MÉTHODES

Quelle statistique pour le *Big Data* ?

Entretien avec Gilbert SAPORTA

Professeur émérite de statistique appliquée
Conservatoire National des Arts et Métiers

Tout le monde s'intéresse au *Big Data*. Le public est de mieux en mieux informé sur les potentialités que les données massives recèlent et sur les dangers que leur utilisation peut comporter. Mais très rares sont ceux qui savent ce qui se cache « sous le capot » des nouvelles méthodes. Statistique et Société a demandé à Gilbert Saporta, qui fait partie de ce petit nombre, d'éclairer autant que possible les non-spécialistes.

Statistique et société : Du point de vue des méthodes, qu'est-ce qui détermine si on est dans un contexte *Big Data* ou non ? Y a-t-il un seuil en nombre d'observations, de variables ? Un seuil par rapport aux capacités mémoire d'un ordinateur ? Que penser du critère de la « vélocité » si souvent avancé ?

Gilbert Saporta : Quand on évoque le *Big Data*, on pense en premier lieu au volume des données, en d'autres termes à la taille du fichier correspondant. Une des premières occurrences de l'expression *Big Data* dans la littérature scientifique est la communication de deux chercheurs de la NASA, Cox et Ellsworth, au congrès SIGGRAPH de 1997[1].

Rappelons tout d'abord que selon les époques la notion de « Big » a beaucoup varié. Qui ne se souvient que la grande taille pour les estimations et les tests commençait à n=30 ? Avec le traitement des recensements, les statisticiens sont confrontés depuis longtemps à des données massives. Comme le rappelle David Donoho[2], cela a conduit Herman Hollerith à inventer la carte perforée et à créer IBM. C'est la technologie qui fixe les limites. Comme le disait déjà à peu près en ces termes John Tukey : « Big », c'est quand cela ne rentre pas dans ma machine. Il n'est donc pas possible de fixer des seuils. Toutefois on parle de données de grande dimension quand le nombre de variables dépasse de beaucoup le nombre d'observations, mais c'est un sujet à part. On parlera donc de *Big Data* quand les données ne peuvent être stockées sur un seul ordinateur (données réparties) et quand les traitements vont nécessiter plusieurs machines (calculs distribués). Le volume est le premier V, d'une série de trois, introduits par le Gartner Group en 2008. Le deuxième V est en effet la vélocité qui renvoie aux flux de données recueillies en temps réel sur le web, ou par des capteurs, tels les objets connectés, les compteurs électriques « intelligents », etc. Les problèmes de stockage et d'algorithmique deviennent alors essentiels car on ne peut tout conserver, et il faut recourir à des méthodes incrémentales qui actualisent les résultats au fur et à mesure de l'acquisition de nouvelles données.

1. Cox M. and Ellsworth D. (1997), Managing Big Data for Scientific Visualization, Exploring Gigabyte Datasets in Real-Time: Algorithms, Data Management, and Time-Critical Design, *Siggraph* 97, Course Notes 4, New York, ACM Press.
2. Donoho D (2015), 50 years of Data Science, *Tukey Centennial workshop* http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Le troisième V, souvent mis en avant, est la variété des types de données recueillies : numériques et qualitatives comme d'habitude, mais aussi des images, des données issues de réseaux sociaux (qui est lié avec qui ?) et non structurées comme des textes.

S&S : Quels sont les traitements qu'on sait faire dans des temps raisonnables sur des bases de données de très grande taille ? Réciproquement, y a-t-il des traitements qu'on ne sait pas faire sur une base de données qui ne tient pas en mémoire d'un ordinateur ?

GS : Tout est une question de moyens et d'approches spécifiques. Quand on peut utiliser par exemple le cloud d'Amazon, ou de puissants réseaux d'ordinateurs, et des programmes conçus pour travailler en parallèle, suivant par exemple le modèle de programmation MapReduce inventé par Google, on peut mettre en œuvre la plupart des traitements statistiques standard. Des bibliothèques libres de programmes dédiées au *Big Data* comme Spark, Scikit-Learn, MLLib contiennent les méthodes favorites des statisticiens : régression linéaire et logistique, classification supervisée et non supervisée, réduction de dimension, mais aussi des algorithmes d'apprentissage comme les forêts aléatoires et les séparateurs à vaste marge (SVM). Le catalogue s'enrichit sans cesse grâce aux travaux de communautés d'utilisateurs car ce sont des systèmes ouverts, tout comme l'environnement R. Par contre, certaines méthodes comme l'estimation de densité multidimensionnelle sont mal adaptées aux données massives.

