *CD cartographiques*, souvent munis de logiciels de lecture (voire de recherche) plus ou moins performants. On trouve dans les supermarchés de la plupart des pays européens des CD qui combinent le réseau routier avec des indications touristiques (en particulier hôtels et restaurants) ; leur prix est modique, et l'on peut en extraire des fonds de carte utilisables à d'autres fins. En France, l'IGN vend (excessivement cher) des CD en mode vecteur utilisables avec les principaux SIG. Les cartes au 1/25000 sont disponibles chez un éditeur privé, par demi-département, à un tarif raisonnable. Des organismes spécialisés commercialisent des photos satellitaires.

Plusieurs services de conservation ont entrepris de numériser des séries de cartes anciennes, des résultats devraient apparaître dans les toutes prochaines années.

Chapitre 7

DONNÉES SPATIALES : ANALYSE

Après un rapide survol du géocodage et de la cartographie, il importe d'examiner les principales méthodes susceptibles d'extraire des informations utiles de la répartition spatiale des phéonomènes historiques. En France même, en dépit de la tradition qui paraît lier l'histoire et la géographie, l'usage de telles méthodes par les historiens reste excessivement limitée. Les efforts liés notamment au Laboratoire de cartographie historique de Jacques Bertin à la VIe Section de l'EPHE ont fait long feu, sans doute en partie en raison du refus des membres de ce laboratoire d'accepter l'idée que l'on ne peut pas tout faire à l'aide de la "graphique" seule. Des revues prestigieuses aussi différentes que *Deutsches Archiv* ou *Past and Present* ne publient jamais la moindre carte. Il faut reconnaître au surplus que, parmi les géographes eux-mêmes, les tenants de "l'analyse géographique", qui entreprennent une analyse fondée méthodiquement sur les statistiques spatiales, se heurtent à de hargneuses résistances. Il reste que des données historiques sérielles munies de coordonnées spatiales sont extrêmement nombreuses, pour toutes les époques. D'un autre côté, les statistiques spatiales sont en plein essor. Les chartistes doivent saisir ce problème à bras le corps, les perspectives sont particulièrement prometteuses.

SOMMAIRE

7.1. LES PRINCIPAUX TYPES DE DONNÉES SPATIALES

7.1.1 les trois catégories de base et leurs variantes.

On considérera ici essentiellement les *données géocodées*, c'est-à-dire repérées dans un espace à deux dimensions positionné à la surface de la terre. En pratique, les mêmes méthodes s'appliquent à toutes les données dans un espace à deux dimensions, définies par des coordonnées orthonormées (plan de fouille par exemple). Pour diverses raisons exposées précédemment, on ramène les coordonnées géographiques sphériques à des *coordonnées planes* (par l'un ou l'autre des systèmes de projection), sans quoi les calculs (quoique possibles) seraient considérablement et inutilement compliqués.

Grosso modo, on considère des objets de dimension 0 (points), 1 (lignes) et 2 (polygones). Jusqu'ici, l'intérêt pour des objets de dimension fractionnaire (les fractals de B. Mandelbrot) est demeuré confiné à des cercles étroits (nous y reviendrons brièvement plus bas).

a) **points**. La notion paraît simple et limpide. Dans certains cas toutefois, on se heurte à des *problèmes de précision et d'échelle* (les deux étant le plus souvent liés). Si l'on travaille disons dans un cadre national ou départemental, le "point" pourra être un cercle d'un kilomètre de diamètre. Il arrive que la nature même de l'objet ne permet pas une précision meilleure que l'hectomètre, par exemple le microtoponyme relevé sur un cadastre. On a tout intérêt à enregistrer les données avec toute la précision possible, on peut toujour simplifier ensuite, l'inverse n'est pas vrai. Essayez de trouver sur le terrain un bâtiment localisé à un kilomètre près...(notamment en zone urbaine). Deux variantes principales : le *point simple*, uniquement caractérisé par ses deux coordonnées (et un numéro ou code) et le *point valué*, c'est-à-dire correspondant à une valeur quelconque, qui peut elle-même renvoyer à deux types : la valeur ponctuelle d'un phénomène continu (e.g. température), ou la taille (effectif) d'un phénomène ponctuel (e.g. nombre de pièces dans un trésor monétaire).

b) **lignes**. La notion se complique déjà dangereusement. Concrètement, les statistiques spatiales (i.e. les logiciels qui les font) ne connaissent que les points ; une ligne est donc *une suite de n points, définissant (n-1) segments*. Là encore, on devra régler la question de la précision, en particulier le nombre de points définissant une ligne quelconque. La ligne peut être orientée (e.g. trajet) ou non (e.g. limite). Elle peut être valuée de multiples manières. On peut considérer *une largeur* (largeur d'une route, d'un mur), ou *l'intensité d'un flux* (cubage, tonnage, nombre de personnes, etc), mais dans ce cas, on est souvent amené à distinguer deux valeurs (le débit d'un cours d'eau est univoque, mais le trafic d'une route est à double sens). Il est rare que le débit d'une rivière ou le trafic d'une route soient constants sur toute leur longueur. Un ensemble de lignes articulées les unes par rapport aux autres constitue un "graphe" au sens mathématique du terme. Ce que l'on peut appeler plus ordinairement un réseau. La "théorie des graphes" forme une branche entière des mathématiques.

c) **polygones**. Paradoxalement, la question est relativement plus simple. On considère une suite de points fermée, et l'on s'intéresse à l'espace ainsi délimité. Les difficultés naissent principalement des discontinuités : soit le polygone est constitué de plusieurs sous-ensembles disjoints (type archipel), soit il est "percé" d'un ou plusieurs "trous" (zone intérieure au polygone qui néanmoins ne lui appartient pas, type lac ou enclave). Les difficultés ici sont plutôt d'ordre abstrait : la surface considérée est-elle vraiment homogène ? les limites sont-elles effectives ou arbitraires par rapport au phénomène étudié (utilisation de l'hexagone pour cartographier des phénomènes du 11e ou du 13e siècle...). La plupart des SIG offrent des moyens plus ou moins simples de subdiviser les polygones, de les regrouper, de les hiérarchiser (problèmes d'emboîtements).

d) **autres types**. On n'examinera pas ici d'autres types plus complexes, dont il importe toutefois de bien connaître l'existence (on peut y avoir recours pour traiter des questions qui, en apparence, n'ont aucun rapport avec l'idée première que l'on s'en fait). Par exemple la *pente* : inclinaison et orientation, repérées en général en chaque point d'une grille plus ou moins fine ; c'est un type d'objet

que l'on peut généraliser au travers de la notion de gradient. Un peu plus complexes, les données de type "*rose des vents*" : en chaque point d'une grille, une intensité dans plusieurs directions.

7.1.2 méthodes de codage et d'enregistrement

Le principe général est élémentaire : des points codés, repérés dans un système orthonormé. La plupart des logiciels parviennent plus ou moins aisément à récupérer des données enregistrées dans une feuille de tableur (eventuellement dans un SGBD ou un SIG). Mais *chaque logiciel, sinon chaque programme, réenregistre les données dans un format propre*. Bien que les points et les polygones constituent des entités assez simples, il existe une multitude de manière de les enregistrer. Dans R même, les diverses bibliothèques spécialisées dans les traitements de données spatiales (elles sont de plus en plus nombreuses et performantes) utilisent chacune un mode d'enregistrement différent, ce qui oblige souvent à écrire de petites fonctions de recodage si l'on veut utiliser des fonctions appartenant à plusieurs bibliothèques pour traiter un même jeu de données. Ce n'est pas bien compliqué, mais c'est une perte de temps. Il vaut mieux le savoir et prévoir le temps nécessaire.

En gros, on se contente le plus souvent de trois colonnes La première avec le code (du point ou du polygone), une colonnne des x et une colonne des y. Chaque polygone étant simplement identifié par tous les points successifs codés identiquement. Cela tient dans un fichier assez peu volumineux. Il faut cependant bien prendre garde que l'information que l'on traite n'est pas constituée par une succession de points, mais par les relations de tous les points, ou de tous les polygones, entre eux. A la différence des données bivariées ordinaires, qui peuvent se présenter dans n'importe quel ordre, ce qui forme la spécificité des données spatiales est que l'on considère la relation géométrique de chaque point (ou de chaque polygone) avec tous les autres : l'information spatiale proprement dite est en réalité logée dans la matrice des distances de tous les points pris par paires. De telle sorte que les traitements statistiques des données spatiales en tant que telles impliquent le plus souvent de traiter non pas n points, mais une matrice de (n x n) distances, c'est-à-dire des objets dont la taille s'accroît très rapidement ; si l'on considère 200 points par exemple, ce qui est assez peu, la matrice comporte 40 000 cases. Ce qui explique que les statistiques spatiales ne soient apparues et ne se soient pas développées avant le moment où des ordinateurs suffisamment puissants ont été couramment disponibles.

7.2. PRINCIPAUX TRAITEMENTS DES DONNÉES PONCTUELLES

7.2.1 le nuage de points : forme et position

On retrouve dans ce cadre bonne partie de ce qui a été dit à propos des graphiques de données bivariées. On devra naturellement prendre garde ici à *utiliser les mêmes unités pour les abcisses et les ordonnées*. Pour des points non valués, la première question est celle de la forme du nuage. On ne saurait trop recommander le calcul des densités sur une grille (KDE, kernel density estimator), suivi par le tracé de "**courbes d'isodensité**" (fonction `contour()`). Le choix crucial d'une largeur de bande (bandwidth) n'a pas de solution optimale définie univoquement, il faut se livrer à une série d'essais successifs.

Lorsque l'on a à faire à un nuage qui semble à peu près homogène, on peut définir *quelques paramètres élémentaires*, qui peuvent être utiles en particulier si l'on souhaite comparer rapidement quelques nuages (ou le même nuage à divers moments) : une position centrale et un indice de dispersion. On se trouve dans une situation assez voisine de celle des distributions univariées. Le **point central** pourra être le *point moyen* (appelé "barycentre"), dont les coordonnées sont la

moyenne des X et la moyenne des Y, ou le *point médian* (dont on a une bonne approximation en déterminant le X médian et le Y médian). Le point médian est le point qui minimise la somme des distances, le point moyen est celui qui minimise la somme des carrés des distances. Le point médian est plus robuste, et si l'on ne doit en retenir qu'un, c'est celui-ci. La comparaison des deux est instructive : *s'il existe un écart sensible*, c'est qu'un certain nombre de points extrêmes (aux bords du nuage) sont regroupés dans une certaine direction.

Il existe plusieurs façons de calculer un **indice de dispersion**. La plus robuste et la plus facile à interpréter est l'*écart médian au point médian* : c'est le rayon du cercle "central" qui regroupe la moitié des points. Bien entendu, dans l'espace géographique considéré, un point peut avoir une situation centrale pour ainsi dire structurelle (type chef-lieu) ; dans ce cas, il peut être utile de calculer l'écart médian à ce point.

7.2.2 traitements géométriques simples : triangulation et tessellation

Etant donné un semis de points, on sait depuis le 19e siècle relier entre eux les points les plus proches de manière à ce que toute la surface concernée soit couverte de triangles non sécants ; c'est la "**triangulation de Delaunay**" (Charles-Eugène Delaunay, 1816-1872) [mathématiquement, ces triangles sont définis par le fait que le cercle circonscrit à chacun d'eux ne renferme aucun des autres points du semis]. Si l'on trace la médiatrice de tous les segments dessinés par cette procédure, chaque point se trouve entouré d'un polygone plus ou moins régulier selon la disposition des points. C'est la "**tessellation" de Dirichlet** (P.-G. Lejeune-Dirichlet, 1805-1859) (ou de Voronoï). On peut en général superposer les deux graphiques, mais si la triangulation de Delaunay est indispensable d'un point de vue analytique (et sert d'ailleurs dans diverses autres occasions), *la tessellation donne le plus souvent un graphique beaucoup plus suggestif*. Les "tesselles" sont beaucoup plus visibles et parlantes que les points eux-mêmes, et donnent une idée à la fois de la densité et de la répartition plus ou moins régulière des points.

On peut en général calculer *la surface des teselles*, et en tirer divers indices statistiques. Prendre garde toutefois aux "effets de bord" : si les triangles de Delaunay sont définis sans ambiguïté, les tesselles situées à la limite du nuage peuvent être "bouclées" de diverses manières (selon la manière dont les programmes sont écrits, les tesselles des bords sont ouvertes, ou fermées par un segment ou un arc de cercle). Ces traitements sont impossibles pour les points superposés. Soit on n'en considère qu'un seul, soit on utilise une fonction qui les déplace très légèrement (`jitter()`).

7.2.3 l'analyse des processus ponctuels ("point process analysis")

Après une considération de la forme globale du nuage et une observation graphique, il est indiqué de procéder à un examen numérique de *la disposition des points les uns par rapport aux autres*. Soit les points sont répartis aléatoirement (la distribution des X et celle des Y sont complètement **aléatoires** : mathématiquement, les deux distributions peuvent être assimilées avec une probabilité suffisante à des distributions "de Poisson") ; soit au contraire les points sont répartis de manière relativement **régulière**, ou inversement **concentrés** en petits paquets : écarts dans un sens ou dans l'autre par rapport à une distribution aléatoire. Une manière assez grossière mais efficace de tester cette situation consiste à analyser la "*distance au plus proche voisin*" (nearest neighbour). Le programme calcule les distances entre tous les points, et retient pour chacun la distance du plus proche. On obtient ainsi une distribution univariée ; la loi d'une telle distribution dans le cas d'une répartion entièrement aléatoire est connue depuis longtemps et relativement simple ; elle ne dépend que de la densité générale des points sur la surface considérée. On peut donc aisément constater, par simple ajustement graphique, si la répartition observée est aléatoire,

régulière ou concentrée.

Une analyse plus complète consiste à examiner la *distribution des k plus proches voisins*. Là encore, on peut comparer distribution observée et distribution théorique.

7.2.4 les points valués

S'il s'agit de dénombrements (nombre d'individus groupés en tel ou tel point), on peut les analyser comme autant de points superposés : l'analyse des densités par la méthode de l'estimateur (KDE) est plus utile que jamais, puisque le graphique ponctuel ne peut donner qu'une image faussée, dans laquelle tous les points ont la même valeur.

S'il s'agit de mesures d'un phénomène continu, la situation est complètement différente, il faut procéder à **une interpolation**, c'est à dire évaluer le phénomène de manière continue à partir des points où il est connu (ou tout au moins entre eux). Depuis les années 60, de très nombreuses recherches ont été menées dans ce domaine, illustrées notamment par les travaux de Georges Matheron, inventeur d'une méthode qu'il baptisa "krigeage". Les traités de statistique spatiale consacrent de longs développements à l'analyse des algorithmes propres à chacune des trois ou quatre grandes familles de méthodes, et à leur comparaison. Dans la plupart des exemples proposés, *les résultats diffèrent peu d'une méthode à l'autre*, et la différence tient surtout à la plus ou moins grande facilité pour introduire des paramètres permettant de "lisser" les résultats. Au demeurant, dans le domaine historique, des phénomènes continus connus ponctuellement ne sont pas fréquents, et ne paraissent pas requérir une interpolation d'une précision exceptionnelle. On choisira la méthode pour laquelle un programme est disponible et d'un usage pas trop complexe.

7.3. PRINCIPAUX TRAITEMENTS DES POLYGONES

7.3.1 analyses préliminaires

Chaque polygone étant défini par la suite de ses sommets, il existe au moins un algoritme simple permettant de calculer avec exactitude sa **surface** et les **coordonnées de son centre de gravité**. (curieusement, divers programmes de "cartographie automatique" utilisent des algorithmes faux : une fois de plus, on ne peut se fier qu'aux programmes dont le listing est public...). Une étude de la distribution univariée des surfaces est un préalable indispensable. La position des centres de gravité est le plus souvent utilisée pour définir l'emplacement des étiquettes ; raison pour laquelle il vaut mieux, si on le peut, vérifier que le centre de gravité est bien à l'intérieur du polygone, ce qui peut très bien ne pas être le cas, que le polygone ait une forme en arc de cercle (par exemple) ou soit constitué de plusieurs éléments séparés.

7.3.2 la discrétisation : principe

On a déjà évoqué le "**découpage en classes d'une variable continue**", d'abord à propos des distributions univariées (la procédure s'impose par exemple pour pouvoir procéder à un test du chi-deux), puis à propos des distributions multivariées, pour lesquelles un codage (disjonctif ou non) est très fréquemment d'une grande utilité. Le découpage d'une distribution continue revient à remplacer les valeurs réelles (données) par un nombre restreint de valeurs calculées (ordinairement centrales, pour le moins "représentatives" de la classe à laquelle est affecté chaque individu) ; on passe donc bien d'une distribution continue à une distribution discrète, d'où ce terme de "discrétisation". Mais pourquoi "discrétiser" et surtout comment ?

Tous les langages un peu évolués (R est dans ce cas) permettent de définir les couleurs dans un nuancier copieux (pour le moins 3 couleurs de base à 256 positions chacune, soit au total plus de

seize millions de couleurs). Si les trois valeurs sont identiques, on va du blanc au gris : on peut donc définir par exemple 256 gris (et ainsi de suite) ; *on pourrait donc imaginer de colorier chaque polygone avec une nuance calculée sur une échelle exactement proportionnelle aux données d'origine* : on aurait à très peu près une représentation continue pour un phénomène lui aussi censé continu. On peut procéder de cette manière. Mais l'expérience montre surabondamment que *de telles cartes sont peu lisibles et difficiles à interpréter*. Or l'objet de la procédure est de tenter de déceler une organisation spatiale du phénomène considéré : on cherche à identifier des disparités, des regroupements, des gradients ou des barrières. Un bon graphique est un graphique qui permet de repérer rapidement des formes et des articulations (nous y reviendrons). Une échelle continue ne facilite pas ce résultat, au contraire. Ce que permettent au contraire des teintes (ou des trames) suffisamment distinctes (s'il s'agit de trames, on suppose qu'elles sont conçues de manière à rendre immédiate la gradation : il n'arrive que trop souvent que ce ne soit pas le cas).

7.3.3 la discrétisation : difficultés

Depuis les années 60 au moins, les géographes et cartographes sont sont avisés que le même jeu de données pouvait, sans aucune tricherie, donner des cartes complètement différentes, induisant des interprétations variées sinon opposées : il suffit de changer le mode de discrétisation. Or il est peu près prouvé qu'il n'existe pas de solution optimale unique à ce problème. L'idée très générale est que *le rendu est bon si l'on "voit quelque chose"* sur ladite carte. Mais alors, si deux discrétisations différentes permettent deux discours opposés, où va-t-on ?

Cette difficulté, réelle et sérieuse, présente l'avantage d'obliger à se poser explicitement la question des critères d'interprétation d'une carte de ce genre, des pièges à déjouer et des méthodes simples à utiliser systématiquement. **Le piège le plus courant tient à la prépondérance des surfaces relatives des polygones comme expression de l'importance des unités géographiques, quel que soit le phénomène considéré**. S'il s'agit de phénomènes naturels ou agricoles, cela peut convenir, sinon on court au n'importe quoi. L'exemple inusable est celui des cartes de géographie électorale, où, par le jeu des surfaces, la Lozère tient le même rôle que la région parisienne. Même quand on est dûment prévenu, il est presque impossible de "rectifier" convenablement cette énorme distorsion. Les journaux tentent souvent de contourner la difficulté en plaçant dans un coin de la page un "carton" de la région parisienne, mais ce n'est qu'un pis-aller. Au demeurant, la même difficulté se pose à toutes les échelles, par exemple au niveau départemental, où un canton de moins de 1000 électeurs couvre plus de surface que l'agglomération du chef-lieu, qui peut en compter 20 ou 50 fois plus.

Les solutions graphiques ne sont pas légion. L'**anamorphose**, qui consiste à dessiner une carte fictive où chaque unité a une surface proportionnelle à l'importance du phénomène considéré (ici, le corps électoral), est très difficile à mettre en œuvre pratiquement, et d'une lecture malaisée, car les déformations sont en général importantes et il devient compliqué d'identifier les unités (toute l'attention est attirée par cette difficulté, et le graphique perd tout intérêt). Une solution plus simple consiste à **placer au centre de chaque unité un symbole graphique** (un carré, ou encore mieux un cercle) **de surface proportionnelle au phénomène, et de ne colorier, ou tramer, que ce symbole**. Abstraitement, cela paraît satisfaisant. Mais concrètement c'est loin d'être idéal : les symboles sont par définition des objets discontinus, alors que l'on cherche à faire apparaître des ensembles, et surtout la vue est d'abord retenue par la taille des symboles, avant qu'elle parvienne à identifier les trames ou couleurs ; ce mode de représentation évite les distorsions, mais n'atteint que très péniblement son rôle de structuration.

Cela noté, et souligné, reste la discrétisation en elle-même. Les principales manières de découper les classes ont été évoquées à propos des distributions univariées : intervalles égaux (souvent la méthode que les programmes retiennent par défaut), effectifs égaux (quantiles),

découpage probabiliste en fonction de la moyenne et de l'écart-type (il en existe encore plusieurs autres), et choix du nombre de classes, en principe entre 3 et 7. Bien entendu, *un examen approfondi de la distribution univariée de la variable considérée est un préalable.* C'est une faute grossière de l'éviter. C'est le seul moyen de déterminer rapidement la forme de la courbe et de vérifier l'éventuelle présence de plusieurs ensembles (courbe plurimodale), ainsi que les valeurs extrêmes. Il est insensé de découper des classes avant d'avoir procédé à cet examen.

C'est à ce point qu'il peut être efficace, sinon indispensable, de faire intervenir une **pondération**, c'est à dire de regrouper les individus en les affectant de "poids" différents, proportionnels à une grandeur que l'on juge pertinente. On pourra ainsi obtenir des classes de poids sinon tout à fait égaux, au moins équivalents, alors que les effectifs (nombres d'individus) seront très différents. Les géographes, surtout sensibles aux surfaces, ont inventé la notion de "courbe clinographique", qui désigne simplement une pondération des unités par la surface qu'elles occupent. Cela peut être efficace dans le cas de polygones de surfaces très différentes, en admettant que la surface soit la base pertinente. Mais on peut utiliser toute autre base, notamment la population, ou toute autre grandeur par rapport à laquelle on calcule des densités ou des proportions. Cette procédure peut permettre de corriger, partiellement, la distorsion due aux surfaces.

L'expérience tend à montrer que, tant que l'on ne dispose que d'outils rudimentaires et assez peu interactifs, la répartition spatiale d'un jeu de données appliquées à des polygones ne peut pas être saisie autrement que par *un grand nombre d'essais, au minimum une douzaine* si le phénomène est assez simple et bien structuré. En augmentant le nombre de classes (quelle que soit la manière dont elles sont définies) ont isole les valeurs extrêmes : sont-elles ou non regroupées, où et comment ? En diminuant au contraire ce nombre, on voit se dessiner, ou non, des limites plus ou moins stables ; toutes ces manœuvres doivent être exécutées deux fois, avec une gamme monochrome et avec une gamme à deux couleurs opposées (type rouge et bleu) : le même découpage produit deux impressions visuelles souvent très différentes. Il arrive souvent que l'on soit amené à repasser plusieurs fois la même série avant de remarquer des éléments de structure.

En résumé :

a) réfléchir intensivement à la nature de la série numérique dont on dispose, à ses rapports avec le phénomène dont elle est censée être un indicateur, et au rapport de ce phénomène avec l'espace ;
b) procéder à deux analyses détaillées des distributions univariées selon la procédure ordinaire (les surfaces des polygones d'une part, le phénomène étudié de l'autre), en tirer éventuellement des seuils manifestes ;
c) produire à l'écran, et sur le papier dès que l'on aperçoit quelque chose, un grand nombre de cartes en faisant varier le nombre de classe et le rendu chromatique.
d) le cas échéant, affiner un résultat en modifiant à la main les limites de classe.
e) dans tous les cas où l'on a repéré une distorsion sensible, il faut autant que possible l'indiquer dans le commentaire joint à la carte, et proposer deux cartes ou davantage, par exemple une carte "classique" et une carte avec des symboles à taille proportionnelle ; la combinaison des deux lectures peut faciliter la prise en compte de la (ou des) distorsion.
f) exposer les résultats en combinant l'analyse numérique (statistique univariée) et l'analyse cartographique (graphique et visuelle). Ne jamais perdre de vue que ce genre de carte n'est pas destinée à fournir une illustration, mais un outil d'analyse et de réflexion (ce qui n'est pas du tout la même chose). Dans tous les cas où ce n'est pas impossible, fournir les données (chiffrées) à côté de la ou des cartes.

7.3.4 un outil de base : la matrice de contiguïté

Lorsqu'une zone est divisée en polygones de manière classique, il n'y a ni vides ni recouvrements. Sauf sur les bords de la zone, chaque segment appartient à deux polygones contigus.

Il n'est donc pas difficile d'*identifier les polygones contigus* en analysant les coordonnées qui les définissent. De cette analyse on tire une **matrice symétrique**, ayant autant de colonnes et de lignes qu'il y a de polygones ; une case comporte un 1 si la ligne et la colonne qui se croisent correspondent à deux polygones contigus, et un zéro dans tous les autres cas. Ce tableau, une fois constitué, permet toute une série de procédures sur les polygones contigus.

7.3.5 le lissage spatial

La manœuvre la plus simple est le lissage, qui est l'équivalent spatial (à de sérieuses nuances près) de la moyenne mobile sur des séries chronologiques. On peut se contenter de *remplacer la valeur observée dans un polygone par la moyenne de cette valeur et de toutes celles des polygones contigus*. Mais on peut aussi introduire *trois types de pondération*. Comme pour le découpage en classes, on peut affecter un "poids" à chaque polygone, qui peut être sa surface, mais aussi toute autre grandeur jugée significative. Une autre possibilité est de tenir compte de la distance. On pourra par exemple calculer le tableau des distances entre le centre de gravité de tous les polygones, et utiliser l'inverse de cette distance comme coefficient de pondération (le résultat, si les polygones ont des tailles très différentes, sera tendanciellement l'inverse du précédent : si l'on pondère par la surface, un grand polygone pèsera bien plus lourd ; si l'on pondère par la distance, un grand polygone, dont le centre de gravité sera plus éloigné, pèsera moins). Mais le plus intéressant consiste à augmenter "l'**ordre de contiguïté**", c'est-à-dire à tenir compte non seulement des voisins directs, mais aussi des voisins des voisins (ordre 2), de leurs voisins (ordre 3), etc., en pondérant d'une manière ou d'une autre. On peut obtenir un résultat voisin en itérant le lissage de premier ordre (c'est à peu près le même principe que pour les moyennes mobiles). En pratique, si les polygones sont nombreux, on a tout intérêt à procéder à une telle itération, et à examiner attentivement le résultat à chaque étape : le lissage est de plus en plus fort, et permet, selon la structure spatiale du phénomène, de mettre en relief, à un ordre non prévisible, une tendance générale (si elle existe).

Bien que la littérature n'y fasse pour ainsi dire jamais allusion, on peut aussi calculer des indices de dispersion, bien utiles pour visualiser l'intensité des écarts entre voisins. On peut calculer la moyenne des écarts entre la valeur d'un polygone et celles des polygones contigus, on peut aussi calculer l'écart-type (ou le coefficient de variation) de la population formée par un polygone et tous ses voisins. Chaque polygone se voit alors représenté selon *une échelle qui traduit l'hétérogénéité locale*. On pourra lisser les résultats, si besoin est, pour voir apparaître sur la carte des zones de plus ou moins grande hétérogénéité à l'échelle locale (c'est-à-dire à l'échelle des polygones examinés).

7.3.6 l'autocorrélation spatiale

Le passage de l'autocorrélation des séries chronologiques à l'autocorrélation spatiale est assez complexe. Pour deux raisons : les points successifs d'une série chronologique sont en général équidistants, ce n'est en général pas le cas des points considérés sur une carte (sauf dans le cas où l'on raisonne sur une grille hexagonale, ce qui n'est pas courant) ; d'autre part, l'ordre de succession chronologique est unique et immédiatement défini, tandis que, dans l'espace géographique, les directions sont multiples (indénombrables en fait) et surtout que cet espace est à peu près toujours anisotrope, c'est-à-dire que, en chaque point considéré, la signification des diverses directions varie plus ou moins fortement. Sauf à compliquer sensiblement les calculs (et même dans ce cas - type "lissage elliptique" - , il n'y a pas de solution vraiment satisfaisante), *on est amené à effectuer les calculs "comme si" l'espace géographique était isotrope*, approximation qui peut être acceptable, mais qui peut ne pas l'être : il faut consacrer à cette difficulté une solide réflexion.

La littérature (et les programmes) ont consacré **deux indices**, celui de Moran (1948) et celui de Geary (1954). Résumons. Dans le test de Geary, on calcule une variance globale en considérant les carrés des différences de valeurs entre tous les individus pris par paires, puis une variance locale,

en examinant uniquement les carrés des différences entre valeurs contiguës. Si la contiguïté ne joue aucun rôle, les deux valeurs seront voisines (rapport proche de 1), sinon il aura attraction ou répulsion. Dans le test de Moran, on calcule l'équivalent d'un coefficient de corrélation linéaire ordinaire (rapport entre la covariation des unités contiguës et variance globale) : on trouve donc zéro s'il n'y a pas de corrélation, des indices positifs si la corrélation est positive, négatifs dans le cas contraire. En général, les deux tests donnent des *résultats analogues*. Les statisticiens les apprécient, car on peut comparer les résultats à ceux que donnerait une distribution aléatoire, et donc calculer des probabilités (situation observée s'écartant plus ou moins d'une dispersion aléatoire). L'intérêt concret est qu'ils permettent des comparaisons. On peut en effet calculer un indice pour une contiguïté du premier ordre, du deuxième, etc., et *construire un graphique qui peut indiquer une variation significative (principe du corrélogramme), c'est-à-dire l'évolution de la "ressemblance" des valeurs en fonction de l'ordre de contiguïté*. Comme d'habitude, plutôt que disserter sur les avantages respectifs de ces deux tests, je conseille de les éxécuter tous les deux et de comparer les résultats.

Dès 1967, N. Mantel a eu l'idée simple de généraliser l'analyse en mettant directement en relation un tableau de distances topographiques (tableau général des distances concrètes entre toutes les paires) et un tableau des écarts entre toutes les paires s'agissant de la variable considérée. On peut bien entendu calculer le coefficient de corrélation linéaire (qui peut donner lieu à une interprétation probabiliste), mais le graphique, à mon avis, s'impose : il suffit, par exemple, de *mettre l'écart topographique en abscisse et l'écart de la variable en ordonnée ; on obtient un nuage bivarié auquel il convient d'appliquer les procédures ordinaires pertinentes pour les distributions bivariées* (en particulier les diverses transformations qui permettent de linéariser un nuage autrement informe, ainsi que le tracé de contours par la méthode du KDE).

Il reste que ces diverses méthodes sont toutes globales : les coefficients concernent une zone donnée prise comme un tout. Cet ensemble de procédures est irremplaçable si l'on veut comparer l'effet des contiguïtés et des distances dans plusieurs zones. On ne doit pas perdre de vue que *l'analyse cartographique des écarts entre chaque polygone (ou point) et ses voisins, complété autant que nécessaire par lissage, peut permettre d'observer une zonation significative des effets de contiguïté, que les coefficients globaux ne permettent en aucun cas d'atteindre.*

7.3.7 les "distances" multivariées

On a considéré jusqu'ici qu'à chaque polygone correspond une variable. Le calcul des "distances" s'effectue alors par division ou soustraction (ou tout autre calcul jugé pertinent). Mais on peut tout à fait envisager de considérer simultanément un ensemble de variables, et l'on peut utiliser l'une ou l'autre des diverses "distances" évoquées dans le chapitre sur les distributions multivariées. Les possibilités sont très nombreuses. Toutes les procédures d'analyse des "distances" (simples) entre polygones que l'on vient d'évoquer s'appliquent de la même manière aux distances multivariées.

7.4. COMPARER DES CARTES

7.4.1 Précautions élémentaires

Une représentation cartographique destinée à l'analyse et à la recherche de structures n'est pas une carte topographique ni une planche d'atlas. D'une certaine manière, c'est même **l'exact inverse**. Les cartes courantes visent à fournir le plus grand nombre possibles d'indications, dans les limites de la lisibilité. Il s'agit d'outils de repérage ponctuel, comparables à des inventaires. La carte de recherche vise tout au contraire à faire apparaître des formes globales. Il est donc tout à fait

contreproductif de tenter d'inscrire sur le même fond de carte plusieurs ensembles distincts ; à l'extrême rigueur deux, à condition que les formes soient simples et déjà bien repérées par ailleurs. On pourra tracer deux contours simples, à l'aide de deux traits fortement distincts, par exemple avec deux couleurs opposées, du type bleu-rouge. Toute superposition de semis de points ou de trames, si astucieusement conçue soit-elle, sera malaisément déchiffrable, voire complètement opaque. ***La règle d'or doit être : un phénomène - une carte***. Ce qui implique, si une structure est composée de plusieurs ensembles ou se présente selon plusieurs états successifs, de **démultiplier les cartes** autant que nécessaire, et jamais l'inverse.

