JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

ANTOINE DE FALGUEROLLES

Discussion et commentaires. Approche graphique en analyse des données. Trop de camemberts tuent le camembert

Journal de la société française de statistique, tome 141, n° 4 (2000), p. 45-49

<http://www.numdam.org/item?id=JSFS_2000__141_4_45_0>

DISCUSSION ET COMMENTAIRES

Approche graphique en analyse des données

Trop de camemberts tuent le camembert

Antoine de FALGUEROLLES [1]

L'article de Jean-Paul Valois vient opportunément réaffirmer l'utilité et le sérieux des méthodes graphiques en statistique appliquée. La statistique graphique est trop souvent considérée comme allant de soi. En conséquence, les méthodes graphiques sont souvent perçues comme les parents pauvres et cachés de la statistique ou, au mieux, comme une petite cerise sur le gâteau d'un modèle mathématique. J'examinerai d'abord le statut des méthodes graphiques en statistique, puis j'évoquerai quelques « grands » graphiques avant de questionner leur futur immédiat.

1. STATISTIQUE GRAPHIQUE – STATISTIQUE

En terme de volume de publications ou de volume d'enseignements, l'avantage est acquis aux méthodes statistiques débouchant sur des formalisations de type mathématique. Une élémentaire recherche bibliométrique montre que les méthodes de régression, d'une lignée aussi ancienne que celle des méthodes graphiques, sont, de nos jours, autrement plus courtisées que ces dernières [2]. Doit-on en déduire que les méthodes graphiques, abouties et figées, sont dépassées ? Il serait d'ailleurs tentant pour le statisticien d'accepter cette opinion. En effet, sauf à accepter les graphiques indigents fournis par des logiciels d'« abattage », l'informatisation de leur construction reste encore assez lourde. Il est de ce fait actuellement beaucoup plus facile de dactylographier les équations d'un modèle, même complexe, que de produire et intégrer un bon graphique dans un texte. Mais la réflexion de Jean-Paul Valois nous invite à penser que les méthodes de statistique graphique valent encore le coup.

---

1. Laboratoire de Satistique et Probabilités, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, e-mail : falguero@cict.fr

2. Il ne s'agit pas là d'opposer régression et méthodes graphiques. En effet, l'évaluation de l'adéquation d'un modèle de régression fait souvent appel à des outils graphiques.

En lisant son article, on peut penser que les méthodes graphiques de la statistique interviennent au moins de trois façons, que chaque statisticien adapte à son projet en combinant sémantique et syntaxe graphiques :

- comme instruments de présentation de données : elles offrent alors des analogues ornés de tableaux de données ;
- comme instruments d'une réflexion : il s'agit alors de graphiques provisoires, d'esquisses de nature éphémère ;
- comme instruments de communication d'informations extraites de données : il s'agit alors de graphiques démonstratifs qui s'adressent à un public plus ou moins formé.

L'article de Jean-Paul Valois montre encore que la statistique graphique est surtout utile lorsqu'elle est voulue au service d'un questionnement de données relatives à un phénomène (classes 2 et 3). Si, paraphrasant Emmanuel Hocquard (1987), on admet que le travail du statisticien appliqué consiste à « transposer une réalité abstraite » (les données) « à une fiction concrète » (le modèle), le bon graphique statistique est alors celui qui facilite ou contrôle cette transposition. Hélas, il semble qu'il n'y ait pas de miracle : une transposition complexe ne peut être décrite par un graphique simple.

2. LE PASSÉ

En étudiant la production de la statistique graphique française destinée au grand public au cours du XIXe siècle, on est d'abord frappé par l'ingéniosité, la diversité et l'indéniable élégance des graphiques produits. On lira par exemple certains des dix-sept volumes des « Albums de Statistique Graphique » publiés annuellement par le Ministère des Travaux Publics (1879-1899) ou les deux volumes de l'« Atlas Graphique et Statistique de la Suisse » publiés par le Bureau de Statistique du Département Fédéral de l'Intérieur Suisse (1897, 1914). Il s'agit essentiellement de documents graphiques relevant de la classe 1.

On peut alors se demander pourquoi ce type de publication s'est pratiquement éteint avec le XIXe siècle. Gilles Palsky (1996, p. 142) mentionne des raisons financières. Mais, un peu décalée dans le temps et à une exception près, cette défaveur s'est aussi appliquée à une autre utilisation des méthodes graphiques : celle du calcul graphique préconisé notamment en France par Léon Lalanne, 1843, 1878 (Dominique Tournès, 2000). L'exception plus durable en était la populaire règle à calcul, tuée par une calculette devenue depuis graphique (au sens de la classe 2).

