Journal de la Société Française de Statistique
Vol. 158 No. 2 (2017)

Éditorial du numéro spécial : Humanités et Statistiques

Title: Editorial of the special issue: Humanities and Statistics

Stéphane Lamassé[1] et Fabrice Rossi[2]

Le mouvement des « humanités numériques » a conduit à une utilisation assez systématique des outils informatiques au sein des humanités. Des approches pionnières comme la Text Encoding Initiative [3] ont ainsi permis la numérisation d'archives textuelles très volumineuses. Ces archives deviennent progressivement multimédia et sont décrites par des métadonnées très riches. En parallèle à ce mouvement de numérisation des objets d'études des humanités, de nombreux chercheurs des différents domaines des humanités ont commencé à appliquer leurs méthodes à des objets naturellement numériques, comme notamment les interactions entre utilisateurs de réseaux sociaux en ligne. Cette tendance générale est aussi amplifiée par le mouvement de l'« open data » qui donne accès à des corpus de données très volumineux, produits ou collectés par des entités publiques (gouvernements à différents niveaux, par exemple). L'emploi de l'expression « Big Data » pourra ainsi se faire au sens littéral dans les humanités.

La disponibilité de données informatiques décrivant les objets d'étude traditionnels des humanités ou les objets numériques produits aujourd'hui, a accru de façon considérable les possibilités d'analyse quantitative dans ces domaines. Il est ainsi possible de s'appuyer sur des outils d'analyse de données, de modélisation statistique, de visualisation, etc. pour explorer les données historiques, géographiques, légales, etc. Dans les cas les plus simples, les outils classiques, développés pour les données tabulaires usuelles, peuvent apporter des éclairages intéressants et rester pertinents et efficaces. Ils ont l'avantage de faire partie intégrante des cursus universitaires des chercheurs des sciences humaines et sociales, ce qui donne à ces derniers une grande autonomie de traitement de leurs données, en particulier quand les méthodes sont disponibles dans des logiciels statistiques classiques.

L'omniprésence des données ne doit pas cependant faire croire que les humanités passent soudainement d'une absence de données au déluge qu'on observe dans certains domaines de l'informatique. Comme le montrent les articles de ce numéro spécial, l'extension du domaine de la donnée se fait tout autant en largeur qu'en profondeur : bien que certains champs voient arriver des données très volumineuses, on observe aussi une quantification de plus en plus fréquente des

[1] LaMOP UMR 8589, Université Paris 1 Panthéon Sorbonne, 1 rue Victor Cousin, 75232 Paris cedex 05.
E-mail : stephane.lamasse@univ-paris1.fr
[2] SAMM EA 4543, Université Paris 1 Panthéon Sorbonne, 90 rue de Tolbiac, 75013 Paris, France.
E-mail : fabrice.rossi@univ-paris1.fr
[3] http://www.tei-c.org/

 ISSN: 2102-6238

humanités. Celle-ci n'engendre pas nécessairement des données volumineuses, au contraire. On se retrouve donc aussi à devoir modéliser des données très parcellaires, avec toutes les difficultés que cela pose en statistique, même pour les méthodes classiques maîtrisées par les chercheurs en sciences humaines et sociales.

En outre, le mouvement de collecte numérique décrit ci-dessus produit des données de plus en plus complexes, ce qui demande le développement de méthodes spécifiques ou, *a minima*, l'utilisation de méthodes mathématiques ou informatiques très récentes. Ceci conduit naturellement à l'émergence de collaborations pluridisciplinaires, associant chercheurs des sciences humaines et chercheurs en visualisation, analyse et/ou modélisation de données. Il est souvent indispensable de procéder par itération en adaptant et en complexifiant progressivement les modèles utilisés pour capturer toute la richesse des données étudiées, ce qui inscrit ces collaborations dans le moyen ou le long terme.

Il existe en France des lieux d'intersections nombreux et importants où peuvent se rencontrer des spécialistes de plusieurs disciplines. Les occasions de lire des études de cas mêlant statistique et sciences humaines sont nombreuses et relativement anciennes. Ainsi par exemple une des premières revues a avoir proposé une rencontre entre mathématiciens et sciences humaines et sociales est sans doute *Mathématiques et sciences humaines*[4] publiée dès 1962 sous l'égide du CAMS (*Centre d'analyse et de mathématiques sociales*[5]) de l'EHESS et de Marc Barbut. Dans chaque discipline des sciences humaines et sociales on trouve des revues qui témoignent d'une activité scientifique dense et vivace que ce soit en démographie, histoire, géographie, sociologie, etc. La quantité de colloques et de congrès est tout aussi importante. L'archéologie, par exemple, organise, depuis 1973 à Birmingham, le congrès CAA (*Computer Applications & quantitative Methods in Archaeology*[6]). Ces rencontres répondaient à la multiplication des fouilles de sauvetage et au recours devenu nécessaire à l'informatique pour acquérir et traiter des données. De façon assez générale, l'informatisation nous semble précéder puis amplifier le recours aux données et l'emploi de méthodes statistiques.