7.4.2 les cartes de liaison

C'est une question de bon sens : si l'on recherche la forme spatiale du lien entre deux phénomènes, il faut d'abord tenter d'établir aussi nettement que possible la forme de l'un et de l'autre, puis imaginer un procédé graphique uniquement destiné à représenter le lien. Le plus souvent, cela passe par des calculs élémentaires : si l'on a à faire à des polygones, la question ne se pose pas ; s'il s'agit de semis de points, on fera en sorte de calculer des grilles de densité constituées de manière strictement identique, de manière à ce que chaque point de l'une corresponde à un point de l'autre. A partir de là, on dispose de deux ensembles numériques appliqués aux mêmes coordonnées, les deux solutions classiques consistent à retenir la différence ou le rapport. Avec un logiciel comme R, cela ne présente aucune difficulté, une instruction élémentaire suffit. On aura toujours intérêt à procéder à divers essais. Le résultat devra par définition comporter **trois cartes, l'état 1, l'état 2, et la carte de liaison**. Le tout conçu de telle manière que la forme de l'écart apparaisse nettement. Devant une telle difficulté, la patience est l'arme principale du chercheur. Car dans presque tous les cas, les valeurs par défaut choisies par le logiciel ne permettent pas d'obtenir des résultats comparables, il faut procéder à la main et par tâtonnements.

7.4.3 l'analyse de la diffusion

Il existe plusieurs grandes familles de processus de diffusion, chacune demandant des outils analytiques et graphiques appropriés. La *diffusion en nappe* est la plus simple (type tache d'huile), des coefficients assez élémentaires permettent d'en rendre compte ; une attention particulière doit toutefois être portée à la traduction graphique : l'expérience montre que des diffusions simples sont rendues incompréhensibles par des graphiques mal conçus. On devra essayer de caractériser la zone -limite, entre ce que l'on pourrait appeler l'infiltration progressive, et le front conquérant nettement marqué.

Un autre modèle est celui du *développement radial*, lorsque le phénomène progresse essentiellement le long de quelques axes à partir d'un centre, produisant ainsi une structure en étoile. L'analyse et le rendu graphiques ne sont pas simples.

Le *modèle hiérarchique* a donné lieu à de très nombreux travaux. A partir d'un point ou d'une zone, le phénomène "saute" en quelques sorte dans un certain nombre de centres secondaires, d'où, s'étant bien implanté, il saute dans des centres de troisième ordre, et ainsi de suite. On peut aboutir à une couverture à peu près uniforme ou à la stabilisation d'une structure bien hiérarchisée.

Ce type de structure renvoie à des **distributions statistiques de type clairement parétien**. D'une manière générale d'ailleurs, les phénomènes de diffusion, quelle qu'en soit la forme, renvoient plutôt à des modèles parétiens (fractals) qu'à des distributions liées à l'univers gaussien. Les recherches, très actives à cet égard dans les sciences de la nature, sont nettement moins avancées dans le domaine des sciences sociales ; on dispose seulement d'études de cas qui ne forment pas encore un tout bien organisé. Il y a beaucoup de recherches à poursuivre dans cette direction.

7.5. PERSPECTIVES

Comme on l'a signalé en commençant, les statistiques spatiales sont de développement récent, très récent : même les dernières synthèses sont tout à fait partielles, des procédures et des programmes nouveaux apparaissent continuellement. On terminera sur deux réflexions générales.

7.5.1 nécessité d'une réflexion abstraite sur les phénomènes considérés

Toute société humaine est située dans l'espace géographique, il est puéril d'imaginer des éléments de sens que l'on croirait si abstraits qu'ils puissent être analysés sans aucune référence aux lieux. La *summa* de saint Thomas est incompréhensible en dehors de la situation de Paris, des réseaux de couvents et de studia des ordres mendiants, de la géographie des diocèses, de la place de Rome, de l'origine des étudiants (liste non limitative). On dispose de répertoires énormes d'objets anciens, médiévaux notamment, répertoriés systématiquement, datés, et bien localisés géographiquement, les filigranes, les monnaies des trésors, les vitraux, etc... Jusqu'à présent, aucune analyse en termes de statistique spatiale n'a été entreprise. Qu'attend-on ?

Dans ce cas plus encore que dans d'autres, une réflexion approfondie sur les structures et leurs relations, les indicateurs, les biais, est une conditio sine qua non de tout progrès. Ce que l'on peut appeler *les "opérateurs spatiaux élémentaires" ne sont pas en nombre indéfini* : étendue, limite, axe, isotropie, homogénéité, densité, équilibre, symétrie, hiérarchie, flux, polarisation. Toute la question est de savoir, à propos de n'importe quel phénomène historique, quels sont les opérateurs les plus pertinent pour penser et analyser **la composante spatiale, toujours présente, et en général bien plus significative que ce que la tradition annonce**. Insistons seulement sur deux points.

La question de *l'échelle d'observation et de la taille du graphique* sont des questions décisives. La notion clé de l'homogénéité se présente de manière complètement différente si l'on raisonne avec des parcelles, des cantons ou des régions. Dans toute la mesure du possible (c'est-à-dire en fonction des données dont on dispose), il est essentiel d'examiner toutes les échelles : le même type de procédure donnera des résultats complètement différents aux diverses échelles, et c'est l'articulation de cette diversité qui constituera le résultat le plus significatif. Bien se souvenir également que la même carte (la même !), selon qu'on l'observe en 30cmx40cm ou en 9cmx12cm ne sera pas du tout lue de la même manière. Des informations bien visibles dans un cas disparaîtront complètement dans l'autre. La taille concrète du graphique joue un rôle trop négligé.

La notion de *réseau recouvre un ensemble très riche et complexe* ; c'est par essence un objet où interviennent simultanément de nombreuses relations, dont les rapports ne sont pas prévisibles. Et c'est bien pourquoi cet objet appartient à l'univers parétien, c'est-à-dire celui des hasards combinés et plus ou moins emboîtés. C'est probablement un des secteurs de la recherche où l'on peut s'attendre aux évolutions et aux progrès les plus importants. Il faut s'y jeter.

7.5.2 vers des méthodes plus souples et plus interactives

La tradition cartographique du dessin à l'encre de chine pèse encore très lourd. Dans ce cadre, la carte est un résultat, pas vraiment un outil. Les statisticiens ont commencé à élaborer de nouvelles procédures de manipulation et de calculs, dès lors que les limites des mémoires électroniques ont reculé au point de permettre sans délai des calculs sur des matrices gigantesques. Mais les mêmes statisticiens ne sont que trop portés à croire que la carte reste de l'ordre du dessin, fût-il désormais informatisé. En physique ou en biologie, la position des points sur un graphique n'a pas de sens propre, les formes doivent recevoir une traduction mathématique pour être définies. Dans les sciences sociales, il en va différemment. L'Irlande est une île, la Hongrie n'a pas de côte,

Anvers n'est pas Venise... **La fixité des données topographiques et l'inertie des structures sociales sont des données spécifiques dont on ne peut pas ne pas tenir compte**. La statistique spatiale des données historiques doit donc faire un *usage très intensif des cartes*, c'est le moyen privilégié d'intégrer ces spécificités. C'est pourquoi, dans les conditions techniques actuelles, il est un peu surprenant de constater le faible développement de méthodes et de programmes plus faciles d'emploi et surtout beaucoup plus interactifs que ceux dont on dispose couramment.

Prenons un exemple simple. A propos de polygones, j'ai évoqué la discrétisation, le lissage, l'autocorrélation. Points sur lesquels on trouve à la fois "de la littérature" et des programmes. Ces derniers sont presque tous conçus pour produire une carte ; étant donné un ensemble de polygones, des données et quelques calculs, on obtient une carte ; si le programme n'est pas trop rudimentaire, il prévoit une série de paramètres, et en modifiant ceux-ci on peut obtenir d'autres cartes. Mais je ne connais pas (il en existe peut-être) de programme qui permette de construire une carte progressivement, en insérant des éléments un par un, pour voir. Si l'on considère par exemple les segments qui relient les centres des polygones contigus (style "triangulation de Delaunay"), tenant compte de la similarité-dissimilarité entre polygones contigus, on peut aisément classer tous ces segments par ordre d'écart croissant, en partant de l'écart le plus faible. On pourrait parfaitement imaginer un programme qui permette de dessiner un par un, interactivement, ces segments, de manière à voir se constituer peu à peu des éléments de "toiles d'araignée" (avec possibilité le cas échéant, de revenir en arrière) ; symétriquement, on pourrait aussi, partant des écarts les plus forts, tracer en gras les limites entre les polygones contigus les plus différents, de manière à essayer de faire apparaître, progressivement, des limites plus ou moins nettes. Dans des cas de ce genre, comme dans beaucoup d'autres, l'interactivité est le moyen le plus économique de parvenir à un résultat graphiquement cohérent, dont on puisse alors rechercher la signification. Programmer ce genre de procédure n'est pas compliqué, *on demande des volontaires*.

Chapitre 8

L'ÉLABORATION DES GRAPHIQUES

La notion de "sémiologie graphique" a été proposée par Jacques Bertin (1918-2002) en 1967. John Wilder Tukey (1915-2000) publiait en 1977 *Exploratory Data Analysis*. Et pourtant, alors même que les outils informatiques capables de transformer en graphiques des données numériques ont connu un développement vertigineux, la situation actuelle demeure plutôt caractérisée par la profusion de graphiques illisibles et l'absence de graphiques dans beaucoup de cas où ils seraient de première utilité.

SOMMAIRE

8.1. LES PRINCIPES DE LA REPRÉSENTATION GRAPHIQUE

8.1.1 bref historique

Pendant longtemps, les graphiques ont été réservés à des usages professionnels, ou tout à fait spécifiques : *dessin industriel, plans d'architectes, cartographie*. Les ingénieurs traçaient des courbes sur du papier millimétré pour représenter des équations complexes, représentations qui permettaient une "solution graphique approchée" dans beaucoup de cas où les calculs directs étaient à peu près irréalisables (système dit des "*abaques*"). Toutes ces opérations requéraient des capacités de dessinateur qui donnaient lieu à des formations spéciales, débouchant sur *des professions elles aussi spécifiques*. Le simple fait de savoir tracer des lettres sur des plans était tout un art.

Dans le même temps, les analyses statistiques recouraient très rarement aux graphiques, parfois réalisés (à la main bien entendu), mais pour ainsi dire jamais publiés. De toute manière, comme on l'a rappelé à plusieurs reprises, les calculs étaient eux aussi effectués à la main, et l'on cherchait donc par tous les moyens à les simplifier. Les "*papiers fonctionnels*" (semi-logarithmique et log-log principalement, mais aussi gausso-arithmétique ou gaussso-logarithmique) permettaient d'éviter des calculs, mais leur usage était limité, la plupart des historiens ignorant tout simplement leur existence.

A présent, la situation paraît presque inversée. Les machines effectuent tous les calculs imaginables, des logiciels d'usage courant permettent de dessiner impeccablement une grande variété de graphiques, et au surplus les "suites bureautiques" intègrent toutes des "fonctions graphiques" qui permettent de traduire en graphiques des séries de chiffres en quelques clics.

Il apparaît néanmoins que **cette situation n'est pas satisfaisante du tout** : l'immense majorité des graphiques ainsi réalisés, que l'on trouve à profusion dans les magazines, un peu moins dans les travaux historiques, reste *inutile*, souvent *illisible*, parfois carrément trompeuse. Comment a-t-on pu en arriver là ?

Ce sont les "économistes", au sens large, qui, dans le domaine des sciences sociales, ont les premiers tenté de traduire les chiffres en graphiques, en particulier les courbes chronologiques. Le mouvement commença dans les années 30 et s'amplifia rapidement après la seconde guerre mondiale. On trouve des réflexions déjà approfondies dans l'ouvrage d'André Piatier (1914-1991), *Statistique descriptive et initiation à l'analyse* (Paris, 1962), et des développements substantiels dans l'ouvrage maintes fois réédité de Gérard Calot (par la suite directeur général de l'INSEE), *Cours de statistique descriptive* (Paris, 1965). Piatier engagea une réflexion sur les bons et les mauvais graphiques. Calot insistait davantage sur les papiers fonctionnels. Le grand livre de Jacques Bertin, *Sémiologie graphique*, parut en 1967. Bertin, lui, était cartographe d'origine et dirigeait depuis les années 50 le "laboratoire de cartographie historique" créé au sein de la VIe Section de l'EPHE. Il réfléchissait plutôt en termes de bonne et de mauvaise carte, mais tenta de généraliser sa réflexion au **problème de la perception visuelle de la structure des données**. Il proposa des méthodes pratiques originales de manipulation pour découvrir des structures d'ordre (les "matrices de Bertin"). Malencontreusement, ni lui ni ses collaborateurs n'ont accepté l'idée pourtant simple et décisive que les graphiques sont étroitement liés aux calculs : il s'engagea de facto dans une impasse. D'une certaine façon, il en alla un peu de même avec J. W. Tukey, pourtant statisticien professionnel, qui proposa des formes de graphiques originales efficaces (le box-plot, alias "boîte-à-moustaches"), mais défendit contre toute raison l'idée que l'on pourrait explorer les données à l'aide d'outils purement graphiques.

Les informaticiens qui développent les "suites bureautiques" ne sont nullement des statisticiens, et l'on ne saurait le leur reprocher. Mais, du coup, les "modèles graphiques" qu'ils mettent en œuvre avec prédilection cherchent bien *plus à produire de l'effet qu'à transmettre de l'information*. Les fausses perspectives, les gammes de couleur (si possible avec dégradés), les

formes étranges visent à étonner et à distraire : il s'agit d'"agrémenter" des tableaux jugés par principe "arides".

Un logiciel ouvert, novateur et profondément pensé comme R tend à rapprocher et à articuler calculs et graphiques. Les statisticiens qui composent le groupe de base (core team) sont manifestement très préoccupés par cette articulation, et l'on peut s'attendre à des progrès sensibles au cours des prochaines années. Il est notable par exemple que R ne propose aucun des ces horribles "graphiques 3D" qui hantent les magazines, et émette une mise en garde quasi comminatoire à l'encontre du camembert (piechart). R comporte des instructions qui permettent de construire des fonctions graphiques plus ou moins interactives, mais il reste encore beaucoup à faire.

Si l'on admet le principe, énoncé au début de ce cours, que, pour un historien, les méthodes statistiques visent à explorer les données pour en découvrir les structures non accessibles en simple lecture, il est pour ainsi dire immédiat que calculs et graphiques sont très étroitement complémentaires. Ce que l'on a essayé de montrer jusqu'ici à chaque occasion. Il n'en reste pas moins *indispensable de se poser globalement les questions les plus générales*, pour repérer plus rapidement les bonnes procédures et éviter les innombrables pièges : quelles sont les contraintes qui pèsent sur un graphique ? à quoi sert un graphique ? en quoi consistent les palettes de bons outils ?

8.1.2 les contraintes qui pèsent sur la représentation graphique

Un terme résume bien, à lui seul, la difficulté : la ***lisibilité***. Un texte imprimé, quels que soient les caractères et la mise en page, demeure lisible si le corps des caractères a au moins 1,5mm. Le reste est seulement une question d'habitude et d'esthétique. Pour un graphique, il en va tout autrement. La signification des objets qui figurent sur un graphique n'est que très modérement conventionnelle. Une ligne peut tout autant désigner un lien qu'une limite. Les symboles et les trames n'ont aucune signification intrinsèque. Il faut donc pouvoir se reporter à une légende, compréhensible autant que possible, pour, ensuite, identifier sur le graphique la position des éléments : ***plus*** *cette opération est longue et* ***moins*** *le graphique est lisible*.

La question la plus simple en apparence, et pourtant la plus souvent négligée, est celle de la *taille brute*. L'œil humain, à 30 ou 40cm (distance normale de lecture) distingue ("résolution") jusqu'à environ 1/10e de mm. Les objets plus petits deviennent flous et indiscernables. Depuis plusieurs années, le nombre des "pixels" sur un écran d'ordinateur s'est stabilisé. Le "pitch" (taille du pixel) varie couramment entre 0,30 et 0,28mm. Techniquement, il serait tout à fait possible de réduire cette taille, mais, ce faisant, on ne pourrait pas afficher beaucoup plus de chose, pour la raison que l'on vient d'évoquer. D'où cette stabilisation. Des symboles distincts lorsqu'ils ont 2mm deviennent *indistincts* lorsqu'ils passent à 0,8mm.

Dans la pratique courante, les chercheurs produisent des graphiques de format A4 (21 x 29,7cm), mais les éditeurs, de livres ou de revues, sous prétexte de "gagner de la place", les réduisent fortement. Un graphique construit dans un rectangle de 20 x 24cm se retrouve mesurer 5 x 8cm. Par ce simple fait, il devient illisible et inutile. Il s'agit d'ailleurs d'une tendance générale chez les "graphistes", qui réduisent de même les photographies à l'état de *vignettes ridicules*. C'est la mode !! Face à cette tendance stupide et destructrice, il revient aux auteurs de faire face à leurs responsabilités : il suffit d'exiger un jeu d'épreuves montées et de refuser fermement le bon à tirer si les graphiques ou les photographies sont illisibles du fait d'une réduction maniaque. Pour donner un ordre de grandeur : **on doit refuser toute "illustration" dont la diagonale est inférieure à 12cm**. C'est un minimum.

Une deuxième contrainte relève principalement, comme la première, des conditions de publication : *possibilité ou impossibilité d'utiliser des couleurs*. Les graphiques se construisent d'abord et avant tout sur un écran, et les écrans monochromes ont à présent disparu. Il serait donc insensé de se priver, dans la démarche de recherche, des possibilités offertes par une utilisation

rationnelle de quelques couleurs. Mais les imprimantes couleur ne sont pas les plus répandues dans les organismes de recherche et de conservation, du moins pour le moment, et surtout il est peu probable que les conditions de publication sur papier des travaux "scientifiques" varient fortement ; les faibles tirages amènent la plupart des éditeurs à exclure a priori la possibilité de la quadrichromie et d'un papier à surface lisse (contrainte qui, notons le en passant, pèse de même sur la plupart des revues et des collections sérieuses d'histoire de l'art, ce qui est tout de même un peu paradoxal). Dans ces conditions, *un graphique parfaitement lisible grâce à l'utilisation de deux ou trois couleurs devient très difficile à réaliser en noir et blanc* : lorsque l'opposition de deux ou trois couleurs joue un rôle clé dans un graphique, il est rarement possible d'obtenir une lisibilité équivalente en noir et blanc, même avec beaucoup d'expérience et de savoir-faire.

Une troisième difficulté relève au contraire presque entièrement de l'auteur du graphique, et provient de **la complication et de la surcharge**. Les graphiques dont nous parlons ici sont des graphiques d'exploration, d'analyse et de communication ; pas des inventaires, comme peuvent l'être des planches d'atlas. La profusion des symboles, l'entrelacement des trames, le barbouillage des couleurs sont autant d'obstacles à la lecture ; si l'on veut être lisible, **il faut élaguer à l'extrême**, et en définitive ne retenir que le strict minimum. Contrairement à ce que l'on pourrait croire intuitivement, la force d'un graphique ne résulte *pas du tout de l'accumulation* d'informations, *mais de la clarté* avec laquelle il rend manifeste une relation jusque là inaperçue. Si malgré tout il semble nécessaire de transcrire une information abondante, la solution consiste à juxtaposer deux, trois ou quatre graphiques simples (pas davantage, sinon les comparaisons entre les graphiques ne sont plus possibles : la marquetterie de graphiques est une erreur) ; naturellement, dans ce cas, la mise en page doit être surveillée de plus près encore que d'habitude.

8.1.3 la finalité spécifique du graphique : **la forme comme indice d'une relation**

Les graphiques transcrivent des données numériques, à l'aide d'unités graphiques élémentaires (points, lignes, trames) disposées dans un plan. Le résultat est *une construction géométrique (plane)*. Par rapport aux données numériques d'origine, il y a toujours une perte d'information. Quel est donc l'intérêt de la procédure ? **L'objectif est de faire apparaître des relations spécifiquement géométriques** : on a vu comment une simple ligne permet de se faire au premier coup d'œil une idée de la forme d'une distribution de type gaussien ; comment une ou deux isolignes permettent de visualiser la forme d'un nuage représentant la relation entre deux distributions (distribution bivariée) ; comment les analyses factorielles permettent de repérer des formes de nuages, des oppositions, des répartitions, des trajectoires ; deux lignes peuvent être parallèles, convergentes, divergentes ; deux nuages peuvent être confondus, partiellement superposés, entièrement disjoints. Deux ensembles peuvent être symétriques, homothétiques, emboîtés.

Contrairement à une série de chiffres, par essence analytique, *le graphique utile fait apparaître une forme reconnaissable et définissable, synthétique ; étant supposé (les exceptions sont rarissimes) qu'une telle forme est l'indice d'une propriété globale de la réalité traduite par les chiffres*. Une fois la forme reconnue, deux démarches complémentaires sont indispensables :
1. examiner l'intensité de la relation traduite par cette forme, ou, ce qui revient plus moins au même, repérer la présence et l'importance relative des anomalies (et/ou des valeurs "aberrantes"), ce qui conduit à déterminer une gradation ou une hiérarchisation des éléments du graphique par rapport à la forme reconnue comme dominante ; 2. revenir aux chiffres, règle d'or, intangible, retour aux chiffres qui n'a lui même de sens que par rapport à la réalité sociale historique que l'on cherche à structurer. Ce qui suppose (il vaut mieux être explicite) une solide connaissance de cette réalité historique d'une part, des principes généraux de l'analyse statistique d'autre part, l'un n'étant pas séparable de l'autre (raison pour laquelle un graphiste qui, en général, ne dispose ni de l'une ni de

l'autre, n'a guère de chance d'aboutir au moindre résultat).

Autrement dit, la construction des graphiques dont nous parlons ici n'a de sens que comme une étape, ou un élément, dans un processus général de structuration, étape qui nécessite la maîtrise d'outils particuliers, mais qui peut faire gagner beaucoup de temps : une forme globale constitue en elle-même une synthèse forte, le plus souvent malaisée à atteindre par d'autres moyens. On comprend aussi pourquoi *l'interactivité est si souhaitable* : étant donné un jeu de chiffres et un type de graphique, on peut, le plus souvent, construire une multitude de graphiques différents en faisant varier tel ou tel paramètre de la construction, et il est rare que l'on tombe immédiatement sur les meilleurs ; on parvient d'autant plus vite au graphique le plus lisible que l'on peut plus aisément faire varier ces paramètres et construire rapidement des séries de graphiques.

8.2. COMMENT PROCÉDER DE TELLE MANIÈRE QU'UN GRAPHIQUE RÉPONDE À SA FINALITÉ ?

On peut, pour clarifier les idées, distinguer une série de caractères et de procédures, qui se recouvrent en partie, mais qui peuvent constituer autant de sujets d'attention précis dans le processus d'élaboration des graphiques :

8.2.1 la taille

On a insisté plus haut sur la nécessité de faire face énergiquement aux putatives "contraintes éditoriales". C'est que *la lisibilité passe d'abord par des caractères dimensionnels*. Il faut réfléchir soigneusement aux relations entre la taille globale du graphique, la taille des éléments de base (type caractères ou symboles) et la taille relative des divers éléments les uns par rapport aux autres et par rapport à la taille globale. L'accessibilité à l'information spécifique du graphique, c'est-à-dire la mise en évidence d'une forme, à quoi peuvent s'attacher hiérarchiquement des éléments complémentaires, suppose impérativement *que tous les éléments soient commodément lisibles* (donc ne descendent pas en dessous d'une taille minimale) et que l'on réalise finalement un graphique dans lequel la hiérarchie et la répartition des tailles facilitent autant que possible la perception des relations qui constituent l'information centrale du graphique.

8.2.2 les parasites

De nombreuses procédures graphiques génèrent des **artefacts**, auxquels on ne prête pas forcément attention parce que l'on y est plus ou moins habitué, mais qui cependant **captent automatiquement le regard**, au détriment des éléments significatifs. Cela tient aux structures même de la perception visuelle. Par exemple, une ligne brisée attire plus l'œil qu'une ligne sinueuse, et une ligne sinueuse plus qu'une ligne droite. Il n'y a rien à faire, c'est ainsi. Dans le cas classique d'une courbe de prix en *dents de scie*, l'œil est automatiquement attiré par les pointes et l'aspect en dents de scie, si bien que la tendance, s'il y en a une, quoiqu'elle ait une signification au moins aussi importante, n'est perceptible qu'au prix d'un effort tout à fait particulier. Dans cette situation, on peut discuter sur l'importance relative des variations à court et à moyen terme (la bonne solution consiste en général à représenter sur le même graphique une courbe de valeur centrale et une courbe d'indice de dispersion, en faisant éventuellement varier la largeur de la fenêtre mobile). Mais dans le cas d'un histogramme, la discussion est sans objet : les "*tuyaux d'orgue*", qui constituent généralement ce genre de diagramme, sont des artefacts à 100%. Et l'œil voit les tuyaux d'orgue bien avant de repérer la forme générale de l'histogramme, bien que cette dernière soit seule significative. Les tuyaux d'orgue sont un des meilleurs exemple de ces parasites artificiels qui brouillent la vision.

Même si l'on ne trace pas les traits verticaux, la forme en *escalier* de la ligne sommitale induit inévitablement l'impression de sauts, elle aussi tout à fait artificielle. C'est le genre de graphique à proscrire purement et simplement. On a signalé au passage le danger des diagrammes "en oursin" pour la représentation des nuages. Ce type de figuration attribue une importance démesurée au centre de gravité, quelle que soit la forme réelle du nuage. C'est un piège de première grandeur.

Le remède le plus général est constitué par **un lissage raisonnable et contrôlé**. La méthode du KDE (*kernel density estimator*, estimateur de densité du noyau de convolution) est une des plus générales, elle est utilisable commodément aussi bien en 1 qu'en 2 dimensions (diagrammes de densité d'une distribution univariée, nuages de points (quelle qu'en soit la signification) ; le seul inconvénient (provisoire, on peut le penser) réside dans l'absence générale d'algorithme de correction des effets de bord. On peut en imaginer plusieurs, il faudrait qu'on les trouve intégrés d'origine aux fonctions qui exécutent le calcul du KDE. Un lissage également très utile est constitué par les *courbes splines* qui, à la place d'une suite de segments produisant autant d'angles (artificiels le plus souvent), dessinent une courbe régulière, les segments droits étant remplacés par des portions de fonctions polynomiales, le plus souvent du 3e degré, auxquelles on impose simplement d'être tangentes à la même droite aux points de jonction. Contrairement à ce qui se passe avec l'utilisation des fenêtres mobiles (médianes mobiles et moyennes mobiles par exemple), les courbes splines passent exactement par les points d'origine, seule diffère l'allure de la courbe qui les joint.

8.2.3 des repères compréhensibles et efficaces

On a souligné que toute carte doit obligatoirement comporter une échelle précise et des éléments suffisants de géoréférencement. La majorité des graphiques comportent deux échelles, il est indispensable que le lecteur puisse sans effort comprendre *la nature et le sens des unités utilisées*. Dans le cas des analyses factorielles, les plans factoriels ne comportent pas d'unité ni d'échelle (ce qui déconcerte beaucoup de lecteurs), mais il ne faut pas oublier d'indiquer de manière univoque le numéro des axes représentés. La question de la grille doit être traitée au cas par cas. Tout dépend de la nature et de la précision des données figurées. Quand cela est possible, une grille non pas noire mais grise permet de fournir un système de repère discret qui ne détourne pas l'attention du phénomène principal. Dans tous les cas d'utilisation de symboles ou de trames, il faut apporter le plus grand soin à la légende, toujours avec le même souci d'une parfaite lisibilité, mais sans que l'emplacement ou la taille accaparent l'attention.

On ne doit pas négliger la possibilité de fournir *quelques repères ponctuels*. Dans le cas d'une distribution uni- ou bivariée plus ou moins gaussienne, indiquer la ou les médianes, soit par un repère sur l'échelle, soit par un trait traversant le graphique, est une procédure simple qui aide la lecture. On peut préférer indiquer de cette manière le premier et le troisième quartile (principe sous-jacent au boxplot). On pourra procéder de même pour telle ou telle valeur ayant une signification particulière.

8.2.4 l'ordonnancement des éléments figurés

C'est sans doute sur ce point que Jacques Bertin a le plus insisté, avec raison. L'ordonnancement concerne en fait deux aspects complémentaires : *la position relative sur le graphique*, et *l'aspect, ordonné ou non, des éléments figuratifs* (trames ou symboles). Une série chronologique ou des données spatiales contiennent un ordre intrinsèque qui n'offre guère de choix. La courbe de densité d'une distribution univariée repose sur un tri croissant (on l'oublie tant cela paraît aller de soi). Les difficultés commencent lorsque les éléments figurés ne comportent pas d'ordre visible a priori, ou en comportent plusieurs qui diffèrent. J. Bertin avait imaginé des procédures manuelles de réarrangement. L'idée était, dans les conditions techniques de l'époque, excellente, mais diverses procédures de calcul permettent d'obtenir des résultats analogues ou

meilleurs : s'il n'y a que deux variables, *l'algorithme des moyennes réciproques*, couplé au graphique de "Bertin-Cibois", donne de très bons résultats ; s'il y en a davantage, *les analyses factorielles constituent l'outil privilégié*. C'est, jusqu'à présent, l'outil le plus efficace et le plus général que l'on ait trouvé pour faire apparaître une structure d'ordre(s) au sein d'un ensemble complexe, le cas le plus spectaculaire étant le *scalogramme*. Ce n'est pas un hasard si les résultats des analyses factorielles se lisent d'abord et avant tout sur des graphiques.

Mais on néglige trop souvent de surveiller de près la gradation visuelle des symboles et des trames. Si ces éléments doivent représenter une structure ordonnée (ordre de taille, ordre d'intensité, ordre chronologique...), il est impératif que cet ordre se lise instantanément, la légende servant seulement à indiquer la valeur des bornes entre classes. C'est une grande banalité de rappeler qu'une palette multicolore est totalement inappropriée pour figurer un ordre, c'est pourtant une des erreurs graphiques les plus communément observées.

8.2.5 les transformations

L'objectif essentiel des graphiques est de permettre de repérer des formes. Or une forme simple se repère beaucoup plus aisément qu'une autre, et s'interprète ensuite bien plus facilement. Une bonne partie des "papiers fonctionnels" n'avait pas d'autre usage. Les logiciels qui permettent la combinaison graphiques-calculs la plus commode, comme R, facilitent considérablement ce genre de procédure, dont il faut faire un usage intensif. En présence d'une distribution dont la courbe de densité est unimodale sans que le mode soit bloqué à une extrêmité, toutes les transformations doivent être essayées jusqu'à ce *que l'on parvienne à produire un graphique symétrique.* Pour toutes les courbes chronologiques, il faut faire par principe au moins un essai avec des *ordonnées logarithmiques*. C'est le seul moyen d'observer et de comparer des pentes et non des valeurs absolues. Certaines transformations sont classiques, comme le remplacement de la courbe des densités brutes par la courbe des fréquences cumulées, ou le remplacment des valeurs absolues d'une courbe chronologique par les différences premières. Lorsque l'on veut comparer des rectangles (parcelles, bâtiments, feuilles de papier ou de parchemin, etc..), la solution courante est d'utiliser longueur et largeur comme abscisse et ordonnée ; l'expérience montre surabondamment qu'il est cent fois plus efficace de calculer la surface et le rapport longueur/largeur (coefficient d'allongement), ce qui donne instantanément à la fois une idée de la distribution des surfaces et de la répartition des allongements en fonction de la surface, qui est en général le critère distinctif le plus pertinent. Les deux séries de données sont strictement équivalentes du point de vue du contenu de l'information, il n'y a ni perte ni gain ; mais d'un point de vue visuel, la seconde configuration est sensiblement plus facile à lire et à interpréter. L'utilisation des rangs à la place des valeurs absolues entraîne une perte plus ou moins importante de précision ; mais il est très fréquent que les graphiques construits sur cette base soient beaucoup plus lisibles.

Il reste que **la transformation la plus efficace est celle qui aboutit à une linéarisation**. Toute équation, simple ou un peu plus complexe, qui permet de transformer une courbe en ligne droite apporte en elle-même une information décisive. On connaît depuis longtemps la "*droite de Henry*" destinée à vérifier le caractère normal ou log-normal d'une distribution, l'utilisation de la distribution des *moyennes conditionnelles* pour tester les distributions parétiennes, proposée par Marc Barbut, gagnerait à être mieux connue et beaucoup plus souvent employée. Ne pas oublier non plus (on y pense trop rarement) que de nombreuses distributions correspondent à une échelle "semi-proportionnelle", intermédiaire entre échelle arithmétique et échelle logarithmique, que l'on peut linéariser grâce à la transformation de Box-Cox.

Une réflexion sur les transformations possibles est toujours fructueuse. Beaucoup de problèmes commencent à s'éclairer lorsque l'on a trouvé une transformation pertinente, qui peut requérir l'invention d'une formule spécifique. C'est dans la manipulation plus ou moins habile et

efficace de ces procédures que se manifestent l'expérience et l'imagination, aliis verbis que l'on distingue le chercheur du technicien.

8.2.6 titre et commentaire

Lorsqu'au cours d'une recherche, on imprime des graphiques (à usage personnel), il est impératif d'y porter un minimum d'indications : nom du fichier de données utilisé, type de transformation effectuée (le cas échéant), indications plus ou moins détaillées sur les avantages et les inconvénients du graphique ; il est vivement conseillé d'indiquer la date, c'est un des moyens les plus simples pour se repérer ensuite dans un dossier qui comporte des dizaines, voire des centaines de feuilles.

Lorsque l'on prépare un graphique pour l'édition (ou la communication, dans une thèse, un article, un ouvrage), il faut trouver *un titre aussi explicite que possible*, et élaborer un commentaire synthétique. L'idée que le graphique vient en appui du texte, et que c'est la lecture de ce dernier qui donne le commentaire, paraît logique, mais elle est fausse en pratique ! *La grande majorité des lecteurs*, au moment de prendre une vue globale de l'article ou de la thèse, vont presque automatiquement s'arrêter sur des graphiques ou des cartes, *avant toute lecture*. Un commentaire très bref (quelques lignes au plus) leur permettra de vérifier immédiatement qu'ils ont bien compris le sens du graphique, ou éventuellement attirera leur attention sur un point névralgique. Le commentaire doit normalement se terminer par une indication de la nature et de la source des données.