J'avancerai deux hypothèses complémentaires.

- La statistique graphique était alors devenue l'expression d'un rêve fou : celui de consigner toutes les données sans chercher à en donner une

interprétation. Ses auteurs se réclamaient en quelque sorte de l'utopie de la carte à l'échelle 1/1 (sur cette utopie, voir l'amusant article de Gilles Palsky, 1999) alors que la masse des données allait en s'accroissant.

- De plus, la multiplicité des types de graphiques utilisés pour représenter des « réalités abstraites » de même nature nuisait à la facilité de leur lecture. De façon abrupte et caricaturale, pourquoi produire tant de variations élaborées du « camembert » de base ? Toutes témoignent de la virtuosité et du raffinement technique de leurs auteurs, mais elles nuisent finalement à la lisibilité du travail fourni en exigeant un réajustement fréquent de leur mode de lecture.

Les dangers des complications inutiles ont été souvent dénoncés, et parfois de façon lapidaire, et le rasoir d'Occam souvent invoqué. Je rappellerai une recommandation formulée par Napoléon Bonaparte à l'adresse de ses ingénieurs géographes : « ... Il faut exprimer toujours de la manière la plus simple comment la chose se peint à l'œil de l'observateur. Il y aura une échelle commune pour tous les dessins... » (cité par Louis Nathaniel Rossel, 1870, p. 130-132).

Napoléon Bonaparte aurait-il su apprécier la carte de sa campagne de Russie produite par Charles Joseph Minard ? Relancée par Edward Tuftee en 1983, cette carte est très souvent citée et étudiée : par exemple, Leland Wilkinson (1999, Chapitre 15, *Semantics*). Elle est aussi très présente sur la « toile » (voir par exemple Michael Friendly, http ://www.math.yorku.ca/SCS/gallery/re-minard.html). Cette carte figurative relève des classes 1 et 3 : elle met en relation clairement, sinon très fidèlement, la fusion des effectifs de la Grande Armée, la distance parcourue et la température et ce, dans son cadre géographique ; elle permet aussi une reconstruction approximative des données qui ont présidé à sa construction.

Comme le rappelle Jean-Paul Valois, un exemple très efficace de graphique est fourni par le diagramme quantile-quantile que Pierre J.P. Henry a imaginé dans le cas gaussien vers 1880 à Toulouse (Pierre Crépel, 1993). Le graphique est très simple (un nuage de points), mais son interprétation est plus abstraite (la pertinence d'une approximation gaussienne d'une répartition empirique [3]).

Ce renforcement discret du contenu mathématique du graphique se retrouve chez Etienne Jules Marey (1879). Le parti pris graphique, « mode d'expression » et « moyen de recherche », semble à première vue total : l'ouvrage ne comporte aucune formule. Mais les graphiques présentés supposent chez leur lecteur une certaine aptitude au traitement quantitatif des données et donc aux mathématiques. C'est encore plus frappant pour le stéréogramme de Luigi Perozzo (1880).

3. D'où le succès de ce type de graphique pour contrôler un modèle de régression.

3. LE PRÉSENT

Les représentations fournies par les méthodes d'analyse factorielles (Jean-Paul Benzecri, 1979) ou de décomposition en éléments singuliers (Ruben K. Gabriel, 1971), *plots et biplots*, sont d'usage très courant. Mais leur apparente simplicité peut être trompeuse car, parfois, contre-intuitive. La lecture de ces graphiques demande donc une certaine formation mathématique de leur public.

Plus conformes à la tradition graphique sont les travaux de Jacques Bertin (1977) de visualisation de matrices pondérées ou les représentations de type mosaïques de tableaux de contingence (John A. Hartigan and Beat Kleiner, 1981). Mais déjà la complexité de la lecture du graphique est le prix à payer pour l'ambition de la méthode : susciter des modèles, percevoir les implications d'un modèle ou des écarts à des modèles.

