JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JEAN-PAUL VALOIS

La représentation graphique des données : des héritages aux pratiques nouvelles

Journal de la société française de statistique, tome 141, n° 4 (2000), p. 93-107

<http://www.numdam.org/item?id=JSFS_2000__141_4_93_0>

RÉPONSE DE JEAN-PAUL VALOIS

La représentation graphique des données : des héritages aux pratiques nouvelles.

RÉSUMÉ

En conclusion et en réponse partielle aux divers intervenants, une synthèse rapide de l'historique des graphiques est mise en regard avec l'évolution de la présentation des documents écrits et notamment avec le choix dominant en faveur d'une présentation linéaire du texte. Le rôle important de l'image pour faire saisir ordres de grandeur ou relations multiples entre variables conduit à s'interroger sur la place de l'image dans la saisie de l'information et sur les contraintes qui en découlent pour l'organisation du document écrit. Parmi les pratiques nouvelles, l'interactivité répond aux besoins d'exploration de problèmes complexes; d'autre part la description de bases de données pose de nouvelles contraintes pour les graphiques traditionnels. Initiée dans les magazines, une place plus grande des images dans les documents écrits se fait jour, au moins pour les textes documentaires ou argumentaires. Un exemple industriel de ces nouvelles pratiques est donné en annexe.

En commençant la réponse qui termine ce dossier, il nous est agréable de remercier vivement les intervenants qui ont accepté d'apporter leur concours à ce débat : Jacques Bertin, William Cleveland, Antoine de Falguerolles, Michael Friendly et Daniel Denis, Dominique Ladiray, Gilles Palsky, Ian Spence, Fernando Tusell et Adalbert Wilhelm. Chacun avec un éclairage spécifique, ils ont conforté certaines intuitions proposées; ils ont apporté des compléments qui enrichissent le débat, et des corrections sur quelques points évoqués trop schématiquement dans l'article initial.

Pour le lecteur statisticien auquel s'adresse cette revue, le point fondamental sous-jacent à ce dossier réside dans l'intérêt d'utiliser les graphiques conjointement aux méthodes numériques. Ce point est relevé particulièrement par Antoine de Falguerollles, Michael Friendly et Daniel Denis, Dominique Ladiray et Fernando Tusell.

Nous nous proposons d'organiser notre réponse (en guise de conclusion toute provisoire) selon deux directions : sous le thème de l'héritage, nous essayons de synthétiser une partie au moins des remarques diverses faites par les commentateurs. Suivant ensuite leurs sollicitations, nous avons choisi de développer trois points : l'interactivité des logiciels, la description de bases de données, et l'évolution du rôle des graphiques dans les présentations écrites; ces trois points ont été regroupés sous l'éclairage de pratiques nouvelles.

1. HÉRITAGES

Rien n'est plus spontané aujourd'hui que de reporter les points dans un repère orthonormé ou de dresser un histogramme. Ces gestes simples sont le résultat d'une longue histoire dont le présent dossier a esquissé les grandes lignes.

Ian Spence apporte d'intéressantes précisions sur le contexte dans lequel William Playfair a ouvert le rideau de la scène des graphiques modernes. L'œuvre de William Playfair a fait date et sert de point de repère, elle constitue la première utilisation des graphiques préfigurant l'intérêt actuel pour l'exploration des données. Comme les statisticiens, les historiens sont confrontés à un choix délicat pour intégrer ou non des individus marginaux : Ian Spence montre que l'œuvre de Playfair n'est peut-être pas significative de son époque. Les contributions précédentes montrent qu'il n'est pas aisé d'avoir une vision exacte du rôle des précurseurs, d'autant que l'émergence des graphiques met en œuvre un riche faisceau de relations entre différentes techniques, approches ou disciplines. En évoquant le milieu culturel qui a permis à Jacques Bertin de marquer un changement d'acte majeur dans la pièce, Gilles Palsky confirme l'intérêt de cette approche, et insiste sur le rôle important joué par les représentations cartographiques.

Outre les influences culturelles, des contraintes techniques sont mises en relief par Ian Spence : Playfair devait graver lui-même ses plaques. De même Tukey dessinait sans ordinateur, Bertin maniait ses réglettes à la main pour obtenir le meilleur tri. Les idées ont ici précédé les facilités techniques. Cela n'empêche pas de souligner, avec Fernando Tusell, les facilités ultérieures offertes par les logiciels en matière de réalisation graphique. La mise à disposition de cette puissance graphique est parallèle à la puissance de calcul aujourd'hui bien répandue, toutes deux étaient il y a quelques décennies encore inespérées voire invraisemblables. Dans cette montée en puissance de l'informatique, les graphiciens n'ont pas toujours été les premiers servis, et Jacques Bertin regrette encore que les circonstances ne lui aient pas permis de développer un programme reprenant l'ensemble de ses « procédures manuelles efficaces ». Notons toutefois des tentatives partielles en ce sens : Caraux (1984), Falguerolles et al. (1997), Chauchat et Risson (1998) ; mais il est vrai que ces réalisations, œuvre de chercheurs convaincus et passionnés, n'ont pas eu la chance d'une diffusion grand public.

Nous proposons de mentionner ici à la lumière de Vandendorp (1999) d'autres circonstances apparemment éloignées de notre sujet mais qui permettront par la suite d'élargir la discussion. Elles ne contredisent pas les différentes circonstances détaillées par nos commentateurs, mais contribuent à enrichir et nuancer le tableau d'ensemble.