Notons enfin que, dans toute la mesure du possible, *il est plus que recommandé de publier les données chiffrées traduites par le graphique en même temps que celui-ci, particulièrement s'il s'agit de données inédites*. Tout lecteur doit pouvoir "refaire" le graphique s'il le juge nécessaire.

8.3. LES OBJETS GRAPHIQUES ÉLÉMENTAIRES

Après avoir passé en revue les principaux axes de réflexion qui doivent être présents dans la tête du chercheur au moment de la confection d'une série de graphiques, il faut reprendre ces questions d'un point de vue plus analytique et plus concret, en examinant les difficultés propres à chaque objet graphique élémentaire, les gammes de solutions classiques, les erreurs à éviter.

8.3.1 le type de graphique

Avant d'entrer dans les détails, il importe de rappeler rapidement les divers types de graphiques existants, en signalant les types rares, et surtout en insistant sur ceux qui doivent être proscrits. Le graphique courant est constitué d'objets divers dont les positions relatives sont fixées par des *coordonnées cartésiennes*. C'est le plan simple de la géométrie ordinaire. Les logiciels graphiques (R notamment) disposent de fonctions simples et puissantes pour fabriquer pour ainsi dire automatiquement de tels graphiques. On peut utiliser cet automatisme, mais il y a lieu néanmoins de s'en méfier. Une saine méthode préconise de *se poser systématiquement la question du choix des unités et des échelles*. Il est souvent indispensable de modifier les valeurs extrêmes (notamment si l'on veut construire plusieurs graphiques comparables), il est presque toujours utile de voir ce que donne une échelle logarithmique.

La plupart des logiciels proposent des graphiques en coordonnées polaires et triangulaires. Les **coordonnées polaires** sont efficaces si l'on a à faire à un phénomène fortement périodique : données mensuelles, hebdomadaires, éventuellement journalières (toutes les heures, toutes les 3h, etc.). Mais le graphique devient vite illisible si l'on superpose plus d'une demi-douzaine de cycles. Les **coordonnées triangulaires** sont appropriées aux comparaisons de répartitions en trois

fractions : à chaque objet doivent correspondre trois pourcentages de total 100. En face de données de ce type, le graphique triangulaire est un outil de classement puissant ; il permet également d'examiner des évolutions, il suffit de relier les points successifs par une trajectoire. La difficulté de lecture provient de l'ambiguïté des échelles : il faut faire en sorte que le lecteur comprenne immédiatement à quel objet correspond chacun des trois axes.

Tout le monde connaît les "*pyramides des âges*", c'est cette relative familiarité qui les rend utiles : bien qu'il s'agisse d'un graphique très spécifique, il n'est pas nécessaire d'expliquer longuement comment il est construit.

Reste **le fatras des graphiques plus ou moins tordus et illisibles**. Jacques Bertin, avant l'apparition de la pseudo-perspective, avait déjà tenté d'établir la liste méthodique des "illustrations muettes". Comparer les portions d'un cercle est impossible. Les objets de base se ramènent en fait à peu près à des triangles, mais disposés différemment, l'œil n'en vient pas à bout ; c'est encore pire dans le cas des "papillons", c'est-à-dire lorsque les portions correspondent à des rayons variables. Le dessin est amusant, mais le résultat entièrement opaque. *Proposer de comparer deux camemberts n'a pas de sens*. **Il est beaucoup plus simple de lire deux colonnes de chiffres ordonnées**. On utilise parfois le demi-cercle, en particulier pour représenter la proportion des divers partis dans une assemblée ; le plus souvent, les effectifs bruts sont indiqués, et dès lors on comprend vite de quoi il s'agit. Mais c'est un cas particulier, et d'ailleurs les logiciels proposent rarement le demi-cercle comme forme standard de graphique. Un autre type d' "illustration muette" est constitué par le graphique "en banderolles", dont il existe de multiples variantes. Des éléments de colonne sont soit empilés soit juxtaposés. Des trames ou des couleurs sont censées permettre de les différencier et en même temps de les comparer. Il faudrait que l'œil puisse comparer simultanément plusieurs séries d'éléments décalés, et le plus souvent non ordonnés. C'est impossible, même avec de la patience. Ce genre de graphique, purement analytique, est impropre à faire apparaître la moindre forme d'ensemble, il est entièrement inutile.

Toute pseudo "troisième dimension" (effet de perspective) doit être sévèrement proscrite. Les éléments se superposent et s'embrouillent, la lecture varie plus ou moins fortement selon le point de vue choisi : il s'agit du meilleur moyen de tromper le lecteur. S'agissant des "camemberts", toutes les études réalisées ont montré que leur perception est très difficile. Avec le "camembert en 3D", on atteint *le sommet de la tricherie* : un secteur identique, selon qu'il est figuré de face ou de profil, n'a ni la même surface ni la même apparence ; 15% vus de face paraissent considérables, vus de profil sont réduits à presque rien... Au demeurant, il s'agit d'un phénomène très général. S'agissant par exemple d'architecture et de bâtiments, une série bien conçue de plans et de coupes est infiniment plus efficace pour comprendre la structure d'un édifice que n'importe quelle "représentation 3D", fixe ou animée. La vue en perspective, comme la photographie, sont utiles pour juger de l'effet global produit par un édifice, selon l'endroit où l'on se trouve ; le point de vue purement perceptif et esthétique n'est pas sans importance, mais *la compréhension de la structure passe nécessairement par des plans et coupes*, qui seuls permettent d'avoir une vue cohérente et homogène des diverses parties et de leur agencement.

Comme dernier exemple d'objet graphique inutile sinon trompeur, mentionnons encore le "*dendrogramme*", produit de l'un des nombreux algorithmes dits de "classification automatique". Comme on l'a rapidement signalé dans le chapitre consacré aux données multivariées, ces algorithmes peuvent avoir un intérêt pratique, mais ne sont d'aucune utilité dans un processus de recherche. Les résultats sont foncièrement instables, et l'interprétation des classements toujours malcommode et très approximative. Les "dendrogrammes" ne permettent pas, de toute manière, de représenter plus de quelques dizaines d'objets ; l'examen des distances entre objets et des regroupements qui en résultent doit se faire par d'autres méthodes.

Le choix de tel ou tel type de graphique en fonction des données disponibles et des structures

que l'on recherche est tout le contraire d'une routine. Il demande de l'expérience et surtout une réflexion constamment en éveil, un sens critique aiguisé.

8.3.2 les caractères typographiques

On doit prêter beaucoup d'attention au choix d'une ou de plusieurs polices. Le critère de choix est bien entendu la lisibilité et non pas l'esthétique. Les caractères les plus simples (genre Arial) sont sans conteste les plus appropriés ; *toutes les fioritures doivent être proscrites.* Comme on l'a déjà noté, il faut tenir compte du fait que le graphique, s'il doit être reproduit, risque fort d'être plus ou moins réduit. Un corps 8, qui sera tout à fait lisible dans un format A4, pourrait bien disparaître après réduction ! Si l'on veut distinguer des catégories d'éléments par des polices différentes, il y a lieu d'être très prudent. Dans un premier temps, il vaut mieux se contenter de jouer sur l'opposition simple entre romain et italique. Si cela est possible, quelques couleurs bien distinctes peuvent faire l'affaire. Le mélange de polices est rarement efficace.

Un point particulier mérite une attention spéciale : la similitude, dans la majorité des polices, entre le 1 (un) et le l (l minuscule) d'une part, et entre le 0 (zéro) et le O (o majuscule). Exemples :

times :	1 l 0 O	*1 l 0 O*
arial :	1 l 0 O	*1 l 0 O*
courier ::	1 l 0 O	*1 l 0 O*
lucida sans :	1 l 0 O	1 l 0 O

Il n'existe malheureusement pas de solution simple. Le zéro barré (Ø) est facile à reconnaître, mais son emploi n'est ni courant ni commode. Il existe des logiciels de dessin de police (on en trouve en shareware), qui, le plus souvent, permettent de modifier une police existante. En partant d'une police simple, type arial, on peut modifier assez facilement deux ou trois caractères (par exemple allonger fortement la barre oblique du 1, rendre le O circulaire et rétrécir encore le zéro, ou le munir d'une barre transversale oblique). Cela fait, on réenregistre la police sous un nom reconnaissable et elle devient utilisable par les logiciels graphiques qui utilisent les polices TrueType. Il faut être particulièrement attentif à ce détail dès que l'on utilise des codes qui mélangent les lettres et les chiffres. (C'est l'une des principales difficultés auxquelles se heurtent les logiciels de reconnaissance de caractères).

8.3.3 les symboles

Bien employés, les symboles sont un outil de visualisation puissant. Utilisés de manière quasi aléatoire (comme on le voit souvent), ils brouillent complètement la perception des formes. On peut jouer sur deux paramètres : la forme du symbole et sa taille. La taille est une question relativement simple. Il faut surtout se souvenir que, *si l'on cherche à obtenir des symboles proportionnels à une grandeur, on doit tenir compte de leur surface, c'est-à-dire quel'on doit tenir compte de la racine carrée de la largeur ou du rayon (ou d'un équivalent).* Cela dit, il faut de toute manière procéder à des essais, et une fois encore penser aux effets de la réduction de la taille globale du graphique.

Sauf cas particuliers, quatre ou cinq formes différentes constituent un maximum si l'on souhaite une perception globale efficace. Si l'on veut représenter des classes d'objets non ordonnées, il faut choisir des symboles nettement différents, et ne prêtant pas à confusion. Par exemple, une croix oblique, un cercle et un triangle. Deux triangles, l'un avec la pointe en bas, l'autre avec la pointe en haut, se distingueront très difficilement. Il reste néanmoins que *des symboles de même taille, si distincts soient-ils par leur dessin, ne forment pas des nuages repérables.* C'est ici que les couleurs sont à peu près irremplaçables. *Si l'on ne peut faire appel aux couleurs, il faut matérialiser*

les nuages par des isolignes ou juxtaposer plusieurs graphiques.

Si au contraire les classes sont ordonnées, il faut trouver, parmi les symboles disponibles, une petite **série nettement graduée**, par exemple un trait horizontal, une croix, une étoile ; là encore, des essais sont indispensables.

8.3.4 les traits et les trames

Les possibilités varient grandement d'un logiciel à l'autre. Même des logiciels graphiques célèbres n'offrent qu'une gamme très restreinte de trames géométriques graduées, voire même pas de gamme du tout. De même pour les traits : on se trouve souvent réduit au seul choix de la largeur. Ici encore, il faut *rechercher la simplicité et la clarté*. Si l'on veut comparer la configuration de deux nuages, le plus efficace est de se contenter de deux *isolignes bien différenciées* ; deux trames plus ou moins superposées ne seront lisibles qu'à condition de les choisir nettement distinctes, mais non couvrantes (type hachures obliques). Bien entendu, tout dépend de la configuration : si plusieurs nuages sont bien séparés, toutes les solutions seront recevables ; les superpositions partielles ou totales requièrent une attention spéciale. De la même manière que pour les symboles (mais en général avec des possibilités de choix bien plus restreintes), il faut faire en sorte que le type de relation entre les divers modes de figuration soit immédiatement perceptible. Contrairement à ce que certains manuels affirment, *l'usage simultané de deux types de trames graduées sur le même graphique (par exemple un trame constituée de hachures et une trame constituée de points) est à peu près illisible*. Ici encore, la couleur est le seul outil vraiment efficace.

8.3.5 l'usage des couleurs

Les évolutions techniques rapides en cours rendent l'usage des couleurs de plus en plus commun. C'est **un indéniable progrès**. Mais il faut garder à l'esprit deux idées principales :
a) on peut toujours améliorer un graphique noir et blanc en utilisant des couleurs, mais il faut surtout savoir comment traduire en noir et blanc un graphique que l'on a d'abord réalisé en couleurs. C'est pourquoi il est nécessaire de bien savoir manier les symboles et les trames en jouant seulement sur la taille et la forme, ce qui est fréquemment très malaisé. Il faut examiner lucidement le résultat : si l'on ne voit rien, ou pas grand chose, il faut laisser le graphique dans son dossier.
b) l'utilisation des couleurs requiert une réflexion spécifique. En dépit du fait que les couleurs correspondent à une variation régulière de longueurs d'onde, les couleurs de l'arc-en-ciel ne correspondent visuellement à aucun ordre. Il est impossible de traduire une hiérarchie quelconque avec rouge-vert-bleu. On peut assez facilement trouver sept ou huit couleurs nettement distinctes et indépendantes de tout ordre. C'est dans cette perspective que les couleurs ont le plus d'utilité : elles permettent de distinguer nettement et à peu de frais un nombre raisonnable de classes non ordonnées. En pratique, il est conseillé de les utiliser de manière redondante : par exemple cinq symboles bien distincts et non hiérarchisés pourront utilement être rendus, si c'est possible, avec cinq couleurs différentes ; le graphique sera ipso facto plus lisible. De même, deux traits différents pourront être colorés de deux couleurs, cela facilitera sensiblement la lecture. Bien entendu, pour la plupart des couleurs, on peut aussi jouer sur les intensités, et obtenir des nuances de rouge comme on produit des nuances de gris. Il peut être intéressant d'utiliser deux palettes, on emploie fréquemment le rouge et le bleu ; on peut, à la rigueur, si le phénomène à représenter s'y prête, ajouter une gamme de vert. Mais il est strictement déconseillé de dépasser cette limite. On tombe très vite dans le *barbouillage, qui constitue le risque majeur* inhérent à l'emploi des couleurs. Celles-ci constituent un outil graphique puissant, mais à la condition d'être utilisées **en petit nombre, avec des tons francs** : considération élémentaire, qui n'est, comme tout lecteur l'aura remarqué, que la ènième formulation du principe unique, celui de la lisibilité.

8.4. LE RÔLE DES GRAPHIQUES DANS UNE STRATÉGIE DE RECHERCHE

On ne doit en aucun cas perdre de vue le présupposé majeur : dans un travail de recherche, **un graphique** (comme d'ailleurs une reproduction) **n'est pas une illustration, mais un outil de visualisation de relations**. Il ne s'agit pas non plus d'un inventaire, le tableau de chiffres ou la carte-atlas sont faits pour cela. Donc, ni distraction ni somme d'informations atomisées. Mais outil de perception synthétique d'une (ou de quelques) relation(s) simple(s). C'est pourquoi les travaux de Jacques Bertin, consacrés aux mécanismes et aux conditions de perception de ces relations, ont eu une importance aussi décisive, et continuent de nourrir la réflexion.

Cependant, depuis les années 60, les conditions techniques, aussi bien de calcul que de dessin, ont été bouleversées : on doit utiliser ces nouvelles possibilités. En se familiarisant avec un logiciel efficace, on se donne les moyens de tester des séries d'hypothèses, en alternant et en **combinant systématiquement les calculs et les graphiques**. Comme l'on part d'ordinaire avec l'hypothèse que les données sont structurées, on doit parvenir à faire apparaître des formes simples qui permettent d'identifier ces relations. Ce résultat est rarement immédiat, mais suppose ordonnancements et transformations numériques.

Lorsqu'une forme est repérée, il importe d'avoir deux réflexes : 1. la forme repérée n'est qu'un indice, il faut "**retourner aux données**", l'évaluation et la mise en lumière exacte de la ou des relations aperçues ne peuvent se faire qu'avec les chiffres, étant entendu que ceux-ci ne sont qu'un indicateur d'une réalité sociale historique, qui constitue, seule, l'objet d'étude ; 2. une forme repérée, c'est bien , plusieurs c'est mieux : une structure est par définition un ensemble de relations articulées. **On ne peut à aucun moment être certain d'avoir épuisé toute l'information potentielle contenue dans un tableau**. Il faut viser, au delà de toute relation simple, une structure. L'expérience montre qu'il faut de la patience. Au demeurant, on retrouve simplement ici un fondement de la "méthode chartiste" : une vérification supplémentaire n'est jamais inutile, plus de précision, un peu d'imagination permettent à tout coup d'enrichir la connaissance scientifique.

Chapitre 9

DISTRIBUTIONS LEXICALES

Les textes et les mots sont la matière première ordinaire du travail de tout chartiste. Or ce sont des objets qui se prêtent remarquablement bien à une multitude de manipulations formelles et statistiques d'une redoutable efficacité. Ce que malencontreusement la quasi totalité des historiens ignorent (encore). Au surplus des matériaux directement utilisables apparaissent en quantité exponentiellement croissante, tandis que les recherches sur les procédures de traitement sont dans une phase de prolifération extraordinaire. Les chartistes ne doivent pas attendre pour monter dans le train en marche !

SOMMAIRE

9.1. CARACTÈRES DE LA SITUATION ACTUELLE

9.1.1 bref historique

L'intérêt pour les mots pris intrinsèquement, en dehors de la suite "naturelle" des textes, n'est pas récent. On considère habituellement que les premiers à se consacrer à ce genre d'activité furent les massorètes, c'est-à-dire les érudits hébraïsants qui compilèrent la **massorah**, liste de tous les mots de la bible hébraïque rangés par ordre alphabétique, comptés et commentés. Par la suite, ce sont les dominicains de Paris qui, au milieu du 13e siècle, établirent la première **concordance de la Vulgate**, à l'occasion de quoi fut établie une division en chapitre universellement adoptée par la suite, et commença un premier travail de "critique", les dominicains s'étant heurtés à un nombre considérable de variantes entre lesquelles ils tentèrent de choisir "la bonne". Ce fut également le moment où triompha en Occident l'ordre alphabétique : les premiers index alphabétiques, issus d'une forme de "fiches", datent du 13e siècle.

Des questions du même genre, corollairement liées aux unités insécables que sont les lettres, préoccupèrent les *cryptographes* (surtout à partir du 17e siècle), puis les créateurs du *télégraphe* (Samuel Morse, 1791-1872), enfin les *sténographes*, parmi lesquels on retient surtout le nom de Jean-Baptiste Estoup, qui publia au début du 20e siècle des *Gammes sténographiques*, et qui mit au point un système sténographique très efficace parce que fondé sur la fréquence observée des lettres et des phonèmes (en français). Il est remarquable que **les problèmes de codage optimal et de cryptographie** continuent de donner lieu en ce moment-même à des recherches intensives.

Notons au passage qu'un problème analogue se pose aux enseignants de langues : comment établir le "vocabulaire de base", par quoi il faut naturellement commencer ? Dès 1939, un chercheur américain, **Paul Bernard Diederich**, publiait à Chicago le résultat d'une compilation manuelle considérable et bien conduite : *The frequency of the latin words and their endings*. Il donnait une liste d'environ 5000 mots, avec leur fréquence observée dans trois bonnes anthologies, une de prose classique, une de poésie classique et une de latin médiéval. Une telle liste, qui aurait pu et dû rendre de grands services, est restée jusqu'à aujourd'hui à peu près totalement ignorée...

9.1.2 difficultés actuelles

L'évolution fut lente et progressive dans ce domaine jusque dans les années 80. Depuis moins d'un quart de siècle, l'apparition puis le développement exponentiel d'internet ont mis à la disposition de tout un chacun une quantité croissante de textes numérisés, au milieu desquels il est devenu fort malaisé de se reconnaître. La capacité à repérer dans cette masse monstrueuse les informations pertinentes en "temps réel", c'est-à-dire jour par jour, est devenue une activité professionnelle à temps plein ("veille technologique"), activité qui repose pour l'essentiel sur l'emploi de logiciels appropriés, c'est-à-dire intégrant des algorithmes de repérage et de recherche capables de travailler à peu près sans intervention manuelle à une vitesse de plus en plus élevée du fait de l'augmentation des masses à traiter. La notion clé est celle d' "*information retrieval*" (IR), plus ou moins liée à un syntagme plus prétentieux (qu'avec un minimum d'esprit critique on pourrait même qualifier de vaguement absurde), le "*knowledge management*" (KM). L'IR est aujourd'hui un enjeu économique de base et suscite par conséquent des investissements considérables, raison pour laquelle la majorité des instituts de "mathématiques appliquées" de par le monde y consacrent une part croissante de leur activité.

Mais l'activité scientifique a ses rythmes, que même un investissement financier massif ne peut pas bouleverser durablement. Les travaux du dernier quart de siècle se caractérisent par un taux de redondance très élevé, l'enfoncement d'une grande quantité de portes ouvertes, un cloisonnement paralysant entre une multitude de "sous-champs" prétendant à l'autonomie, diverses formes de

rétention de soi-disant "secrets de fabrication" ; dans le fatras produit, il est difficile d'identifier les innovations réellement efficaces, et l'on ne voit pas apparaître quoi que ce soit qui ressemblerait à une synthèse, même provisoire. A ma connaissance du moins (je peux me tromper), il n'existe pas d'ouvrage, technique ou de vulgarisation, qui présenterait de manière accessible et équilibrée l'état actuel des savoirs et des techniques de manière un tant soit peu organisée et pondérée. Pour le moment, ce que certains proposent d'appeler l'« **industrie de la langue** » produit des bénéfices et des brevets, mais pas un corps de connaissances vraiment novateur ; on ne peut que procéder, au prix de pas mal de temps et d'empirisme, à une exploration plus ou moins aléatoire des sites des universités et des centres de recherche, dans lesquels on trouve pourtant, à l'improviste, mais de plus en plus souvent, des articles et des mémoires exposant des observations et des propositions d'algorithmes qui peuvent s'avérer utiles à la recherche historique, au prix d'un très sérieux effort de réflexion et d'adaptation..

9.1.3 sources : les textes numérisés librement accessibles

Le fait le plus notable est la mise en ligne, dans des sites accessibles sans limitation, de textes "anciens" en quantité croissante. Par "ancien", on entend ici, pour simplifier, tout texte édité avant la première guerre mondiale. **Ces textes, jusqu'à preuve du contraire, sont libres de tout droit, appartiennent à l'héritage culturel de l'humanité, et peuvent être utilisés sans restriction par n'importe qui**. Dans les années 90, quelques éditeurs ***cupides*** ont entrepris de privatiser une partie de ce patrimoine. Ayant assuré la numérisation de ces textes, ils les ont reproduits sur des CDroms, qu'ils ont commercialisés à des tarifs extravagants (certains, pas tous). D'autres ont imaginé de faire payer l'accès par internet à ce genre de fichiers.

Heureusement, dès ce moment, des intellectuels plus soucieux de diffusion culturelle, ont entrepris de mettre en ligne, de façon complètement libre, des textes numérisés de manière coopérative. La première grande initiative de ce genre, qui mérite vraiment d'être connue et soutenue, est le « *projet Gutenberg* ». De grandes institutions universitaires ou de conservation, des associations spécialisées, ont largement pris le relais et l'on voit aujourd'hui apparaître sur internet une quantité croissante de textes de toutes époques et en toutes langues, en accès libre. En France, le site « *gallica* » de la BNF offre déjà une importante quantité de matériaux. Le site « *bibliotheca augustana* » de l'université d'Augsburg propose une série impressionnante de classiques dans les principales langues européennes, et comporte de nombreux textes médiévaux et de la renaissance. L'université de Navarre donne les œuvres complètes de *Thomas d'Aquin*. Le site « *www.thelatinlibrary.com* » propose l'essentiel de la littérature antique et des textes médiévaux en quantité croissante (ainsi qu'une profusion de liens variés renvoyant à des sites utiles à tout chercheur latiniste). Notons cependant que même les moteurs de recherche les plus puissants n'indexent pas de manière optimale ces ensembles. Il faut souvent de nombreux détours avant de parvenir au site pertinent.

Cette liste n'est à aucun égard exhaustive, et ne risque pas de l'être avant longtemps : le trait le plus frappant de la situation est précisément la croissance du nombre de sites et, conséquemment, de la quantité de textes accessibles de cette manière. Sans faire preuve d'un optimisme exagéré, et sauf bouleversement imprévisible des règles de fonctionnement d'internet, on peut raisonnablement estimer que l'essentiel des textes "littéraires" "anciens" seront ainsi disponibles d'ici une dizaine d'années. Reste, pour un historien, la question centrale des textes "diplomatiques" ou "d'archives". De ce côté, la situation est nettement moins brillante. Il serait infiniment souhaitable que les historiens, au premier chef les chartistes, prennent une claire conscience du problème et mettent en œuvre les moyens susceptibles de combler dans les meilleurs délais cette lacune gênante (et qui pourrait le devenir de plus en plus).

9.1.4 sources : une profusion de « working papers »

Ainsi qu'on l'a dit plus haut, la plupart des centres qui se consacrent à la "computational linguistic", à l'"information retrieval", au "natural language processing" et autres curiosités du même style font eux-mêmes un usage massif d'internet et tentent d'y apparaître sous leur meilleur jour, ce qui les amène à mettre en ligne une partie significative de leur production. Comme il se produit fatalement une certaine circularité, les moteurs de recherche, eux-mêmes issus de ces travaux, les indexent assez soigneusement. A condition donc d'utiliser les mots clefs adéquats, un usage patient de ces moteurs de recherche permet d'avoir accès à une masse considérable d'écrits relatifs à ces matières. La qualité, l'intérêt, la clarté, l'utilité de tous ces papiers sont excessivement variables. Neuf sur dix des documents que l'on télécharge et imprime apparaissent, après une lecture souvent éprouvante, comme inutilisables. Comme toujours, les titres sont en général d'une aide médiocre. Heureusement, on tombe de ci de là sur des textes pertinents et compréhensibles qui peuvent aider à élaborer un peu mieux les méthodes disponibles dans ce domaine complexe et foisonnant de la statistique lexicale. De toute manière, il n'existe pas de choix manifestement meilleur. Les nombreuses revues liées à ce domaine proposent un matériau tout aussi inégal et dispersé.

9.2. LES UNITÉS DE BASE

9.2.1 rappel élémentaire : langue et discours

Depuis les travaux de Saussure et de nombreux autres, il apparaît à la fois classique et logique de distinguer langue et discours. Le *discours est concret*, c'est ce que l'on perçoit directement, sous forme écrite ou orale ; la *langue*, parfois définie comme un ensemble de règles (définition trop restrictive), appartient au domaine du "*virtuel*". Depuis longtemps, la statistique lexicale n'a eu aucune peine à établir une correspondance canonique : le discours correspond à l'échantillon observé, tandis que la langue est la "population parente" d'où est tiré l'échantillon. On a pris l'habitude de distinguer ainsi le "vocabulaire" d'un texte ou d'un corpus, considéré comme un échantillon d'un putatif "lexique" de la langue.

Cette manière d'aborder la réalité n'est pas sans intérêt. L'intuition et/ou l'expérience tendent à montrer que tous les textes possibles ne sont que des réalisations partielles et limitées d'une réalité qui les dépassent et les englobe. Tout le monde a pu constater que, même dans les plus gros dictionnaires, ceux qui se veulent les plus exhaustifs, "**il manque des mots**" : il n'y a pas d'exception. En interrogeant des corpus gigantesques (disponibles en ligne), on est continuellement surpris de noter que tel ou tel mot, ou tel ou tel emploi d'un mot, ne sont pas représentés.

Pourtant, l'analogie a des limites, sur lesquelles les statisticiens aussi bien que les linguistes et lexicographes pourraient sûrement s'interroger encore avec profit. Car les axiomes de la théorie des probabilités, que tout mathématicien considère comme les fondements de la statistique, comportent l'idée que "population" signifie ensemble clos, en général sommable et supposé homogène. Au minimum clairement définissable. Or la notion de langue ne répond à aucun de ces critères. Il est extrêmement risqué (on y reviendra bientôt) d'assimiler "la langue" (quelle qu'elle fût ou quelle qu'elle soit) à une "population parente" au sens des statisticiens. Les plus lexicographes, les spécialistes de sémantique historique, de sociolinguistique, de pragmatique, ont tous fait le même constat : "la langue" est une réalité floue, mouvante, hétérogène.

Il reste que l'opposition saussurienne, quoiqu'il soit téméraire de la transposer sans précaution dans les termes de la statistique classique, demeure incontournable. Dans les années 80 et 90, des spécialistes de statistique linguistique ont cru résoudre la difficulté en décidant d'écarter entièrement la notion de "langue" et en se fondant exclusivement sur des "corpus". Ce déni de réalité ne pouvait aboutir qu'à des mécomptes. Une grande partie des travaux fondés sur cette orientation ont inévitablement abouti à des résultats faux.

9.2.2 occurrence, forme et lemme

Nous arrivons ici au cœur du sujet : quel est l'individu de base, au sens de la statistique ? *Les praticiens oscillent entre deux positions* : ou bien se contenter de toute suite de lettres entre deux blancs, à l'exclusion des signes de ponctuation et autres signes typographiques, ou bien rattacher chaque mot observé à une entrée de dictionnaire, en accolant si possible à chaque mot ainsi défini une série d'indications concernant sa position (numéro d'ordre), sa forme (morphologie), sa fonction (syntaxe).

Pour simplifier : tant que les dépouillements demeurèrent manuels, le rattachement à une entrée de dictionnaire allait presque de soi. Les lexicographes procédaient ainsi, par définition pourrait-on dire, mais il s'agissait plutôt de grapillage que de dépouillements ; la plus grande entreprise jamais réalisée sur une telle base est sans doute l'*index thomisticus*, conçue et dirigée au lendemain de la seconde guerre mondiale par **le père Roberto Busa**, qui réussit à dépouiller l'intégralité des textes attribués à Thomas d'Aquin, et publia les résultats sous forme d'une série d'une soixantaine d'in-folio à la fin des années 70. La procédure consistant à rattacher chaque mot du texte à une entrée de dictionnaire est appelée **lemmatisation** ; on distingue alors le lemme (objet indexé et donnant ensuite lieu à comptages), et les formes concrètes dudit lemme, telles qu'elles apparaissent dans le texte dépouillé. Dans ce cadre, on admet l'équivalence de toutes ces formes, qui comptent pour autant d'occurrences dudit lemme.

La diffusion de l'informatique puis surtout de la micro-informatique rendirent l'avantage aux formes brutes. La plupart des langages de programmation permettent de traiter des "chaînes de caractères", donc d'identifier sans faute les blancs et les signes de ponctuation, et d'aboutir sans aucune intervention à un découpage des formes, que l'on peut ensuite, tout aussi simplement, classer, indexer, compter de toutes les manières imaginables. Pendant de longues années, la bataille a fait rage entre partisans et adversaires de la lemmatisation. Depuis un certain temps déjà, de nombreux chercheurs ont tenté des expériences analogues : opérer une analyse statistique du même corpus avec **et** sans lemmatisation ; pour les langues "courantes" (anglais, français, et même latin), la plupart de ces expériences ont abouti à la même conclusion : pour un texte d'une longueur minimale (quelques milliers de mots), les deux procédures aboutissent à des conclusions statistiques très proches. Ce qui a vivement conforté les partisans des formes brutes.

9.2.3 étiquetage

Cependant, dès la fin des années 80, les techniciens de l'information retrieval (e tutti quanti) se sont aperçus qu'*une forme brute pouvait être très ambiguë*, correspondre de facto à des lemmes bien distincts ou, plus couramment, renvoyer à des sens du même "mot" complètement différents, ce qui constitue un obstacle grave pour une indexation efficace. L'examen du "contexte" (que l'on abordera dans le prochain chapitre) apparut rapidement comme un impératif, mais la simple proximité graphique s'avéra vite insuffisante ; l'extrême utilité d'une introduction des relations syntaxiques devint bientôt manifeste. D'où la création (toujours en cours) d'algorithmes permettant d'identifier automatiquement les fonctions ; ce que l'on appelle en anglais "*morpho-syntactical tagging*", en français *étiquetage morpho-syntaxique*. La lemmatisation a donc ressurgi, mais plus complexe même, dès lors que l'objectif est de distinguer non seulement les formes (flexions et

autres) mais aussi, pour un même lemme, les "sens" considérés comme différents (je n'évoque pas ici la question de savoir à partir de quand et selon quels critères un algorithme doit décider que deux emplois renvoient à deux "sens" distincts du même mot).

L'industrie de la langue a besoin d'aller de plus en plus vite. Durant plus d'une décennie, se sont affrontés les tenants des "grandes machines", utilisant pour leurs analyses des corpus énormes, préétablis, de règles de combinaison et d'emploi pour chaque mot, et les partisans des algorithmes "**unsupervised**", autoadaptatifs, c'est-à-dire capables de repérer ces règles sans aide externe, en s'adaptant "on the fly" à chaque corpus. Cette dernière catégorie d'algorithmes a rapidement gagné en efficacité, atteignant des "taux de réussite" en pratique tout à fait suffisants (du genre 97%). Sensiblement moins coûteux, autoadaptatifs, infiniment plus rapides, les algorithmes "unsupervised" paraissent avoir gagné la partie (impression personnelle). Un certain nombre de principes qui sont à la base de ces algorithmes sont connus, et d'ailleurs discutés. Mais ni les programmes ni a fortiori les listages ne sont disponibles : les enjeux économiques priment. *Si l'on souhaite, comme cela paraît raisonnable, passer à l'étape du "unsupervised morpho-syntactical tagging" du latin médiéval, il faudra probablement que des médiévistes prennent eux-mêmes l'affaire en mains*.

9.2.4 fréquences absolues et relatives

Si l'on laisse provisoirement de côté cette question de l'étiquetage, et que l'on suppose avoir résolu le choix entre formes et lemmes, on se trouve devant une suite simple de chaînes de caractères séparées par des blancs. Un tri alphabétique suffit pour tout ranger en ordre (généralement ascendant, et direct, c'est-à-dire classant selon l'ordre de la première, deuxième, etc. lettre). Les formes ou lemmes étant ainsi regroupés, il est immédiat d'obtenir le nombre total (nombre total d'occurrences du texte, alias longueur du texte) et l'effectif en valeur absolue de chaque unité. Autrement dit, sa fréquence absolue dans tel texte ou tel corpus. Connaissant l'effectif total du texte ou du corpus, une division donne l'effectif relatif. Ces grandeurs ultra-simples (en apparence) sont *le matériau de base* pour des calculs qui peuvent atteindre une redoutable complexité.

