Data visualisation et enseignement de la statistique au travers d'exemples historiques en R

Jonathan El Methni [1]

[1] *Université Paris Descartes, Sorbonne Paris Cité, Laboratoire MAP5, UMR CNRS 8145, 75006 Paris, France*
jonathan.el-methni@parisdescartes.fr

Résumé. Cette communication a pour but de présenter des exemples de grands moments de la visualisation de données mettant en évidence l'impact historique qu'ont pu avoir les statistiques, l'histoire de la statistique dans l'Histoire ainsi que le soulèvement de diverses questions éthiques et les liens tissés avec d'autres disciplines. Ces exemples seront illustrés par des graphiques statistiques réalisés avec le logiciel R à partir de jeux de données historiques contenus dans le package HistData. On mettra en correspondance les graphiques statistiques obtenus et leurs homologues historiques. Ces travaux permettent d'aborder la statistique sous un nouvel angle pédagogique ainsi que d'enrichir et d'illustrer des enseignements de statistiques.

Mots-clés. Visualisation de données, histoire de la statistique, enseignement, R, package HistData.

Abstract. The purpose of this communication is to present examples of data visualization highlighting the historical impact of statistics, the history of statistics in History, the raising of various ethical issues and links with other disciplines. These examples will be illustrated by statistical graphics made with the R software from historical data sets contained in the HistData package. The statistical graphics obtained will be matched to their historical counterparts. This work makes it possible to approach statistics in a new educational context and to enrich and illustrate lessons in statistics.

Keywords. Data visualization, history of statistics, education, R, HistData package.

1 Cheminement et questionnement

Ce travail est le fruit de mon expérience, c'est pourquoi je commencerai par expliquer le cheminement qui a mené à cette communication. A la suite de mon recrutement en tant que maître de conférences en 2014 à l'Université Paris Descartes, j'ai enseigné au département STatistique et Information Décisionnelle (STID) de l'Institut Universitaire de Technologie (IUT) Paris Descartes. Mes cours s'adressent à de jeunes étudiants de deuxième année en formation initiale. Il y a une diversité dans les thèmes de cours que

je dispense, qui vont de la statistique descriptive au modèle linéaire (régression simple et multiple, anova, ancova). Afin de motiver ce public j'ai cherché à illustrer mes cours à l'aide d'exemples issus de lois physiques. En régression linéaire simple on peut citer l'exemple de la loi d'Ohm afin d'estimer la valeur de la résistance, celui de la loi de Galilée permettant d'estimer la constante de gravitation universelle ou encore la loi de Hubble dans le but d'estimer l'âge de l'Univers.

Partant du fait qu'il y a de plus en plus de données disponibles et accessibles (données socio-économiques, data-journalisme, mobilité, transports, sport, etc.) le département STID a ouvert en 2015 un Diplôme Universitaire (DU) s'intitulant DU DataViz. Ce diplôme qui s'adresse à des personnes en formation continue porte sur la visualisation et l'aide à l'interprétration des données. D'une durée courte de 150 heures et conciliable avec une activité professionnelle, il vise des étudiants de niveau licence 3. L'équipe pédagogique est mixte, composée à la fois d'universitaires et de professionnels issus du monde socio-économique. La data visualisation a pour principal objectif d'explorer des données brutes et de les traduire en information interprétable à l'aide de représentations graphiques. Cette discipline est avant tout un outil d'analyse et de compréhension, qui offre également la possibilité d'engager des stratégies, de faciliter la prise de décision, voire d'innover mais surtout de communiquer et de transmettre.

La première année j'ai donné un cours plutôt "classique" avec quelques contextualisations historiques. Or, ce public, plus âgé (que ceux de la formation initiale de l'IUT), a été très intéressé et j'ai eu à répondre à des questions concernant les enjeux sous-jacents aux techniques statistiques présentées. Leurs questionnements étaient du type : Dans quel but cette technique a-t-elle été développée ? Quel était le contexte historique ? Dans le but d'enrichir mes cours avec des données historiques j'ai découvert un site entièrement dédié à la visualisation de données. Ce site[1] est l'oeuvre de Mickael Friendly professeur de psychologie à York University. Il est à l'origine du Milestone Project. Un projet sur l'histoire de la visualisation de données qui a donné lieu à diverses publications (voir Friendly (2007)). Ce site regorge de liens vers des livres, des galeries de visualisation de données avec relecture de graphiques historiques, des cours, des articles, et des liens vers R et SAS. En particulier il renvoie vers un package de R, développé par les créateurs du site, s'intitulant HistData. On y trouve des jeux de données historiques ainsi que des exemples de visualisation de données possibles à réaliser. A l'aide de ce package et du logiciel R j'ai pu reformuler mes cours en y ajoutant une dimension historique et illustrative.