S&S : Dans l'univers *Big Data*, le nombre de données étant tellement important, tout test statistique devient significatif. Faut-il oublier les tests statistiques et les intervalles de confiance ? Et qu'utilisera-t-on pour les remplacer ?

GS : En effet. Tout écart à une hypothèse nulle devient significatif : déjà avec 10 000 observations, ce qui n'est pas du *Big Data*, un coefficient de corrélation égal à 0.02 est déclaré significativement non nul au risque 5 % bilatéral. Est-ce utile ? Bien sûr que non, et ne parlons pas de millions de données ! Tous les coefficients d'un modèle deviennent alors « significatifs », mais de quoi ? En plus, et ce n'est paradoxal qu'en apparence, les tests d'adéquation rejettent tous les modèles usuels, car ils sont trop simples pour exprimer de grandes masses de données. Une autre forme d'inférence doit être mise en œuvre, liée à la notion de reproductibilité ou de généralisabilité des résultats.

La théorie de l'apprentissage statistique, développée par Vladimir Vapnik et Alexei Cervonenkis, donne des bornes pour la différence entre la performance en apprentissage et la performance sur de nouvelles données, dans le cadre prédictif. Cette théorie permet de qualifier, en fonction d'une mesure de complexité, les modèles qui généralisent bien.

L'application de cette théorie n'étant pas toujours aisée, on recourt souvent au procédé suivant : ayant séparé aléatoirement les données disponibles en deux sous-ensembles, on vérifie sur le deuxième sous-ensemble si les résultats obtenus sur le premier restent valables. On pourra procéder à plusieurs séparations pour étudier la variabilité et éviter des cas trop particuliers. Cette pratique, voisine de la validation croisée, est particulièrement bien adaptée aux données massives. On l'attribue généralement au machine learning qui l'a systématisée pour éviter les phénomènes de surajustement, mais elle est bien plus ancienne. Dès 1941, le psychométricien Paul Horst dans un chapitre intitulé

> « *L'utilité d'une procédure de prédiction n'est pas établie lorsqu'on a trouvé qu'elle prédisait correctement sur l'échantillon original ; l'étape suivante nécessaire doit être son application à au moins un second groupe. Ce n'est que si elle prédit correctement sur des échantillons ultérieurs que la valeur de la procédure peut être considérée comme établie* »[3].

S&S : Y a-t-il toujours un modèle dans un traitement Big Data ? Explicite ou sous-jacent ? Que penser des déclarations de ceux qui disent qu'on doit « arrêter de modéliser » ?

GS : Il faut s'entendre sur ce que l'on appelle modéliser ! S'il s'agit de modèles génératifs, c'est-à-dire de modèles en général simples, et interprétables dans le langage du champ d'application, censés décrire le mécanisme qui a engendré les données, la réponse est clairement non. Aucun modèle simple ne peut représenter de grandes masses de données. Le célèbre aphorisme de George Box « tous les modèles sont faux, certains sont utiles » s'applique parfaitement.

En *Big Data*, on utilise abondamment des modèles prédictifs, mais au sens d'algorithmes, sans chercher à mimer le processus génératif que l'on considérera inconnu. Le seul critère est la capacité de prédire de nouvelles observations. On peut d'ailleurs prouver certains résultats en apparence paradoxaux : ainsi de la remarque de V.Vapnik : « On obtient parfois de meilleurs modèles en évitant délibérément de reproduire les vrais mécanismes »[4]. Dans le même ordre d'idée, Shmueli indique :

> « *La significativité statistique joue un rôle mineur, ou pas de rôle du tout, pour établir la performance en prédiction. En fait, il arrive parfois qu'on obtienne une meilleure précision de la prédiction en retirant des variables en entrée ayant de petits coefficients, même s'ils sont statistiquement significatifs* »[5]

J'avais abordé en 2008 dans une conférence invitée au congrès Compstat[6], la distinction entre *modèles pour comprendre et modèles pour prédire*, sans avoir connaissance de l'article exceptionnel de Leo Breiman de 2001[7] sur les deux cultures de la modélisation statistique dont je recommande vivement la lecture, ainsi que celle de la conférence de David Donoho citée plus haut, à l'occasion du centenaire de la naissance de John Tukey, qui reprend largement l'article de Breiman.