L'œuvre de Playfair se situe à un moment charnière dans l'histoire de la lecture, alors que la place du lecteur devient plus active, et qu'une tendance à la lecture sélective commence à apparaître. Le succès mitigé de Playfair au XVIIIe siècle, souligné par Ian Spence, peut être mis en rapport avec une tendance promue par les éditeurs : ils ont fait prévaloir à cette époque, et de façon continue depuis lors, une évolution vers le modèle du texte technique qui nous est familier aujourd'hui. La priorité accordée à une lisibilité

optimale a conduit successivement à une codification de l'orthographe et de la ponctuation, à une rédaction impersonnelle et sans effets de rhétorique, à une exigence de cohérence qui conduit à incorporer au texte tous les éléments qui en font un lieu de signification autonome et indépendante.

Le but recherché a été de privilégier le confort de saisie linéaire du texte par le lecteur. L'agréable difficulté que l'on rencontre parfois pour s'extraire d'un bon roman montre à quel degré d'efficacité est parvenue la 'machine textuelle' (selon l'expression de Vandendorp). La prédilection pour le fil continu a induit un mouvement d'épuration de la mise en page, qui contraste avec la gestion de l'espace dans les manuscrits et dans les incunables du début de l'imprimerie, où les gloses encadraient le texte principal et maintenaient une constante proximité spatiale avec lui.

Dans le même temps, différents procédés se sont mis en place pour favoriser la tabularité, c'est-à-dire la possibilité pour le lecteur d'accéder à des fragments dans l'ordre qu'il choisit. Alors que des essais variés avaient encore cours au milieu du XVIII^{e} siècle, c'est la tabularité fonctionnelle (titres, table des matières, alinéa) qui a été privilégiée par les imprimeurs, et non la tabularité visuelle (coexistence sur une même page de divers éléments textuels ou iconographiques).

Le recours à l'image, et pour ce qui nous concerne à la représentation graphique des données, se justifie dans les deux cas suivants.

- Lorsqu'il s'agit de transmettre une information sur les ordres de grandeur. Les canaux de perception visuelle sont alors plus efficaces que ceux dédiés au langage (Deheane *et al.*, 1998,1999). L'objectivité des graphiques, discutée par Dominique Ladiray, doit être appréciée en tenant compte des effets de la présentation sélective ('plans visuels') qui induit un ordre de perception (Valois, 1993) ; sans reprendre la discussion de l'article initial, nous sommes reconnaissants à Gilles Palsky d'indiquer des antécédents historiques (dès le XVII^{e} siècle) pour cette idée dont il souligne l'importance.

- Lorsqu'il s'agit d'aider à percevoir des relations complexes dans le domaine multivarié. La puissance d'une présentation simultanée des informations a été montrée par les travaux d'ergonomie (voir entre autres Carswell et Wickens, 1990).

Dans quel sens peut-on dire qu'il y a lecture du graphique ? « Il y a lecture au sens où l'usager interprète des signes, fait des choix et produit du sens en mettant des données en relation avec un contexte... Pour qui accepterait pleinement la possibilité de lire des images, la tentation serait grande de mettre en place une machine à lire qui soit aussi efficace en cette matière que pour le texte » (Vandendorp, *ib.*). Il nous semble que les travaux de Jacques Bertin, ou ceux qui s'inspirent de son œuvre, correspondent à une telle tentative. Nous avons précédemment noté que les travaux de neurophysiologie suggèrent de guider visuellement le lecteur en fonction de la question posée, en utilisant les signes comme un « rail sémantique », selon l'expression employée par Vandendorp pour la mise en page des textes. Les règles pratiques proposées par Bertin ont pu être utilisées pour discuter de la mise en page (Martin,

1989), ce qui confirme des points communs entre lecture des textes et lecture des graphiques.

2. LES PRATIQUES NOUVELLES

L'importance des possibilités interactives offertes par les logiciels actuellement disponibles est soulignée par Adalbert Wilhelm et Fernando Tusell. Une interrogation sur le rôle des graphiques dans l'étude des bases de données est formulée par William Cleveland, Dominique Ladiray et Fernando Tusell. Nous élargirons la discussion sur la place des graphiques dans les documents, en écho aux remarques ou questions d'Antoine de Falguerolles sur l'avenir des graphiques.

2.1 L'interactivité

Un certain nombre de logiciels proposent aujourd'hui des graphiques interactifs. Si l'on prend l'exemple du logiciel SAS, trois solutions différentes sont proposées :

- soit une réalisation assistée de graphiques traditionnels (module Graph-n-Go et Enterprise Guide disponibles à partir de la version 8),
- soit un module commercialisé dès les années 80 sous le nom de SAS/INSIGHT, qui propose une réalisation assistée plus rudimentaire (pas de titre sur les figures), mais met en œuvre un lien dynamique entre les graphiques,
- soit la possibilité de munir les graphiques d'un lien hypertexte.

D'autres logiciels proposent aujourd'hui l'une ou l'autre des formules, mais la réalisation assistée des graphiques demeure la plus répandue (EXCEL, COREL CHART...). La programmation objet permet de redonner la priorité aux données et de déclencher à la demande des formes graphiques diverses à partir d'une même combinaison de variables. Par contre, le lien dynamique entre les graphiques n'est encore proposé que par quelques éditeurs de logiciels statistiques.