A travers cette communication je vais présenter cinq exemples qui me paraissent pertinents car ils mettent en évidence l'impact historique qu'ont pu avoir les statistiques, l'histoire de la statistique dans l'Histoire ainsi que le soulèvement de diverses questions éthiques et les liens tissés avec d'autres disciplines (permettant dans certains cas le développement de ces dernières). J'aimerais montrer à travers ces exemples quelques grands moments de la visualisation de données et l'incidence qu'elle a pu avoir et qu'elle a

[1] http://www.datavis.ca

toujours. J'utiliserai plus particulièrement le package HistData mais également le package Guerry de R pour un cas particulier.

2 De l'utilité de la statistique au travers d'exemples historiques

Cette partie a pour but de mettre en avant cinq cas historiques de l'utilisation et de l'utilité de la statistique. Tous ces exemples seront illustrés par de nombreux graphiques réalisés avec R et mis en correspondance avec leurs homologues historiques.

Le premier cas qui nous intéressera sera celui de Wiliam Playfair (1759–1823), ingénieur et économiste écossais. Playfair fut un des pionniers de la représentation graphique de données. Il est crédité de l'invention des séries chronologiques, des histogrammes et des diagrammes circulaires. Il a développé un des classiques de la visualisation de données (voir Playfair (1821)) concernant l'évolution du salaire hebdomadaire d'un "bon mécanicien" et celle du prix du blé de 1565 à 1821. Par ce travail, il a voulu montrer que le pouvoir d'achat d'un "bon mécanicien" n'avait jamais été aussi élevé qu'en 1821.

Le deuxième cas (certainement le plus connu) que l'on présentera sera celui de Charles Joseph Minard (1781–1870), ingénieur civil français. Minar fut l'un des premiers à utiliser des graphiques appliqués au génie civil et aux statistiques. Sa carte figurative des pertes successives en hommes de l'armée française dans la campagne de Russie en 1812–1813 (voir Minard (1844)) est considérée comme un (si ce n'est le) chef d'oeuvre de visualisation de données. En effet cette carte en deux dimensions intègre et synthétise parfaitement pas moins de six niveaux d'informations, elle donne la chronologie des évènements, la localisation et l'itinéraire de l'armée indiquant les points de séparation et de regroupement des unités, pertes humaines de l'armée (particulièrement sensibles lors de la traversée de la Bérézina) ainsi que les variations de la température de l'air au cours de la retraite des troupes de Napoléon Bonaparte.

On s'intéressera par la suite à Florence Nightingale (1820–1910), infirmière britannique. De même que ces deux prédécesseurs Nightingale fut une pionnière dans l'utilisation des statistiques dans le domaine de la santé et plus particulièrement dans la représentation visuelle de l'information. A la suite de la guerre de Crimée (1853–1856), elle se mit à utiliser une version améliorée des diagrammes circulaires de Playfair dans le but d'illustrer les causes saisonnières de mortalité des patients de l'hôpital militaire qu'elle gèrait. Ses diagrammes (voir Nightingale (1857)) des causes de mortalités dans les armées de l'Est ont montré que la plupart des soldats anglais morts durant la guerre de Crimée l'ont été de maladie plutôt que de blessures ou d'autres causes. Ses rapports sur la nature et les conditions de soins médicaux permirent aux membres du parlement de réaliser l'ampleur

du désastre et menèrent à une réforme médicale. Ces derniers n'auraient probablement pas pu lire ou comprendre des rapports statistiques traditionnels.