Certains modèles sont d'interprétation aisée, comme les arbres de décision, d'autres sont plutôt des boîtes noires, comme les réseaux de neurones, les forêts aléatoires, le boosting etc. Il est courant de combiner les prévisions de différents modèles (linéairement ou non) plutôt que de rechercher le meilleur modèle : ce sont les meta-modèles ou modèles d'ensemble qui remportent souvent les compétitions.

Arrêter de modéliser (au sens des modèles génératifs) renvoie à la tribune provocatrice de Chris Anderson (2008) sur la fin de la théorie qui prétendait :

> « *Les péta-octets nous permettent de dire : « la corrélation suffit ». On peut s'arrêter de rechercher des modèles. On peut analyser les données sans hypothèses sur ce que cela pourrait montrer. On peut injecter les chiffres dans les plus grandes grappes d'ordinateurs que le monde ait jamais vues, et laisser les algorithmes statistiques trouver des configurations là où la science en est incapable* »[8]

3. « *The usefulness of a prediction procedure is not established when it is found to predict adequately on the original sample; the necessary next step must be its application to at least a second group. Only if it predicts adequately on subsequent samples can the value of the procedure be regarded as established* ». Horst, P., Wallin, P. C., Guttman, L. C., Wallin, F. B. C., Clausen, J. A., Reed, R. C. et Rosenthal, E. C. (1941), The prediction of personal adjustment : A survey of logical problems and research techniques, with illustrative application to problems of vocational selection, school success, marriage, and crime. *Social science research council*.
4. « *Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms* ». V.Vapnik (2006), *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer
5. « *Statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, even if they are statistically significant, results in improved prediction accuracy* » Shmueli G. (2010), To explain or to predict? *Statistical Science*, 25, 289–310
6. Saporta G. (2008) Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322
7. Breiman L. (2001) Statistical modeling: The two cultures. *Statistical Science*, 16 199–215
8. « *Petabytes allow us to say: « Correlation is enough. » We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot* ». C.Anderson (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired*, http://www.wired.com/2008/06/pb-theory/

La suite lui a donné tort et renvoie à la confusion entre corrélation et causalité. Ainsi de l'exemple souvent cité de la prévision de la grippe saisonnière avec Google Flu Trends. Vers 2010 des chercheurs de Google eurent l'idée de corréler les requêtes concernant des mots clés liés à l'apparition de la grippe (fièvre, etc.). Ils obtinrent ainsi un indicateur précurseur de l'épidémie, en avance sur les déclarations des médecins. Les prévisions furent excellentes jusqu'en 2012 mais le système se détraqua ensuite (surestimation) et fut abandonné dans sa version d'origine.

La moralité de l'histoire est que si on peut prédire sans comprendre, il faut prendre garde aux changements : toute méthode de prévision suppose que le futur ressemblera à ce que l'on connait déjà… Comprendre pour mieux prédire a de l'avenir !

S&S : Au fait, le *machine learning*, c'est quoi ?

GS : Pour faire simple, le *machine learning* désigne pour l'essentiel une famille d'algorithmes d'apprentissage supervisé, c'est à dire où on cherche à prédire une réponse en procédant par amélioration successive du prédicteur au fur et mesure qu'il traite de nouvelles données, en comparant valeur prévue et valeur vraie. L'ancêtre de ces méthodes est le perceptron de Rosenblatt né en 1958 des travaux en 1943 de McCulloch et Pitts sur les neurones artificiels. Les modèles dépendent d'un grand nombre de paramètres et pour bien apprendre, il leur faut en effet de très nombreuses données, d'où le lien avec les Big Data. Le *machine learning* est aussi un domaine de recherches très actif regroupant informaticiens, mathématiciens et statisticiens.

S&S : Analyser beaucoup de données ne signifie pas que l'on est exhaustif. Les données massives semblent souvent recueillies sans plans d'échantillonnage ou de sondage. Ne risque-t-on pas d'avoir des données massives mais biaisées car recueillies sur une sous-population uniquement (les utilisateurs de smartphone par exemple) ?

GS : Tout à fait ! Quantité ne veut pas dire forcément qualité et les précautions usuelles s'imposent. Il faut par exemple disposer de variables de calage pour redresser les résultats si on veut qu'ils soient en accord avec des études préalables.