Nous n'avons pas abordé cette question dans le dossier d'ouverture, car la plupart du temps les logiciels grand public ou commerciaux ne proposent guère de formes graphiques réellement innovantes. La richesse de la panoplie offerte n'est souvent qu'apparente, elle résulte surtout d'une variété d'effets visuels (couleurs, pseudo-relief), qui sont des variantes des procédés de report les plus conventionnels.

La contribution d'Adalbert Wilhelm fournit les éléments de réflexion pour aborder la question de l'interactivité des graphiques. Nous ne rappellerons ici qu'un point clef : le lien dynamique consiste à répercuter sur un graphique les effets d'une action effectuée sur un autre graphique juxtaposé sur l'écran ; par exemple, on clique sur une barre d'histogramme et le lien dynamique

repère dans les autres graphiques (par la taille, la couleur ou la police) tous les individus englobés dans cette tranche de valeurs.

Il n'est certes pas commode d'illustrer sur une page imprimée, fixe par définition, un comportement dynamique et actif de l'utilisateur, mais pour les lecteurs non familiers de cette technique, la figure 1 (voir Annexe en fin de cette réponse) est proposée comme illustration du propos et des suggestions d'Adalbert Wilhelm.

L'auteur confesse faire un large usage de cette technique. La coexistence de graphiques variés permet de visualiser certains au moins des aspects d'un contexte multidimensionnel, comme le souligne William Cleveland : en exhibant les relations entre individus, le lien entre graphiques permet de « pénétrer » dans ce monde multidimensionnel, c'est-à-dire de suggérer des éléments de réponse aux questions que l'on se pose. Ce type d'environnement graphique muni de liens dynamiques serait probablement peu utile pour aborder des questions inférentielles monovariées. Le succès de ces méthodes provient de l'extension actuelle de l'analyse des données à des situations dans lesquelles les réponses aux questions s'enchaînent à la façon d'une énigme policière, et ne peuvent pas toujours être décrites par une mise à plat des variables dans un seul tableau, comme on le pratique dans les méthodes factorielles, qui supposent un raisonnement simultané portant sur l'ensemble des variables. Cette remarque rejoint donc la réflexion de Fernando Tusell sur le défi que représentent les problèmes industriels complexes aussi bien pour les graphiques que pour la statistique traditionnelle.

Une autre forme d'interactivité est réalisée par les liens hypertextes disponibles par exemple dans le langage HTML. Ces liens ont été initialement conçus pour des relations textuelles, d'où leur nom, ils sont aujourd'hui largement utilisés sur Internet. Il est désormais possible de les appliquer aux éléments d'un graphique. On peut ainsi cliquer sur une barre d'histogramme pour avoir accès à une explicitation de la sous-population correspondante, par affichage d'un tableau auxiliaire. Les points d'un graphique bivarié peuvent également être « sensibilisés ». En cliquant sur tel point d'une carte, l'utilisateur va faire apparaître un graphique complémentaire ; pour le domaine de données choisi pour nos figures, on peut accéder immédiatement aux courbes de production de chaque puits en cliquant sur son symbole cartographique. L'interactivité est ici uniquement formelle, puisque l'utilisateur accède à une série d'images graphiques préalablement stockées telles des cartes postales ; il n'y a pas production d'une image nouvelle comme dans le cas du lien dynamique. Le lien hypertexte appliqué aux graphiques change uniquement le mode d'accès aux images : l'utilisateur les consulte dans l'ordre des questions qu'il se pose, par exemple en vérifiant si des points voisins sur la carte comportent les mêmes tendances, au lieu d'être obligé de rechercher ces informations dans un fichier d'images rangé par ordre alphabétique.

Le lien hypertexte appliqué aux graphiques, et le lien dynamique entre graphiques ont pour caractère commun de mettre l'utilisateur en situation de juxtaposer un nombre plus ou moins grand de graphiques, qu'il peut explorer à tour de rôle sans ordre imposé. Cette démarche trouve un antécédent dans

la présentation dite 'mosaïque' proposée d'abord par les magazines, et qui s'est étendue aux livres à partir des années 60 (Vandendorp, *ib.*) ; elle est en rupture complète avec la priorité traditionnelle donnée à la saisie linéaire du texte. L'interactivité des environnements logiciels prolonge et exacerbe alors une situation de dialogue, que Vandendorp estime implicite dans tous les textes argumentaires.

2.2 Le traitement de bases de données

L'un des défis proposés actuellement à l'analyse de données est l'extraction d'information à partir de bases de données importantes. Il s'agit là d'un axe de recherche dans lequel les techniques graphiques se doivent d'apporter leur contribution, aux côtés d'autres techniques (Cleveland, 2001). Nous nous bornerons ici à quelques remarques rapides. Le nombre très élevé d'individus (plus de 20 000 pour l'exemple donné en Annexe, figures 2 et 3) implique des contraintes pour les graphiques même les plus conventionnels ; nous examinerons le cas du report bivarié et de l'histogramme.

Dans le report bivarié des individus, se produisent de nombreuses superpositions ; la perception visuelle du nuage, qui ne tient compte que de la juxtaposition des points, et non de leur superposition, est de ce fait inappropriée (Annexe, figure 2). Il apparaît indispensable de traiter un tel graphique en courbes de densité de points (détails dans l'article initial ou dans Valois et Grun Réhomme, 2001) ; cette technique permet de restituer les lignes dominantes du nuage de points.