9.2.5 classements possibles

On peut faire naturellement apparaître cette même liste sous quatre formes élémentaires :
- l'ordre alphabétique direct ;
- l'ordre alphabétique inverse (en prenant les lettres à partir de la dernière, ce qui, s'agissant par exemple des formes en latin, permet de regrouper tous les mots ayant même terminaison) ;
- ordre de fréquence croissant ;
- ordre de fréquence décroissant (le plus usuel : on commence par la forme la plus fréquente, et ainsi de suite ; ce tri aboutit à donner à chaque forme un rang, il y a bien entendu de plus en plus d'ex-aequo, que l'on range le plus souvent, par commodité, dans l'ordre alphabétique ; on peut bien entendu compter les ex-aequo, et connaître ainsi le nombre de formes de fréquence 1, de fréquence 2, de fréquence 3,.....).

9.3. OBSERVATIONS EMPIRIQUES UNIVERSELLES

9.3.1 la croissance indéfinie du vocabulaire

Les manipulations élémentaires que l'on vient de décrire en quelques lignes semblent presque simplistes. Aussi longtemps que l'on a dû les exécuter à la main, elles représentaient un

travail incroyablement long et fastidieux : les textes traités de cette manière sont restés excessivement rares. L'essentiel du temps disponible était absorbé par les manipulations aux dépens du traitement final. La situation est complètement inversée. On peut constituer des listes de ce genre à raison de plusieurs dizaines par heure. Dès lors, des régularités massives sont apparues, définitivement incontestables.

La plus importante et la plus lourde de conséquences, en dépit de son caractère anodin, est ***la croissance indéfinie du vocabulaire lorsque le texte s'allonge***. Pour des auteurs peu prolixes, le corpus comporte quelques dizaines de milliers de formes, pour d'autres, on peut atteindre quelques millions. Les corpus traités à l'heure actuelle par les plus gros moteurs de recherche sont (non vidi) de l'ordre d'un trillion de mots (combien de zéros ??) Le nombre de formes ou de lemmes croît nettement moins vite que le nombre des occurrences (effectif total), mais il ne cesse jamais de croître. C'est un *phénomène a priori étrange et contrintuitif*. Mais c'est seulement la prise en compte complète de cette observation qui fonde toute statistique lexicale réaliste.

Cette observation décisive entraîne au moins deux conséquences d'une portée considérable :
a. **le "lexique" n'est pas dénombrable**. Si l'on admet qu'un corpus peut croître indéfiniment, cela signifie que le vocabulaire va aussi croître indéfiniment. C'est une observation extrêmement ennuyeuse pour tous les tenants de l'application des statistiques "classiques" aux faits lexicaux. Le lexique, loin de répondre à la définition axiomatique d'une population statistique, est un "ensemble ouvert", non dénombrable. C'est tout simplement un objet fractal, analysable dans le cadre parétien et non dans le cadre gaussien traditionnel.
b. **la fréquence relative de la plupart des formes et lemmes n'est pas stable**, mais dépend avant tout de la longueur du corpus à partir de laquelle on la calcule. Aucune démonstration n'est nécessaire pour les hapax : si N est l'effectif total, la fréquence relative de tout hapax est 1/N. Si N croît indéfiniment, 1/N décroît indéfiniment. Cela est clair pour toutes les fréquences faibles et moyennes. Cela l'est de moins en moins au fur et à mesure que l'on se rapproche des formes les plus fréquentes.

9.3.2 la prépondérance des hapax

Dans n'importe quel texte, que l'on considère formes ou lemmes, *les hapax sont beaucoup plus fréquents que les formes de fréquence 2*, elles mêmes beaucoup plus fréquentes que les formes de fréquence 3, και τα λοιπα. Les hapax représentent couramment 40 à 50% des formes d'un corpus, il semble (non vidi) que cette fréquence atteigne 60% et plus dans les corpus gigantesques évoqués plus haut. Cette observation complète la précédente. Si le vocabulaire était une population fermée, l'allongement d'un corpus devrait amener un taux de répétition croissant des formes et/ou des lemmes, et donc une diminution, au moins relative, des hapax. Or c'est le contraire que l'on observe.

9.3.3 stabilité relative de la forme la plus fréquente, problème des « mots-outils »

A l'autre extrêmité, c'est-à-dire avec le mot (forme ou lemme) le plus fréquent, il apparaît en règle générale que la fréquence relative est stable. Dans la plupart des langues européennes, les mots les plus fréquents ne sont ni des substantifs ni des verbes, mais des "particules", qui peuvent être, selon les cas, des prépositions, des articles, des pronoms, des conjonctions, certains adverbes : ce que, d'un terme vague, contesté tant par les linguistes que par les statisticiens, on appelle ordinairement des "*mots-outils*". Ces mots arrivent à jet continu dans tout corpus, et c'est précisément cet aspect de "jet continu" qui entraîne une fréquence relative tout à fait stable ou à peu près stable.

Cette observation purement numérique faite, on ne doit pas cacher que ces mots causent de gros tracas. Car on n'arrive à les définir ni du point de vue du sens ou de la grammaire, ni du point de vue statistique. Statistiquement, aucune césure perceptible ne se manifeste. Le passage est

insensible. Les praticiens empiristes, à propos de l'anglais, se contentent de dire qu'il est sans conséquence de "couper" à peu près n'importe où entre le 30e rang et le 200e. Pourtant, comme on le verra plus bas, ces mots très fréquents ont une fréquence tendanciellement inférieure à ce qu'une distribution parfaitement régulière et homogène laisserait attendre. *On ne sait pas trop pourquoi*. Si l'on considère les 50 formes les plus fréquentes dans un traité de saint Augustin, on trouve quelques substantifs...

9.3.4 variété des langues

Même si l'on restreint l'examen aux langues européennes, on est frappé par les différences fondamentales quant au traitement des lemmes. Il existe au moins deux échelles de différentiation : la présence plus ou moins forte de variations morphologiques, le degré plus ou moins élevé d'agglutination. *L'anglais est extrême sur les deux échelles* : aucune agglutination, des variations morphologiques minimes. Le castillan agglutine quelques pronoms, l'allemand forme des pléthores de "mots composés", le sommet semble atteint par le finlandais, qui de cette manière fabrique avec une grande liberté des mots de 40 ou 50 lettres. La morphologie française est plus complexe que celle de l'anglais mais le latin est meilleur, et le russe vraiment excellent. Une troisième échelle pourrait également être considérée, selon la plus ou moins grande signification de l'ordre des mots : des principes d'analyse applicables au français ou à l'anglais sont, de ce point de vue, inutilisables pour le latin. Il est facile de voir que l'universelle tentation des linguistes et des lexico-statisticiens de *s'imaginer qu'un algorithme applicable à l'anglais (contemporain) vaut en tous temps et en tous lieux est une plaisanterie qui peut coûter cher*. C'est pourtant le présupposé implicite de l'écrasante majorité des travaux disponibles...

9.4. LES RÉGULARITÉS STATISTIQUES FONDAMENTALES

9.4.1 les outils simples

On trouve sur internet plusieurs **logiciels de concordance en freeware** (de diffusion et d'utilisation libres, mais non pas libres au sens de la GPL, donc pas open source). Partant d'un texte sans autres caractères de contrôle que les CR/LF, ces programmes assurent le découpage, et des sorties triées ad libitum. Surtout, ils fournissent des concordances, c'est-à-dire des listes d'occurrences, le plus souvent triées alphabétiquement, mais où chaque occurrence, au lieu d'apparaître isolée, est munie de son contexte, c'est-à-dire des quelques mots qui la précèdent et des quelques mots qui la suivent. On peut ordinairement choisir le nombre de mots avant et après à conserver, ainsi que l'ordre de présentation : on a en général la possibilité d'un tri alphabétique de l'occurrence immédiatement antérieure ou immédiatement postérieure. Une telle sortie est appelée concordance par référence aux "concordances bibliques" établies selon ce principe depuis le 13e siècle au moins, comme on l'a signalé plus haut. De telles listes, extraordinairement longues à constituer à la main, sont à présent produites de manière automatique et instantanée ; leur consultation en lecture simple est toujours d'une extrême utilité, faisant apparaître des rapprochements et des régularités qu'aucune lecture suivie ne permet. *Si l'on veut étudier un texte de manière un tant soi peu approfondie, c'est une faute de ne pas procéder à la constitution et à l'examen d'une concordance*. Les programmes élémentaires de concordance sont plus simples à manipuler qu'un traitement de texte ou un tableur.

A cet égard, on doit faire trois remarques complémentaires :

a. les listes produites par ces programmes sont en général données sous un format ultra simple, sans caractère de contrôle ; elles sont récupérables très facilement par un traitement de texte, un tableur,

ou surtout un logiciel de statistique : le matériau ainsi manipulable est d'une invraisemblable richesse ;
b. les découpages, tris et concordances sont des opérations d'une très grande simplicité ; la taille actuelle des unités centrales (CPU) et la vitesse des microprocesseurs ont fait disparaître tous les problèmes pratiques qui hantaient encore les programmeurs de programmes pour micro-ordinateurs dans les années 80. Les textes les plus longs traités par les historiens dépassent exceptionnellement quelques millions d'occurrences. En fait, la difficulté ordinaire est plutôt inverse, résultant de la brièveté des textes, quelques centaines de mots seulement pour une charte ou une lettre, par exemple, ce qui nécessite des précautions à propos desquelles la réflexion n'est guère avancée.
c. les logiciels disponibles, verouillés, comportent rarement toutes les possibilités que l'on souhaiterait. **Il serait assez simple de programmer un tel logiciel dans n'importe lequel des langages actuellement usuels, et de le diffuser sous la GPL. L'avantage décisif d'une telle entreprise serait de pouvoir progressivement enrichir le logiciel de fonctions supplémentaires, en commençant sans doute par une aide à la lemmatisation** (je ne connais aucun programme libre assurant une telle fonction, qui paraît pourtant simple à programmer).

9.4.2 les précurseurs

La réflexion pratique sur les comptages ne paraît guère antérieure à ce que firent les sténographes à partir de la fin du 19e siècle. On cite diverses recherches sur les nombres des espèces et des familles (animales et végétales) réalisées par Willis et le grand statisticien U. Yule dans les années 20. On attribue généralement une influence décisive à l'enseignement du mathématicien américain (hétérodoxe) Norbert Wiener (1894-1964), professeur au MIT, connu surtout comme le fondateur de la cybernétique.

9.4.3 C.E. Shannon et la « théorie de l'information »

Tout laisse penser qu'un tournant majeur fut réalisé par une publication due à Claude Elwood Shannon (1916-2001), "A mathematical theory of communication" (*The Bell System Technical Journal*, 27-1948) (ce texte est disponible en accès libre sur internet, lecture conseillée, ne serait-ce que d'un point de vue historique).

C. Shannon, ingénieur, travaillait sur des problèmes de codage et de transmission de données. Il partit du principe élémentaire que l'« information » est fondamentalement une question de fréquences relatives. Il eut l'extrême clairvoyance, à un moment où les recherches sur les distributions de fréquences étaient embryonnaires, de comprendre qu'une appréhension efficace de ces fréquences supposait l'emploi quasi exclusif des logarithmes, plus précisément des logarithmes en base 2. Il est naturellement impossible de résumer ici cette théorie en quelques lignes.

La formule fondamentale dont tout le reste dérive est :

$H(X) = -\Sigma\ p_i \log(p_i)$ (en choisissant $\log(p_i)$ comme $\log_2(p_i)$)

Dans cette formule, p_i représente simplement la probabilité de l'élément i au sein de n éléments formant un tout. L'« information » est définie comme l'entropie d'un système, qui est une mesure du degré d'incertitude d'un système par rapport à tous ses états possibles. Le signe moins devant la formule est seulement là pour obtenir une grandeur positive, attendu que les logarithmes de nombres inférieurs à 1 (cas des probabilités, comprises entre 0 et 1) sont par définition des nombres négatifs.

H(X) est définie comme l'entropie du système X (qui peut par exemple être une phrase ou un texte). *La formule de Shannon a des propriétés très remarquables* :
- elle s'annule lorsqu'un des états est certain et les autres impossibles (information nulle !)
- pour un nombre donné d'états, elle est maximale lorsque ces états sont équiprobables, et augmente avec le nombre des états, ce qui est parfaitement conforme à l'intuition et à l'observation ;

- elle est additive, c'est-à-dire que lorsque plusieurs systèmes indépendants se trouvent réunis en un seul, leurs entropies s'ajoutent.

L'avantage décisif des logarithmes en base 2 est de fournir une définition aussi élégante que robuste de l'« unité d'information ». Lorsque l'on considère le "système" le plus simple, constitué seulement de deux états équiprobables (type pile ou face, 0 ou 1), on trouve en effet :

$H(X) = -(1/2 \log(1/2) + 1/2 \log(1/2) = 1$

L'unité d'entropie ainsi définie est appelée unité binaire ou bit. Or toute l'informatique est fondée sur l'emploi d'éléments électroniques à 2 états, justement définis comme bits. On saisit donc, au moins intuitivement, pourquoi les formules de Shannon sont la base de toutes les recherches ultérieures de la "théorie des ordinateurs" et finalement de l'informatique en général. Mais elles ont une portée nettement plus large, indépendamment des ordinateurs : on ne peut pas faire l'économie de leur connaissance si l'on veut raisonner sur les fréquences lexicales.

9.4.4 G.K. Zipf et la « loi rang-taille »

Presque au même moment, un autre chercheur américain, George Kingsley Zipf (1902-1950) publia, en 1949, un ouvrage d'allure un peu étrange, mêlant des découvertes mathématiques de première importance à des considérations générales plus ou moins loufoques. Ce mélange freina regrettablement la diffusion de ses découvertes. L'essentiel tient dans la "loi de Zipf" ou "loi rang-taille", formule d'une extrême simplicité. *Si l'on range les mots d'un corpus dans l'ordre de fréquence décroissant, et que l'on considère donc la suite bivariée rangs-fréquences, le produit rang x fréquence est une constante, ou suit éventuellement une simple fonction linéaire.*

Des recherches et des discussions sans fin ont eu lieu, et continuent activement, à propos de cette "loi". A partir du moment où l'on a disposé d'un nombre indéfini de corpus de toutes tailles, des observations systématiques sont devenues possibles. Une conclusion, d'apparence peu satisfaisante, s'impose manifestement : si l'on considère les choses "de loin" ou "en gros", cette loi est effectivement tout à fait universelle ; si l'on regarde "de près", les ajustements sont pour la plupart médiocres, les distributions observées s'écartant plus ou moins nettement de la distribution théorique correspondant à la "loi de Zipf".

D'une certaine manière, cette loi traduit mathématiquement un fait d'observation : dans tout corpus, on trouve quelques mots très fréquents, un nombre important de mots de fréquence moyenne et une multitude indénombrable de mots rares. G.K. Zipf, sur la base des quelques données disponibles à la fin des années 40, a donné de cette observation une formulation mathématique simple qui, en dépit de ses imperfections, demeure une base incontournable de toute étude des fréquences. **C'est une loi de type parétien**.

9.4.5 la correction de B. Mandelbrot et la caractérisation des fractales

Un apport décisif fut donné par le mathématicien français Benoît Mandelbrot, dès une première publication (dans les *Comptes rendus de l'Académie des Sciences*) en 1951. Travaillant à partir des publications de Wiener, Shannon, Zipf, il proposa une "correction" à la loi de Zipf. En introduisant *un paramètre supplémentaire*, il obtint un bien meilleur ajustement aux distributions observées, permettant surtout d'obtenir un ajustement acceptable pour les fréquences les plus élevées qui, comme on l'a déjà signalé, ont le plus souvent des fréquences nettement inférieures à ce qu'elles seraient si elles "respectaient" la loi de Zipf.

Il reste que l'importance majeure des travaux de B. Mandelbrot tient surtout à la formulation théorique étayant et justifiant l'extrême généralité de ce genre de distribution, issues d'un "univers parétien" et non gaussien, d'où la création (controversée, mais qui tend à se généraliser) de *la notion de fractales ou objets fractals*.

Dans les années 90, B. Mandelbrot est revenu brièvement sur les questions lexicales, et a

proposé une vue beaucoup plus homogène des résultats de Shannon, Zipf et de lui-même. Il reste qu'il est paradoxal (et apparemment peu de gens paraissent avoir fait cette observation) que B. Mandelbrot n'ait jamais tenté de traiter le problème clé de l'instabilité des fréquences en fonction de la longueur du corpus, qui est un des problèmes concrets les plus résistants (et gênants) auxquels se heurtent les chercheurs.

9.4.6 persistance d'erreurs grossières

Bornons nous ici à un simple rappel. Dans les années 70 et 80 (voire 90) de nombreux manuels de statististique "lexicale" ou "linguistique" ont énoncé des suites de contrevérités grossières, résultant de l'utilisation irréfléchie et fallacieuse des lois "classiques" au matériel lexical (lois binomiale, de Poisson, hypergéométrique, notamment). De tels travaux sont nuls et non avenus. Des "laboratoires" ayant pignon sur rue ont élaboré des logiciels (parfois payants et vendus !!) fondés sur l'utilisation de ces formules erronées. Il y a lieu de faire preuve ici d'une extrême prudence et d'un sérieux esprit critique.

Le remplacement dogmatique de la "langue" par des "corpus" n'est qu'une variante de cette erreur. L'instabilité fondamentale des fréquences relatives en fonction de la longueur du corpus ruine ab ovo toute tentative d'utiliser directement les fréquences relatives observées dans n'importe quel corpus pour analyser d'autres textes ou des sous-parties du corpus.

9.4.7 convergences et diffusion lentes

La disponibilité désormais massive de textes numérisés a donné un coup de fouet à toutes les recherches dans ce domaine, encore que les travaux historiques paraissent malheureusement en recul, non seulement en fréquence relative, mais même en nombre absolu. Il serait grand temps d'envisager de profiter intelligemment des avancées récentes, ce qui nécessitera bien entendu de sérieuses adaptation à un matériau spéfique. Seuls des historiens peuvent procéder à une telle mise au point.

Les problèmes de codage et de compression, à la suite des découvertes de Shannon, n'ont pas cessé de produire des avancées significatives (algorithme de *Huffman* dès les années 50, puis de *Lempel-Ziv-Welch* dans les années 70, que l'on utilise toujours, cf l'apparition récurrente de l'acronyme LZW dans une multitude de logiciels de compression ou de traitement d'images). La plupart de ces travaux sont fondés sur **des recherches liées à la structure arborescente des fréquences**, ce qui confirme (on l'a rarement remarqué, encore moins souligné) que la notion d'arbre hiérarchique est probablement le fondement logique principal de toutes les distributions de type parétien, fondement simple qui devrait permettre à terme de réunifier enfin statistiques gaussiennes et parétiennes (on en est encore loin).

Bien entendu, les praticiens de l'information retrieval ont redécouvert la roue, et sont inévitablement retombés sur la "loi de Zipf" et sur les logarithmes. Des synthèses accessibles tardent à se manifester, mais ce n'est pas une raison de ne pas tenter de repérer les avancées disponibles, qui sont susceptibles d'apporter aux études historiques une aide décisive. On doit noter que de jeunes chercheurs, complètement inconnus, ont proposé récemment (et continuent) des analyses et des découvertes très prometteuses, s'agissant en particulier de **formules mathématiques beaucoup plus conformes aux distributions observées** et aux divers paramètres permettant de les caractériser et donc de les comparer (e.g. Andras Kornai, Victor Lavrenko,...).

Considérations finales

On a seulement, ici, survolé quelques notions de base sur les distributions des fréquences lexicales. C'est un domaine en pleine évolution, relativement complexe. Mais des avancées irréversibles ont eu lieu, et les praticiens développent des algorithmes "qui marchent". C'est aux historiens eux-mêmes de traiter les problèmes historiques, en utilisant avec une extrême circonspection des découvertes empiriques à peu près exclusivement liées aux structure lexicales et sémantiques de l'anglais contemporain.

Chapitre 10

SÉMANTIQUE ET FORMALISATION

Un des plus grands écrivains du vingtième siècle, Jorge Semprun, qui manie au moins trois langues avec la même maestria, raconte un épisode survenu dans le camp de Buchenwald où il avait été déporté, quelques jours après l'arrivée des troupes américaines :

En avril 1945, dans le réfectoire d'un baraquement français de Buchenwald, le 34, nous avons déclamé Char et Aragon, mon copain Taslitzky et moi. Soudain, au moment où Boris me récitait à pleine voix un poème d'Aragon à la gloire du Guépéou, un hurlement s'est fait entendre, l'interrompant.

Nous avons tourné la tête.

Un vieux déporté français était assis au bout de la table. Nous n'y avions pas prêté attention, dans le feu de nos récitations. Il mangeait... Il mangeait sérieusement : méticuleusement...Il avait donc hurlé pour attirer notre attention. Et il l'avait obtenue.

- abscons, vos poètes ! criait-il à tue-tête. Esbroufeurs, tortionnaires de la langue!...

Nous attendions la suite, apitoyés par son grand âge et la détresse coléreuse de ses yeux transparents. La suite est venue soudain. D'une voix qu'il a enflée, qui a retenti dans l'espace vide du réfectoire, il nous a déclaméLes Châtiments. Plus précisément le passage sur Waterloo, morne plaine. Il s'est dressé, pour finir, a mimé l'arrivée de Blücher sur le champ de bataille, à la place de Grouchy. Il a commandé d'un geste ample du bras l'ébranlement de la Garde impériale, «tous ceux de Friedland et ceux de Rivoli portant le noir colback ou le casque poli », entrant dans la fournaise de Waterloo.

(*L'écriture ou la vie*, édition Folio, pp. 234-236). On pourrait tout aussi bien citer les réflexions sur Kafka (*ibidem*, pp. 337-341) : il vaudrait mieux lire tout ce roman autobiographique. Un lecteur intelligent et capable de réflexion y apprendra bien plus de sémantique qu'en lisant des dizaines d'ouvrages abscons de linguistique, anglo-saxons ou autres. Je ne saurais d'ailleurs trop conseiller de lire René Char ou Apollinaire, Hölderlin et Rilke pour ceux qui en sont capables. Tout médiéviste qui ne veut pas mourir idiot devrait se livrer à un petit exercice hautement instructif : choisir une page de saint Augustin, se procurer une ou deux "traductions" en français ou dans une autre langue, et chercher les fautes. Le résultat est **toujours** surprenant ; les approximations et les fantaisies fleurissent comme paquerettes au printemps, quand ce ne sont pas de médiocres fautes de construction. "Tortionnaires de la langue" : on peut prêter à cette étonnante expression une très vaste portée. Dans l'épisode où elle apparaît, elle renvoie, au moins, à trois "plans de réalité" imbriqués mais bien distincts. C'est une structure que perçoit intuitivement n'importe quel lecteur **cultivé** contemporain de Jorge Semprun. Mais à qui fera-t-on croire qu'une quelconque "traduction" de saint Augustin, dans une situation analogue, puisse transposer quoi que ce soit, à supposer que le traducteur ait saisi l'imbrication spécifique des réalités de tel ou tel passage ? La sémantique pose de redoutables problèmes, la sémantique historique est une affaire infiniment plus difficile. La quasi-totalité des historiens traitent cette question en la niant purement et simplement. Ce devrait être un sujet de préoccupation central pour tout chartiste. Je tenterai seulement ici de dresser un tableau résumé de ce que l'on peut espérer des procédures de formalisation, qui connaissent actuellement des développements inattendus.

SOMMAIRE

10.1. POSITION DU PROBLÈME

10.1.1 des champs d'investigations hétérogènes et dispersés

En dépit de quelques évolutions toutes récentes (création en 2003, à Göttingen, d'une collection "Historische Semantik" chez Vandenoek und Ruprecht), le moins que l'on puisse dire est que la sémantique historique n'a jamais constitué un véritable pôle d'études, et qu'une tentative, même tout à fait superficielle et rapide, pour faire le tour des activités qui pourraient ou devraient concourrir aux progrès de cette discipline oblige à **parcourir des espaces très variés**, totalement dépourvus de connexions réciproques et extérieurs les uns aux autres.

Une telle situation est dommageable pour les études historiques : les historiens, français en particulier, n'ont pour ainsi dire aucune notion de philologie, ne connaissent rien à la linguistique ou à la sociologie, et n'ont aucune idée de la manière dont fonctionnent les systèmes permettant de trouver des informations sur internet (pour autant qu'ils se servent de cet outil). Ces disciplines travaillent de manière tout à fait isolée, si bien que la même difficulté est abordée plusieurs fois de suite, mais nécessite la réitération des mêmes efforts, qui ainsi ne sont pas cumulatifs ; en même temps, des erreurs, bien identifiées ici, demeurent prévalentes là. Au total, *un grand gaspillage, et un tableau qui semble peu attractif*.

Mais on ne doit pas s'arrêter à cette première constatation. Il faut tenter de mieux définir les obstacles et identifier les possibles bénéfices.

10.1.2 la tragique faiblesse du cadre théorique

Les disciplines qui ont à faire prioritairement au sens des mots, des objets, des pratiques, se comportent à peu près toutes comme si cette notion de "sens" allait de soi. Tout le monde sait de quoi il s'agit, il n'y a pas lieu de s'y attarder. Si bien que les seuls à s'en préoccuper (en général modérement) sont les philosophes et les logiciens, qui précisément ne traitent d'aucun objet concret, et parlent donc du "sens" comme d'une entité déconnectée de toute réalité, ce qui ne peut aboutir à rien de vraiment sensé, c'est-à-dire à des discours abscons dénués de toute pertinence pour les praticiens des sciences sociales ou de la manipulation des textes disponibles sur internet....

Or quelques minutes de réflexion suffisent pour s'apercevoir qu'*il s'agit tout au contraire d'une notion difficile, ambiguë, susceptible de donner lieu à des développements complètement contradictoires*. Et l'on peut, ici encore bien moins qu'ailleurs, se contenter d'un catalogue de définitions dépareillées. Provisoirement, et en simplifiant à l'excès, on peut **distinguer deux pôles, eux-mêmes dédoublés**.

D'un côté, ce que l'on pourrait appeler **le pôle du "bon sens", c'est-à-dire le sens commun, contemporain ordinaire**. La notion apparaît alors comme une propriété intrinsèque d'un énoncé ou d'un comportement : telle phrase "a du sens" ou n'en a pas. Cette position, un peu affinée et développée abstraitement, renvoie à deux variantes que l'on désigne traditionnellement (sans doute à tort), comme la position platoniciennne et la position aristotélicienne ; dans le premier cas, toute chose a un sens intrinsèque, en soi, qui lui préexiste de toute éternité (les "idées" platoniciennes), dans le second cas, on privilégie l'observation, d'où l'on pense déduire une signification tout aussi intrinsèque et presque aussi éternelle.

Cette manière de voir a été progressivement contestée depuis la fin du 17e siècle, et la réflexion scientifique du 19e a abouti à l'idée qu'aucun sens n'est possible ni pensable en dehors d'**un système de repères, c'est-à-dire d'un ensemble cohérent de relations**, par rapport auxquelles tel objet ou telle action occupe une position ou effectue un mouvement, en quoi consiste le sens, qui n'est rien d'autre. C'est à ce second pôle que se rattachent tous les scientifiques qui ont consacré un minimum de temps à réfléchir sur les bases conceptuelles de leur discipline (qui ne sont

pas forcément majoritaires). Il faut toutefois souligner qu'à ce pôle s'opère également une séparation, entre ceux qui affirment que tout sens réside dans les représentations et/ou la langue (c'est la base de tous les développements "postmodernes" américains) et ceux qui pensent que la société humaine constitue un tout, indissolublement matériel et idéel, dans lequel les pratiques de la langue jouent un rôle majeur, mais toujours subordonné à une logique sociale d'ensemble, dont les "pratiques discursives" ne constituent qu'une partie, dont l'autonomie, réelle, est tout à fait limitée.

10.1.3 les historiens dans le désert de la pensée

Durant tout le 19e siècle, à de rarissimes exceptions près (essentiellement des historiens de l'Antiquité), les historiens européens ont méthodiquement adhéré au premier pôle. De deux choses l'une : ou bien un texte est véridique (il a du sens, et un sens correct) ou bien il n'en a pas (légende, fantasme, "forgerie") et bien entendu, dans ce cas, ne mérite que **la corbeille à papier**. Une telle vision fonde toute la "critique historique" du 19e siècle, parée du titre pompeux de "discrimen veri ac falsi". Le 20e siècle n'a pas connu d'évolution sérieuse, en tout cas en France. Dans les premières années du 21e siècle, une majorité de médiévistes professionnels y croit toujours fermement.

D'un point de vue strictement empirique, de tels présupposés génèrent des difficultés inextricables. Pour la simple raison que le "bon sens" est la chose du monde la moins bien partagée, et qu'une part considérable des textes médiévaux peut tomber dans l'une ou l'autre catégorie selon l'idée que l'on se fait de ce qui est compatible avec ledit "bon sens". En pratique, et sans jamais le dire, les historiens (au premier chef les médiévistes, mais tous les autres ont copié) ont élaboré **un modus vivendi vicieux, pour ne pas dire diabolique** : des groupes se sont constitués de facto (jamais de jure), chacun s'occupant des textes considérés par lui comme "acceptables" et les plus "intéressants", et abandonnant le reste aux autres groupes. Cahin-caha, la société médiévale s'est ainsi trouvée saucissonnée en fragments disjoints, assaisonnés de saveurs étranges, variées, et au surplus instables. Un seul point assure le consensus : la civilisation médiévale était d'abord et avant tout ***varietas***, et toute tentative pour lui trouver une logique générale est unaniment vouée aux gémonies.

Cette manière de procéder est un obstacle radical au progrès des connaissances ; quitte à subir le sort évoqué à l'instant, j'ai de moins en moins de doute sur la nécessité d'abandonner ce découpage grotesque ; je trouve absurde et inacceptable de continuer à passer le ***scalpel*** au milieu des chartes et des chroniques, selon que l'on désigne telle phrase ou tel épisode comme "forgé" ou "légendaire". En tout état de cause, les fantasmes et les "croyances" jouent dans toute société un rôle considérable, et l'historien qui déclare a priori devoir ne pas les prendre au sérieux dénonce ainsi lui-même sa radicale incompétence. Une charte ou une chronique avaient, sinon pour le copiste, en tout cas pour le rédacteur, un sens général ; le parchemin était dispendieux ; *c'est un a priori intenable de s'imaginer que le clerc médiéval était un sot, ou qu'il couvrait son parchemin de balivernes pour les enfants. Le métier de l'historien consiste à prendre au sérieux tous les documents et à s'efforcer d'établir dans quelles circonstances ils ont été conçus et exécutés, par qui, pourquoi*. Il faut partir de l'idée que **les documents étaient cohérents et qu'ils avaient un sens** (quitte à parvenir à démontrer le contraire, mais je ne connais pas d'exemple probant). Cette définition du métier aboutit immédiatement à ce qui ressemble à une catastrophe pratique : l'immense majorité des documents anciens, si on ne les charcute plus au scalpel, apparaissent pleins de bizarreries, sinon complètement obscurs. Je dois avouer que cela, loin de me déplaire, m'amuse énormément : une expérience malheureusement déjà longue m'a montré que, dans cette voie, on découvre beaucoup de choses et l'histoire, loin des récitations et des litanies, devient une activité intellectuelle enthousiasmante.

Ces quelques lignes devraient suffire pour laisser percevoir que la réflexion sur la notion de "sens", sur les présupposés qu'elle implique et sur les réorientations qu'elle suggère, est un enjeu

majeur, sinon l'enjeu principal, pour l'avenir des études historiques, de l'histoire médiévale notamment, mais des autres tout autant. Si les méthodes de la sémantique historique sont quasi inexistantes, il est impératif de s'employer énergiquement à les développer.

10.1.4 quelques caractères spécifiques des textes anciens

La quasi-totalité des textes antérieurs à 1800, sinon même à 1914, présentent une série de caractères propres, tant du point de vue intrinsèque (nature du matériau) que des problèmes intellectuels qu'ils soulèvent et des contraintes pratiques auxquelles se heurte leur étude. On est pour ainsi dire obligé, dans les circonstances actuelles (provisoires et labiles), de raisonner largement en opposition avec le matériau sur lequel travaille l'armée innombrable au service des "industries de la langue", i.e. les "pages internet" avant tout.

a) ces textes sont en *quantité finie* (par définition), et cette quantité n'est pas susceptible de s'accroître (corpus fermé) ;

b) la plupart des textes, au moins jusqu'au 16e siècle, présentent des *variantes "graphiques"* importantes voire très importantes ; la "normalisation" des éditeurs s'est effectuée sans méthode ;

c) dans toute l'Europe, jusqu'au 16e siècle au moins, tous les individus sachant lire et écrire étaient capables d'*écrire deux langues, et en pratique en parlaient au moins trois* ;

d) ces langues n'étaient pas juxtaposées (comme dans l'Europe actuelle), mais *hiérarchisées et profondément imbriquées* ;

e) l'étude de ces textes n'est soumise à aucune contrainte de vitesse ("temps réel") ;

f) ce sont des langues vraiment mortes, on ne peut donc disposer d'aucun "expert", au sens des informaticiens ; les historiens, en dépit de toute "familiarité", n'utilisent pas ces mots de manière pratique, les textes (transparents pour ceux qui les ont écrits jadis) ne le sont plus pour personne ;

g) à l'inverse de ce qui se passe dans les langues actuelles, et de ce que croient d'ailleurs presque tous les historiens, ce sont **les mots les plus fréquents, ceux qui, en apparence, semblent le plus "aller de soi", qui posent en fait les problèmes les plus redoutables (noms concrets, type *domus*, *terra*, *vinea* ; verbes courants, type *dare*, *sedere*).**

De cette petite liste (non limitative !), on peut retenir synthétiquement trois problèmes :

* une réflexion sur les principes de la sémantique historique et sur les caractères spécifiques de ce que pouvait être "le sens" dans l'Europe ancienne est une tâche prioritaire ; c'est un préalable fondamental à toute tentative (nécessaire) de formalisation ;
* la dispersion des pratiques et des expériences dont on peut essayer de tirer parti ne doit pas décourager, il faut essayer d'entreprendre un premier tour d'horizon ; je donnerai donc ici, contrairement aux chapitres précédents, un certain nombre d'indications bibliographiques ;
* on peut raisonnablement espérer pouvoir s'inspirer des procédures élaborées récemment par les informaticiens ; il faut donc tenter de comprendre en quoi elles consistent, tout en sachant qu'il faudra tout reprendre et adapter ; dans la situation actuelle, les logiciels disponibles, compatibles avec les méthodes d'une sémantique historique qui reste pour l'essentiel à inventer, sont inexistants, il faudra tout inventer.