J'évoquerai également un problème connexe : les données massives sont souvent récoltées automatiquement dans un but qui n'est pas nécessairement statistique : je pense ici aux capteurs installés dans de plus en plus d'équipements comme des véhicules, des fauteuils, les caméras de surveillance. Ces données sont souvent « sales » et les prétraitements sont essentiels.

S&S : Les données massives posent de façon cruciale la question de leur ouverture. Il y a *Big Data* aussi parce qu'on peut récolter des données facilement un peu partout sur le web, qu'elles soient publiques ou privées et les articuler les unes aux autres. Cela nous semble transformer assez profondément le métier de statisticien public dans la mesure où cela rend la frontière entre les données publiques et privées plus floue. Qu'en pensez-vous ?

GS : Bien qu'universitaire, je suis avec attention les évolutions de la statistique publique. A la suite du mémorandum de Scheveningen adopté en 2013 par l'assemblée des directeurs généraux des instituts nationaux de statistique européen, un plan *Big Data* a été adopté par les acteurs du système statistique européen et en particulier d'Eurostat. Son ancien directeur général, Walter Radermacher, a souvent évoqué la « Statistics 4.0 » et la fin de l'usine à enquêtes comme modèle des INS. Le plan *Big Data* a pour but de préparer le Système Statistique Européen à intégrer des sources de données massives dans la production des statistiques officielles. Ces sources nouvelles viennent compléter les registres et données administratives qui peuvent être déjà très volumineuses mais sont mises à jour lentement. Des expériences concluantes ont été menées concernant l'utilisation des données d'opérateurs de téléphonie mobile pour

améliorer les statistiques du tourisme et de la mobilité, la transmission des données de caisses de supermarché pour l'indice des prix à la consommation, l'utilisation des offres d'emploi sur Internet pour actualiser les enquêtes emploi etc. Il est clair que la mesure du commerce électronique passe par l'accès à des données issues des sites de vente en ligne. La rapidité de mise à jour, le volume traité et les économies réalisées sont des arguments essentiels, mais il ne faut pas sous-estimer la difficulté d'utiliser des données externes.

L'utilisation de sources de données privées soulève différentes questions : quelle en est la représentativité quand il y a plusieurs opérateurs de téléphonie ? comment s'assurer de la véracité des données (un quatrième V souvent évoqué) quand les statisticiens publics n'en contrôlent pas la production, et qu'il n'y a pas de vérité de terrain ? Quelle est la pérennité de ces sources privées qui peuvent se tarir selon le bon vouloir des entreprises ou leur propre pérennité ? La qualité des sources est également très variable selon la précision des recueils : caméras, réseaux sociaux, commerce électronique. Tout cela pose des questions contractuelles : quel modèle économique pour une coopération entre des services publics et des entreprises privées motivées par le profit, et légales (obligation de transmission).

L'accréditation de sources de données *Big Data* pour leur utilisation en statistique officielle est d'ailleurs documentée par Eurostat.

Le métier de statisticien public va en effet être amené à changer profondément, tout d'abord pour des raisons essentiellement technologiques : les statisticiens publics devront monter en compétences sur les aspects informatiques, s'approprier de nouvelles méthodologies d'analyse (*machine learning*) et être encore plus vigilants sur la protection des données et l'éthique. Vu la rapidité avec laquelle les données et les outils évoluent, les statisticiens publics doivent sortir de leur tour d'ivoire : tous ne deviendront pas des data scientists, mais ils devront collaborer avec les data scientists et ingénieurs du privé et les chercheurs universitaires. Le travail en équipe multidisciplinaire sera une nécessité car un même individu ne peut cumuler toutes les compétences. On qualifie de « *iStatisticien* » ce nouveau métier[9].

S&S : En quelques mots, quelles principales innovations dans la formation des jeunes pensez-vous devoir être introduites pour répondre à la nouveauté du *Big Data* ?

GS : Les jeunes statisticiens doivent être mieux formés aux nouvelles technologies informatiques évoquées dans la deuxième question de cet entretien, connaitre les principes des systèmes de gestion NoSQL (Not only SQL) et savoir coder dans des langages comme Python. Sur le plan méthodologique, les méthodes computationnelles et d'apprentissage doivent être enseignées avec les méthodes de l'analyse statistique multivariée. Dans le cadre de la grande dimension, les méthodes de régularisation doivent trouver leur place à côté des grands classiques que sont les moindres carrés et le maximum de vraisemblance.