Le nombre très élevé d'individus issu de bases de données importantes induit des particularités par rapport à l'analyse statistique traditionnelle. Quand un nuage de points se prête à une modélisation par régression, il est d'usage de considérer de façon indifférenciée les points écartés de la tendance centrale, qu'ils soient au-dessus ou en-dessous de la droite (d'où le recours aux moindres carrés). On suppose de même que les individus ont une distribution suffisamment homogène tout au long de la droite de régression. L'annexe, figure 2, montre qu'en présence d'une base de données importante, ces deux aspects doivent être pesés soigneusement : il peut arriver que les individus marginaux soient en nombre suffisant pour former çà ou là des sous-groupes consistants et dignes d'intérêt, et par ailleurs la population peut se répartir de façon inhomogène le long de la tendance principale. Ces détails de répartition ne peuvent qu'être négligés pour des populations de faible effectif (une centaine de points par exemple), mais peuvent faire l'objet d'un examen complémentaire si les sous-groupes comportent chacun plusieurs centaines voire quelques milliers d'individus : Cleveland (2001) voit là une caractéristique de l'approche "Data Mining" des grandes bases de données, alors que l'analyse de données traditionnelles met l'accent sur un résumé des tendances principales. Fernando Tussel agrée sur ce point en voyant dans ce défi l'une des clefs de l'analyse de bases de données importantes.

Certaines utilisations de l'histogramme se trouvent mises en difficulté quand on traite de telles bases de données. L'histogramme en lui-même n'est pas en cause ; contrairement au report bivarié, il convient parfaitement pour des

lots de données très volumineux. Mais si l'on distingue une sous-population en notant sa contribution à chaque classe par une couleur différente (des exemples sont donnés en figure 1), la lisibilité de cette opération devient faible voire nulle si la sous-population examinée ne représente qu'une faible partie de l'effectif global. La comparaison à la population principale, telle que réalisée en figure 1, n'est alors pas efficace visuellement, et l'on peut conseiller de recourir à des reports dont l'effet visuel est indépendant de l'effectif, par exemple : courbe de fréquence cumulée, plot des quantiles, ou boîtes de dispersion normalisées présentées dans le dossier d'ouverture; un exemple d'utilisation de cette dernière représentation est détaillé en annexe, figure 3.

Nous agréons avec les limites soulignées par Fernando Tusell concernant la représentation des données par une enveloppe : celle-ci ne peut à l'évidence jouer qu'un rôle transitoire – qui peut s'avérer suffisant dans quelques situations – mais une représentation du détail des individus (tel que présenté en figure 3 de l'annexe) est nécessaire dès que l'on cherche à analyser le détail des relations multivariées. Nous retrouvons dans cette succession (approche d'ensemble, analyse plus détaillée) une démarche décrite par Jacques Bertin. L'utilisation de formes graphiques différentes à chaque stade converge avec les remarques de Michael Friendly et Daniel Denis sur la nécessité de penser les graphiques en tenant compte à la fois de la grammaire des signes et de la relation des graphiques à la question posée. En élargissant ce dernier point, c'est la place des graphiques dans le document qui retiendra notre attention pour terminer.

2.3 Le rôle des graphiques et la place des graphiques

Reléguée il y a quelques décennies au rôle d'illustration auxiliaire du texte, l'image en devient un partenaire indispensable; le mot 'partenaire' a également été retenu par Tukey que nous rejoignons sur ce point : les graphiques ne sont pas seulement des partenaires dans la rédaction (présentation finale), mais véritablement des outils de raisonnement tout au long de l'étude. On peut concéder à Jacques Bertin que l'implication d'autres personnes peut être moindre vis à vis de la construction graphique; néanmoins son œuvre repose sur le postulat que les règles de perception – et donc d'efficacité de présentation – sont générales. En outre, le statisticien lui-même est impliqué le premier dans la recherche de sens à travers les données, plusieurs de nos commentateurs le soulignent avec force.

Outil pour découvrir du sens dans les données, le graphique doit alors être utilisé de façon itérative. La question n'est plus de produire un graphique pour illustrer in fine un résultat, mais au contraire d'imaginer au préalable le graphique approprié pour apporter des éléments de réponse à la question posée : il faut ensuite manipuler les données pour obtenir le graphique désiré. En plein accord avec Jacques Bertin (discussion orale), cette pratique fait s'effondrer la limite traditionnelle entre graphique de travail et graphique de communication : réalisé dans ces conditions, un tel graphique fait parler les données en rapport avec la question posée, et la version finale, dite de communication, ne différera que par un ajustement ultime des titres

et graduations. Dans sa contribution, Jacques Bertin résume cette position en plaidant pour que la graphique permette « la recherche, la réflexion et la décision ». En montrant dans l'article initial l'intérêt possible de formes graphiques méconnues ou simplement sous-utilisées, nous n'avons pas voulu promouvoir tel graphique comme nouvelle panacée ; nous avons simplement cherché à montrer que des formes graphiques inhabituelles peuvent s'avérer parfois plus efficaces que les formes conventionnelles, et qu'à des questions et situations variées devaient répondre des formes graphiques diversifiées.