4. LE FUTUR IMMÉDIAT

Quelle évolution pour les méthodes de statistique graphique ? Ce n'est pas l'objet central de l'article de Jean-Paul Valois. Mais on aimerait avoir son avis à ce sujet. Certes les moyen informatiques, mêmes élémentaires, permettent d'introduire un élément dynamique dans la graphique statistique (Erich Neuwirth, 2000). Les coordonnées parallèles d'Alfred Inselberg (1985, 1996) sont aussi d'une grande efficacité dans ce contexte interactif. Par ailleurs, la visualisation de données individuelles connaît un certain renouveau. Un exemple est fourni par l'étonnant « crayon de Lexis » de Brian Francis et Mike Fuller (1996). A ce jour, les méthodes graphiques classiques s'accommodent donc bien de « petits ensembles de données ». Mais l'étude de grands ensembles de données appelle sans doute à un profond renouvellement des méthodes graphiques et des méthodes de formation à leur utilisation. J'entends ici la formation à leur construction (pixelisation, courbes de Peano,...) et leur lecture dans l'environnement des nouvelles technologies de communication. La « toile » fourmille de sites témoignant de travaux en cours auxquels les moteurs de recherche donnent accès : par exemple, un recensement des graphiques d'exploration de base de donnée (Daniel A Keim, 1997), un site à visées pédagogiques inactif depuis 1999 http ://www.agocg/ac.uk /, ... pour ne pas en citer des dizaines ! J'espère, qu'à la suite du travail de Jean-Paul Valois, ce foisonnement suscitera de nombreux autres articles.

Remerciements : Je voudrais dire ma profonde gratitude à Stephen M. Stigler pour m'avoir permis de découvrir l'ouvrage emblématique de Kaemtz ainsi que son annexe rédigée par Lalanne. Mes remerciements vont aussi, bien sûr, à mes confrères les « chevaliers des albums de statistique graphique » pour leurs encouragements et leurs conseils précieux. Philippe Besse, au cours d'utiles discussions, m'a aussi apporté une aide certaine.

RÉFÉRENCES COMPLÉMENTAIRES

Album de Statistique Graphique (1879, ..., 1887) : Ministère des Travaux Publics. Paris : Imprimerie Nationale.

Atlas Graphique et Statistique de la Suisse / Graphisch-statistischer Atlas der Schweiz (1897, 1914) : Bureau de statistique du département fédéral de l'intérieur. Bern : Buchdruckerei Stämpfli Cie.

[cité par Valois] Jean-Paul BENZECRI (1979)

[cité par Valois] Jacques BERTIN (1977)

Pierre CRÉPEL (1993) : Henri et la droite de Henry. *Matapli*, n° 36, 19-22.

Brian FRANCIS and Mike FULLER (1996) : Visualisation of event histories. *Journal of the Royal Statistical Society*, Series A, 159, 301-308.

Ruben K. GABRIEL (1971) : The biplot-graphic display of matrices with application to principal components analysis, *Biometrika*, 54, 453-467.

John A. HARTIGAN and Beat KLEINER (1981) : Mosaics for contingency tables. *Computer science and statistics : Proceedings of the 13th Symposium on the Interface*, W.F Eddy (ed.). New-York : Springer Verlag. 268-273.

Emmanuel HOCQUARD (1987) : *Un privé à Tanger*. Paris : POL.

[cité par Valois] Alfredo INSELBERG (1985)

Alfred INSELBERG (1996) : *Visual data mining with Parallel coordinates*. Workshop on strategies for data analysis. Antony Unwin, chairperson. Universität Augsburg.

Daniel A. KEIM (1997) *Visual techniques for exploring databases* Invited tutorial, International Conference on Knowledge Discovery in databases.

(http ://www.dbs.informatik.uni-muenchen.de/ daniel/KDDtutorial.ps)

Léon LALANNE (1843) : Appendice contenant la représentation graphique des tableaux numérique dans l'ouvrage de L. F. Kaemtz. Par exemple dans la traduction anglaise de C.V. Walker : *A complete course of meteorology*. London : Hippolyte Ballière (1845).

Léon LALANNE (1878) : *Méthodes graphiques pour l'expression des lois empiriques ou mathématiques à trois variables avec des applications à l'art de l'ingénieur et à la résolution des équations numériques d'un degré quelconque*. Paris : Imprimerie Nationale

[cité par Valois] Etienne Jules MAREY (1879)

Erich NEUWIRTH (2000) Spreadsheets as tools for statistical computing and statistical education. *COMPSTAT, Proceedings in Computational Statistics*. J.G. Bethlehem and P.G.M. van der Heijden (editors). Heidelberg : Physica-Verlag, 131-138.

[cité par Valois] Gilles PALSKY (1996)

Gilles PALSKY (1999) : Borges, Carroll et la carte au 1/1. *Cybergeo*, n° 106

(http :// www.cybergeo.presse.fr/revgeo.htm).

[cité par Valois] Luigi PEROZZO (1880).

Louis Nathaniel ROSSEL (1871) : *Abrégé de l'art de la guerre*, Paris : Lachaud. 1871 p. 130-132.

[cité par Valois] Edward R. TUFTE (1983)

Dominique TOURNÈS (2000) : Pour une histoire du calcul graphique. Revue d'histoire des mathématiques, Tome 6, Fascicule 1, 127-161.

[cité par Valois] Leland WILKINSON (1999)