Il faut que les étudiants aient accès à de grandes bases de données pour s'exercer. La participation à des compétitions comme Kaggle (https://www.kaggle.com/), Datascience (https://www.datascience.net/fr/home/), Challenge data (https://challengedata.ens.fr/fr/home) doit être encouragée et mieux, faire partie intégrante des cursus, à l'instar de la formation continue « Data Science pour l'actuariat » de l'Institut des actuaires.

Je suis optimiste sur l'évolution en cours des masters et des écoles de statistique en France car les enseignants-chercheurs et les directions ont bien saisi les enjeux et ont les compétences nécessaires. Par contre d'autres formations devraient faire leur *aggiornamento*, telles les

9. http://tietotrendit.stat.fi/mag/article/153/

licences d'économie et gestion où les cours de statistique datent souvent d'avant la révolution numérique. Je recommande à ce sujet la lecture de l'article de Hal Varian, célèbre professeur de microéconomie, devenu chef économiste de Google, qui encourage ses collègues économistes à aller voir du côté des algorithmes de *machine learning* et ne pas se contenter des classiques régressions linéaire et logistique[10].

S&S : Le grand public connait déjà le terme *Big Data* et certaines utilisations pratiques du Big Data. Comment expliquez-vous cela alors que la recherche en *Big Data* semble en être à ses débuts ?

GS : On ne peut pas dire cela. Même si « *Big Data* » n'est pas toujours dans les mots-clés, la recherche en *Big Data* est active depuis plus de 10 ans tant du côté des informaticiens que des statisticiens, et ce n'est pas que de la technologie.

Le traitement des flux de données, de la très grande dimension suscite des travaux très pointus, publiés dans de grandes revues. L'analyse des réseaux sociaux avec leurs immenses graphes a ressuscité l'intérêt pour la théorie des graphes et conduit à des travaux originaux.

L'apprentissage profond (*deep learning*) dont la dénomination date de 2006, a renouvelé les problématiques des réseaux de neurones et conduit à des réalisations spectaculaires en apprentissage supervisé (reconnaissance d'images, intelligence artificielle) car on peut maintenant entrainer des modèles comportant des milliers de paramètres grâce aux bases de données massives.

Le Big Data permet de poser de manière nouvelle le problème de la causalité : la *National Academy of Sciences* américaine a organisé en 2015 un colloque passionnant intitulé « Drawing Causal Inference from Big Data » réunissant entre autres : Léon Bottou, Peter Bühlmann, Michael Jordan, Judea Pearl, Bernhard Schölkopf, Hal Varian[11]. La même année, la revue Political Science and Politics publiait un cahier spécial avec 8 articles sur « Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science? ».

Du côté de la statistique officielle, cela bouge aussi très vite et de nombreux travaux de recherche appliquée sont présentés dans les conférences NTTS (New Techniques and Technologies for Statistics) organisées par Eurostat, qui rassemblent tous les deux ans plusieurs centaines de participants.

Je conclurai sur deux citations. La première est de George E.P. Box :

> « *Les statisticiens, comme les artistes, ont la mauvaise habitude de tomber amoureux de leurs modèles* »[12]

Et la seconde de W.Edwards Deming :

> « *Les données scientifiques ne sont pas collectées pour des musées ; elles sont collectées comme une base pour une action. S'il n'y a rien à faire avec des données, il n'y a aucune utilité à les collecter. Le but ultime de la collecte des données est de fournir une base pour l'action, ou une recommandation pour une action. L'étape intermédiaire entre la collecte des données et l'action est la prédiction.* »[13]

10. H.Varian (2014) Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28, 2, 3–28
11. Le site http://www.nasonline.org/programs/sackler-colloquia/completed_colloquia/Big-data.html contient les vidéos des conférences dont certaines ont été publiés en 2016 dans les PNAS (volume 113, n°27)
12. « *Statisticians, like artists, have the bad habit of falling in love with their models.* »
13. « *Scientific data are not taken for museum purposes; they are taken as a basis for doing something. If nothing is to be done with the data, then there is no use in collecting any. The ultimate purpose of taking data is to provide a basis for action or a recommendation for action. The step intermediate between the collection of data and the action is prediction.* » Deming W.-E.(1942), On a Classification of the Problems of Statistical Inference, *Journal of the American Statistical Association*. Le Dr. Deming a écrit cet article alors qu'il travaillait pour le Bureau du Census des États-Unis.