Nous proposons maintenant d'étendre la réflexion au rôle respectif du texte et des graphiques. Une tendance s'affirme pour déléguer à l'image une part croissante des données descriptives et référentielles. Les travaux de Benzecri, tout autant que de Tukey ou de Bertin, ou plus récemment de Wolde (références dans le dossier d'ouverture), se sont inscrits dans cette évolution, en mettant la représentation graphique au cœur de la démarche intuitive de perception de l'information utile, dans le dialogue implicite que mène le statisticien avec les données. L'autorité du visuel pour la collecte de faits semble actuellement le fait d'une évolution culturelle (Gauthier, 1996) dans laquelle s'insère le travail du statisticien. Concrètement, pour les rapports d'études utilisant l'analyse des données, une tendance est de faire jouer désormais aux images un rôle majeur dans la transmission d'informations, voire dans la collecte de sens par l'utilisateur. En environnement industriel, cette tendance est nettement amorcée ; elle est accélérée par la facilité actuelle de maniement et de transmission des documents iconographiques, alors qu'à l'inverse la rédaction d'un texte technique en respectant des normes formelles strictes s'avère longue et coûteuse. Le texte se concentre alors sur la définition du contexte et des articulations majeures (conditions de réalisation, structuration des conclusions). On ne peut en effet envisager de tout déléguer à l'image ; Vandendorp, à la suite de bien d'autres auteurs, insiste sur l'importance d'une légende évocatrice. Les articulations permettant de donner du sens doivent être clairement perçues par le lecteur, qui doit en particulier savoir par où commencer la collecte des traits significatifs, où l'arrêter, dans quel ordre en établir les rapports, bref identifier clairement ce que l'image doit lui dire.

Si ces conditions sont respectées, l'image peut jouer un rôle plus grand que par le passé, les moyens techniques et les habitudes culturelles peuvent aujourd'hui donner leur plein épanouissement aux intuitions de Playfair soulignées par Ian Spence : « Un lecteur habitué à la richesse de l'information fournie par les graphiques s'attendra à ce qu'on y ait recours chaque fois que c'est possible » (Vanderdorp, *ib.*).

La place nouvelle des graphiques aboutit à concevoir des écrits comportant une surface équivalente de texte et d'image, comme illustré dans l'Annexe ci-après. Cela permet une mise en regard systématique des illustrations et du commentaire correspondant. Gunter (1994) a noté cette qualité dans l'ouvrage de Cleveland (1993). Comme pour nombre de formes graphiques, on peut trouver dans la littérature ancienne des antécédents à une telle disposition : Vandendorp (ib, p. 60) cite un ouvrage de 1632 entièrement conçu pour être

vu en double page, au point que la numérotation ne prenait en compte qu'une page sur deux.

On peut conclure cet aspect du dossier en notant que la coexistence texte/image se fait actuellement au bénéfice du texte, qui donne son sens à l'image. L'évolution en cours remet en cause l'équilibre antérieurement admis qui confinait les images à un rôle d'illustration subalterne. Nous ne savons pas actuellement jusqu'où ira cette tendance. L'intérêt souligné par Kosslyn (1994) d'intégrer des éléments de légende dans l'image elle-même peut conduire à une intégration encore plus poussée. Cependant on ne peut intégrer à l'image que quelques mots descriptifs ou suggestifs, les développements logiques devront être fournis séparément, idéalement juste à côté de l'image. Il est possible d'imaginer pour un avenir proche des documents hypertextes qui délivrent les arguments abstraits sous forme d'éléments sonores, que l'on déclencherait en cliquant sur des icônes appropriées : mais il est vraisemblable qu'un tel document hybride ne conviendra que dans des cas particuliers, en outre il ne conviendrait qu'au stade de la présentation finale. On peut penser en revanche que les liens hypertextes appliqués à l'enchaînement de graphiques ont de beaux jours devant eux, du moins dans les contextes où la transmission par la seule voie électronique s'avère possible et suffisante.

Soulignons que cette tendance ne concerne pas tous les documents écrits, et que les documents s'adressant à un public spécialisé et reposant sur un argumentaire abstrait – et a fortiori les textes de type récit – maintiendront leur spécificité. Mais on peut penser qu'elle s'affirmera au contraire dans les textes techniques argumentaires plus concrets ou à vocation plus descriptive.

RÉFÉRENCES COMPLÉMENTAIRES

CARAUX G. (1984), Réorganisation et présentation visuelle d'une matrice de données numériques, un algorithme itératif, *Revue de Statistique Appliquée*, XXXII, (4), 5-23.

CHAUCHAT J.H., RISSON A. (1998), Bertin's Graphics and Multidimensional Data Analysis, in Visualization of Categorical Data, Blasius J. et M. Greenacre M., ed., Academic Press, chap. 3, 37-45.

CLEVELAND W.S. (2001), Visualizing Very Large Internet Traffic Databases, XXXIII^{e} Journées de Statistique, Nantes, 14-18 mai 2001.

DE FALGUEROLLES A., FRIEDRICH F., SAWITZKI G. (1997), A tribute to J. Bertin's graphical data analysis, in Advances in Statistical Software, Bandilla W. et Faulbaum F., ed., 6, Lucius & Lucius, Stuttgart, 11-20.

VALOIS J.-P., GRUN-REHOMME M. (2001), Boîtes de dispersion et relations multivariées, XXXIII^{e} Journées de statistique, Nantes, 14-18 mai 2001.

VANDENDORP C. (1999), Du papyrus à l'hypertexte, essai sur les mutations du texte et de la lecture, Paris, La Découverte (Cahiers libres), 268 pp.

ANNEXE : COMMENTAIRES SUR LA FIGURE 1

Problème posé

Sur 4000 roches issues d'un champ pétrolier à gaz plissé en dôme, ont été mesurées la porosité (Phi) et la perméabilité (K) qui conditionnent le comportement des hydrocarbures. On connaît la cote relative ZF de chaque prélèvement entre la base (0) et le sommet (100) de la formation géologique auquel il appartient, 6 formations ont été reconnues (0 à 5). La question est de définir des relations entre Phi et K.