10.2. LES PRATIQUES TRADITIONNELLES

10.2.1 les glossaires

Dans le fil de la tradition médiévale des gloses marginales, puis des gloses classées alphabétiquement à partir du 14e siècle, apparaît Charles Du Cange, et ses deux fameux glossaires, latin (1678) et grec. Des mots "difficiles", avec exemples et commentaires plus ou moins

explicatifs. On cite ensuite La Curne de Sainte Palaye, puis Godefroy, qui n'a pas été remplacé. L'obligation de consulter les deux parties du Godefroy est suffisamment fatigante pour que l'on réfléchisse une seconde à cette structure : la seconde partie n'est arrivée qu'après, l'idée toujours dominante étant de fournir des gloses pour les mots n'ayant plus cours. D'ailleurs en Allemagne, les frères Grimm entreprirent un Wörterbuch où figuraient tous les mots allemands depuis les origines : en fait, on se trouve devant la même attitude, qui considère les mots *en dehors de tout contexte défini chronologiquement*. Tous ces auteurs, de grand mérite, n'ont pas entrepris la moindre réflexion sur les rapports entre langue et société.

10.2.2 piétinements du 19e siècle

L'invention de la philologie en Allemagne ne parvint en France qu'avec retard. Mais sous ce terme se plaçaient principalement la phonétique historique et les recherches d'étymologie. Sans le dire de façon toujours explicite, la plupart des auteurs s'imaginaient que le sens d'un mot à un moment donné dépendait principalement des ses antécédents ("etymon"). Une telle idée, **pour l'essentiel fausse** (voir, entre dix mille, l'exemple flamboyant de fauteuil < faldistoel - Faltstuhl), reste, aujourd'hui même, prépondérante. On considère gnéralement que le premier auteur ayant énoncé un ensemble d'idées constitutives de la sémantique historique fut Karl Reisig (*Vorlesungen über die lateinische Sprachwissenschaft*, Leipzig, 1839). Il est hautement significatif que Reisig ait voulu ordonner le sens des mots selon un ordre qu'il considérait comme étant simultanément logique et chronologique : l'évolution comme expression de la logique pure !

10.2.3 les fondateurs

A peu près au même moment, un allemand et un français tentèrent de manière beaucoup plus réaliste de jeter les bases d'une histoire du vocabulaire. Il faut connaître **Hermann PAUL, *Principien der Sprachgeschichte*, Halle, 1886**, et **Michel BRÉAL, *Essai de sémantique. Science des significations*, Paris, 1897**. Une grande partie des éléments de théorie des changements sémantiques qu'ils mirent en place était empruntée à l'arsenal des définitions de la rhétorique classique (les tropes). Bréal fut un vrai pionnier, en insistant notamment sur le rôle de la polysémie, sur le rôle du contexte et des vocabulaires spécialisés. Diverses remarques montrent une intuition nette de l'importance des fréquences et de leurs évolutions.

On peut rappeler accessoirement que ce fut aussi le moment des premières enquêtes de géographie linguistique, en fait surtout orientées sur la géographie phonétique, mais recueillant dans le même temps un important matériau lexical.

10.2.4 pérennité des méthodes de la lexicographie traditionnelle

Paul, Bréal et d'autres, tentant brièvement d'indiquer ce qu'ils entendaient par "le sens" d'un mot, donnèrent des définitions purement atomistiques, dans le droit fil de la plus ancienne tradition, même plus ou moins mâtinée de considérations sociologiques : le sens d'un mot est l'ensemble des sens ressortant de tous ses usages... ***Atomisme de la théorie du sens et illusion de la transparence*** allaient naturellement de pair. Erreurs partagées par tous les lexicographes. La plupart des grandes entreprises lexicales du 19e et du 20e siècle reposèrent à peu près sur les mêmes procédures : collecte, pifométrique, d'un nombre aussi élevé que possible d'exemples copiés sur des fiches (i.e. un bout de phrase comportant le mot vedette avec une référence plus ou moins précise) ; installation de ces fiches dans des boîtes, avec classement alphabétique ; récupération par le "lexicographe" du paquet de fiches correspondant à tel ou tel vocable ; selon le nombre de fiches disponibles, tri plus ou moins fin "en fonction du contexte" ; tentative de résolution des "cas obscurs" également "en fonction du contexte" (une bonne partie de ces fiches "obscures", résistant à l'analyse, étant *discrètement* remises dans la boîte).

Les variantes principales résidaient (et résident encore) dans la manière d'opérer la mise en ordre de chaque article. Certaines entreprises prirent le parti d'un système inspiré d'une sorte de logique abstraite intangible (sens abstrait, sens concret, sens individuel, sens collectif, isolé, en composition, au propre, au figuré, etc). D'autres, encore plus prosaïquement, décidèrent de reprendre avec le minimum de changement les classements de tel ou tel dictionnaire antérieur ; le sommet de l'absurde étant par exemple atteint avec la consigne de reprendre les sens du *Gaffiot* pour un dictionnaire de latin médiéval (procédé breveté et garanti pour faire disparaître méthodiquement toute relation entre latin médiéval et société médiévale). Chacun sait que F. Niermeyer, qui réalisa l'essentiel de son *Lexicon minus* entièrement seul, était juriste et raisonnait en juriste ; ses sens sont les distinctions d'un juriste du 20e siècle appliquées à l'Europe médiévale...

En dépit de ces variantes, la plupart de ces dictionnaires historiques se caractérisent principalement par le fait qu'ils sont très peu historiques. Les distinctions, imposées a priori, sont presque toujours arbitraires et incongrues. Les évolutions ne sont notées qu'exceptionnellement (pour des dictionnaires qui le plus souvent couvrent plusieurs siècles). Les indications de fréquences sont inexistantes (seule exception, le *TLF*, *Trésor de la langue française*, conçu dès l'origine avec support informatique, aujourd'hui en ligne sur internet). L'atomisme le plus dérisoire règne partout en maître : pour chaque exemple, il faut trouver un sens et un seul, et fournir autant que possible une ***traduction*** en un seul mot (l'emploi d'une périphrase est ressenti comme un échec). L'illusion puérile d'un stock de sens éternel n'est mise en cause par personne, corrélat presque obligé de l'atomisme autosatisfait.

10.3. DÉVELOPPEMENTS AU VINGTIÈME SIÈCLE

10.3.1 Jost Trier et la théorie des champs sémantiques

Le tournant du 20e siècle, le seul dans le domaine qui nous occupe, réside dans l'œuvre de Jost Trier. Dès l'après-guerre, des philologues allemands commencèrent à montrer les faiblesses (litote) de la théorie traditionnelle (Weisgerber, Dornseif). Il revint à un philologue engagé dans des recherches ethnographiques de proposer le premier une reconstruction adaptée à l'ensemble des transformations théoriques de toutes les sciences sociales à partir de la fin du 19e siècle. (***Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes.Bd.1 Von den Anfängen bis zum Beginn des 13. Jahrhunderts*, Heidelberg, 1931**). Jost Trier introduisit d'un seul coup dans la sémantique tous les principes de l'analyse structurale, en parvenant d'entrée de jeu à articuler ces principes avec la prise en compte de l'évolution (ce que la linguistique structurale "à la Saussure" est loin d'avoir réalisé). Les paires fondamentales, langue-discours, synchronie-diachronie constituent le cadre, tandis que l'orientation principale est donnée par l'idée que ***l'objet d'étude est constitué par des ensembles lexicaux organisés et non par des mots isolés, intrinsèquement vides***. L'analyse du champ des capacités intellectuelles en moyen-haut-allemand demeure un modèle quasi insurpassé. Dès le départ était soulevée la difficulté la plus grave, par la distinction entre Sinnbezirk (zone de sens) et sprachliches Feld (champ lexical). L'hypothèse fondatrice était celle d'une correspondance assez étroite, d'où résultait l'idée simple et forte qu'une analyse structurale-historique d'un champ sémantique donnait un accès presque direct à l'évolution d'une partie du système de représentation de la société considérée.

A bien des égards, cette innovation était parallèle à celle de Roman Jakobson, élaborant la phonologie (structurale) par opposition à la phonétique (purement descriptive). La phonologie a connu un succès considérable (et mérité), tandis que les champs sémantiques sombraient dans un oubli à peu près général.

Trier eut peu d'élèves. On trouve un ensemble de textes importants et une bibliographie soignée dans un recueil paru en 1973 : Lothar SCHMIDT (éd.), *Wortfeldforschung. Zur Geschichte und Theorie des sprachlichen Feldes*, Darmstadt, 502 p. En France, les travaux de Trier (à ma connaissance jamais traduits) furent cités et commentés par deux linguistes, Georges Mounin et surtout Pierre Guiraud. Mais, là non plus, sans beaucoup de suite.

Par la suite, entre les années 60 et 80, un petit groupe de germanistes et de médiévistes tenta de mettre sur pied un secteur qu'ils intitulèrent *Bedeutungsforschung* (Friedrich Ohly, Max Wehrli, Hennig Brinkmann) ; ils optèrent d'ailleurs pour une pratique moins conceptualisée. Leur succès ne fut guère meilleur. En France, on ne peut citer qu'un tout petit nombre de travaux, très isolés et sans postérité malgré une très grande qualité, comme Jean CASABONA, *Recherches sur le vocabulaire des sacrifices en grec, des origines à la fin de l'époque classique*, Aix-en-Provence, 1966 (référence que je dois à J.P. Vernant)ou Régine ROBIN, *Histoire et linguistique*, Paris, 1973. Ni l'un ni l'autre ne citent les travaux allemands, dont ils paraissent pourtant extrêmement proches.

La nature de ce blocage est malheureusement facile à identifier : de tels travaux défiaient les distinctions traditionnelles, institutionnelles et fossilisées, entre littérature, linguistique et histoire. Les représentants de ces divers champs se liguèrent spontanément et efficacement pour faire barrage à une telle mise en cause. La validité et l'extrême intérêt des orientations proposées par Trier demeurent entières.

10.3.2 la socio-linguistique

A partir des années 50, en Grande-Bretagne surtout, mais aussi aux Etats-Unis, se développa une autre approche, essentiellement synchronique, mais remettant elle aussi vivement en cause l'autoenfermement de la linguistique : l'examen des rapports entre pratiques linguistiques et classes sociales. Pierre Bourdieu et son groupe contribuèrent à faire connaître ces travaux en France qui, bien entendu, laissèrent les linguistes de marbre. En français, Basil BERNSTEIN, *Langage et classes sociales. Codes socio-linguistiques et contrôle social*, Paris, 1975 ; William LABOV, *Sociolinguistique*, Paris, 1976. On peut citer aussi un recueil utile (non traduit), Pier Paolo GIGLIOLI (éd.), *Language and Social Context. Selected Readings*, London, 1972. Ensemble d'analyses précises, étayées, élaborées théoriquement, pour montrer que les faits de langue sont intrinsèquement des faits sociaux. Enfoncer des portes ouvertes ? pour filer la métaphore, on pourrait dire que ce sont en effet des portes largement ouvertes, mais par lesquelles personne ne passe. Tout historien a beaucoup à apprendre de ces ouvrages.

10.3.3 l'ethno-linguistique

Le succès de l'ethno-linguistique s'explique de la même manière, par des raisons symétriques : il est rapidement apparu que c'était seulement une branche de l'ethnologie, et tout le monde en fut rassuré.

On attribue généralement la mise sur pied de cette discipline à Benjamin Lee Whorf, ethnologue américain, spécialiste des amérindiens, et qui commença à publier des travaux sur diverses langues amérindiennes à partir de 1935 (repris dans B.L. WHORF, *Language, Thought, and Reality. Selected Writings of BLW*, New-York, 1956). On parle couramment de "l'hypothèse de Whorf", qui consiste à poser qu'il existe une relation directe entre une société, sa langue et sa représentation du monde. Des recherches particulièrement actives eurent lieu tout au long des années 50, 60, 70. L'intérêt pour ces travaux semble avoir nettement faibli depuis. Mais le corpus constitué est impressionnant. On peut citer un gros recueil assez large et représentatif : Dell HYMES (éd.), *Language in Culture and Society. A Reader in Linguistics and Anthropology*, New-York, 1964. Ou encore celui d'Edwin ARDENER (éd.), *Social Anthropology and Language*, London, 1971.

Les français tinrent leur place, par exemple Bernard Pottier ou Geneviève Calame-Griaule (*Langage et cultures africaines. Essais d'ethno-linguistique*, Paris, 1977).

On attribue à Floyd Lounsbury la méthode dite d'« analyse componentielle », qui est une forme de décalque sur un champ sémantique des principes de la phonologie de Jakobson (je simplifie). L'intérêt de ce type de pratique réside naturellement dans l'indication que la formalisation structurale permet de découvrir des éléments d'information décisifs, que l'on ne peut pas atteindre autrement.

Ce survol excessivement cursif est seulement là pour signaler des pistes de lecture et de réflexion, et amène à trois observations :
1. à partir des années 30 naquirent et se développèrent des champs de recherche conceptuellement très proches (un objet, le rapport entre langue et société, un outil principal, l'analyse structurale) ;
2. ces divers champs n'eurent cependant à peu près aucun rapport pratique, et furent en partie laminés (surtout dans le monde anglo-saxon) par la montée de l'idéologie postmoderne : si "tout est langue", à quoi bon examiner les relations entre langue et société ?
3. les historiens, qui, particulièrement ceux qui travaillent sur des périodes anciennes, ont à faire journellement à des textes incompréhensibles, auraient énormément à apprendre de ces expériences ; jusqu'à présent, les contacts sont demeurés ultraconfidentiels. C'est infiniment regrettable, mais rien n'est perdu, il suffirait de s'y mettre.

10.4. LES NOUVELLES TECHNIQUES : ÉTAT DE L'ART

10.4.1 une succession rapide et d'apparence désordonnée

Ici, l'histoire ancienne débute en 1945. On a évoqué précédemment les noms de N. Wiener, C. Shannon, G.K. Zipf. Il s'agissait de chercheurs ayant une solide formation mathématique, fortement intéressés par les constructions générales abstraites. Toutes proportions gardées, ils apparaissent comme des théoriciens.

Il s'agit à présent de jeter un coup d'œil du côté de la préhistoire de ce que, depuis une dizaine d'années tout au plus on a baptisé avec une emphase de très mauvais aloi les "nouvelles technologies", qui sont tout juste des techniques (quels que soient les coups de force institutionnels auxquels leurs tenants ont pu se livrer dans certains milieux universitaires).

Pour simplifier, on peut distinguer trois étapes :
a- **MT** (**machine translation**), l'époque de la traduction automatique, en gros 1945-1965. Les outils disponibles nous semblent rétrospectivement dérisoires, et ils étaient en effet très limités. Mais ils constituaient pourtant une nouveauté considérable et ont fait naître des espoirs démesurés. L'idée de traduction automatique semble avoir été évoquée nettement dès 1945. Les chercheurs des années 50 entreprirent de mettre sur pied un système complet et unifié d'« atomes de sens » pas trop nombreux, dont les combinaisons permettraient de rendre n'importe quel énoncé dans une sorte de métalangage formalisé universel. De nombreuses observations pertinentes furent faites, quelques simulations, mais on atteignit jamais la moindre réalisation (malgré de considérables apports financiers vers 1960, à l'époque dite de la "guerre froide").
b- **AI** (**artificial intelligence**), l'époque de l'intelligence artificielle et des systèmes experts (1965-1985 environ). Les moyens de calcul s'étant considérablement renforcés, les premiers langages informatiques virent le jour et se répandirent. Si les machines acceptaient un langage, on pouvait donc essayer de les rendre "intelligentes", en leur inculquant suffisamment de règles. On entreprit donc la constitution de bases d'expertise et de thesaurus gigantesques. Dans le domaine historique, quelques thesaurus furent constitués, qui ne fonctionnèrent jamais. (Plus tard, on décida dans

certains cas de se contenter d'une indexation beaucoup plus simple, et certaines opérations arrivèrent à termes, mais n'eurent pas de postérité du fait de l'obsolescence accélérée des matériels et de choix techniques aléatoires). Dans le domaine industriel ou médical, ce genre de perspective produisit d'intéressants prototypes, mais au prix d'un labeur considérable, alors même que l'évolution technique rendait le produit déjà insuffisant au moment de sa mise en service.
c- **IR** (**information retrieval**), depuis 1985 environ, recherche de l'information (je ne suis pas sûr de la "traduction"). Les années 80 virent l'apparition des microordinateurs, leur expansion fulgurante, l'accroissement exponentiel de leurs capacités (toujours en cours), puis la naissance et l'invasion planétaire d'internet, quasi instantanée à l'échelle historique. Invasion suivie d'une marche vers l'infini de la matière (textes, images, sons) rendue ainsi disponible. La plupart des ambitions abstraites antérieures furent définitivement balayées au profit de n'importe quel logiciel susceptible de faire face avec quelque efficacité à cette invasion. Toute l'information est sur le net, mais il faut la trouver, et la trouver tout de suite. Il n'est plus du tout question d'interroger des experts, il faut des algorithmes qui trouvent l'information. Les machines ayant des capacités (presque) illimitées, comment utiliser cette puissance matérielle pour trouver, c'est-à-dire cibler et sélectionner ce qui pourrait être ***pertinent*** ?? Les mots d'ordre se déclinent désormais ainsi : data mining, classification automatique, veille technologique, et l'adjectif magique est « unsupervised » : ça doit marcher tout seul.

Curieusement, reconnaissons-le, *ça ne marche pas trop mal*. Mais comment est-ce donc possible ?

10.4.2 essayons de comprendre (un peu)

C'est encore plus curieux quand on essaye de comprendre. En moins d'une quinzaine d'années, des équipes d'informaticiens ont **redécouvert** tous les vieux principes de la lexicographie et de la sémantique (sans s'en apercevoir, bien entendu). L'outil de base de la vieille lexicographie, **le contexte, est devenu l'arme absolue de l'information retrieval**, écartant complètement toute notion concurrente. En soi, cela n'est pas bouleversant. Mais ce qui devient subitement vraiment curieux pour l'historien ou le philologue, c'est que, dans le cas présent, ça marche tout seul, et que cela résoud chemin faisant des problèmes bien plus embrouillés que ceux auxquels se frottait (avec une efficacité très moyenne, comme on l'a rappelé plus haut) la lexicographie traditionnelle. Raison pour laquelle les historiens auraient peut-être quelque avantage à comprendre ce qui se passe derrière les écrans.

Dans les conditions présentes je n'ai ni la place ni surtout les connaissances nécessaires pour exposer ce sujet même superficiellement. Je dois me contenter de signaler quelques mots clefs qui apparaissent à jet continu dans tous les textes (proliférants) qui abordent ces questions. Je retiens principalement quatre acronymes (les anglophones préfèrent cette expression à "abréviation")
*** POS tagging**, i.e. part-of-speech tagging, soit : étiquettage des catégories de mots ; bien entendu unsupervised. Problème : comment trouver, sans dictionnaire, la catégorie grammaticale d'un mot inconnu ? Apparemment, des algorithmes assez simples y arrivent.
*** MSD**, i.e. morpho-syntactical descriptor, c'est-à-dire un peu plus compliqué que dans le cas précédent, puisque cette fois, il faut préciser (selon les langues) le genre, le nombre, la fonction et/ou le cas, temps, mode, etc. Question et réponse : les mêmes.
*** WSD**, word sense disambiguation. pas de traduction disponible à ce jour. Ici, "on brûle" : les informaticiens ont redécouvert l'opposition entre discours et langue ; ils n'ont que des mots, beaucoup de mots, mais ils doivent produire une information, quels que soient les mots. Il doivent donc régler, à la seconde, les problèmes, examinés depuis vingt-cinq siècles, de la polysémie et de la synonymie. Et il y a en effet des algorithmes qui le font assez bien.
*** LSA**, latent semantic analysis. Comme disait l'Anabase, *Thalassa, Thalassa* ! Le must actuel, c'est

l'analyse sémantique latente (avec un centre mondial à l'université de Boulder - Colorado). Ceux-là sont tout de même un peu plus sophistiqués, ils ont redécouvert l'analyse factorielle, et ils s'en sont même aperçus. Et c'est ce qui marche le mieux. En français (c'est moins cuistre), *ouf !!* nous retrouvons subitement un terrain connu. A ceci près que les tableaux traités ne sont pas du tout les tableaux bruts, mais des tableaux copieusement transformés à coups de logarithmes, pour tenir compte approximativement des biais colossaux provoqués par les distributions de Zipf (là encore, une redécouverte, qui sonne comme une solide confirmation).

Ce qui conduit à trois observations globales opposées : **1.** la rupture entre la troisième phase et les deux précédentes est forte : on passe de considérations générales sans effet concret à des algorithmes qui fonctionnent effectivement et convenablement ; et cette rupture correspond directement à l'abandon à peu près général des données a priori et des classements préétablis et fixes au profit de mises en ordre souples, résultant du traitement des données, et donc capables de s'adapter en permanence aux évolution rapides des corpus examinés ; **2.** bien que les programmes ne soient pas disponibles, les principaux algorithmes sont à peu près publics, tout uniment parce qu'une condition primordiale de leur efficacité tient à leur simplicité ; **3.** comme on l'a indiqué en commençant, les corpus des historiens (comme d'ailleurs les desiderata) sont notablement différents de ceux du net et de ses utilisateurs, il faudra de toute manière procéder à des remaniements profonds des procédures si l'on veut éviter de confondre la *summa* de Thomas d'Aquin et le *Wall Street Journal*.

10.5. POSSIBLES TRANSPOSITIONS : BILAN / PERSPECTIVES

10.5.1 évolutions récentes chez quelques historiens

Des développements substantiels ont eu lieu en Italie, que je connais très mal, mais sur lesquels il faut s'informer (Nevio Zorzetti, Andrea Bozzi). En Allemagne, comme on l'a signalé tout au début, la notion de *historische Semantik*, qui n'avait jamais disparu, produit des travaux utiles. Le petit manuel de **Gerd FRITZ, *Historische Semantik*, Stuttgart, 1998**, quoique partiel et partial, comporte beaucoup d'informations sous une forme condensée (lecture indispensable). On doit également souligner l'importance d'un volume à visée proprement historique : Rolf REICHARDT (éd.), *Aufklärung und historische Semantik. Interdisziplinäre Beiträge zur westeuropäischen Kulturgeschichte*, Berlin, 1998. Réflexions de niveau élevé de Ralf KONERSMANN, *Der Schleier des Timanthes. Perspektiven der historischen Semantik*, Frankfurt, 1994.

10.5.2 "histoire et informatique"

Sous cette dénomination, on trouve, notamment dans le domaine francophone, des travaux d'intérêt très variable, qui ont malheureusement en commun de prétendre contre tout bon sens de s'ériger en champ autonome : un outil technique (d'ailleurs en général mal maîtrisé) ne saurait en aucun cas constituer le fondement d'une activité proprement scientifique.

Depuis la fin des années 60, des "centres" se sont constitués, en général autour d'une personnalité, dans une université bien précise, visant la manipulation du lexique d'un domaine particulier, à l'aide de procédures en général assez élémentaires, parfois complètement erronées. De tels centres, voués aux langues anciennes ou médiévales se sont développés à Liège et Louvain ; des centres visant plutôt le vocabulaire français se sont constitués à Montréal, à Nancy, à Nice. Le seul groupe à visée nettement historique (que je connaisse) est celui qu'a constitué Maurice Tournier à l'ENS de Saint-Cloud. On doit à M. Tournier d'utiles travaux sur des documents des 19e et 20e siècle, et des chercheurs travaillant avec lui ont produit des résultats méritoires sur les textes de

l'époque de la Révolution française. M. Tournier est l'inventeur de la notion de ***lexicogramme***, notion particulièrement utile et pertinente, abandonnée en cours de route pour des raisons qui m'échappent.

Lorsque l'on essaye de prendre une vue globale de l'activité de ces centres, on est frappé 1. de l'absence à peu près complète de coopération, chacun s'est efforcé de réinventer la roue dans son coin ; 2. du caractère verouillé et non-disponible des programmes mis au point, dont on connaît tout au plus le nom, on ne sait même pas approximativement comment ils fonctionnent ; 3. de l'absence générale de réflexion globale abstraite sur les questions sémantiques fondamentales, non plus que sur les propriétés statistiques de base des distributions lexicales, absences d'où ont résulté des erreurs létales. Vus de l'extérieur, tous ces travaux ne paraissent guère susceptibles de déchaîner l'enthousiasme des historiens.

10.5.3 questions de portée générale

Les notions-outils les plus communs et les plus usuels posent toujours des problèmes graves, et des incertitudes complètes demeurent sur des aspects fondamentaux.

a) corpus. Le rapport langue-corpus, dont on a vu que les développements pratiques les plus récents font implicitement un usage massif, n'est pas une question que l'on peut prétendre écarter, comme l'ont fait la majorité des groupes cités au paragraphe précédent. Or cette notion même de corpus oppose de facto de redoutables problèmes techniques, du simple fait du caractère de ses structures statistiques essentielles (évoquées au chapitre précédent). Les informaticiens qui écrivent des programmes qui marchent se réfèrent à peu près tous à la "loi de Zipf", mais chacun a sa petite recette emprique pour en contrôler les effets. Le groupe de l'université de Leipzig a sa formule, le groupe du Colorado (LSA) a la sienne (avec forte présence de logarithmes) ; d'autres cherchent plutôt du côté des recettes éprouvées de la "statistique non-paramétrique", en traduisant tout en rangs, et en normalisant les distributions des rangs. L'idée qu'une analyse unsupervised est d'autant plus efficace que le corpus traité est plus homogène est exprimée partout, c'est le bon sens même. Mais comment traduire cette observation en pratique ?

b) dictionnaires. En informatique, la notion de dictionnaire a un sens tout à fait spécial : il s'agit d'une simple liste alphabétique, le plus souvent sur une seule colonne, moins souvent deux ou trois (listes munies de caractères). Cette notion, centrale dans les premières phases de l'évolution informatique, est devenue un simple outil *fluide*, reconstruit en permanence. Ce qui est d'une certaine manière rassurant pour les historiens (la notion-épouvantail de thesaurus a perdu toute sa superbe), mais ne laisse pas de poser des problèmes concrets qu'il faudra résoudre, car ils ne font pas partie de ceux que traitent les techniciens d'internet : l'omniprésence des variantes graphique, à elle seule, impose presque inévitablement le recours à une forme plus ou moins complète de lemmatisation ; or aucun des "groupes" cités plus haut n'a pris la peine de diffuser le moindre programme libre et ouvert d'aide à la lemmatisation. C'est une question informatique assez élémentaire, qu'il faudra traiter de toute urgence.

c) contexte et cooccurrences (en anglais collocations). Cela reste la question essentielle, une réflexion de base est un desideratum de premier rang. D'une part, une réflexion générale sur les variétés et leurs propriétés : contexte étroit, contexte large, domaine de pertinence, corpus spécialisés, etc. Comment se définissent concrètement et numériquement ces divers niveaux, comment les reconnaître de manière formalisée ? D'autre part, en quoi consistent les caractères statistiques des cooccurrences, dont on sent bien intuitivement qu'ils se situent à un degré de complexité supérieur à celui des distributions lexicales simples (dont on a vu de loin les difficultés qu'elles soulèvent déjà). Ainsi qu'on l'a signalé à l'instant, on trouve de-ci de-là quelques "recettes", il faudrait les examiner posément, les comparer, les comprendre.

d) formalisation morpho-syntaxique. Cette question est concrètement liée à celle de la

lemmatisation, mais abstraitement tout à fait différente. L'informatique "qui marche" n'utilise le tagging que comme un outil un peu accessoire pour la disambiguation. La fonction d'un mot dans une phrase n'est pas indépendante de son sens. Les informaticiens tirent de cette règle des conséquences minimales. Mon expérience d'historien la plus récente m'amène à penser que les historiens devraient au contraire en tirer des conséquences majeures : l'examen manuel d'une concordance brève (quelques centaines d'occurrences, mot-clé **KWIC, key-word in context**) devient extraordinairement efficace à partir du moment, et seulement à partir du moment, où les mots du contexte ont été classés selon leur fonction et leurs caractères grammaticaux. Pour un corpus de 100 ou 200 occurrences d'un lexème bien identifié, on peut opérer cette formalisation "à la main" en quelques heures. Le POS tagging ne paraît pas utiliser cette potentialité. Pourquoi ne pas essayer ? Comme on l'a déjà noté, les mots les plus difficiles et les plus intéressants sont les mots les plus ordinaires. En interrogeant des "bases de données textuelles", on récupère **des milliers d'occurrences**. C'est un matériau d'une extraordinaire richesse, un ***matériau clé pour l'avenir de la sémantique historique*** : aujourd'hui, personne ne sait quoi en faire, et on n'en fait donc rien.....

10.5.4 problèmes spécifiquement historiques

L'idée que toute "traduction" d'un texte médiéval soit impossible ou nécessairement plus ou moins fausse fait bondir **de rage** beaucoup de collègues.

C'est pourtant un fait empiriquement démontrable et vérifiable. Cela tient à une raison très simple, d'une portée incalculable : ***l'organisation des relations sémantiques dans une langue médiévale présente divers caractères propres totalement opposés aux caractères correspondants dans les langues contemporaines*** :

* le *multilinguisme* étrangement hiérarchisé des populations médiévales sachant lire et écrire entraîne eo ipso des ensembles de relations très importantes et très difficiles à reconstituer, en tout cas **sans aucun équivalent possible** ;
* le caractère fondamentalement non-historique de la représentation du monde dans l'Europe ancienne entraînait un usage général du latin comme d'une langue courante, hors de toute perspective évolutive : tous les mots latins (classiques) étroitement liés aux structures propres de la société antique continuaient d'être utilisés, sans plus de correspondant clair ; d'où des usages souvent en apparence incohérent de ce que je propose d'appeler les "***mots latins vides***", qui,autant que je sache, n'ont jamais donné lieu à une réflexion organisée ;
* les structures de base du système des représentations de l'Europe ancienne étaient complètement différentes des nôtres. (On peut s'en apercevoir même sans avoir lu Whorf). Du coup, des relations de base, essentielles, telles que ***la polysémie et la synonymie, se présentent dans les textes (au moins jusqu'au 18e siècle) d'une manière complètement différentes (sinon opposée) à ce qu'elles sont dans les systèmes des langues contemporaines***. L'idée que l'on pourrait trouver un quelconque équivalent est en dessous de la puérilité. Il y a là un champ de recherche immense, inexploité, dont la plupart des historiens ne semblent pas même soupçonner l'existence. Chartistes, au travail !!!

CONSIDÉRATIONS FINALES

Les notions clés, anciennes, de **contexte et de cooccurrences** demeurent plus que jamais les outils fondamentaux de la sémantique. Leur emploi massif et efficace par l'informatique la plus récente ne peut plus laisser planer le moindre doute sur l'intérêt majeur, pour les historiens, de reprendre à fond ce chantier, en tenant compte, enfin, de nombreux travaux particulièrement pertinents en dépit de l'obscurité dans laquelle les inerties et les barrières universitaires les ont laisser choir. Jost Trier devrait être une lecture de base de tout apprenti historien et de tout futur

conservateur.

Cette perspective est d'autant plus intéressante et urgente que la possibilité d'**écrire des programmes** permettant de mettre en œuvre toutes les hypothèses formelles est indiscutable. Les machines et les langages actuels (notamment les langages dits "de script", particulièrement adaptés aux traitements des chaînes de caractères) offrent des outils puissants d'emploi simple.

Qu'attend-on ?

Chapitre 11

STATISTIQUE LEXICALE ET ÉRUDITION

Après avoir examiné les caractères spécifiques des distributions lexicales et survolé les questions de cooccurrence, de contexte et de champ sémantique, nous nous trouvons en mesure d'essayer de voir en quoi des techniques appropriées fondées sur la statistique lexicale sont, ou pourraient être, susceptibles d'apporter une aide particulière aux historiens, notamment dans les domaines traditionnels de l'identification, de l'attribution et de la datation. Disons-le immédiatement : le potentiel, aujourd'hui encore presque inexploité, est considérable.