Le cas suivant concernera André-Michel Guerry (1802–1866), statisticien et juriste français. Il est considéré (avec Adolphe Quetelet) comme le fondateur de la "statistique morale", discipline à l'origine du développement de la criminologie, de la sociologie et des sciences sociales. On s'intéressera particulièrement à deux de ses cartes choroplèthes de France (voir Guerry (1833)). Elles représentent les départements français coloriés selon le nombre de crimes contre les personnes pour la première et selon les atteintes à la propriété pour la seconde. Le but recherché était d'apporter une réponse cartographique aux questions sociales de l'époque : Est-ce que le niveau d'instruction et de criminalité sont liés ? Guerry souhaitait alors faire un lien entre deux variables.

Enfin notre dernier exemple sera celui de Sir Francis Galton (1822 –1911), anthropologue, explorateur, géographe, inventeur, météorologue, proto-généticien, psychométricien et statisticien britannique. Sir Francis Galton, cousin de Charles Darwin, cherchait à faire le lien entre la théorie de la sélection naturelle et la recherche en statistique. Il défendit la théorie de l'évolution, en se proposant de montrer qu'elle permettait des prévisions susceptibles d'être vérifiées. Ses études portèrent sur la transmission de caractères héréditaires (voir Galton (1869)), de ce fait il est considéré comme le fondateur de l'eugénisme. En 1885, travaillant sur l'hérédité, il chercha à expliquer la taille des enfants en fonction de celle de leurs parents. Il en conclura, à l'aide d'un graphique que l'on présentera, que bien qu'il y ait une tendance pour les parents de taille élevée à avoir des enfants de taille élevée et pour les parents de petite taille à avoir des enfants petits, la taille moyenne des enfants nés de parents d'une taille donnée avait tendance à se rapprocher de la taille moyenne de la population générale. En d'autres termes, la taille des enfants nés de parents inhabituellement grands ou petits se rapprochait de la taille moyenne de la population. Dans les termes de Galton, il s'agissait d'une "régression vers la médiocrité", d'où l'origine du mot régression en statistique.

3 Conclusions et perspectives

En conclusion, il me semble primordial de contextualiser l'apparition et/ou l'usage de méthodes statistiques. A l'aide des packages HistData et Guerry de R nous avons cette possibilité. Les données disponibles abordent une très grande diversité de thèmes et de disciplines : sociologie, physique, commerce, psychologie, médecine, militaire, épidémiologie, cartographie, biologie, etc. Il est à mes yeux essentiel de faire ce travail historique, de plus ce dernier soulève de nombreuses questions éthiques, qui peuvent faire écho à l'actualité, telle que l'émergence du Big Data ou la cartographie de réseaux sociaux.

Je me donne comme perspective d'insérer un maximum d'exemples historiques dans mes cours ou tout du moins les contextualiser le plus possible. C'est déjà le cas dans mon cours de modèle linéaire où je fais travailler les étudiants sur les données historiques de Sir Francis Galton. Ceci leur offre la possibilité de reproduire les graphiques de nos illustres prédécesseurs. Sous mon impulsion et celle de Xavier Sense, enseignant dans le DU DataViz, nous comptons monter un cours de visualisation de données dans un contexte historique. Nous sommes en train de penser le cours afin que les étudiants puissent aborder la statistique sous un nouvel angle pédagogique. Ils pourront dans un premier temps mettre en oeuvre des techniques statistiques vues en cours et par la suite développer leurs propres outils de visualisation de données.

Bibliographie

[1] Friendly, M. (2007) A Brief History of Data Visualization. In Chen, C., Hardle, W. & Unwin, A. (eds) *Handbook of Computational Statistics: Data Visualization*, Springer–Verlag, Vol. **III**, Ch. 1, 1–34.
[2] Galton, F. (1869). Hereditary Genius: An Inquiry into its Laws and Consequences. *London: Macmillan.*
[3] Guerry, A–M. (1833). Essai Sur La Statistique Morale de la France. *Paris: Crochard.*
[4] Minard, C–J (1844), Tableaux graphiques et cartes figuratives. *Bibliothèque numérique patrimoniale des ponts et chaussées.*
[5] Nightingale, F. (1857). Mortality of the British Army. *London: Harrison and Sons.*
[6] Playfair, W. (1821). Letter on our Agricultural Distresses, *Their Causes and Remedies. London: W. Sams.*