LE MONDE DES DONNÉES

SFdS

HISTOIRE(S) DE(S) DONNÉES NUMÉRIQUES

Jean-Jacques Droesbeke
Catherine Vermandele

edp sciences

Histoire(s) de(s) données numériques

Histoire(s) de(s) données numériques

JEAN-JACQUES DROESBEKE
CATHERINE VERMANDELE

Préface d'Emmanuel Didier

edp sciences

17, avenue du Hoggar – P.A. de Courtabœuf
BP 112, 91944 Les Ulis Cedex A

Composition et mise en pages : Patrick Leleux PAO

Imprimé en France
ISBN (papier) : 978-2-7598-2201-0
ISBN (ebook) : 978-2-7598-2213-3

À Alain Desrosières,
qui savait si bien commenter les données
et cultiver l'amitié

« … Les histoires sont le meilleur moyen
d'élever la vie au-dessus de la médiocrité du quotidien. »

Gilles Legardinier
Complètement cramé

SOMMAIRE

UNE NOUVELLE COLLECTION D'OUVRAGES DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

L'article premier des statuts de la Société Française de Statistique (SFdS) stipule que cette société savante « *a pour but de promouvoir l'utilisation de la statistique et ses développements méthodologiques, d'assurer la représentation de ceux qui la pratiquent, l'enseignent et y effectuent de la recherche, de coopérer avec les autres organisations concernées. Elle se propose en particulier de faciliter les échanges entre statisticiens travaillant dans les administrations, les entreprises et les établissements d'enseignement ou de recherche* ». Le deuxième article précise que « *Les moyens d'action de l'association consistent en particulier en l'organisation de réunions et de congrès, l'édition de publications et en l'attribution de prix, médailles et récompenses* ».

La SFdS propose deux types de publications : des revues et des ouvrages. Citons d'abord les premières.

Le ***Journal de la Société Française de Statistique*** (http://journal-sfds.fr) publie, après évaluation par des spécialistes, des articles, en français ou en anglais, dédiés à des recherches méthodologiques ou appliquées dans tous les champs de la statistique. Il vise un niveau académique élevé. Il souhaite encourager la publication d'articles émanant de jeunes chercheurs ainsi que des numéros spéciaux

thématiques dressant l'état de l'art sur des sujets spécifiques afin de fournir un outil de référence pour la communauté des chercheurs académiques et industriels.

La revue ***Statistique et Enseignement*** (http://www.statistique-et-enseignement.fr) publie, après évaluation par des spécialistes, des contributions relatives à l'enseignement de la statistique (niveaux scolaires ou universitaires), à la formation extra-scolaire dans cette discipline, et à sa popularisation « grand public ».

Statistique et Société (http://statistique-et-societe.fr) est un magazine trimestriel publié par la Société Française de Statistique pour quiconque s'intéresse à l'analyse et l'interprétation de controverses et de débats dans lesquels intervient la statistique. C'est une publication d'intérêt général pour les statisticiens, les utilisateurs de statistiques, et tous ceux que la méthodologie statistique intéresse. Ce n'est pas un journal de recherche, et les articles, s'ils font l'objet d'un travail éditorial, ne sont pas soumis à révision par des pairs.

Enfin, la revue ***CSBIGS*** (www.csbigs.fr*)* – *Case Studies in Business, Industry and Government Statistics* – a pour objectif de publier en anglais des études de cas présentant des applications de la statistique principalement dans les champs des trois grands secteurs économiques et de la statistique officielle. Les finalités sont de promouvoir l'utilisation de nouvelles techniques statistiques en illustrant leur mise en œuvre dans des cas d'application, de fournir des cas à visée pédagogique pour les enseignants, de permettre à des consultants de confronter des pratiques au savoir universitaire, d'inciter la communauté à interagir sur des cas en proposant des études ou des approches alternatives.

À côté de ces revues, la Société Française de Statistique publie des ouvrages, essentiellement au sein de quatre collections. La première d'entre elles propose les ouvrages issus des ***Journées d'étude en statistique***[1], publiés chez Technip. Une deuxième collection, intitulée

1. Voir Droesbeke (2017).

Pratique de la statistique, est publiée en collaboration avec les Presses universitaires de Rennes. Une troisième collection, ***La statistique autrement,*** a pour objectif de favoriser la compréhension de la statistique et de son enseignement dans la société de notre époque. Les ouvrages de cette collection sont publiés par les éditions Technip à Paris.

Enfin, la dernière collection créée par la Société Française de Statistique s'intitule ***Le monde des données***. Les ouvrages de cette collection sont écrits à l'intention de ceux qui ne connaissent pas suffisamment cette discipline et son langage pour accéder aux ouvrages diffusés habituellement sur le marché. L'ouvrage que vous tenez entre les mains est le premier livre de cette nouvelle collection. S'adressant à un public très large, il témoigne de la volonté de la Société Française de Statistique de s'ouvrir davantage vers la société civile.