SOMMAIRE

11.1. TOPONYMIE

11.1.1 bref historique

Le terme d'onomastique paraît remonter en français au 16e siècle. Dès cette époque, de nombreux érudits munirent leurs publications d'index des noms propres. Comme on l'a rappelé dans le chapitre précédent, la lexicographie prit son essor à partir du 17e. Mais il fallut attendre le lent développement (en France) de la philologie dans la première moitié du 19e siècle pour voir apparaître les premières réflexions sur les noms de lieux. A cette époque, le "**celtisme**" était très largement dominant, et il jouait encore un rôle central dans les travaux du premier historien qui se fût vraiment intéressé à ces questions, Henri d'Arbois de Jubainville (1827-1910). On repère *l'adjectif "toponymique" en 1853 et le substantif "toponyme" en 1876.* Cet intérêt eut une conséquence très heureuse, le démarrage de la grande entreprise des *Dictionnaires topographiques* des départements : les premiers, sauf erreur, furent ceux de l'Eure-et-Loir (1861), de l'Yonne (1862), des Basses-Pyrénées (1863). Ces ouvrages furent presque tous le résultat du labeur des archivistes. On ne peut que regretter que cette entreprise, qui n'a pas couvert la moitié des départements, soit (presque) au point mort depuis un demi-siècle, sans que la Direction des Archives paraisse s'en émouvoir...

En France, la toponymie fut établie sur des bases philologiquement à peu près claires par l'enseignement d'Auguste Longnon (1844-1911) à la IVe Section de l'EPHE (ses travaux furent publiés après sa mort, *Les noms de lieux de la France*, Paris, 1920-1929). On lui doit également l'impressionnante série des *Pouillés des provinces de France*, conçue comme matériel toponymique, mais qui rend bien d'autres services. Cet enseignement fut poursuivi par Albert Dauzat (1877-1955), dont les ouvrages sont encore considérés comme des références. Dans le domaine francophone, on doit encore citer les noms d'Auguste Vincent, Paul Lebel, Charles Rostaing.

En Allemagne, les travaux fondateurs furent ceux de Ernst Förstemann, *Altdeutsches Namenbuch*, 1856 ss (*1. Personennamen, 2. Orts- und sonstige geographischen Namen*). Ils sont toujours réédités. A la fin du 19e et au début du 20e siècle, diverses tentatives furent faites pour tirer de la toponymie des arguments en faveur de telle ou telle théorie ethnique (tel toponyme, tel suffixe, seraient caractéristiques des Bavarois, des Alamans, etc) : travaux de Wilhelm Arnold et Sigmund Riezler notamment. Ernst Gamillscheg apporta une importante contribution à la toponymie du nord de la France. La dernière grande synthèse est celle d'Adolf Bach (*Deutsche Namenkunde. 1. Die deutschen Personennamen*, Heidelberg, 1943 ; *2. Die deutschen Ortsnamen*, Heidelberg, 1953). En Allemagne, la toponymie demeure un domaine très actif, ce qui n'est plus guère le cas en France (voir l'article "Ortsnamen" du *Lexikon des Mittelalters* ; il n'y a pas d'article "toponymie" dans le *Dictionnaire du Moyen Age*, de C. GAUVARD, A. de LIBERA, M. ZINK). En France, toponymie et microtoponymie sont du ressort de la philologie et spécialement de la dialectologie. On doit souligner l'importance et **le caractère très novateur des travaux de Gérard TAVERDET**, en particulier sa *Microtoponymie de la Bourgogne*, 12 tomes, Dijon, 1989-1993.

Les historiens français, au premier chef les médiévistes, sans avoir jamais procédé à un examen critique méthodique, sont plus ou moins parvenus à l'idée que les fameuses "couches" toponymiques (ligure, celte, gallo-romaine, germanique), même si l'on arrive à les distinguer, n'apportent en fait aucune information utilisable, ni quant aux déplacements de population, ni quant à la densité du peuplement à telle ou telle époque. Ces deux points sont avérés, le second en particulier. Mais cela n'enlève rien au fait que le relatif abandon de la toponymie par les historiens est très regrettable : il suffit de feuilleter un cartulaire pour constater l'attristante proportion de toponymes non identifiés, et l'on aimerait bien, de toute manière, comprendre, selon les époques et les zones, quels ont été les mécanismes sociaux par lesquels tels toponymes sont apparus ou ont disparu.

11.1.2 les sources

Pour toutes les localités importantes, et à l'échelle européenne, on peut toujours utiliser l'*Orbis latinus* de Johann Georg Theodor Graesse (Dresden, 1861) ; ce répertoire, numérisé, est **accessible gratuitement en ligne**, ce qui rend des services. On peut aussi, le cas échéant, jeter un coup d'œil dans la vieille *Topobibliographie* d'Ulysse Chevalier. Là où ils existent, les dictionnaires topographiques départementaux constituent un outil de travail fondamental.

L'Institut Géographique National a collecté tous les toponymes présents sur les cartes au 1/25000, et annonce avoir ainsi une liste de 1700000 toponymes. Mais quelques comparaisons suffisent pour montrer que, malgré leur relative richesse, ces cartes sont très incomplètes. En France, la base essentielle de documentation, la seule qui couvre de manière à peu près homogène tout le territoire, est constituée par les informations contenues dans les ***États des sections du "cadastre napoléonien"*** (établi en gros entre 1810 et 1840). On ne saurait assez répéter que les plans cadastraux eux-mêmes ne comportent qu'une partie de cette information. Il semble que 3000000 de microtoponymes soit un ordre de grandeur acceptable. Jusqu'à présent, *l'accès à cette information est très malaisé* : dans chaque département, les AD conservent (presque toujours) un exemplaire de l'État des sections de chaque commune, grand registre in-folio manuscrit, de manipulation délicate, de lecture parfois difficile ; les parcelles sont classées par "sections" et l'ordre numérique dans chaque section dépend des méthodes des topographes qui établirent ce cadastre. Il n'existe aucune récapitulation, il faut donc parcourir des dizaines de feuillets pour chaque commune. En principe, un second exemplaire identique est conservé dans chaque commune. Des campagnes de numérisation des plans ont été entreprises dans un certain nombre de départements, il serait hautement souhaitable que l'on numérise également les microtoponymes, en identifiant chacun d'entre eux à l'aide d'un géoréférencement simple (type NTF ou RGF93). Si une telle opération était menée à l'échelle nationale, on disposerait d'un instrument de travail d'une efficacité sans précédent. Une telle opération serait bien *plus simple, plus rapide et moins onéreuse* qu'une numérisation de plans (les deux opérations ne constituent pas une alternative !) ; la diffusion de cette information pourrait se faire facilement, en ligne et/ou sur CDrom.

Tous les historiens savent la quantité énorme de toponymes et microtoponymes que recèlent tous les documents fonciers qui s'étalent en gros entre le 14e et le 18e siècle, censiers, terriers et plans-terriers. Dans ce domaine, presque tout reste à faire.

11.1.3 caractères numériques

Les analyses numériques des toponymes - microtoponymes sont, à l'heure actuelle, rarissimes. La seule liste complète que je connaisse est celle des noms de communes formées d'un nom de saint (hagiotoponymes), liste donnée par J.-L. Lemaître (*Sources et méthodes de l'hagiographie médiévale*, Paris, 1993, pp. 194-195). Il s'agit d'une distribution de type parétien (apparence indiscutable d'une loi de Zipf-Mandelbrot) : 238 Saint-Martin, 171 Saint-Jean, et 649 hapax à l'autre extrêmité. Cette observation est tout à fait représentative de la situation générale : ***la distribution des toponymes est de type parétien***, type de distribution qui entraîne des conséquences de première importance quant aux procédures numériques applicables (et surtout : non applicables !). Sommes, moyennes, pourcentages sont instables par construction et ne peuvent pas être utilisés, à l'inverse de ce que beaucoup pensent pouvoir faire sans réfléchir. On ne saurait trop insister sur ce point, jusqu'ici complètement méconnu.

11.1.4 perspectives d'analyse

Durant les dernières décennies, des tentatives discrètes ont indiqué des voies intéressantes. Une recherche collective sur les microtoponymes traduisant des défrichements a abouti à montrer ce

qu'un balayage méthodique d'un champ délimité pourrait apporter. On voit immédiatement ce qu'un corpus vaste et homogène permettrait de mettre en lumière. Les travaux de Gérard Taverdet, signalés plus haut, réalisés manuellement, montrent bien ce qu'un travail systématique sur l'ensemble du matériel microtoponymique d'une région peut faire apparaître : un grand nombre de cartes font voir des séries remarquables de répartitions géographiques. On pressent ce que des procédures statistiques ad hoc permettraient de montrer, en particulier du point de vue des corrélations, positives ou négatives (opérations qu'il est impossible d'éxécuter à la main).

Schématiquement, on peut *distinguer trois types de recherches* :
a. des *analyses géographiques*, généralisant les méthodes traditionnelles de la géographie dialectale, en examinant méthodiquement les distances entre toponymes identiques, analogues ou complémentaires, en faisant apparaître les corrélations locales, en affinant la signification des toponymes par une mise en relation avec les autres données spatiales disponibles ;
b. des *analyses chronologiques* plus méthodiques, aussi bien en utilisant plus rationnellement les matériaux présents dans les corpus des diverses périodes (ce qui devrait permettre de préciser plus nettement les dates d'apparition de telle forme, de tel suffixe, de tel type de désignation), qu'en réexaminant sur une base plus solide les fameuses "couches" ;
c. en essayant, par combinaison de divers types de sources, de *faire apparaître des modalités de transformation*, tant ponctuelles (tel ou tel type) que globales (transformation du **système toponymique** lui-même). A cet égard - on y reviendra - , il serait temps de se départir de l'idée naïve d'une opposition permanente entre "noms communs" et "noms propres", tout autant que de l'idée d'un emboîtement tout aussi permanent entre "toponymie" (= lieux habités) et microtoponymie (= lieux-dits). Pourquoi et quand se fixent des syntagmes comme "moulin neuf" ou "maison blanche", et à quel moment, dans quelles conditions ces expressions deviennent-elles des "toponymes" ? On ne saurait en aucun cas perdre de vue que la moitié au moins des toponymes (toutes catégories confondues) résistent de facto à toute recherche d'étymologie. On n'a jamais tenté d'utiliser de manière systématique (donc statistique) ce matériau délaissé.

Plus généralement, lorsqu'il s'agit d'autre chose que de pures listes (rarissimes avant le 19e siècle), les toponymes apparaissent toujours dans des contextes, c'est-à-dire liés à d'autres termes, à valeur topographique ou non (Nicolas HURON, *Termes de topographie urbaine dans les actes des rois de France 840-987*, Paris, 1990), ainsi qu'à une terminologie de définition ou d'énonciation (in loco qui dicitur, villa nuncupante, etc) ; les toponymes sont d'abord un sous-ensemble du champ sémantique de l'espace, il y serait urgent de reprendre toute la question sous cet angle.

11.2. ANTHROPONYMIE

11.2.1 historique

En Allemagne, les grands ouvrages d'onomastique traitaient tant les anthroponymes que les toponymes (Förstemann, Bach). En France, on note au contraire un net décalage : *l'anthroponymie est apparue plus tardivement* ; si la notion de "toponymie" date des années 1850, "anthroponymie" apparaît seulement en 1938. Dauzat et Rostaing tentèrent de trouver une étymologie à la plus grande partie des patronymes français. Comme le fait remarquer D. Geuenich dans l'article "Personennamenforschung" du *Lexikon des Mittelalters*, cette question de l'étymologie est plutôt une affaire de dilettantes. Dans les catalogues de bibliothèque, la vedette "anthroponymie" déclanche une avalanche de titres du genre armoriaux, nobiliaires et autres productions "généalogistes". Même pour de simples identifications, l'utilité de tels ouvrages est mince.

Des travaux ont été consacrés à l'onomastique dans l'œuvre de la plupart des "grands auteurs", à commencer par ceux de l'Antiquité. Hans von KAMTZ, *Homerische Personennamen :*

sprachwissenschaftliche und historische Klassifikation, Göttingen, 1982. David Roy BAILEY, *Onomasticon to Cicero's treatises*, Stuttgart, 1996. En fait, c'est tout le système de dénomination qu'il faut considérer : Heikki SOLIN, *Namenpaare : eine Studie zur römischen Namengebung*, Helsinki, 1990. Olli SALOMIES, *Adoptive and polyonymous nomenclature in the Roman Empire*, Helsinki, 1992. Les médiévistes sont confrontés à la question difficile (et non encore résolue) du passage du système de dénomination romain au système du haut Moyen Age, complètement différent. Dieter GEUENICH, Wolfgang HAUBRICHS, Jörg JARNUT (éds), *Nomen et gens : zur historischen Aussagekraft frühmittelalterlicher Personennamen*, Berlin, 1997 ; ID, *Person und Name : methodische Probleme bei der Erstellung eines Personennamenbuches des Frühmittelalters*, Berlin, 2002.

Citons au passage des répertoires onomastiques de la littérature médiévale : Louis-Ferdinand FLUTRE, *Table des noms propres avec toutes leurs variantes figurant dans les romans du Moyen Age écrits en français et actuellement publiés ou analysés*, Poitiers, 1962. André MOISAN, *Répertoire des noms propres de personnes et de lieux cités dans les chansons de geste françaises et les œuvres étrangères dérivées*, Genève, 1986. Christopher W. BRUCE, *The Arthurian name dictionary*, New-York, 1999. Les travaux d'orientation vraiment historique sont rares : Pascual MARTINEZ SOPENA (éd.), *Antroponimia y sociedad : sistemas de identificación hispano-cristianos en los siglos IX a XIII*, Santiago de Compostela, 1995 ; Antoine VALLET, *Les noms de personnes du Forez et confins (actuel département de la Loire) aux XIIe, XIIIe et XIVe siècles*, Paris, 1961. Friedhelm DEBUS (éd.), *Stadtbücher als namenkundliche Quelle*, Mainz, 2000.

Si quelques auteurs ont eu l'idée qu'il existe, dans chaque société et à chaque époque, *un véritable système de dénomination des individus*, cette idée semble avoir ***très rarement entraîné une enquête large de tout le champ sémantique concerné***, et les indispensables analyses statistiques sont demeurées confidentielles.

11.2.2 analyses numériques

Le travail le plus clair est **un long article du sociologue Jacques Maître sur les noms de baptême en France** : "Les fréquences des prénoms de baptême en France", *L'année sociologique*, 3-1964, pp. 31-74. On peut y joindre Michel TESNIERE, "Fréquence des noms de famille", *Journal de la société de statistique de Paris*, 116-1975, pp. 24-32. Pierre DARLU, Anna DEGIOANNI, Jacques RUFFIÉ, "Quelques statistiques sur la distribution des patronymes en France", *Population*, 3-1997, pp. 607-633.

Un point général est hors de discussion : **toutes les distributions d'anthroponymes sont de type parétien**, Jacques Maître conclut à la possibilité d'un ajustement sur une loi de Zipf-Mandelbrot. Ce qui implique que l'usage des statistiques gaussiennes dans ce domaine produit inévitablement *des séries d'erreurs graves*. On en prendra deux exemples.

Louis Duchesne est l'auteur, entre autres, des *Fastes épiscopaux de l'ancienne Gaule*, 3 vol., Paris, 1894-1915. Si l'"hypercritique" de cet auteur a été notée, ce fut sans analyser le substrat des hypothèses de cet érudit, racine de ses erreurs. L. Duchesne s'imaginait les évêques mérovingiens comme des bourgeois de la fin du 19e siècle, attentifs à ce qu'une confusion entre deux individus fût impossible. Pour lui, le nom était simplement le signe de l'identité, à peu près au sens que le code civil ou la police donnent à cette notion. Il fut ainsi conduit à écarter plus ou moins brutalement de ses propres listes tel ou tel personnage, en arguant du fait qu'il portait le même nom qu'un évêque d'un diocèse voisin, à peu de distance chronologique. Double erreur : d'abord le nom au 6e siècle n'avait que de lointains rapports avec le nom à la fin du 19e et, surtout, les noms des évêques de cette époque, comme de n'importe quel autre groupe, étaient répartis selon une distribution parétienne, qui entraînait la présence d'une importante quantité d'homonymes, ce qu'attestent d'ailleurs les listes de souscripteurs des "conciles mérovingiens".

Monique Bourin a organisé et publié une série de rencontres *Genèse médiévale de l'anthroponymie moderne*, 4 vol en 6 t., Tours, 1989-1997. L'objectif, indiqué clairement d'entrée de jeu, était de préciser les modalités du passage, aux 10e-12e siècles, du système "à un nom" du haut Moyen Age au système moderne "à deux noms". Les enquêtes réunies concernent la France et la péninsule ibérique. Au plan statistique, on se bornera à deux remarques : 1. les listes fournies (en nombre d'ailleurs modeste) sont presque toutes tronquées ("les noms les plus fréquents", "palmarès") ; 2. à de multiples reprises sont calculés des pourcentages, et lorsque les effectifs bruts sont indiqués, il apparaît le plus souvent que lesdits effectifs sont sensiblement différents. Du premier point découle inévitablement qu'aucune étude des distributions ne peut être effectuée sur ces bases puisque, comme on l'a rappelé dans un chapitre précédent, les distributions parétiennes sont tout autant caractérisées par les hapax que par les fortes fréquences. A propos du second point, on rappellera de nouveau que, les effectifs "globaux" de ces distributions n'étant pas dénombrables, les pourcentages dépendent fortement de l'effectif considéré et sont instables par construction : on ne peut leur accorder la moindre valeur. L'ensemble des considérations numériques contenues dans ces quatre volumes part de prémisses fausses, toutes les analyses sont entièrement à reprendre sur d'autres bases. Au demeurant, on est frappé, en lisant la conclusion générale, de retrouver pour ainsi dire inchangées les observations de départ : on est passé, entre le 10e et le 12e siècle, d'un système "à un nom" à un système "à deux noms", selon des rythmes et des chronologies présentant quelques variantes selon les régions...

On voit là deux exemples de ce que l'on avait énoncé dans le chapitre d'introduction : en matière numérique, le "bon sens" est le pire des conseillers, il vaut mieux pas de statistique du tout qu'une statistique naïve, inévitablement trompeuse.

11.2.3 distinguer les personnes

L'ethnologue le moins averti sait que, dans son enquête de terrain, il doit s'efforcer de relever, pour le plus possible d'individus, *les termes d'adresse et les termes de désignation*. Dans la vie courante d'une population "traditionnelle", il s'agit de deux situations bien distinctes, et les termes utilisés ne se recouvrent que très partiellement. Même sans adhérer aux théories linguistiques dites "pragmatistes", tout le monde sait très bien que, dans notre société même, où sont mises en œuvre des procédures de repérage d'une complexité jamais atteinte auparavant, *le même individu est appelé ou désigné, selon les circonstances, de multiples manières*. Et c'est cet ensemble qui doit être pris en compte par le sociologue ou par l'historien et non pas le seul "numéro de Sécurité Sociale", pourtant parfaitement univoque et suffisant. La nécessité manifeste d'éviter les "erreurs sur la personne" est, dans son principe, quasi universelle ; mais, dans la pratique, elle relève toujours d'une situation concrète spécifique : et les spécificités dépendent directement de l'organisation et du fonctionnement de la société considérée. Autrement dit, la possibilité de confusion est propre à telle ou telle configuration sociale ; avant de s'employer à reconstituer des évolutions, on doit tenter de comprendre comment tel "système d'identification" (comportant l'ensemble des éléments évoqués à l'instant) correspondait fonctionnellement à tel "système social" : c'est le cadre de réflexion minimal. Une longue liste de professions relevée dans l'état-civil du 19e siècle est de type purement parétien (A. GUERREAU, "A propos d'une liste de dénominations professionnelles dans la France du XIX^{e} siècle", *Annales E.S.C.*, 48-1993, pp. 979-986).

On ne saurait trop insister sur le fait que toute statistique historique ne peut être qu'un instrument parmi d'autres au service d'une réflexion rationnelle sur telle ou telle société passée. Dans l'état actuel, les médiévistes sont incapables d'expliquer aussi bien le système d'identification du haut Moyen Age que celui du bas Moyen-Age. Ce qui constitue un indice hautement significatif de l'extrême faiblesse (litote) des représentations que les historiens se font aujourd'hui des structures

de la société médiévale. Des outils statistiques congruents avec les structures analysées peuvent être d'un grand secours, voire indispensables. A condition que l'on en comprenne la logique ; dans le cas présent, *il faut commencer par saisir suffisamment comment tout système d'information correspond aux énoncés de Shannon, Zipf, Mandelbrot et quelques autres*. En croyant pouvoir contourner cette difficulté, on tombe dans le n'importe quoi.

11.3. NOTES GÉNÉRALES SUR L'ONOMASTIQUE

Il importe, avant de se pencher sur un autre groupe de problèmes, de reformuler de manière synthétique quelques remarques concernant l'ensemble des "noms propres" :
a. les noms propres dérivent à peu près tous, à l'origine, de noms communs ; selon les langues et les sociétés, ce sens d'origine est encore perceptible, peut être restitué, ou est devenu strictement opaque. En pratique, *cette différence n'est pas pertinente*. Simplement, toute société humaine dispose d'un sous-système linguistique qui lui permet de ne pas confondre les objets qu'elle juge non-permutables : système de repérage et d'identification dont la complexité est en gros proportionnelle à la complexité de la société elle-même
b. il faut surtout se souvenir que ce système a comme fonction de base de s'adapter aux diverses situations qui nécessitent son emploi. *Les noms propres sont toujours employés dans un contexte*, ce qui tendrait à étayer l'idée que les anthroponymes ne sont qu'une partie du champ sémantique de l'idenfication, tandis que les toponymes sont un sous-ensemble du champ sémantique de l'espace. Il n'est certes pas inutile de dresser des index de noms propres d'un cartulaire ou d'une chronique. Mais il peu probable qu'une analyse ne prenant en compte que ces vocables, à l'exclusion de leur contexte, puisse parvenir à des conclusions utiles.
c. *le paradoxe apparent résultant de la constatation que les noms propres sont toujours répartis selon des distributions de Zipf-Mandelbrot provient précisément de la non prise en compte de ce dernier point*. En réduisant toponymie ou anthroponymie à des listes de termes uniques, on se heurte immédiatement à une aporie : **si le système fonctionnait sur cette base laminée**, la confusion serait générale. Or toute expérience montre que ce n'est pas le cas. Dans toute situation concrète, des éléments que l'on pourrait dire "atomiques" sont combinés selon des règles de hiérarchie et d'association qui permettent de transmettre une information compréhensible. On en vient à se demander dans quelle mesure ce que nous appelons des "noms propres" ne sont pas d'abord des signes incluant une part suffisamment forte d'arbitraire pour les faire reconnaître instantanément comme des "marqueurs" de la non permutabilité des objets auxquels ils s'appliquent, avant même d'être utilisés comme des outils de désignation (représentation du sens commun).

11.4. ATTRIBUTIONS

11.4.1 un problème de masse

Sans aller chercher plus loin, il suffit de considérer la liste des œuvres prises en compte dans le dictionnaire de Blaise ou celui de Niermeyer pour avoir une première idée du nombre de textes considérés comme "pseudo-X" ou complètement anonymes. (Le répertoire *Clavis patristica pseudepigraphorum medii aevi* a commencé de paraître en 1990, cinq volumes parus en 2003. Sur les diverses méthodes d'attribution, Jacques BERLIOZ et al., *Identifier sources et citations*, Turnhout, 1994). Sans compter les innombrables mentions "authenticité douteuse" (je ne rappelle que par provocation les cas désespérés célèbres, tel texte de Grégoire le Grand, tel texte

d'Abélard...). Notre notion d'« auteur » n'existait pas au Moyen Age, et une infinie quantité de textes ont circulé sans nom d'auteur ou avec une attribution inexacte. Si le jugement que l'on porte sur ce genre de pratique a évolué, la pratique elle-même a perduré, et risque fort de se perpéturer longtemps encore. Le nombre de personnes célèbres qui signent des ouvrages qu'elles n'ont pas écrits ne semble pas en diminution. Inutile d'insister : la question touche toutes les périodes et tous les genres.

11.4.2 principes de formalisation

Le principe de base est élémentaire : il revient à **combiner deux constatations**. 1. il est possible, dans certaines conditions, de repérer *les « constantes chiffrées du discours » caractéristiques de tel ou tel auteur* ; 2. en utilisant des procédures numériques ad hoc, on peut *calculer la « distance » séparant ou rapprochant un texte anonyme ou douteux d'un ensemble de textes clairement attribués*.

Ce qui revient en pratique à mettre en relation ce que nous avons vu à propos de "distributions lexicales" et de "distributions multivariées". Le principe est simple, mais on ne peut pas dire que les méthodes soient éprouvées.

11.4.3 des expériences encore peu nombreuses

Une incertitude forte entoure encore la question principale : quels sont donc les éléments stables qui sont spécifiques d'un auteur ? A ma connaissance, une réponse s'impose : nous n'en savons rien. En gros, on pourrait distinguer trois catégories d'informations :
a. *des indices généraux*, le plus connu étant l'indice de "richesse du vocabulaire". Nous avons vu précédemment que, jusqu'à aujourd'hui, on ne sait pas calculer un indice qui soit stable quelle que soit la longueur du texte/corpus considéré. Il reste que l'on peut toujours faire un calcul simple sur des fragments de même longueur et tenter une comparaison. Il faudrait commencer par opérer des essais nombreux et variés sur des textes bien attribués pour définir, au moins empiriquement, ce que l'on peut attendre d'une telle procédure. Je ne connais pas de telle série de tests.
b. des indices portant sur *des proportions non lexicales*, par exemple fréquences des cas, des formes verbales, proportions du texte dans les principales et les subordonnées, longueur des phrases, etc. Quelques calculs de ce genre tendent à montrer que de tels indices varient fortement selon le type de texte : prose ou vers, roman ou théâtre, etc. Ce que l'on conçoit aisément : la deuxième personne risque d'être plus fréquente dans une pièce de théâtre que dans un article de journal.
c. *des indices purement lexicaux*. Ici, une série de choix doivent être faits : les "formes" ou les "lemmes", tous les mots, seulement les plus fréquents ou seulement les plus rares ?

Une recherche intéressante est celle de Sylvie MELLET, "La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte ?", *Médiévales*, 42-2002, pp.13-26. Des textes de classiques latins ont été soumis à une série de tests. Il en ressort assez nettement que les textes lemmatisés ou non lemmatisés donnent à peu près les mêmes résultats (assez bons), mais que les indices calculés sur des fréquences grammaticales sont beaucoup plus incertains. Quelques essais personnels me laissent supposer que *les 50 mots les plus fréquents permettent d'obtenir des regroupements pertinents*, même avec des textes traitant de sujets différents. Je serais assez tenté de penser que les "mots-outils" (quelle que soit la manière dont on les définit) pourraient constituer un des marqueurs les plus efficaces, dans la mesure où ils dépendent peu du genre ou du sujet. Il faudrait entreprendre des séries de tests suffisantes. C'est à cette conclusion qu'est parvenu Éric Brian, qui a réussi avec cette méthode à distinguer deux auteurs dans un ouvrage français de la fin du 18e siècle (il a utilisé 122 mots-outils pour sa discrimination : E. BRIAN, "Moyens de connaître les plumes. Études lexicométriques", in *Moheau. Recherches et considérations sur la population de la France (1778)*, Paris, 1994, pp. 383-396).

La méthode en apparence la plus simple consiste à procéder à une analyse des correspondances sur un « tableau lexical entier », c'est-à-dire simplement un tableau d'effectifs, croisant les formes (par exemple en ligne) et les textes ou fragments (en colonnes). En principe, deux distributions analogues correspondront à des points voisins, et inversement. En prime, on verra apparaître les mots les plus caractéristiques de tel ou tel fragment. L'avantage principal de cette procédure est qu'elle est exécutable sans intervention du chercheur sur les listes (voir par exemple le très intéressant chapitre "Analyse discriminante textuelle" dans Ludovic LEBART & André SALEM, *Statistique textuelle*, Paris, 1994, pp. 241-282).

Ici, je dois énoncer **une très forte mise en garde** : une telle procédure n'est utilisable que si les colonnes (= les textes traités) ont des effectifs à peu près équivalents. Dans ce cas, les résultats sont acceptables. Au contraire, **si l'on considère simultanément des textes de longueurs variables (on voit des tableaux où un seul texte a un effectif supérieur à tous les autres réunis), le résultat est faux par construction** : comme on l'a montré et déjà répété, les effectifs relatifs varient sensiblement en fonction de la longueur du texte, et l'analyse factorielle compare alors des objets qui ne sont pas du tout comparables. La catastrophe est garantie. Malheureusement, je n'ai jamais lu une telle mise en garde où que ce soit, et vu, de ci de là de longs commentaire d'analyses factorielles inconsistantes... La plus élémentaire précaution consisterait au moins à découper un texte plus long que les autres en fragments, de manière à rétablir l'équilibre numérique entre les colonnes.

Il reste que l'on est en droit de se demander si l'on peut légitimement traiter par analyse factorielle ces tableaux lexicaux entiers en considérant simplement les effectifs bruts. Quoi qu'ait pu écrire tel ou tel, les analyses factorielles reposent peu ou prou sur l'hypothèse sous-jacente de distributions normales ou subnormales. La question est donc : *serait-il possible d'appliquer aux tableaux de fréquences brutes des transformations qui tendent plus ou moins efficacement à "corriger" la forme parétienne des distributions ?* Nous revenons ici à ce qui a été dit à propos des diverses procédures de transformation, capables de produire des distributions centrées et symétriques. Il me paraît utile de résumer ici succinctement la proposition des initiateurs de la LSA (latent semantic analysis) ; voir notamment Thomas K. LANDAUER, Peter W. FOLZ & Darell LAHAM, "An Introduction to Latent Semantic Analysis", *Discourse Processes*, 25-1998, pp. 259-284 (texte disponible sur internet). Ces auteurs font subir au tableau d'effectifs une double transformation :

a. on remplace chaque effectif x par x1 = log (x+1) (ce qui permet de retrouver 0 dans les cases à effectif nul) ;

b. on applique la formule de Shannon à chaque case, ce qui veut dire que 1. on calcule l'entropie de chaque mot dans sa colonne (texte) (-p log(p)), 2. on fait la somme des entropies de chaque ligne (ce qui permet d'évaluer la "quantité d'information" véhiculée par le mot en question), 3. on pondère la valeur x1 de chaque case par cette somme en ligne (simple division). Au total, il y a donc une double pondération, qui, théoriquement, affecte à chaque case une valeur proportionnelle à la "quantité d'information" qu'elle paraît contenir. Un tel calcul se programme en quelques lignes. Il faudrait effectuer des tests assez nombreux, dans des cas variés, pour pouvoir examiner empiriquement l'effet d'une telle procédure. **Tout est à faire !!**

Comme dans toute analyse factorielle, on doit procéder à une série de choix pratiques : quels sont les éléments actifs et les éléments supplémentaires ? Dans le cas d'une recherche d'attribution, faut-il comparer le texte indéterminé à une série de textes d'un même auteur, ou de deux auteurs (voire davantage), et en quelle proportions ? De nombreuses séries de tests apparaissent comme une tâche prioritaire.

11.5. DATATIONS

11.5.1 une question permanente, symétrique de la précédente

La difficulté est peut-être encore plus lancinante que la précédente. Chacun, selon ses expériences, a eu l'occasion de rencontrer des séries de documents mal datés, ou pas datés du tout. Il est inutile d'insister. Mais il faut bien voir d'entrée de jeu que la recherche vise ici des éléments en partie complémentaires des précédents : quels sont les indices subissant une évolution plus ou moins régulière, que l'on puisse repérer et sur lesquels on puisse donc se fonder pour ordonner des séries ? Il est important de bien saisir que **l'on tend ainsi à classer tous les indices disponibles en trois groupes** : ceux qui, dans des conditions données, sont à peu près **stables** (pour un auteur, cela paraît plausible, pour une période, une zone, c'est peut-être moins net) ; ceux qui, dans un laps de temps donné, subissent une **évolution tendancielle** à peu près régulière ; enfin ceux qui dépendent essentiellement des **circonstances** et du contexte. A ma connaissance, une telle décomposition n'a jamais été théorisée. Ce qui d'ailleurs se comprend : pour y songer, il faut être historien et avoir aussi des notions suffisantes de philologie et de statistique. On voit immédiatement qu'une telle question comporte une part technique importante, mais qu'elle a de fortes implications sur la réflexion à propos des rapports entre langue et société, dans une perspective historique (= fonctionnement / évolution).

11.5.2 l'empire du scalogramme

Une pratique désormais à peu près balisée consiste à produire des scalogrammes à partir de suite de textes (ou de corpus) à peu près homogènes se suivant à intervalles réguliers. Jacques Guilhaumou a fait œuvre de pionnier en traitant une série de sous-ensembles de textes du *Père Duchêne* au milieu des tempêtes révolutionnaires ("L'historien du discours et la lexicométrie", *Histoire & Mesure*, 1-1986, pp. 27-46). D'autres ont saisi des séries d'éditoriaux de tel ou tel journal récent. Une pratique maintenant courante consiste à appliquer cette méthode à des discours politiques assez formalisés (genre déclarations d'investiture d'un gouvernement). Les **tableaux lexicaux entiers** ont produit *dans tous les cas un très beau scalogramme* (alias "effet Gutmann"). Autrement dit, les "points-textes" sont situés le long d'une courbe arquée régulière, et se suivent dans l'ordre chronologique, avec peu d'irrégularités. Le principal bénéfice de ces analyses est de permettre de repérer les formes les plus caractéristiques de chaque période ou, encore mieux, les formes qui apparaissent et celles qui disparaissent à chaque étape. Insinuons-le : il s'agit rarement de révélations ; si l'on connaît l'évolution politique dans le milieu considéré, on ne fait guère mieux que retrouver ce que l'on savait depuis longtemps.