L'objectif de ce numéro spécial est de capturer un instantané de ce foisonnement en proposant des exemples de collaborations récentes fructueuses entre sciences humaines et sociales d'une part et statistique d'autre part. Bien entendu, un numéro spécial ne peut rendre qu'un reflet partiel et incomplet de ces collaborations, mais avec six articles, il donne un aperçu des pratiques courantes (emploi avancé de la statistique) et des développement en cours (nouvelles méthodologies, nouveaux modèles). Il illustre aussi les différentes modes d'appropriation ou de création de méthodes statistiques dans les SHS. En faisant relire chaque article par un(e) expert(e) de chacune des disciplines concernées (statistique d'une part et la science humaine concernée d'autre part), nous avons essayé de faire en sorte que les articles soient accessibles aux chercheurs statisticiens et aux chercheurs des humanités.

L'article de A. Disser, M. L'Héritier, P. Dillmann et A. Arles, « Le chantier de la tour de Mutte à Metz : regards sur la production du fer en Lorraine à la fin du Moyen Âge. Fouille de données,

[4] https://msh.revues.org/
[5] http://cams.ehess.fr/
[6] http://caaconference.org/

 ISSN: 2102-6238

analyses prédictives et traitement spatial des données. », qui ouvre ce numéro, illustre la maîtrise par des archéologues d'un ensemble de techniques d'analyse statistique des données qu'on pourrait présenter comme classiques : analyse en composantes principales, classification ascendante hiérarchique et régression logistique. Du point de vue archéologique, l'article s'intéresse au développement de la sidérurgie lorraine à la fin du Moyen Âge, étudié grâce à des mesures sur les renforts ferreux, et les scellements en plomb associés, utilisés lors de la construction du beffroi de la ville de Metz. La méthodologie développée est assez sophistiquée et s'appuie sur une démarche classique en sciences humaines mais aussi en visualisation d'information : préférence est donnée aux méthodes non supervisées (ou semi-supervisées) qui mettent le spécialiste au cœur de l'analyse. Ici par exemple, on réalise une classification des scellements à partir de leur composition chimique sans introduire dans l'analyse leur position sur le beffroi. Celle-ci est utilisée par les experts pour valider *a posteriori* la classification en jugeant de l'uniformité « géographique » des classes. Une approche supervisée est aussi rejetée dans la suite de l'article afin d'éviter des problèmes liés à la représentativité des données de référence en archéologie. Cet article souligne une difficulté majeure en statistique qui se manifeste souvent sans les sciences historiques, par exemple : la production et la collecte des données n'est que sous le contrôle partiel des chercheurs. Contrairement à certaines situations en médecine, en agronomie ou dans l'industrie où on peut mettre en place un plan d'expérience contrôlé, les données de référence utilisées ici ont en partie été produites par un processus historique. En appliquant naïvement des méthodes statistiques on pourrait être amené à mélanger ou confondre le processus de production et de sélection des données avec le phénomène qu'on souhaite étudier à partir de ces données.

Dans « La mobilité inter-entreprises des migrants de Tunisie en région parisienne dans les Trente Glorieuses. Quelques outils statistiques au service d'une démarche historienne. », A.-S. Bruno montre l'importance de la prise en compte de la structure des données dans leur analyse. Les modèles multi-niveaux avec effets aléatoires sont maintenant assez classiques en statistique et on connaît leur importance dans de nombreuses applications, comme l'analyse de survie par exemple. Ces modèles sont peu à peu adoptés dans d'autres champs notamment ici dans l'analyse biographique. Il s'agit d'étudier des trajectoires professionnelles pour mesurer et comprendre les déterminants de la mobilité inter-entreprise. La structure hiérarchique des données est claire car on analyse la « survie » d'un individu dans une entreprise. De ce fait, on dispose en général de plusieurs trajectoires pour un même individu ce qui impose l'utilisation d'un modèle multi-niveaux. Ce dernier permet de faire ressortir un certain nombre de déterminants pour la mobilité entre entreprises, comme par exemple le secteur d'activité. Cet article montre comme le précédent l'autonomie acquise par les chercheurs des humanités pour l'utilisation des outils statistiques. Il illustre en outre l'évolution des pratiques permise par cette autonomie : au début des années 2000, l'approche classique consistait à utiliser un modèle de Cox classique sur des données réduites à un seul épisode par individu, réduisant d'autant le volume de données réellement utilisées. L'utilisation de modèles multi-niveaux permet maintenant d'intégrer toutes les données dans l'analyse ce qui constitue un apport important dans ces champs d'étude où les données restent peu volumineuses.