Après une étude préalable menée par Emmanuel Didier, rédacteur en chef de la revue *Statistique et Société*, Jean-Jacques Droesbeke, président de la cellule Publications de la SFdS et Catherine Vermandele, rédactrice en chef de la revue *Statistique et Enseignement*, nous avons signé une convention de publication avec EDP Sciences que nous remercions chaleureusement ici.

Puissent les lectrices et les lecteurs trouver dans cette nouvelle collection ce qu'ils recherchent pour mieux entrer dans le monde des données qui s'ouvre de plus en plus aux citoyens que nous sommes.

Gérard Biau
Président de la Société Française de Statistique
Octobre 2017

PRÉFACE
LES DONNÉES ET LA VIE

Comme beaucoup, vous êtes passionnée ou passionné par l'explosion de données numériques à laquelle nous assistons aujourd'hui, qui nous promet des avancées économiques, sociales et culturelles de tous ordres. Et en même temps, vous restez dubitative ou dubitatif sur les façons de vous approprier tous ces nombres, de reprendre le contrôle sur ce nouveau vocabulaire parfois surprenant, sinon abscons.

Et bien vous êtes ici entre de bonnes mains ! Ce livre a tout ce qu'il faut pour faire évanouir vos doutes. Il vous offre les moyens de remettre en perspective et dans leur contexte d'usage ces données, et ainsi de les juger avec aplomb et confiance.

Vous êtes entre de bonnes mains d'abord à cause de la carrière des auteurs. Tous deux sont de grands experts en nombres : professeurs de statistique à l'Université libre de Bruxelles, ils sont réputés pour leurs travaux sur l'inférence ou la modélisation. Mais ces titres seraient bien insuffisants pour écrire le livre qu'ils nous proposent ici, car ce n'est pas assez de savoir les mathématiques, il faut aussi savoir les rendre accessibles. Or nos deux auteurs sont habités par cette passion depuis longtemps. Ils sont tous deux membres de la Société Belge de Statistique et de la Société Française de Statistique. Catherine Vermandele est de

surcroît passionnée par l'enseignement et la pédagogie de la statistique, à tel point qu'elle dirige la revue *Statistique et Enseignement*. Quant à Jean-Jacques Droesbeke, membre de l'Institut international de statistique, il est aussi, depuis fort longtemps, passionné par l'histoire de la statistique – il est co-auteur du « Que sais-Je ? » qui porte ce titre. Enfin, ces auteurs se sont associés avec un dessinateur de bandes dessinées qui illustre le texte de façon fort plaisante, produisant ainsi un effet de récréation bienvenu. Pédagogie, histoire, illustration, voilà l'écrin dans lequel les auteurs nous présentent aimablement les données qui autrement pourraient nous impressionner.

Qu'ils me permettent d'ajouter qu'ils ont aussi la grande qualité d'être Belges. « *Et alors ?* », pourrait-on me demander. Et bien, sans tomber dans le culturalisme, je peux tout de même dire que mes amis belges sont pour la plupart simples, dénués de prétention, drôles et très pertinents – et ceux-ci ne font pas exception.

Vous êtes entre de bonnes mains aussi pour la composition du livre qu'ils vous proposent. Les auteurs l'ont structuré en de très brefs chapitres, la plupart du temps de cinq ou six pages, ce qui facilite la lecture. Les chapitres sont regroupés en dix parties, dont l'ordonnancement donne une direction générale à l'ouvrage, qui est chrono-thématique : on part des plus anciennes données retrouvées (à Sumer) et on arrive au monde contemporain, et à chaque étape on aborde une question propre aux données (qu'est-ce que la corrélation, un événement rare, un bon graphique, etc.). Chaque période est l'occasion de poser une question. Or une grande liberté que nous propose ce livre est qu'il n'est pas nécessaire de le lire linéairement. Comme dans *La vie mode d'emploi,* le roman magistral de Georges Perec, on peut lire les chapitres dans l'ordre que l'on veut. Assurément, ces auteurs sont du genre à nous procurer des marges de manœuvre, et non l'inverse !

Vous êtes encore entre de bonnes mains à cause de l'argument général de l'ouvrage. On a parfois tendance à penser que les données numériques constituent une langue à part, et même un monde séparé, secret, auquel n'ont accès que ceux qui, après de pénibles efforts, se

sont habitués à parler la langue hermétique des formalismes mathématiques. Mais ce livre nous prouve par A + B que cet argument est faux. Au contraire, il nous montre que les données font intimement partie de notre vie de tous les jours et participent à toutes les sphères de nos activités. Il n'y a pas de séparation entre le quotidien et les données : celles-ci sont partout et partout on trouve des données. Les auteurs recourent à un très grand nombre d'historiettes, qui peuvent être prises à une source précise (on y rencontre Pythagore, Bernoulli, Cassini, Newton, Quetelet, Minard et bien d'autres), ou à la culture commune (l'invention du jeu d'échec), ou même inventées de toute pièce (on assistera à un débat syndical ou à une partie de Cluedo), pour mettre les données dans leur contexte et montrer la variété des situations où elles peuvent prendre sens. Ils utilisent aussi la littérature – en particulier le recueil de Queneau intitulé *Cent mille milliards de poèmes* – montrant ainsi implicitement que Pascal avait tort lorsqu'il opposait l'esprit de synthèse, littéraire, et l'esprit de géométrie. Ils pointent la valeur heuristique des paradoxes, si souvent utilisés en matière de mathématiques. Bref, leur texte nous montre que les données sont au cœur de la vie, qu'elles sont vivantes et joyeuses, de mille et une manières. Ils s'inscrivent ainsi dans la belle tradition initiée par Émile Borel lorsqu'il avait écrit *Les probabilités et la vie*.