Observations

Le report bivarié montre une dispersion décourageante (sous-figure supérieure). La deuxième sous-figure examine les prélèvements pour lesquels K a la valeur minimale (inférieure au seuil de détection, d'où l'alignement des points sur l'horizontale), ils ont été entourés à la souris. Le lien dynamique repère leur contribution dans les histogrammes (partie noire des barres) : ils sont omniprésents, mais très largement dominants dans la formation 0.

Dans la troisième sous-figure, quelques clics de souris ont permis de ne conserver à l'écran que les prélèvements appartenant à la formation 1. La dispersion des mesures Phi, K est encore très importante. L'histogramme des cotes relatives est parcouru en cliquant successivement sur les barres ; pour les cotes supérieures à 80 % (les points correspondants sont ici en symboles gras), une majorité des points se répartit autour d'une ligne parallèle à la bissectrice. Pour les autres cotes (points plus petits), les prélèvements se concentrent en partie gauche ou supérieure du report (Phi,K).

Pour valider cette impression, on dresse les courbes d'isodensité de points pour la formation 1 (on a limité la représentation à 75 % des valeurs formant le cœur de la population dans chaque graphique Phi, LogK). Les prélèvements des cotes inférieures (sous-figure inférieure gauche) se cantonnent majoritairement près de l'axe vertical. Les échantillons sommitaux (sous-figure centrale) se répartissent majoritairement près de la bissectrice, selon une relation que l'on pourrait modéliser par régression. Cette tendance s'accentue sur les prélèvements situés à moins de 2 km du cœur de la structure (losanges en sous-figure inférieure droite).

Interprétation

La formation 0 sous-jacente au réservoir est quasi-imperméable, ce qui explique l'accumulation du gaz dans les formations sus-jacentes.

Des métadonnées externes permettent d'interpréter les particularités de la formation 1 : des perforations y sont décrites dans sa partie supérieure, liées à une karstification ancienne, elles augmentent conjointement la porosité et la perméabilité, particulièrement au cœur de la structure. Dans le reste de cette formation dolomitique, la roche massive réagit en fracturation (élévation de K alors que Phi reste très faible).

Les relations spatiales en x, y, z expliquent donc la dispersion des mesures Phi, K.

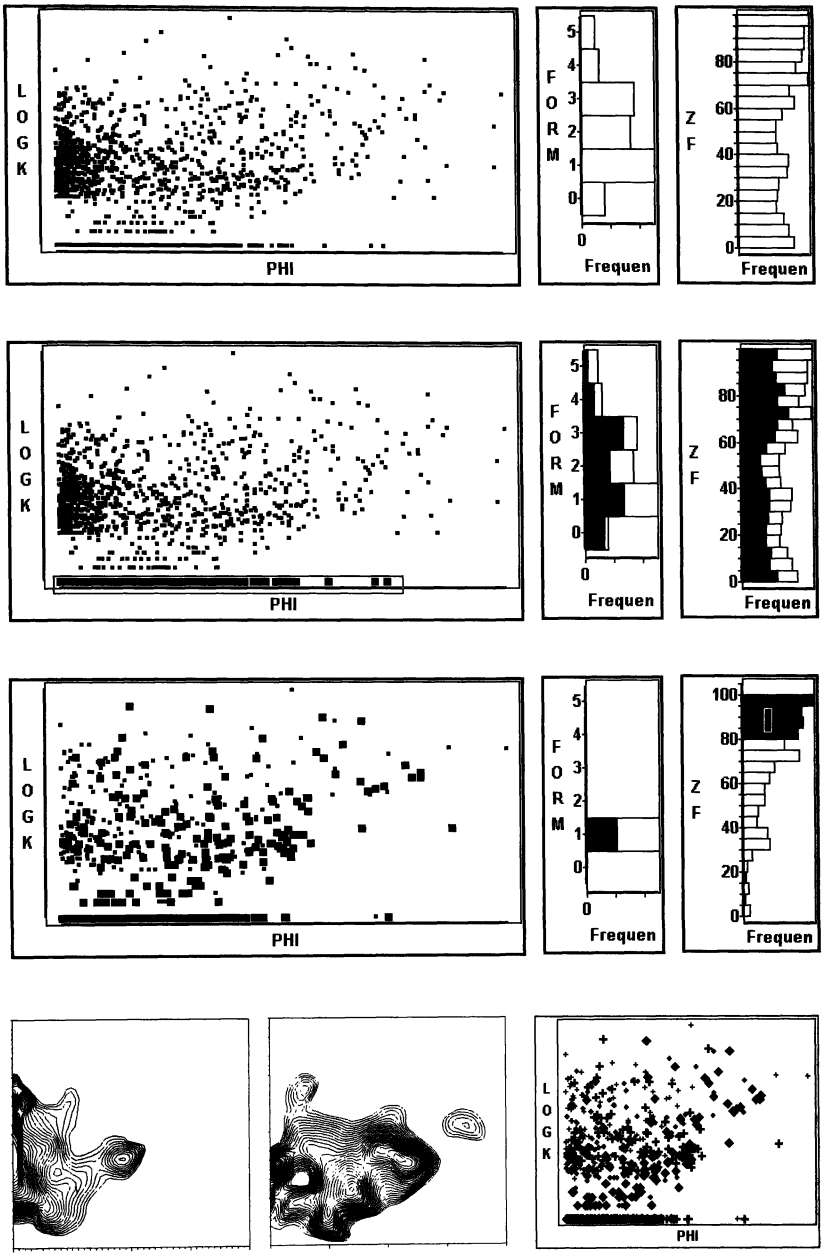

FIG 1. — Possibilités offertes par le lien dynamique entre graphiques, illustration sur un cas industriel.