Si l'on examine les caractères numériques de ces tableaux, on s'aperçoit que, dans tous les cas, les textes ou corpus mis en œuvre sont de longueur équivalente. Les sujets sont à peu près homogènes, seule la thématique évolue progressivement. Les intervalles sont eux aussi à peu près constants, et la durée totale considérée dépasse rarement une vingtaine d'années (souvent bien moins). Pour simplifier : au sein d'un ensemble globalement stable, seuls quelques mots-clefs sont remplacés par d'autres, avec des rythmes divers, mais de manière tendanciellement uniforme. C'est un acquis qu'il importe de connaître. A partir de là, **deux questions** peuvent être soulevées : 1. peut-on, de manière assez fiable, inverser la question ; c'est-à-dire connaissant une suite ainsi datée et chaînée, repérer la position relative d'un nouvel élément, appartenant au corpus mais dont on ignore la date ? 2. quelles sont les limites pratiques d'une telle méthode, c'est-à-dire à partir de quel degré d'hétérogénéité la méthode commence-t-elle à flancher (durée, espace, genre et thématique) ?

11.5.3 variété des cas à traiter

La difficulté centrale demeure : **comment traiter des textes de longueurs différentes, en particulier de petits textes ?** Pour donner un ordre de grandeur, la plupart des exemples que je connais renvoient à des colonnes (partie de corpus) d'effectifs compris en 10000 et 50000 formes. Peut-on imaginer traiter des chartes (moins de 1000 mots), des lettres ou des sermons (moins de 3000 mots) ? Des transformations du type de celle de la LSA évoquée plus haut peuvent-elles constituer une parade ?

D'un autre côté, beaucoup de textes mal datés ou réputés indatables (type vies de saints mérovingiens) présentent, apparemment, une sensible hétérogénéité intrinsèque : copie de passages entiers , "centonisation", interpolations, etc (Nicholas BROUSSEAU, "Lemmatisation et traitement statistique : de nouveaux instruments pour la critique diplomatique ? Le cas des diplômes pseudo-originaux au nom de Louis le Germanique", *Médiévales*, 42-2002, pp. 27-44). Il est indéniable que certains types de textes très formalisés (notamment en vers), surtout quand existe une volonté de ne pas s'écarter d'un modèle (idéal médiéval !), peuvent a priori sembler rebelles à tout traitement de ce genre. A ma connaissance, il n'y a eu aucune expérience sérieuse, encore moins de démonstration. Les expériences auxquelles fait allusion le paragraphe précédent montrent une méthode "qui marche". Tant que l'on n'aura pas établi assez clairement les limites de validité de cette méthode, on ne pourra pas aller bien loin. Nous en sommes exactement au stade où de gros corpus sont assez commodément disponibles (c'est récent), et où les machines les plus courantes ont acquis des capacités de traitement largement au-dessus de ce qui est indispensable pour traiter des corpus de plusieurs millions de mots (c'est récent également). Dans ces conditions, c'est aux historiens de procéder à toutes les expériences imaginables, personne ne le fera à leur place.

11.5.4 une figure spécifique : classer les manuscrits d'une œuvre

Tout chartiste connaît cette question fondamentale de l'érudition classique : comment mettre le meilleur ordre possible au sein d'un ensemble de manuscrits du même texte ? Tout le monde connaît les noms de Lachmann et de Bédier. On cite moins, et c'est une grosse erreur, les travaux de **Dom Quentin, de Dom Froger et de l'abbé Duplacy**. Ces bons pères ont entrepris depuis les années 20 une réflexion sur la formalisation de la comparaison des variantes. Un colloque organisé en 1978 a permis de faire un point assez complet sur ces méthodes, d'où ressort clairement la supériorité de l'analyse factorielle des correspondances (Jean IRIGOIN & Gian Piero ZARRI (éds), *La pratique des ordinateurs dans la critique des textes*, Paris,1979). J'ai proposé un exemple et quelques réflexions complémentaires : A. GUERREAU & Marie-Anne POLO, "Classement des manuscrits et analyses factorielles : le cas de la *scala coeli* de Jean Gobi", *Bibliothèque de l'École des Chartes*, 154-1996, pp. 359-400. Une formalisation bien conçue et l'utilisation de méthodes factorielles simples permettent de regrouper aisément un très grand nombre de manuscrits en se contentant d'examiner méthodiquement quelques dizaines de "lieux variants" (sur cette question, on a dépassé le stade des expériences...)

CONSIDÉRATIONS FINALES

En toute clarté ressortent deux conclusions strictement liées : 1. des techniques archiclassiques comme la datation, l'identification, l'attribution, le classement des manuscrits peuvent être fortement renouvelées par l'utilisation de méthodes appropriées de statistique lexicale ; 2. l'efficacité de procédures de ce genre est totalement dépendante du réalisme des statistiques employées, qui ne peuvent en aucun cas appartenir aux "statistiques classiques" car, dans ce

domaine, toutes les distributions sont de type parétien.

Accessoirement, ceci : un tel ***renouvellement des sciences auxiliaires***, riche d'un considérable potentiel, ne pourra venir que des historiens eux-mêmes.

Chapitre 12

CALCULS ET MESURES AVANT LE SYSTÈME MÉTRIQUE

Une large proportion des documents d'archives (depuis le 3e millénaire avant notre ère) utilise des notations numériques : beaucoup étaient d'ailleurs essentiellement conçus pour conserver une trace tangible de dénombrements. C'est pourquoi tout historien devrait avoir des notions assez détaillées sur les systèmes de décompte, de mesure et de calcul employés dans les civilisations qu'il étudie. Or c'est exactement le contraire que l'on observe. L'enseignement de l'histoire, même à son niveau le plus élevé, n'aborde ces questions qu'exceptionnellement, de manière latérale et fragmentée. Numération et calculs appartiennent à «l'histoire des sciences », tandis que la métrologie historique n'est qu'une sous-partie peu prisée de « l'histoire des techniques ». Or même les illettrés savent compter (jusqu'à un certain point), et l'on ne voit guère comment pourrait fonctionner un quelconque système de mesures qui ne s'appuierait pas sur des méthodes de calcul, si primitives soient-elles. Je donnerai ici quelques informations générales sur l'évolution en Europe depuis la fin de l'Antiquité, en essayant de faire ressortir les articulations majeures.

SOMMAIRE

12.1. NOMBRES ET NUMÉRATION

12.1.1 universalité du dénombrement

La somme de Georges Ifrah, *Histoire universelle des chiffres*, fournit le meilleur point de départ (utiliser la deuxième édition, Paris, 1994, beaucoup plus développée que la première ; la civilisation médiévale est traitée de manière insuffisante et contestable, mais on trouve cependant beaucoup d'informations, dans un cadre général indispensable que l'on ne rencontre pas ailleurs).

Les préhistoriens ont découvert des *os entaillés de signes* qui ne peuvent pas s'interpréter autrement que comme des dénombrements, os remontant, pour les plus anciens, à la période aurignacienne (environ 40000 ans avant notre ère). Des fouilles au Proche- et Moyen-Orient ont permis de découvrir de grandes quantités de *petits objets plus ou moins "standardisés"*, dont certains au moins sont presque sûrement des jetons de calcul, cela à partir du 9e ou 8e millénaire avant notre ère. La numération et les calculs sont donc attestés bien avant l'apparition de l'écriture. *Les plus anciens documents écrits (sumériens et égyptiens, entre 3200 et 3000 avant notre ère)* comportent presque tous des indications numériques, traduisant des systèmes de numération déjà fortement structurés.

L'historien doit être capable de **ne jamais séparer deux perspectives** (habituellement tout à fait distinctes) : d'une part, la classique "*histoire des systèmes de numération*" (qui s'intéresse à peu près exclusivement aux numérations écrites, sous leur aspect le plus sophistiqué) et, d'autre part, *l'analyse sociale de la répartition des capacités pratiques de dénombrement et de calcul*, complètement inégalitaire dans toutes les sociétés, la nôtre comprise. Le désintérêt pour ce second aspect, historiquement fondamental, est à l'origine de nombreuses apories de tous ordres.

12.1.2 la calculatrice universelle la plus ancienne : la main

Le "comput digital" est une réalité à peu près universelle. Mais il y a beaucoup de manières de "compter sur ses doigts". Pour l'Europe médiévale, le texte le plus célèbre est le premier chapitre du *De ratione temporum* de Bède le Vénérable (vers 725), intitulé "de loquela digitorum", chapitre qui fut recopié indépendamment durant tout le Moyen Age. Bède y expose brièvement mais clairement une méthode, très probablement d'origine égyptienne, qui permet de représenter sur une seule main, en pliant diversement les doigts, les nombres de 1 à 99 (donc de 1 à 9999 avec les deux mains). De nombreux manuscrits médiévaux et des ouvrages imprimés des 15e et 16e siècles donnent, sur des tableaux figurés, la position des doigts correspondant à chaque nombre.

Mais cette méthode, qui permet de représenter les nombres, peut tout au plus aider dans les calculs, ce n'est pas une méthode de calcul. Diverses recherches, notamment ethnographiques, ont montré qu'il existe diverses possibilités, combinant les doigts et des tables de multiplication connues par cœur, d'opérer des multiplications très importantes. Les recherches dans ce domaine paraissent (trop) peu développées.

12.1.3 les tailles (Kerbhölzer)

Un article récent de Ludolf Kuchenbuch fait le point sur cette question très importante ("Pragmatische Rechenhaftigkeit ? Kerbhölzer in Bild, Gestalt und Schrift", *Frühmittelalterliche Studien*, 36-2002, pp. 469-490, avec illustrations). Comme on l'a rappelé plus haut, l'usage d'entailles sur des objets allongés pour représenter des nombres remonte à la préhistoire. Le système extrêmement commun au Moyen Age des tailles est un peu plus complexe. On utilisait en général **un bâton de bois, fendu en deux dans le sens de la longueur**. Lorsque l'on voulait faire une entaille, on plaquait les deux morceaux l'un contre l'autre, puis chaque partie (vendeur / acheteur, percepteur / comptable) gardait l'un des deux morceaux. Aucune des deux parties ne pouvait plus

modifier son demi-bâton (sinon en rejoignant à nouveau les deux parties), qui servait ainsi de *mémoire numérique inaltérable*. Ce système très simple fut par exemple employé par les sheriffs anglais pour rendre leurs comptes au roi jusqu'au 18e siècle. Il fut utilisé par une infinité de commerçants dans toute l'Europe jusqu'au début du 20e siècle. Les encoches étaient le plus souvent de deux ou trois types, servant à distinguer par exemple 1, 5 et 10.

On a conservé très peu d'exemplaires, et cela est facile à comprendre : ces objets n'étaient utiles que pour autant qu'ils avaient une valeur probatoire liée (le plus souvent, mais pas toujours) à une opération de crédit. Le crédit éteint, ils perdaient toute signification et n'étaient pas réutilisables, ***sinon comme bois de chauffage***... L'absence d'attestation avant le 11e siècle me semble peu probante.

12.1.4 le système de numération romain

La numération romaine paraît le résultat d'une histoire déjà fort longue. Elle hérita en partie du système grec, mais surtout du système étrusque (mal connu). Dans la classification générale des systèmes de numération, elle *appartenait au groupe le plus ancien, celui des systèmes additifs*. La base était décimale, et chaque ordre (1, 10, 100, etc) était figuré par un signe propre. Le système était complété par des signes intermédiaires (5, 50, 500) et compliqué plutôt que simplifié par l'emploi partiel de la soustraction (ix, xc, etc). C'était un système lourd et peu maniable, notamment pour les grandes valeurs, pour lesquelles les signes conventionnels ont d'ailleurs varié.

Les Grecs, puis les Hébreux, ont utilisé à partir de l'époque hellénistique un autre système, *fondé sur l'utilisation de l'alphabet*, chaque lettre étant affectée d'une valeur, dans l'ordre "naturel". Ce système permettait une écriture plus compacte, mais peu lisible. Dès l'origine, il a donné lieu à des **spéculations** (de plus en plus embrouillées) sur la "valeur numérique" des mots ordinaires, et en particulier des noms propres (le sommet final ayant sans doute été atteint au Moyen Age par les cabbalistes). Ces systèmes, connus seulement au Moyen Age de quelques lettrés chrétiens, ont coexisté avec le système romain au moins jusqu'au 17e siècle (voire bien plus longtemps dans certaines conditions).

A partir du 12e et du 13e siècle, les comptables occidentaux ont modifié partiellement le système romain classique, en y *ajoutant des éléments de système "hybride"*, c'est-à-dire faisant intervenir des produits en plus des additions ; ces produits concernaient en particulier les valeurs 20, cent et mille, ce qui donnait des écritures comme VI^{XX}, IX^{C}, $XXIX^{M}$, ou, combinées,

$XVIII^{M}$ VII^{C} IV^{XX} II...

12.1.5 significations intrinsèques des nombres

Tout le monde connaît les trois personnes de la trinité, les quatre évangiles, les sept jours de la création, les dix commandements, etc. Aujourd'hui, plus personne n'y prête attention : simples nombres. La civilisation médiévale voyait les choses de manière radicalement différente. La Bible est remplie de nombres, certains livres même (*Apocalypse*) comportant des spéculations de type grec (voir ci-dessus). Les **Pères de l'Église, au premier rang saint Augustin**, ont abondamment parsemé leurs œuvres de considérations sur la signification de ces nombres et sur les (innombrables !) correspondances que l'on peut établir entre elles ; ces considérations sont une partie décisive de ce que l'on convient d'appeler la "lecture allégorique" du texte sacré. A cet égard, on ne saurait en aucun cas oublier que les divers "modes de lecture" étaient fortement hiérarchisés (proprio sensu), et que le "sens allégorique" était l'un des plus valorisés : le sens tiré de cette lecture avait une valeur très supérieure à celui tiré des lectures historique et morale. Pour le dire autrement, et de manière un peu anachronique, **les nombres, pour les hommes du Moyen Age, avaient une valeur "ontologique", un sens intrinsèque**. Tout à fait autre chose que ce que nous voyons simplement comme la "suite des entiers naturels". Il s'agissait d'*êtres créés par dieu*, recelant un

sens fort. Bien entendu les clercs, et surtout les plus instruits d'entre eux, avaient sur ces sens des vues très étendues, dont ne disposait pas le vulgaire. Mais même le paysan carolingien illettré connaissait les entités citées dans la première phrase de ce paragraphe, et savait pertinemment qu'elles n'étaient pas sans rapport avec son propre salut personnel. **Gradation, sans aucun doute, opposition, sûrement pas**.

Les gloses sur les Pères n'ont cessé de s'étendre du 8e au 12e siècle. A ce moment-là, Hugues de Saint-Victor tenta une première systématisation, qui fut reprise et amplifiée par *un petit groupe de cisterciens*, dont les textes constituent une vraie somme, comme cette époque en constitua dans tous les domaines essentiels (Odon de Morimond, Guillaume d'Auberive, Geoffoi d'Auxerre ; ces textes importants n'ont été édités que très récemment, par le danois Hanne Lange). Une présentation assez simple est fournie par Heinz MEYER, *Die Zahlenallegorese im Mittelalter. Methode und Gebrauch*, München, 1975, un ouvrage plus complet est celui de Heinz MEYER & Rudolf SUNTRUP, *Lexikon der mittelalterlichen Zahlenbedeutungen*, München, 1987 (1015 pages). On ne comprend rien à la civilisation médiévale si l'on ignore cette forme de représentation.

On ne saurait, pour clore provisoirement cette question, insister assez sur la valeur centrale et fondatrice du UN. *L'**unité**, dans tous les sens possibles du terme, était l'un des piliers, peut-être le principal, de la civilisation médiévale*. L'unité est à la fois la racine et le tout. Toute idée de subdivision était foncièrement négative. Comme on va le voir un peu plus loin, les clercs médiévaux ont eu rapidement à leur disposition des outils qui leur auraient permis de décomposer cette unité. Cela était strictement hors de question, c'eût été de l'autodestruction.

12.1.6 usages proprement savants

Jusqu'au 13e siècle, les clercs ont utilisé un manuel d'arithmétique et un seul, l'*arithmetica* de Boèce. C'est un texte d'une lecture pour nous très difficile. De Boèce également, ils utilisèrent autant la *musica*, qui ne comporte pas moins de calculs que le précédent. Toute la *musique médiévale était fondée sur une réflexion sur les nombres et sur les proportions*. Réflexion complexe et raffinée (l'Occident médiéval n'inventa pas par hasard la première notation musicale au 10e siècle). A côté de la musique, l'astronomie, liée à la fixation des dates liturgiques (comput). L'auteur principal est ici Bède le Vénérable. On ne saurait surévaluer l'importance du lien entre comput digital et fixation de l'ordre liturgique.

12.2. CALCULS

12.2.1 l'héritage antique : l'abaque

Les Grecs héritèrent eux-mêmes la "table à compter" d'une histoire fort longue. Ils utilisèrent d'ailleurs une assez grande variété d'outils (Alain SCHÄRLIG, *Compter avec des cailloux. Le calcul élémentaire sur l'abaque chez les anciens Grecs*, Lausanne, 2001). Une surface munie de lignes et/ou de colonnes, sur laquelle on déplaçait des jetons ; lignes ou colonnes représentant les ordres successifs de la base (1, 10, 100, 1000, etc). L'outil permettait assez facilement les additions et soustractions, les multiplications (à l'aide de produits partiels) étaient déjà plus compliquées. Les divisions étaient vraiment difficiles. Les jetons s'appelaient des *calculi*. Pour les divisions, on utilisait comme on pouvait des tables, la plus connue étant le *liber calculi*, de Victorius d'Aquitaine (milieu du 5e siècle). Une description de ce genre de procédure par des comptables experts est donnée par Richard Fitz-Neel dans son fameux *Dialogus de scaccario* (vers 1178), qui décrit les pratiques comptables de la royauté anglo-normande.

12.2.2 la numération indienne

Comme le rappelle G. Ifrah, les indiens furent les premiers à *combiner trois innovations*, dont la réunion seule constitue vraiment un saut :
* neuf signes numériques arbitraires, détachés de toute signification visuelle ;
* une numération par position stricte ;
* un usage méthodique et raisonné du zéro.

Cette mise au point date de la fin de la civilisation Gupta, c'est-à-dire aux alentours des 5e-6e siècles de notre ère. Un évêque mésopotamien y fit une allusion nette au 7e siècle. Les "chiffres indiens" arrivèrent au Maghreb et en Andalousie dès le 8e siècle. La première grande synthèse en dehors des Indes fut celle d'Al-Khuwarismi (vers 820-850).

12.2.3 la très lente pénétration du "calcul avec des chiffres"

Gerbert d'Aurillac (futur Sylvestre II) découvrit l'existence des chiffres indiens dans le nord de l'Espagne à la fin du 10e siècle. Son innovation consista seulement à remplacer les *calculi* anonymes et identiques par des jetons marqués des chiffres indiens de 1 à 9 (jetons appelés *apices*). Cela ne simplifiait que très modestement l'usage de l'abaque traditionnelle. Ce fut seulement en Espagne, à la fin du 12e siècle, que furent traduits-rédigés en latin des résumés des travaux d'Al-Khuwarismi (*liber Ysagogarum* et *liber alchorismi*). La numération par position et les méthodes de calcul révolutionnaires qu'elle permettait furent popularisées dans les milieux des facultés des arts par les écrits d'Alexandre de Villedieu et John of Holywood (Johannes de Sacrobosco), tous deux du début du 13e siècle. Les chiffres étaient alors tracés sur des tables recouvertes de sable ou de poussière (pour effacement facile). De plus grande portée fut le *liber abaci* de Leonardo Fibonacci (da Pisa), 1202. Celui-ci préconisait la plume et l'encre, et son traité comportait une part importante d'algèbre, surtout orientée vers des problèmes commerciaux. L'Italie, et l'Italie seule, fut la terre de développement de ces méthodes, en particulier aux 14e et 15e siècle (grande synthèse de Luca Paccioli en 1478). Le mouvement fut naturellement lié à l'usage croissant du papier et à l'essor du commerce et de la banque italiens. Ce fut seulement dans la seconde moitié du 15e que le reste de l'Europe fut vraiment atteint. On doit cependant signaler que les autorités de Florence, en 1299, interdirent l'usage des chiffres dans la présentation des bilans comptables. *Les nombres restaient les nombres, les chiffres servaient tout au plus au calcul.* (Excellente synthèse de Guy BEAUJOUAN, "Nombres", in Jacques LE GOFF & Jean-Claude SCHMITT (éds), *Dictionnaire raisonné de l'Occident médiéval*, Paris, 1999, pp. 834-844).

12.2.4 le basculement

Le système médiéval subit discrètement ses premières atteintes entre la fin du 15e siècle et l'orée du 17e. On attribue au français Nicolas Chuquet *la première utilisation des nombres négatifs (vers 1485)* : c'est seulement à ce moment que le zéro devint un élément important du système de numération. L'italien Gerolamo Cardano imagina au milieu du 16e la possibilité de la racine carrée d'un nombre négatif (ce que nous appelons *les nombres complexes*).

Le pas décisif fut sans doute franchi par l'ingénieur flamand Simon Stevin, en 1585, lorsqu'il publia *De Thiende* (la dîme). Ce fut en effet **le premier à proposer un mode d'écriture des "décimales"**. Ce faisant, sans probablement s'en apercevoir, il abattait la prépondérance exclusive de l'unité. Un rôle clé revint à l'écossais John Napier (alias Neper), qui inventa les logarithmes et publia les premières tables en 1614. L'aboutissement de cette évolution se trouva chez G.W. Leibniz (1646-1716), qui mit au point à la fin du 17e siècle **(1684) le calcul infinitésimal**, c'est-à-dire une formalisation claire et rigoureuse des calculs sur l'infiment petit et l'infiniment grand. A partir de là, pour la première fois dans l'histoire de l'humanité, il devenait possible de manipuler correctement le continu. *Là on peut faire débuter la science moderne, en aucun cas antérieurement.* On ne saurait

manquer de noter que Stevin, Napier et Leibniz étaient protestants.

En français, le terme de "graduation" date de la fin du 17e siècle, et l'usage régulier d'instruments de mesure courants qui en portent ne paraît guère antérieur. A ma connaissance, *le Moyen Age ne nous a légué aucune "règle graduée"*, et l'on n'en possède pas une seule représentation. Un objet qui nous semble des plus anodins, et pour ainsi dire "naturel", ne l'était pas autant que ce que le "bon sens" laisse croire.

On ne doit toutefois pas se méprendre à propos de l'idée de "généralisation de l'usage des chiffres arabes à partir du 17e siècle". La fraction de la population qui les connaissait et était capable de s'en servir demeura longtemps minime. Même parmi les commerçants et les administrateurs, qui avaient le plus besoin de faire des calculs, les "quatre opérations" restaient pénibles. En témoigne bien le succès énorme et non démenti avant un 19e siècle bien avancé de "tables" de toutes sortes, tables qui donnaient le résultat de quantité de calculs (le terme "barême", attesté en 1803, provient de François Barrême(1638-1703), qui publia un traité d'arithmétique comportant une multitude de tables, traité qui eut un succès fabuleux durant tout le 18e siècle et jusqu'en 1821... sans compter des adaptions diverses dans des langues étrangères).

On cite à juste titre la phrase de Montaigne, "je ne sais compter ni à jet ni à plume", qui témoigne bien de la capacité (incapacité) de calcul d'un des meilleurs intellectuels de la fin du 16e siècle.

12.3. MESURES

12.3.1 caractères généraux des systèmes de mesure anciens

Une réflexion générale préalable est indispensable, même si elle ne peut résulter que d'une assez longue pratique de recherche dans le domaine dit de la "métrologie historique". Je retiendrai trois points principaux :

a. tous les systèmes de mesure anciens étaient fondamentalement **adaptés aux possibilités de dénombrement et de calcul** dans la société considérée. L'idée de mesure sans calcul est une sottise, même si c'est le présupposé implicite de la plupart des ouvrages de "métrologie historique". D'où l'*intérêt fondamental de 2 et des puissances de 2*, tant il est vrai que le vulgaire, illettré, comptait essentiellement en multipliant et en divisant par deux. Inversement, on ne peut guère douter de la signification sociale forte d'un rapport entre unités du type 9,5 (largement attesté) : un tel rapport engendrait ipso facto un monopole de manipulation et de compréhension au bénéfice de la frange sachant calculer.

b. tous les systèmes de mesure anciens étaient constitués d'**unités adaptées aux objets et aux ordres de grandeur**. Le pied et la perche étaient deux unités de longueur, c'est *un anachronisme dangereux d'imaginer l'un comme le multiple ou le sous-multiple de l'autre*, même si le rapport entre les deux unités était parfaitement fixé à tel endroit et à tel moment. Encore une fois, il faut se faire à l'idée que l'unité médiévale ne se divisait pas. Ce qui d'ailleurs justifiait complètement la pratique, courante durant des siècles, de modifier plus ou moins arbitrairement le rapport entre telle et telle unité (ce qui, en pratique, modifiait bien sûr l'une ou l'autre, dans le sens d'une adaptation aux besoins de tel ou tel groupe).

c. des deux points précédents ressort un troisième, qui n'en conserve pas moins une certaine autonomie, et sur lequel il faut insister : **les unités de mesure et les procédures de mesure étaient d'une précision éminemment variable, adaptée bien sûr aux capacités de calcul, mais surtout au contexte**, c'est-à-dire au rapport entre les objets et les groupes sociaux les manipulant. La notion de précision ou d'exactitude, qui nous semble une entité abstraite générale, n'existait pas en soi, mais n'était qu'une conséquence du contexte de la mesure considérée. En gros, pour le dire

schématiquement, pour toute période antérieure au 17e siècle, *toute "donnée numérique", dans la source ou calculée, comportant plus de trois "chiffres significatifs" doit être considérée avec la plus extrême méfiance* (à l'exception bien entendu des données monétaires purement comptables). Deux chiffres significatifs sont courants, trois possibles ; au-delà, l'historien doit se justifier (des cas existent, ils sont rares et méritent un examen spécial).

12.3.2 les principales catégories de mesures

Les mesures anciennes, à l'opposé du "système métrique", ne formaient pas système. Bien entendu, des correspondances de fait existaient (par exemple entre poids et volume de céréales), mais on ne s'est préoccupé que tardivement (extrême fin du Moyen Age) de les expliciter ; inversement, des correspondances spécifiques déterminaient en principe certaines mesures (comme les unités de surface des champs définies par une unité de volume de grains correspondant à l'ensemencement de ladite surface), mais en réalité ces *correspondances étaient très élastiques et variables*. Notons immédiatement que la métrologie historique ne se réduit en aucun cas aux tables de correspondance entre unités du système métrique et unités de la fin du 18e siècle (comme le croient encore beaucoup d'érudits), tout simplement parce que les mesures anciennes n'ont pas cessé d'évoluer (ce qui ne retire rien au caractère classique de la bibliographie de Paul BURGUBURU, *Essai de bibliographie métrologique universelle*, Paris, 1932, qui fournit notamment un répertoire assez complet de tous les fascicules publiés dans chaque département entre 1795 et les années 1840 pour donner les équivalences entre mesures anciennes et système métrique).

Indiquons ici également l'intérêt majeur des ouvrages de Ronald Zupko, qui a rassemblé une masse énorme de références textuelles, classées alphabétiquement ; c'est aujourd'hui un point de départ obligé pour toute recherche de métrologie historique sur les périodes anciennes (*French weights and measures before the Revolution : a dictionnary of provincial and local units*, Bloomington, 1978. *Italian weights and measures from the Middle Ages to the nineteenth century*, Philadelphia, 1981. *A Dictionnary of weights and measures for the British Isles : the Middle Ages to the twentieth century*, Philadelphia, 1985).

a. monnaies et poids

On part en général de la ***libra*** *romaine* de 327g (divisée en 12 unciae). L'importance cruciale de cette unité, sa restauration par Charlemagne, puis la possibilité de la restituer à peu près à l'époque contemporaine, sont dues à son usage comme base des systèmes monétaires, romain (puis carolingien). En fait, il n'existait *pas un système de poids, mais au moins trois*, correspondant à trois méthodes de pesée (trois outils) : le trébuchet pour les pesées monétaires et de métaux précieux, les balances courantes (à fléaux égaux ou romaine), pour peser les denrées usuelles, particulièrement le pain (parfois la viande, souvent la cire), balances dont la précision et la fiabilité étaient bien plus vagues, enfin les balances pour objets lourds, qui furent souvent construites dans les villes et affermées (jouissant d'un monopole lucratif). Dès le 11e siècle, on voit intervenir le ***marc***, qui varie fortement d'une zone à l'autre, d'un usage à l'autre, de même que le poids des monnaies se mit à varier de plus en plus sensiblement (utile tableau des principaux marcs connus dans Etienne FOURNIAL, *Histoire monétaire de l'Occident médiéval*, Paris, 1970, pp. 161-168 ; donne impavidement 6 chiffres significatifs...). En France, le marc dit de Troyes fut utilisé par les ateliers royaux (244,7g). On peut noter (parce que ce n'est pas intuitif) que l'on a souvent observé que le pain était vendu à prix fixe, le poids variant en fonction du prix des grains.

b. capacités (volumes)

Ici règne la cohue la plus totale et la plus obscure. Certaines mesures étaient utilisées pour les liquides (vin surtout, mais aussi huile, bière). Tous les rapports possibles ont existé entre les

divers types de bouteilles et les innombrables types de tonneaux. Les variations ont souvent été causées par des mesures fiscales (le vin étant un des principaux supports des impôts indirects) : si une ville décidait par exemple d'appliquer une taxe d'un huitième sur le vin, les débits de boisson étaient tenus de vendre toujours la bouteille au même prix, mais avec un contenu diminué d'un huitième. Au contraire, certains tonneaux de grosse contenance, utilisés pour le commerce à longue distance, ont pu demeurer stables très longtemps. Signalons qu'au 19e siècle encore, il était admis que le tonneau bordelais pouvait varier entre 224 et 228 litres (tolérance de ±1%).

Pour les solides, essentiellement les grains (parfois les noix et autres), le système était compliqué par le choix de la forme du récipient, du double fait du possible tassement et du choix entre "mesure rase" et "mesure comble". Plus le récipient était large et mince et plus le comble pouvait être important. Les rapports entre unités étaient fréquemment du type : la mesure x vaut 16 mesures y, 15 rases et une comble. Il arrivait souvent que le même terme fût utilisé pour le vin et pour le blé, mais avec une contenance complètement différente (modius).

c. longueurs

Les romains employèrent à peu près partout le **pied de 29,5cm**. Celui-ci fut également restauré par Charlemagne, et survécut largement jusqu'au 12e siècle, particulièrement dans les constructions. Les commerçants étaient surtout intéressés par la longueur des pièces de tissu. Tout changeait d'une ville à l'autre, et souvent d'un type de tissu à un autre. Les manuels de marchands à partir de la fin du Moyen Age y consacrèrent naturellement une très large place. Pour les longueurs également, on avait fréquemment des unités bien distinctes pour les tissus, les bâtiments et les divers types de terrains (champs, vignes, prés).

L'analyse métrologique des bâtiments anciens est restée longtemps inefficace, cantonnée chaque fois à un seul bâtiment. En considérant de manière méthodique des séries de bâtiments du 6e au 15e siècle, je suis parvenu à établir un ensemble simple de "règles" permettant de définir quelles sont, dans un bâtiment médiéval, les dimensions à prendre en compte, à partir desquelles on peut retrouver la longueur de l'unité qui a servi à l'implanter (« L'analyse des dimensions des édifices médiévaux. Notes de méthode provisoires », in Nicolas REVEYRON (éd.), *Paray-le-Monial, Brionnais-Charolais, le renouveau des études romanes*, Paray-le-Monial, 2000, pp. 327-335, avec bibliographie). Cette possibilité, si elle se confirme, ouvre un vaste champ à la **reconstitution des évolutions des unités de longueur employées dans les constructions, dans toutes les régions d'Europe, pour tout le Moyen Age**...(récemment, des recherches portant sur la fin du Moyen Age ont permis de confronter des bâtiments encore existants avec des contrats de construction : les principales hypothèses que j'ai élaborées paraissent confirmées ; voir le numéro d'*Histoire & Mesure*, 16-2001, "Mesurer les bâtiments anciens", sous la direction de Philippe Bernardi).

Signalons au passage un piège facile à comprendre : l'implantation d'un édifice du culte était le plus souvent réalisée par un ou plusieurs clercs, qui utilisaient pour cela les unités de longueur et les méthodes de calcul dont ils disposaient. Ensuite intervenaient des maçons, des charpentiers, qui pour beaucoup d'entre eux étaient illettrés et n'utilisaient pas les mêmes unités. C'est notamment pourquoi on trouve souvent une discordance entre les dimensions d'un édifice et la taille de tel ou tel élément (épaisseur de mur, largeur ou profondeur de tel ou tel élément de décor). Notons enfin que les clercs médiévaux qui ont écrit sur les dimensions des édifices ont toujours considéré séparément, et différemment, longueur, largeur et hauteur. Pour utiliser un zeste de pédanterie, on peut dire que l'espace médiéval n'était pas isotrope ; les dimensions n'étaient pas des grandeurs abstraites, mais toujours des caractères de tel ou tel objet, nettement défini. *Thomas d'Aquin a réfuté la possibilité d'une ligne infinie (seul dieu est infini).*

d. surfaces

Les surfaces cultivées étaient évaluées soit en fonction du temps de travail nécessaire (journal, ouvrée), soit en fonction de la quantité de grains nécessaire à l'ensemencement (coupée), soit (fréquemment) à l'aide de termes dont l'étymologie et la signification nous échappent (resce, andain, bonnier...). Les dimensions des terrains étaient souvent mesurées (nombreuses chartes antérieures à l'an mille), mais ces dimensions ne servirent que bien plus tard à calculer (approximativement) des surfaces (pas avant le 13e siècle). L'extrême difficulté des multiplications (a fortiori des divisions) est certainement la cause principale. Des unités de surface définies à partir d'unités de longueur (perche carrée, toise carrée, etc) ne se répandirent qu'avec une extrême lenteur et étaient encore loin d'être d'un usage systématique au 18e siècle, même si, à cette époque, des équivalences avaient le plus souvent été calculées (se reporter au chapitre sur la géodésie et l'arpentage).