Vous êtes enfin entre de bonnes mains parce que les auteurs savent qu'il n'y a pas à proprement parler une « révolution » numérique. Ils nous montrent l'épaisseur du socle qui a été construit depuis des siècles pour qu'aujourd'hui quelque chose de nouveau apparaisse, certes, mais certainement pas quelque chose d'inouï, d'inexplicable ou d'incommensurable comme on nous l'assène parfois. Non, les données prennent appui sur une longue et passionnante histoire de la statistique et de ses outils qui nous permettent de leur donner sens, que l'on soit spécialiste ou pas… du moment que l'on a lu ce livre !

Emmanuel Didier
Rédacteur en chef de *Statistique et Société*

AVANT-PROPOS

Chaque jour nous apporte son lot de données numériques, c'est-à-dire d'informations chiffrées. Certaines sont souriantes, d'autres mystérieuses, d'autres encore terrifiantes… Parce qu'elles se rattachent toujours à des situations, à des objets, à des images, à des sentiments qui leur transmettent leurs caractéristiques. Certains journaux ont leur *chiffre du jour*. Vous apprenez le taux d'accroissement du chômage du mois dernier ; avec une ou deux décimales, il fait encore plus mal. Le nombre de grippés prévu pour la semaine prochaine fait froid dans le dos. Celui du nombre de tués la semaine dernière est révoltant. Il est des données moins ravageuses : cinquante couples se sont mariés au même endroit samedi dernier, quel embouteillage dans cette petite ville. En une vie, le Français moyen aura bu quinze barriques de vin rouge ; on imagine le mal de tête… Il y en a pour tous les goûts. Certaines données sont très utiles, d'autres sont bien futiles. Et plus elles sont précises, plus elles semblent crédibles. À notre époque, elles sont tellement nombreuses qu'on n'a plus vraiment le temps de les regarder de près, de les analyser ou tout simplement d'en comprendre l'intérêt. Il faut dire que ceux qui nous les transmettent ne le font pas toujours dans les règles de l'art. Bref, Monsieur et Madame Toulmonde se sentent souvent déstabilisés par

cet afflux de données. Deux attitudes extrêmes en résultent souvent, allant de : « *Oh ! Moi, les chiffres ! Ils sont de toute façon tous faux* ! » à : « *Vous avez vu ce nombre ? Je vais vous expliquer pourquoi rien ne fonctionne en ce moment !* ».

Rendre les données plus proches de nous, leur donner un sens (quand c'est possible !), éviter de mal les présenter… est de plus en plus nécessaire. Il ne faut pas grand-chose pour y arriver. C'est ce que nous avons déjà tenté de démontrer dans un livre publié en 2016 dans la collection *La statistique autrement* de la Société Française de Statistique et destiné en priorité à ceux qui sont chargés de les recueillir, de les analyser, de les diffuser ou d'enseigner les méthodes utilisées à cet effet. Notre but était d'apporter notre petite contribution à tous ceux qui souffrent autant que nous de cette situation.

Mais « à raconter ses maux, souvent on les soulage », écrit Pierre Corneille dans *Polyeucte*. C'est pourquoi nous avons voulu écrire un autre livre, moins technique et construit selon une autre logique. Ce nouvel ouvrage a pour objectif de vous faire rentrer, lectrices et lecteurs, dans le monde des données en suivant un fil conducteur qui raconte son évolution, ses découvertes, ses caractéristiques. Nous avons donc choisi de vous sensibiliser à l'importance des données numériques dans notre vie quotidienne en racontant une histoire essentielle au moyen de quelques histoires emblématiques. La plupart d'entre elles sont vraies, les autres pourraient l'être. Il en est de surprenantes, d'autres bien navrantes. Toutes sont issues de notre volonté de montrer que la façon dont des données numériques ont été utilisées et parfois maltraitées dans le passé doit nous éclairer sur la manière actuelle d'en recueillir, de les fabriquer, de les analyser, de les interpréter.

Insistons sur ce point. Ces histoires ont pour seule ambition de faire réfléchir à la manière d'aborder des données numériques dont la sécheresse de la présentation voisine souvent avec une imprécision dans le langage ou les moyens de communication utilisés. D'autres histoires auraient pu être évoquées mais l'exhaustivité n'était pas