Contrairement aux deux premiers, les articles suivants du numéro spécial sont des productions associant chercheurs des humanités et chercheurs en statistique. De ce fait, les méthodes qui y

 ISSN: 2102-6238

sont développées sont en général de conception plus récentes en statistique, voire complètement nouvelles et motivées par les problèmes soulevées par les humanités.

L'article « Un problème clé de la paléodémographie : comment estimer l'âge au décès ? » de H. Caussinus, L. Buchet, D. Courgeau et I. Séguy, illustre ce type de collaboration inter-disciplinaire. L'objectif du travail est d'estimer la distribution des âges de décès sur un site archéologique à partir de mesure osseuses obtenues sur les squelettes du site (ici des mesures crâniennes). Statistiquement, il s'agit d'abord d'estimer les probabilités conditionnelles de l'âge de décès sachant les mesures osseuses (qu'on suppose ne pas changer au cours du temps long). Ceci se fait de façon simple à partir d'une collection de référence pour laquelle on dispose des âges de décès réels et des mesures osseuses (les auteurs montrent tout de même que l'état actuel des collections de référence rend un peu hasardeuse l'utilisation des probabilités conditionnelles estimées sexe par sexe). On pourrait ensuite naïvement se contenter d'appliquer la règle de Bayes pour estimer la distribution des âges de décès après avoir effectué les mesures osseuses sur les crânes étudiés. Les auteurs montrent qu'il n'en est rien et qu'il faut passer par une approche bayésienne pour produire une estimation satisfaisante. Se pose alors le problème du choix de la distribution *a priori* des âges de décès. Ce point est discuté en détail dans l'article. Il illustre parfaitement la démarche que nous évoquions plus haut, faite d'interaction entre les spécialistes statisticiens et des humanités, d'itération autour d'un problème et de développement d'une solution sophistiquée et spécifique.

P. Lanos et A. Philippe abordent eux aussi des problèmes de temporalité avec une approche bayésienne dans « Hierarchical Bayesian modeling for combining dates in archeological context ». Le problème étudié dans cet article est la combinaison de dates produites par différentes méthodes de mesures objectives mais bruitées (radiocarbone, luminescence, magnétisme, etc.). Les auteurs introduisent un modèle hiérarchique complexe qui intègre les différentes sources de bruit, d'abord au niveau de la date elle-même, puis de la courbe de calibration (qui indique comment une date donnée produit une mesure physique) et enfin sur la mesure physique elle-même. La complexité du modèle conduit à utiliser un algorithme MCMC sophistiqué pour estimer la distribution *a posteriori* des paramètres (les distributions conditionnelles ne sont pas simulables directement). Pour permettre aux archéologues de s'approprier cette méthode, les auteurs ont implémenté l'outil ChronoModel [7], un logiciel libre multi-plateformes qui masque la complexité de l'inférence dans le modèle (même si les diagnostiques de l'algorithme MCMC restent disponibles pour les experts). La restitution des résultats sous forme graphique nous semble particulièrement adaptée dans ce contexte.

Soulignons que la sophistication croissante des modèles statistiques pour les données des humanités, bien illustrée dans les deux articles que nous venons de présenter, pose naturellement la question de l'appropriation de ces modèles par les chercheurs des humanités. Il ne semble pas réaliste de supposer que ces chercheurs pourront à court terme développer eux-mêmes de tels modèles (la maîtrise des algorithmes MCMC n'est pas universelle chez les statisticiens eux-mêmes, par exemple). On trouvait déjà dans l'article d'A.-S. Bruno un recours à un logiciel spécifique aux modèles de Cox multi-niveaux (un *package* pour le logiciel R). Si les modèles de

[7] https://chronomodel.com/

 ISSN: 2102-6238

Cox sont aussi présents dans d'autres logiciels de statistiques (peut être plus faciles d'accès que R), ce n'est en général pas le cas des méthodes les plus récentes. D'où l'importance d'initiatives comme le logiciel ChronoModel qui réduisent considérablement le coût d'acquisition de ces modèles par les non spécialistes, tout en présentant les résultats de façon exploitable même quand on ne maîtrise pas complètement l'inférence statistique sous-jacente.