12.3.3 quelques considérations générales sur l'évolution en Europe

On ne connaît pas parfaitement les mesures romaines, en particulier on ignore jusqu'à quel point les mesures universelles que nous connaissons à peu près ont pu éliminer d'autres mesures locales.

Ces mesures se disloquèrent en grande partie aux 6e-7e siècle, mais furent partiellement restaurées à l'époque de Charlemagne et se maintinrent plus ou moins longtemps (évolution somme toute assez parallèle à celle de la "minuscule caroline"). Le choix d'une livre de 20 sous constitués de 12 deniers d'argent (1 livre = 240 deniers) fut particulièrement pertinent et durable (fin vers 1970). Cette longévité exceptionnelle est sans aucun doute due aux **qualités de 240** : 16 x 3 x 5. 240 est divisible quatre fois par deux, par trois et par cinq, on ne peut guère trouver de nombre qui facilite autant les divisions dans des sociétés qui maîtrisaient très mal la division.

Les évolutions divergentes des deniers à partir de la fin du 11e siècle imposèrent d'abord des systèmes d'équivalence assez compliqués, ce ne fut que très lentement que se dégagea la notion (pour nous assez étrange) de "monnaie de compte" qui ne triompha vraiment qu'au 15e siècle.

La plupart des autres unités de mesure subirent des évolutions parallèles, aboutissant dès le 13e siècle à une mosaïque indescriptible. *Les villes surtout, certains princes, se préoccupèrent d'établir des étalons publics, servant de référence obligée dans l'aire contrôlée par ladite ville (souvent de petite taille)*. C'est ce phénomène général qui justifie le titre d'allure a priori un peu bizarre du principal ouvrage français de métrologie historique (Armand MACHABEY, *La métrologie dans les musées de province et sa contribution à l'histoire des poids et mesure en France depuis le treizième siècle*, Paris, 1962).

Dans ce capharnaum, **chaque groupe tentait d'imposer ses intérêts**, bien différents selon qu'il s'agissait de producteurs, de rentiers, de commerçants. On doit ici se référer à la synthèse magistrale de Witold KULA, *Les mesures et les hommes*, Paris, 1984 (original polonais 1970). Il est un peu dommage que Kula ait si peu fait intervenir les possibilités "sociales" de calcul ; on peut également se demander jusqu'à quel point le caractère essentiellement local de chaque ensemble d'unités de mesure ne participait pas d'une logique sociale générale qui tendait à individualiser au maximum chaque petit espace.

A partir du 16e siècle, et surtout du 17e, les commerçants firent peu à peu prévaloir la nécessité de mesures stables et de portée générale. Le gouvernement de Louis XIV prit à la fin du 17e siècle une série de décisions allant dans ce sens, notamment la définition du fameux **"pied du roi" (32,45cm)**. Mais *les méthodes de calcul et les instruments de mesure n'étaient pas encore parvenus à un degré suffisant d'élaboration pour soutenir une réforme globale*. Le système métrique décimal fut l'une des principales décisions des révolutionnaires, les conquêtes napoléonniennes lui assurèrent une première diffusion européenne.

En dépit des avantages extraordinaires de la combinaison de la numération avec virgule décimale et d'un système de mesures unifié reposant lui aussi sur la base décimale, l'introduction de ce système fut extrêmement lente. Ce fut seulement en 1840 qu'en France même son enseignement fut imposé dans les écoles à l'exclusion de tout autre. Les scientifiques du monde entier l'adoptèrent en général dès le milieu du 19e siècle. Mais chacun connaît la résistance que les pays anglo-saxons continuent de lui opposer, sinon en théorie, du moins en pratique.

La civilisation contemporaine considère les nombres comme des entités abstraites, de caractère artificiel, manipulables sans la moindre restriction. La notion de précision est une valeur cardinale de notre société. L'historien doit savoir que tout cela est récent et qu'avant le 19e siècle, la situation se présentait de manière tout à fait différente, largement inverse. Les nombres étaient des êtres créés par dieu, et en pratique fort difficiles à manipuler. Nous retrouvons ici des *éléments fondamentaux de la sémantique historique* : aucun objet n'a de sens en dehors de la société qui l'utilise. Tout ce qui pouvait donner lieu à mesure (dimensions de toutes sortes) état propriété des objets, et appartenait donc à ces objets, en dehors desquels il perdait tout sens. Il apparaît ici, encore une fois, que toute considération numérique ancienne, quelle qu'elle soit, n'est analysable que dans le cadre d'une réflexion sur le sens, dont seuls des historiens exercés sont capables.

NOTES POUR LA SUITE

Que l'on étudie des bâtiments ou des tessons, des sépultures, des monnaies, des chartes, des traités de théologie, des chroniques, des prés, des impôts, des procès, des fresques ou des séquences de films, dans tous les cas des procédures statistiques appropriées peuvent rendre de grands services, et permettre des observations impossibles sans elles. Il y a bien sûr des variantes, mais la majorité de ces procédures sont applicables aux objets historiques les plus divers.

La raison en est simple : la recherche historique proprio sensu vise l'étude de l'évolution des sociétés du passé, et rien d'autre. Or toute société est une structure, c'est-à-dire un ensemble articulé de relations. Ce sont cette structure et ces relations qui constituent le sens de tout objet historique, quel qu'il soit. La réflexion statistique, en dehors même de tout calcul (voir chapitre 1.3), implique une explicitation formalisatrice de ces relations : voilà pourquoi le gain est assuré. Dans tous les cas.

Ces objets historiques, comme on l'a rappelé, comportent des caractères d'une grande généralité : outre qu'ils sont tous situés dans un temps et un espace définis, ils sont le plus souvent fortement imbriqués les uns dans les autres (donc difficiles à cerner), et notre connaissance en est toujours plus ou moins lacunaire. Caractères assez différents de ceux des objets traités le plus ordinairement par les statistiques. Ce qui implique trois mouvements complémentaires : 1. élaguer (des chapitres de manuels entiers sont inutilisables sinon nuisibles) ; 2. constituer des procédures nouvelles (en jouant particulièrement sur le couple adaptations/expériences) ; 3. viser, à terme, une relative homogénéité de ce nouvel ensemble, en comblant des lacunes, en déterminant tant les limites que l'extension possible de chaque procédure.

Le présent canevas ne laisse que trop apparaître son insuffisance. Les objectifs immédiats vont presque de soi : munir chaque chapitre d'une série d'exemples concrets, rédiger diverses "fiches techniques" (s'agissant notamment de quelques points importants en matière de calculs ou d'informatique), et tenter d'apporter progressivement des compléments et des développements nouveaux dans les domaines où le besoin semble le plus se faire sentir : analyse des réseaux (théorie des graphes), interactivité démultipliée dans les analyses chronologiques et spatiales, recherche de procédures plus complètes et plus efficaces pour l'analyse des distributions parétiennes, constitution d'une véritable boîte à outils pour la sémantique et l'analyse des textes. L'espace des possibles n'est pas mesurable.

BIBLIOGRAPHIE

On trouvera seulement, dans la liste qui suit, la référence à des textes et à des ouvrages qui, à moment ou à un autre, m'ont été utiles. Dans les conditions actuelles, il est important de savoir que l'on peut trouver sur le net une grande quantité de matériel intéressant dans la perspective de ce cours. On peut en effet récupérer des "tutoriels", de qualités diverses, en français, en allemand, en anglais. Je signale tout spécialement le site de statistique écologique de Lyon (http://pbil.univ-lyon1.fr/ADE-4) qui, outre le logiciel open source très performant ADE-4, contient également un cours de statistique (en français !) riche et détaillé (relatif non seulement aux chapitres 2, 3 et 4 du présent cours, mais aussi au chapitre 7, ce qui est bien plus rare). Une recherche attentive (i.e. persévérante et fastidieuse...) permet aussi de découvrir des textes utiles sur des points précis, sur lesquels on souhaite connaître « l'état de l'art ». Il est préférable d'archiver tout ce que l'on trouve, car les sites internet sont très volatiles.

« Le hasard », *Pour la science hors-série*, avril 1996
« Les nombres », *La Recherche*, numéro spécial 278-1995
« Mesurer les bâtiments anciens », *Histoire & Mesure*, 16-2001, sous la direction de Philippe BERNARDI
ALLAIS Maurice, « Fréquence, probabilité et hasard », *Journal de la société de statistique de Paris*, 1983, pp. 70-102 et 144-221
ARDENER Edwin (éd.), *Social Anthropology and Language*, London, 1971
ARMATTE Michel, « Robert Gibrat et la loi de l'effet proportionnel », *Mathématiques, informatique et sciences humaines*, 33-1995, pp. 5-34
BACH Adolf, *Deutsche Namenkunde. 1. Die deutschen Personennamen*, Heidelberg, 1943 ; *2. Die deutschen Ortsnamen*, Heidelberg, 1953
BAILEY David Roy, *Onomasticon to Cicero's treatises*, Stuttgart, 1996
BARBUT Marc et FOURGEAUD Claude, *Eléments d'analyse mathématique des chroniques*, Paris, 1971
BARBUT Marc, « Des bons et des moins bons usages des distributions parétiennes en analyse des données », *Histoire & Mesure*, 1988, pp. 111-128. « Distributions de type parétien et représentation des inégalités », *Mathématiques, informatique et sciences humaines*, 106-1989, pp. 53-69. « Note sur l'ajustement des distributions de Zipf-Mandelbrot en statistique textuelle », *Histoire & Mesure*, 4-1989, pp. 107-119 ; « Une remarque sur l'expression et l'ajustement des distributions de Zipf-Mandelbrot en statistique textuelle », *Mélanges André Lentin*, Paris, 1996. « Une famille de distributions : des parétiennes aux antiparétiennes. Applications à l'étude de la concentration urbaine et de son évolution », *Mathématiques, informatique et sciences humaines*, 141-1998.
BEAUJOUAN Guy, « Nombres », in LE GOFF Jacques et SCHMITT Jean-Claude (éds), *Dictionnaire raisonné de l'Occident médiéval*, Paris, 1999, pp. 834-844
BÉGUIN Michèle et PUMAIN Denise, *La représentation des données géographiques. Statistique et cartographie*, Paris, 1994
BENZECRI Jean-Paul et collab., *L'analyse des données. I. La taxinomie. II. L'analyse des correspondances*, 2 vol., Paris, 1973. *Pratique de l'analyse des données. 3. Linguistique et lexicologie*, Paris, 1981. *Pratique de l'analyse des données. 5. En économie*, Paris, 1986.
BERLIOZ Jacques et al., *Identifier sources et citations*, Turnhout, 1994
BERNSTEIN Basil, *Langage et classes sociales. Codes socio-linguistiques et contrôle social*, Paris, 1975
BERTIN Jacques, *Sémiologie graphique. Les diagrammes, les réseaux, les cartes*, Paris, 1967 ; *La graphique et le traitement graphique de l'information*, Paris, 1977

BESSON Jean-Louis (éd.), *La Cité des chiffres, ou l'illusion statistique*, Paris, 1992
BEST Heinrich et MAN Reinhard (éds), *Quantitative Methoden in der historischsozialwissenschaflichen Forschung*, Stuttgart, 1977
BEST Heinrich et SCHRÖDER Wilhelm Heinz, « Quantitative historische Sozialforschung », in MEIER Christian et RÜSEN Jörn (éds), *Historische Methode*, München, 1988, pp. 235-266
BESTEK Andreas, *Geschichte als Roman : narrative Techniken der Epochendarstellung im englischen historischen Roman des 19. Jahrhunderts, Walter Scott, Edward Bulwer-Lytton und George Eliot*, Trier, 1992
BOCQUET Jean-Pierre et REVERSEAU Jean-Pierre, « Estimation de la stature de la classe féodale d'après les armures du XVI^{e} siècle », *Ethnologie française*, 1-1979, pp. 85-94.
BRÉAL Michel, *Essai de sémantique. Science des significations*, Paris, 1897
BRIAN Éric, « Moyens de connaître les plumes. Étude lexicométrique » in VILQUIN Éric (éd.), *Recherches et considérations sur la population de la France par M. Moheau*, Paris, 1994, pp. 383-396
BROUSSEAU Nicholas, « Lemmatisation et traitement statistique : de nouveaux instruments pour la critique diplomatique ? Le cas des diplômes pseudo-originaux au nom de Louis le Germanique », *Médiévales*, 42-2002, pp. 27-44
BRUCE Christopher W., *The Arthurian name dictionary*, New-York, 1999
BUNDE Armin et HAVLIN Shlomo (éds), *Fractals in Science*, Berlin, 1994
BURGUBURU Paul, *Essai de bibliographie métrologique universelle*, Paris, 1932
BUSSE Dietrich, HERMANNS Fritz, TEUBERT Wolfgang (éds), *Begriffsgeschichte und Diskursgeschichte : Methodenfragen und Forschungsergebnisse der historischen Semantik*, Opladen, 1994
CALAME-GRIAULE Geneviève, *Langage et cultures africaines. Essais d'ethno-linguistique*, Paris, 1977
CALOT Gérard, *Cours de calcul des probabilités*, Paris, 1964 ; *Cours de statistique descriptive*, Paris, 1965
CARCASSONNE Charlotte et HACKENS Tony (éds), *Statistics and Numimatics. PACT - 5*, Strasbourg, 1981
CARCASSONNE Charlotte, *Méthodes statistiques en numismatique*, Louvain-la-Neuve, 1987
CASABONA Jean, *Recherches sur le vocabulaire des sacrifices en grec, des origines à la fin de l'époque classique*, Aix-en-Provence, 1966
CAUVIN Colette, REYMOND Henri et SERRADJ Abdelaziz, *Discrétisation et représentation cartographique*, Montpellier, 1987
CHAUNU Pierre, « L'histoire sérielle. Bilan et perspectives », *Revue historique*, 494-1970, pp. 297-320
CIBOIS Philippe, *La représentation factorielle des tableaux croisés et des données d'enquête : étude de méthodologie sociologique* (thèse), Paris, 1980. *L'analyse factorielle*, Paris, 1983 (1991^{3}) ; *L'analyse des données en sociologie*, Paris, 1984 (1990^{2}). « Introduction à la méthode Tri-deux », *Informatique et Sciences humaines*, 70-71, 1986, pp. 5-13. « Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence », *Bulletin de méthodologie sociologique*, 40-1993, pp. 43-63
COQ Dominique et ORNATO Ezio, « Les séquences de composition du texte dans la typographie du XV^{e} siècle. Une méthode quantitative d'identification », *Histoire & Mesure*, 2-1987, pp. 87-136
COUTROT Bernard et DROESBEKE Fernand, *Les méthodes de prévision*, Paris, 1984
CRESSIE Noël A., *Statistics for Spatial Data*, New York, 1993
DARLU Pierre, DEGIOANNI Anna, RUFFIÉ Jacques, « Quelques statistiques sur la distribution des patronymes en France », *Population*, 3-1997, pp. 607-633
DAUPHINÉ André et VOIRON-CANICIO Christine, *Variogrammes et structures spatiales*, Montpellier, 1988

DAUPHINÉ André, *Chaos, fractales et dynamiques en géographie*, Montpellier, 1995
DEBUS Friedhelm (éd.), *Stadtbücher als namenkundliche Quelle*, Mainz, 2000
DESROSIERES Alain, *La politique des grands nombres. Histoire de la raison statistique*, Paris, 1993
DIEDERICH Paul Bernard, *The frequency of the latin words and their endings,* Chicago, 1939
DILKE Oswald, *Les arpenteurs de la Rome antique*, Sophia Antipolis, 1995 (or. ang. *The Roman Landsurveyors*, 1971)
DJINDJIAN François, « Nouvelles méthodes pour l'analyse spatiale des sites archéologiques », *Histoire & Mesure*, 5-1990, pp. 11-34. *Méthodes pour l'archéologie*, Paris, 1991
DOCKES Pierre et ROSIER Bernard, *Rythmes économiques. Crises et changement social, une perspective historique*, Paris, 1983
DUCASSE Henri (éd.), *Panorama 1985 des traitements de données en archéologie*, Juan-les-Pins, 1985
DUPAQUIER Jacques, « Statistique et démographie historique. Réflexions sur l'ouvrage d'A. Croix, *Nantes et le pays nantais au XVI^e siècle* », *Annales E.S.C.*, 30-1975, pp. 394-401
FERNIE Eric, *Romanesque architecture : design, meaning and metrology*, London, 1995
FLEURY Michel et HENRY Louis, *Nouveau manuel de dépouillement et d'exploitation de l'état-civil ancien*, Paris, 1965 (1985[3])
FLOUD Roderick, *An Introduction to Quantitative Methods for Historians*, London, 1973, 2^e éd. 1979, trad. allemande, Stuttgart, 1980
FLUTRE Louis Ferdinand, *Table des noms propres avec toutes leurs variantes figurant dans les romans du Moyen Age écrits en français et actuellement publiés ou analysés*, Poitiers, 1962
FÖRSTEMANN Ernst, *Altdeutsches Namenbuch*, 1856 ss (*1. Personennamen, 2. Orts- und sonstige geographischen Namen*)
FOTHERINGHAM Stewart et ROGERSON Peter (éds), *Spatial Analysis and GIS*, London, 1994)
FRANKHAUSER Pierre, *La fractalité des structures urbaines*, Paris, 1994
FREY Louis, « Besançon : La Porte Noire. Carrés et diagonales », *Mathématiques, informatique et sciences humaines*, 105-1989, pp. 27-62. « Genèse d'une théorie », *Mathématiques, informatique et sciences humaines*, 39-2001, pp. 5-32
FRITZ Gerd, *Historische Semantik*, Stuttgart, 1998
FURET François, « L'histoire quantitative et la construction du fait historique », *Annales, E.S.C.*, 26-1971, pp. 63-75, repris dans LE GOFF J. et NORA P. (éds), *Faire de l'histoire. I*, Paris, 1974, pp. 42-61
GARNIER Bernard, HOCQUET Jean-Claude et WORONOFF Denis (éds), *Introduction à la métrologie historique*, Paris, 1989
GECKELER Horst, *Zur Wortfelddiskussion*, München, 1971. *Strukturelle Semantik und Wortfeldtheorie*, München, 1982[3]
GENET Jean-Philippe, « La mesure et les champs culturels », *Histoire & Mesure*, 2-1987, pp. 137-169
GENET Jean-Philippe, « Matrices, genres, champs : une approche sur le long terme », in VAILLANT A. (éd.), *Mesure(s) du livre*, Paris, 1992, pp. 57-74. « The dissemination of Manuscripts Relating to English Political Thought in the Fourteenth Century », in JONES M. et VALE M.G.A. (éds), *England and her Neighbours 1066-1453*, London, 1989, pp. 217-237
GERVERS Michael (éd.), *Dating undated medieval charters*, Woodbridge, 2000
GEUENICH Dieter, HAUBRICHS Wolfgang, JARNUT Jörg (éds), *Nomen et gens : zur historischen Aussagekraft frühmittelalterlicher Personennamen*, Berlin, 1997. *Person und Name : methodische Probleme bei der Erstellung eines Personennamenbuches des Frühmittelalters*, Berlin, 2002
GIGLIOLI Pier Paolo (éd.), *Language and Social Context. Selected Readings*, London, 1972
GILLE Bertrand, « Prolégomènes à une histoire des techniques », in ID. (éd.), *Histoire des*

techniques, Paris, 1978, pp. 3-118
GOY Joseph et LE ROY LADURIE Emmanuel (éds), *Les fluctuations du produit de la dîme : conjoncture décimale et domaniale de la fin du Moyen Age au XVIII^e siècle*, Paris, 1972
GREW Raymond et HARRIGAN Patrick J., « L'offuscation pédantesque. Observations sur les préoccupations de J.-N. Luc », *Annales E.S.C.*, 41-1986, pp. 913-922
GUERREAU Alain et POLO de BEAULIEU Marie-Anne, « Classement des manuscrits et analyses factorielles. Le cas de la *Scala coeli* de Jean Gobi », *Bibliothèque de l'École des Chartes*, 154-1996, pp. 359-400
GUERREAU Alain, « Analyse factorielle et analyses statistiques classiques : le cas des Ordres Mendiants dans la France médiévale », *Annales E.S.C.*, 36-1981, pp. 869-912 ; « Observations statistiques sur les créations de couvents franciscains en France, XIII^e-XV^e siècles », *Revue d'histoire de l'Eglise de France*, 70-1984, pp. 27-60. « Analyse statistique des finances municipales de Dijon au XV^e siècle. Observations de méthode sur l'analyse factorielle et les procédés classiques », *Bibliothèque de l'École des Chartes*, 140-1982, pp. 5-34. « Climat et vendanges (XIV^e-XIX^e siècles) : révisions et compléments », *Histoire & Mesure*, 10-1995, pp. 89-147.
« L'analyse des dimensions des édifices médiévaux. Notes de méthode provisoires », in REVEYRON Nicolas (éd.), *Paray-le-Monial. Brionnais-Charolais. Le renouveau des études romanes*, Paray-le-Monial, 2000, pp. 327-335. « L'évolution du parcellaire en Mâconnais, env. 900-env. 1060 », in FELLER Laurent, MANE Perrine, PIPONNIER Françoise (éds), *Le village médiéval et son environnement. Etudes offertes à J.-M. Pesez*, Paris, 1998, pp. 509-535. « Mesures du blé et du pain à Mâcon (XIV^e-XVIII^e siècles) », *Histoire & Mesure*, 3-1988, pp. 163-219.
« Notes statistiques sur les jardins de Saint-Flour (XIV^e siècle) » in BIGET Jean-Louis (éd.), *Les cadastres anciens des villes et leur traitement par l'informatique*, Rome, 1989, pp. 341-357.
« Pourquoi (et comment) l'historien doit-il compter les mots ? », *Histoire & Mesure*, 4-1989, pp. 81-105. « Raymond Aron et l'horreur des chiffres », *Histoire & Mesure*, 1-1986, pp. 51-73.
« A propos d'une liste de dénominations professionnelles dans la France du XIX^e siècle », *Annales E.S.C.*, 48-1993, pp. 979-986
GUILHAUMOU Jacques, « L'historien du discours et la lexicométrie », *Histoire & Mesure*, 1-1986, pp. 27-46
GUIRAUD Pierre, « L'évolution du style de Rimbaud et la chronologie des *Illuminations* » in *Problèmes et méthodes de la statistique linguistique*, Dordrecht, 1959, pp. 127-138. *La sémantique*, Paris, 1972[7]. *Le Jargon de Villon ou le gai savoir de la Coquille*, Paris, 1968. *Le Testament de Villon ou le gai savoir de la Basoche*, Paris, 1970
HECHT Konrad, *Maß und Zahl in der gotischen Baukunst*, Hildesheim, 1979
HEFFER Jean, ROBERT Jean-Louis et SALY Pierre, *Outils statistiques pour les historiens*, Paris, 1981
HENRY Louis et BLUM Alain, *Techniques d'analyse en démographie historique*, Paris, 1988[2]
HOFFMAN Philip T., « Un nouvel indice de la productivité agricole : les baux de Notre-Dame de Paris, 1450-1789 », *Histoire & Mesure*, 6-1991, pp. 215-243
HURON Nicolas, *Termes de topographie urbaine dans les actes des rois de France 840-987*, Paris, 1990
HYMES Dell (éd.), *Language in Culture and Society. A Reader in Linguistics and Anthropology*, New-York, 1964
IFRAH Georges, *Histoire universelle des chiffres*, Paris, 1994
IMBERT Gaston, *Des mouvements de longue durée Kondratieff*, Aix-en-Provence, 1959
IRIGOIN Jean et ZARRI Gian Pierro (éds), *La pratique des ordinateurs dans la critique des textes*, Paris, 1979
IRSIGLER Franz (éd.), *Quantitative Methoden in der Wirtschafts- und Sozialgeschichte der Vorneuzeit*, Stuttgart, 1978
KAMTZ Hans von, *Homerische Personennamen : sprachwissenschaftliche und historische*

Klassifikation, Göttingen, 1982
KENDALL Maurice George et STUART Alan, *The Advanced Theory of Statistics*, dernière édition, 3 vol., 1977-1983 (première : 1943-1946)
KLAPISCH Christiane et DEMONET Michel, « "A uno pane e uno vino" La famille rurale toscane au début du XV^e^ siècle », *Annales E.S.C.*, 27-1972, pp. 873-901
KONERSMANN Ralf, *Der Schleier des Timanthes. Perspektiven der historischen Semantik*, Frankfurt, 1994. *Kritik des Sehens*, Leipzig, 1997
KUCHENBUCH Ludolf, « Pragmatische Rechenhaftigkeit ? Kerbhölzer in Bild, Gestalt und Schrift », *Frühmittelalterliche Studien*, 36-2002, pp. 469-490
KULA Witold, *Les mesures et les hommes*, Paris, 1984 [or. polonais 1970]
LABOV William, *Sociolinguistique*, Paris, 1976
LALOIRE Jean-Claude, *Méthodes de traitement des chroniques. Statistiques et prévisions de ventes*, Paris, 1972
LANDAUER Thomas K., FOLZ Peter W. & LAHAM Darell, « An Introduction to Latent Semantic Analysis », *Discourse Processes*, 25-1998, pp. 259-284 (texte disponible sur internet)
LEBART Ludovic et SALEM André, *Statistique textuelle*, Paris, 1994
LEBLOND Hervé, « Recherches métrologiques sur des plans de bastides médiévales », *Histoire & Mesure*, 2-1987, pp. 55-88
LEFEBVRE Jacques, *Analyses statistiques multidimensionnelles*, Paris, 1976
LEVY-LEBOYER Maurice, « L'héritage de Simiand : prix, profit et termes d'échange au XIX^e^ siècle », *Revue historique*, 493-1970, pp. 77-120
LONGNON Auguste, *Les noms de lieux de la France*, Paris, 1920-1929
MACHABEY Armand, *La métrologie dans les musées de province et sa contribution à l'histoire des poids et mesure en France depuis le treizième siècle*, Paris, 1962
MAIRESSE Jacques (éd.), *Estimation et sondages. Cinq contributions à l'histoire de la statistique*, Paris, 1988
MAITRE Jacques, « Les fréquences des prénoms de baptême en France », *L'année sociologique*, 3-1964, pp. 31-74
MANDELBROT Benoît, « On the theory of word frequencies and on related markovian models of discourse », in JAKOBSON Roman (éd.), *Structure of language and its mathematical aspects*, Providence, 1961, pp. 190-219. « Les constantes chiffrées du discours » in André MARTINET (éd.), *Le langage*, Paris, 1968, pp. 46-56. *Les objets fractals. Forme, hasard et dimension*, Paris, 1995. *Fractales, hasard et finance*, Paris, 1997.
MARTINEZ SOPENA Pascual (éd.), *Antroponimia y sociedad : sistemas de identificación hispano-cristianos en los siglos IX a XIII*, Santiago de Compostela, 1995
MASSONIE Jean-Philippe, « Introduction à la théorie de la mesure », *Histoire & Mesure*, 3-1988, pp. 7-18
MELLET Sylvie, « La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte ? », *Médiévales*, 42-2002, pp.13-26
MEUVRET Jean, « Les données démographiques et statistiques en histoire moderne et contemporaine », in SAMARAN Charles (éd.), *L'histoire et ses méthodes*, Paris, 1961, pp. 893-936
MEYER Heinz et SUNTRUP Rudolf, *Lexikon der mittelalterlichen Zahlenbedeutungen*, München, 1987
MEYER Heinz, *Die Zahlenallegorese im Mittelalter*, München, 1975
MOISAN André, *Répertoire des noms propres de personnes et de lieux cités dans les chansons de geste françaises et les œuvres étrangères dérivées*, Genève, 1986
OHLER Norbert, *Quantitative Methoden für Historiker. Eine Einführung. Mit einer Einführung in die EDV von Hermann Schäfer*, München, 1980
OHLY Friedrich, *Ausgewählte und neue Schriften zur Literaturgeschichte und zur Bedeutungsforschung*, Stuttgart, 1995. *Schriften zur mittelalterlichen Bedeutungsforschung*,

Darmstadt, 1977
ORNATO Ezio, « L'exploitation des sources narratives médiévales dans l'histoire du climat : à propos d'un ouvrage récent », *Histoire & Mesure*, 3-1988, pp. 403-449. *La face cachée du livre médiéval*, Roma, 1997
PARISOT Jean-Paul et LAMBERT Georges, « Les "années" et la rotation de la terre », *Histoire & Mesure*, 1-1986, pp. 119-146
PAUL Hermann, *Principien der Sprachgeschichte*, Halle, 1886[2]
PIATIER, André, *Statistique descriptive et initiation à l'analyse*, Paris, 1966[2]
PUMAIN Denise et SAINT-JULIEN Thérèse, *L'analyse spatiale. Les localisations*, Paris, 1997. *Les interactions spatiales. Flux et changements dans l'espace géographique*, Paris, 2001
REICHARD Rolf (éd.), *Aufklärung und historische Semantik : interdisziplinäre Beiträge zur westeuropäischen Kulturgeschichte*, Berlin, 1998
REISIG Karl, *Vorlesungen über die lateinische Sprachwissenschaft*, Leipzig, 1839
ROBIN Régine, « Fief et seigneurie dans le droit et l'idéologie juridique à la fin du XVIII[e] siècle », *Annales historiques de la Révolution française*, 43-1971, pp. 554-602. « Le champ sémantique de *féodalité* dans les cahiers de doléances généraux de 1789 », *Bulletin du centre d'analyse du discours de l'Université de Lille*, 2-1975, pp. 61-86. *Histoire et linguistique*, Paris, 1973
ROSA Guy, « Comptes pour enfants. Essai de bibliométrie des livres pour l'enfance et la jeunesse (1812-1908) », *Histoire & Mesure*, 5-1990, pp. 343-369
ROUANET Henry, LE ROUX B., BERT M.-C. et BERNARD J.-M., 1. *Procédures naturelles*, Paris, 1987 ; 2. *Analyse inductive des données*, Paris, 1990 ; 3. *Analyse des données multidimensionnelles*, Paris, 1993
ROUANET Henry, LECOUTRE Marie-Paule, BERT Marie-Claude, LECOUTRE Bruno, BERNARD Jean-Marc, *L'inférence statistique dans la démarche du chercheur*, Berne/Berlin, 1991
SAINT-AUBIN Jean-Paul, *Le relevé et la représentation de l'architecture*, Paris, 1992.
SAINT-JULIEN Thérèse, *La diffusion spatiale des innovations*, Montpellier, 1985
SALOMIES Olli, *Adoptive and polyonymous nomenclature in the Roman Empire*, Helsinki, 1992
SALY Pierre, *Méthodes statistiques descriptives pour les historiens*, Paris, 1991
SANTOS DOMINGUEZ Luis Antonio et ESPINOSA ELORZA Rosa Maria, *Manual de semantica historica*, Madrid, 1996
SAPORTA Gilbert, *Probabilités, analyse des données et statistique,* Paris, 1990
SAPOVAL Bernard, *Universalités et fractales*, Paris, 1997
SCHÄRLIG Alain, *Compter avec des cailloux. Le calcul élémentaire sur l'abaque chez les anciens Grecs*, Lausanne, 2001
SCHMIDT Lothar (éd.), *Wortfeldforschung. Zur Geschichte und Theorie des sprachlichen Feldes*, Darmstadt, 1973
SELVIN Hanan, « Durkheim, Booth and Yule : the non-diffusion of an intellectual innovation », *Archives européennes de sociologie*, 17-1976, pp. 39-51
SHANNON Claude Elwood, « A mathematical theory of communication », *The Bell System Technical Journal*, 27-1948 (disponible sur internet)
SIMIAND François, *Statistique et expérience. Remarques de méthodes*, Paris, 1922
SOLIN Heikki, *Namenpaare : eine Studie zur römischen Namengebung*, Helsinki, 1990
TAVERDET Gérard, *Microtoponymie de la Bourgogne*, 12 tomes, Dijon, 1989-1993
TESNIERE Michel, « Fréquence des noms de famille », *Journal de la société de statistique de Paris*, 116-1975, pp. 24-32
THOME Helmut, *Grundkurs Statistik für Historiker*, 2 vol., Köln, 1989-1990
TRIER Jost, *Der deutsche Wortschatz im Sinnbezirk des Verstandes*, Heidelberg, 1931. *Zur Wortfeldtheorie*, Berlin, 1973
TUKEY John Wilder, *Exploratory data analysis*, Reading Mass., 1977
UPTON Graham J. et FINGLETON Bernard, *Spatial Data Analysis by Example*, New York, 1985-

1989
VALLET Antoine, *Les noms de personnes du Forez et confins (actuel département de la Loire) aux XIIe, XIIIe et XIVe siècles*, Paris, 1961
VENABLES William et RIPLEY Brian, *Modern applied statistics with S*, New-York, 2002[4]
VESSEREAU André, *La statistique*, Paris, 1947 (plusieurs dizaines de rééditions…)
VOIRON Christine, *Analyse spatiale et analyse d'images*, Montpellier, 1995
WEHRLI Max, « Der mehrfache Sinn. Probleme der Hermeneutik », in *Literatur im deutschen Mittelalter. Eine poetologische Einführung*, Stuttgart, 1984, pp. 236-270
WHORF Benjamin Lee, *Language, Thought, and Reality. Selected Writings of BLW*, New-York, 1956
ZAJDENWEBER Daniel, *Hasard et prévision*, Paris, 1976
ZIPF George Kingsley, *Human Behavior and the Principle of Least-effort*, Cambridge Mass., 1949
ZUPKO Ronald, *French weights and measures before the Revolution : a dictionnary of provincial and local units*, Bloomington, 1978. *Italian weights and measures from the Middle Ages to the nineteenth century*, Philadelphia, 1981. *A Dictionnary of weights and measures for the British Isles : the Middle Ages to the twentieth century*, Philadelphia, 1985

Table des matières