ANNEXE : COMMENTAIRES SUR LA FIGURE 2

Origine des données, problème posé

Nous utilisons ici une base de données comportant près de 20000 réservoirs pétroliers mondialement répartis. Les deux graphiques montrent l'accroissement de la pression (à gauche) et de la température (à droite) en fonction de la profondeur (axe vertical, les plus grandes profondeurs sont vers le bas du graphique). On cherche à vérifier si les lois physiques d'accroissement de ces deux variables avec la profondeur se retrouvent dans la base de données.

Observations

La sous-figure supérieure comporte le report intégral des individus. Pour la pression (à gauche), on est tenté au vu du graphique de s'interroger sur la linéarité de la relation : le nuage de points paraît légèrement coudé. L'évolution de la température (à droite) en fonction de la profondeur fournit un nuage plus dispersé, on pourrait être tenté de supposer que deux tendances parallèles expliquent la largeur du nuage, en se fondant sur une impression due à la partie inférieure du nuage.

Les courbes de densité de points infirment ces deux impressions. Pour la relation pression-profondeur (sous-figure inférieure gauche), le graphique révèle une relation linéaire rigoureuse pour plus de la moitié de la population (la signification du niveau de valeur des courbes ne peut être rendue lisible à cette échelle, pour raison de lisibilité, la première courbe tracée englobe 35 % de la population). A forte profondeur, quelques réservoirs s'écartent de cette tendance dominante, avec des pressions plus élevées.

Dans la sous-figure inférieure droite, l'ensemble du nuage s'organise selon la relation attendue d'accroissement de la température avec la profondeur ; mais il n'apparaît pas de relation linéaire aussi nette que dans le cas de la pression. Dans la partie principale du nuage, on peut distinguer deux sous-ensembles, mais aucun d'eux ne correspond à la dichotomie que l'on avait supposée au vu du report brut des individus. Un petit groupe de réservoirs se distingue à forte profondeur avec des températures anormalement fortes.

Interprétation, commentaire

La relation entre pression et profondeur répond à une loi physique, on la trouve bien exprimée dans cette base de données. L'évolution de la température n'est pas aussi régulière, on sait par ailleurs que cette évolution dépend des gradients géothermiques qui varient notablement selon les régions du globe.

Pour la méthodologie des graphiques bivariés, appliqués à une base de données importante, il apparaît que dans le cas d'un lot de points très nombreux, l'impression dégagée de la forme du nuage peut être erronée si l'on effectue un report traditionnel point par point, elle est sur cet exemple influencée par les groupes d'individus marginaux. Des précautions particulières sont donc à prendre pour obtenir une vision correcte de la répartition des individus, les courbes de densité de points s'avèrent ici un outil très efficace.

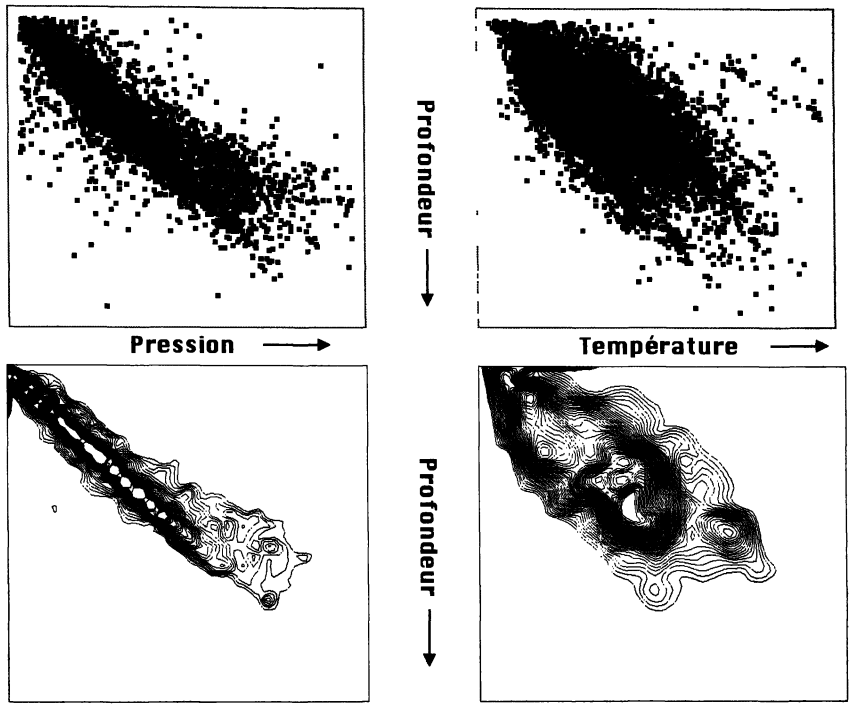

FIG 2. — Report bivarié sur des variables d'une base de données.