L'innovation statistique dans les deux articles précédents réside avant tout dans l'utilisation d'une approche bayésienne. Cette approche reste sous représentée en statistique, les approches fréquentistes étant plus présentes, en particulier dans les cursus destinés aux non spécialistes. Dans l'article « Markov and the Duchy of Savoy : segmenting a century with regime-switching models », J. Alerini, M. Olteanu et J. Ridgway utilisent un cadre fréquentiste classique dans lequel ils développent deux modèles génératif entièrement nouveaux. Il s'agit ici de traiter une série temporelle à valeurs entières dans laquelle on trouve un nombre très important de valeurs nulles (les données historiques comptent le nombre de textes législatifs portant sur la logistique militaire publiés par mois entre 1559 et 1661 dans le Duché de Savoie). Les études historiques précédentes plaident pour la présence d'un changement de régime dans ces données, avec l'entrée du Duché de Savoie dans un cycle de guerres autour de 1610. Cependant, les méthodes classiques de détection de rupture et de modélisation de données avec changement de régime ne donnent pas des résultats concluants sur ce corpus. Ces échecs motivent ainsi une collaboration poussée entre statisticiens et historiens en vue de la construction de modèles de changement de régime capables de s'adapter à la spécificité des données historiques, ici l'excès de zéros. Les résultats obtenus sont très intéressants du point de vue historique, notamment parce que les deux modèles développés donnent des informations complémentaires : l'un met en relief la transformation du Duché de Savoie, l'autre identifie de façon fine certaines périodes historiques.

Ce numéro spécial se termine par l'article « Génération de graphes aléatoires par échanges multiples d'arêtes » de L. Tabourier, J.-P. Cointet et C. Roth. Contrairement aux autres contributions qui mobilisent des données historiques ou archéologiques, cet article est issue d'une collaboration en sociologie. Il s'intéresse aux interactions sociales modélisées par des graphes (les réseaux sociaux au sens propre du terme). Les graphes étant des objets mathématiques riches et complexes, il est parfois difficile en pratique de déterminer si une structure particulière observée dans un graphe réel doit être considérée comme un bruit ou au contraire comme un indice important sur les interactions et les acteurs du graphe. En effet des contraintes structurelles simples (réseaux de parenté, par exemple) peuvent induire indirectement des structures qu'on pourrait attribuer à un graphe particulier alors qu'elles seront présentes dans l'ensemble des graphes respectant ces contraintes. Les auteurs proposent dans cet article une méthode générale de type MCMC capable d'engendrer des graphes aléatoires qui respectent des contraintes arbitraires, ouvrant ainsi la voie à l'identification de contraintes structurantes et à des tests d'hypothèses empiriques sur les graphes. La méthode proposée est appliquée à l'analyse de réseaux scientifiques de collaboration (co-signature d'articles).

Le panorama donné par ce numéro spécial ne peut qu'être incomplet, ne serait-ce que par la couverture thématique limitée des sciences humaines qu'il propose, en se focalisant sur l'histoire au sens large, avec une ouverture vers la sociologie. Le numéro porte aussi en majeure partie sur

 ISSN: 2102-6238

des problèmes liés à la temporalité et sur des données essentiellement numériques. Pour prendre un exemple parmi de nombreux autres, le traitement des données textuelles n'a pas du tout été abordé ici alors qu'il représente un enjeu central pour les humanités. Il aurait ainsi été intéressant de voir comment les outils relativement récents comme les *topic models* [8] sont peu à peu intégrés dans la pratique courante des chercheurs des humanités.

Malgré ses limites, ce panorama met bien en évidence différents mouvements et enjeux. Nous noterons par exemple qu'aucune des contributions à ce numéro ne porte sur des données volumineuses (et encore moins sur les données massives du « big data »). Au contraire, l'un des problèmes majeurs des humanités est le caractère limité des données. Cela est particulièrement important dans le cas des sciences historiques pour lesquelles l'acquisition de plus de données est souvent impossible. D'où l'importance des approches bayésiennes d'une part, et de la prise en compte de toutes les données, même quand elles sont dépendantes statistiquement, d'autre part. Nous noterons aussi l'importance des logiciels dans l'autonomisation des acteurs. Les données des humanités appellent des modèles en général complexes et récents. Dans les premières phases de leur développement, ces modèles ne peuvent être conçus et mis au point que dans des collaborations inter-disciplinaires. Ces collaborations conduisent dans le meilleur des cas à la production de logiciels relativement facile d'accès (comme ChronoModel déjà cité) qui permettent leur utilisation en autonomie (relative) par des chercheurs des humanités.

Nous remercions les auteurs pour leur contribution et pour leur patience, l'édition de ce numéro ayant pris plus de temps que nous l'estimions naïvement au départ, en négligeant les contraintes qu'impliquent l'inter-disciplinarité. Nous remercions aussi les rapporteurs pour leur relectures attentives et leur remarques pertinentes, ainsi que pour les efforts consentis pour étudier la partie des articles qui ne relevait pas de leur domaine d'expertise. Enfin, merci à l'éditeur en chef, Gilles Celeux, pour son suivi constant, ses conseils et ses encouragements.

[8] http://www.cs.columbia.edu/~blei/topicmodeling.html

 ISSN: 2102-6238