ANNEXE : COMMENTAIRES SUR LA FIGURE 3

Origine des données, problème posé

On souhaite comparer les trois réservoirs du gisement étudié en figure 1 aux gisements mondiaux, puis aux 170 gisements de son pays d'origine, en utilisant la base de données utilisée pour la figure 2 (20 000 réservoirs).

Observations

Les variables prises en compte sont (de gauche à droite) : trois estimations de perméabilité (K1, K2, K3), trois estimations de porosité (P1, P2, P3), la salinité de l'eau de gisement (SA), la température maximale (TE) et la pression (PR). Dans les figures, on a limité les traits aux percentiles 10 et 90.

En sous-figure supérieure, les variables autres que la porosité ont subi une transformation logarithmique pour réduire les effets de variance. La sous-figure gauche représente la base mondiale. Malgré une normalisation des valeurs entre le minimum et le maximum de la base (sous-figure centrale), les écarts du pays d'origine (à droite) par rapport à la base mondiale ne sont pas bien mis en évidence : ils restent faibles au regard des effets de variance.

En sous-figure centrale, on a transformé les variables en substituant aux valeurs le rang des individus. Les boîtes se répartissent alors régulièrement sur les limites approximatives 25, 50, 75 % (sous-figure centrale gauche). Les réservoirs du pays d'origine, exprimés dans ce repère et avec ce codage, montrent des écarts notables pour la première estimation de K (à gauche) puis pour les 3 estimations de la porosité. Salinité et température ont une distribution proche de la base mondiale, la pression (à droite) a une médiane moindre que celle de la base mondiale, tout en ayant une gamme de valeurs comparable.

Les boîtes formées en sous-figure centrale droite sont reprises dans la sous-figure inférieure, sur laquelle on a superposé les trois réservoirs du gisement décrit en figure 1, en suivant les conventions d'INSELBERG (1985) pour les axes parallèles. L'un des réservoirs a des valeurs de perméabilité très fortes, situées dans le dernier décile à la fois de la base mondiale et du pays d'origine. La faiblesse des porosités dans le pays d'origine culmine dans les trois réservoirs étudiés. Deux des trois réservoirs ont des valeurs très faibles de salinité.

Commentaire

Sans nous attarder ici sur l'interprétation métier des observations effectuées, notons la capacité de ce type de représentation à fournir des observations multivariées à partir d'une base de données. On peut noter la grande souplesse pour analyser des sous-ensembles, en insistant sur le fait que la visibilité reste très bonne, même quand on compare des sous-ensembles de très faible effectif à des ensembles beaucoup plus vastes.

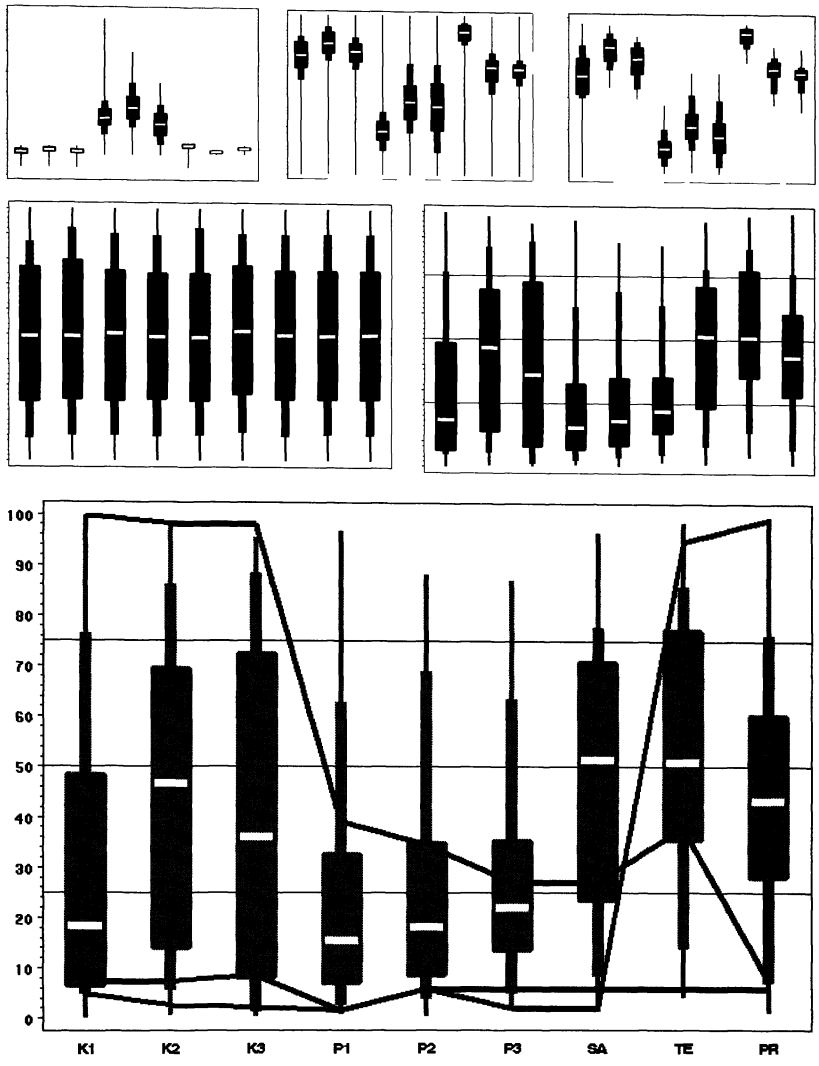

FIG 3. — Boîtes de dispersion appliquées à une base de données.