DATA VISUALISATION ET ENSEIGNEMENT DE LA STATISTIQUE AU TRAVERS D'EXEMPLES HISTORIQUES EN R

Jonathan EL METHNI[1]

TITLE

Data visualization and teaching statistics through historical examples in R

RÉSUMÉ

Cet article a pour but de présenter des exemples de grands moments de la visualisation de données mettant en évidence l'impact historique qu'a pu avoir la diffusion des statistiques, l'histoire de la statistique dans l'Histoire, ainsi que les liens tissés avec d'autres disciplines. Ces exemples seront illustrés par des graphiques 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 statistique.

Mots-clés : *visualisation de données, histoire de la statistique, enseignement, R, package HistData.*

ABSTRACT

The purpose of this paper is to present examples of data visualization highlighting the historical impact of statistics, the history of statistics in History and links with other disciplines. These examples will be illustrated by 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 statistical lessons.

Keywords: *data visualization, history of statistics, education, R, HistData package.*

1 Introduction et contexte

Le présent article est le fruit d'une expérience personnelle, c'est pourquoi nous commencerons par expliquer le cheminement qui a mené à sa réalisation. Le département STatistique et Informatique Décisionnelle (STID) de l'Institut Universitaire de Technologie (IUT) de l'Université Paris Descartes forme des jeunes étudiants en statistique. Les étudiants peuvent tout aussi bien suivre une formation initiale qu'une formation en alternance. Ils abordent au cours de leurs cursus une grande diversité de thèmes statistiques.

Le principal problème pédagogique auquel sont confrontés les enseignants concerne la raison d'être de leurs cours. En effet, les étudiants ne cessent, et cela chaque année, de

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

demander quel est l'intérêt du cours qu'ils suivent ? Est-ce qu'il existe des cas concrets où ce cours est utile ? Peut-on avoir accès à des données ainsi que des exemples de problématiques illustrant ce cours ? Nous essayons du mieux possible de contextualiser nos cours et les divers thèmes abordés. Parmi les cours dispensés, celui concernant la visualisation de données offre la possibilité d'apporter des réponses pédagogiques basées sur une contextualisation historique.

1.1 La visualisation de données ou data visualisation

La visualisation de données est l'ensemble des méthodes ou techniques de représentation de valeurs, d'effets ou de phénomènes, visant à mieux comprendre, assimiler et interpréter ces derniers. Le mot visualisation dérive de l'anglais « visualize » qui trouve son origine dans le latin « video » signifiant « être capable de voir ». Ainsi la vision n'est qu'un intermédiaire souvent synthétique et rapide entre l'information visualisée et le cerveau. A ce dernier échoit le rôle de l'assimilation ainsi que de l'interprétation.

La visualisation des données s'appuie essentiellement sur la spatialisation, les formes et les couleurs ainsi que la dynamique pour représenter de l'information appréhendable et interprétable. Cette approche est bien évidemment limitée ne serait-ce que par les limites induites par le cerveau. Celui-ci est capable de différencier la saturation ou la luminosité d'une même couleur mais peine à tenir compte d'une échelle ordinale imposée aux différentes couleurs. L'expérience pédagogique suivante illustre parfaitement ces propos. Il s'agit de présenter le tableau de données suivant et de demander aux étudiants d'en tirer une information majeure :

Jours	1	2	3	4	5	6	7	8	9	10
Atelier A	3452	2865	4125	3807	1245	3912	1985	4007	2963	3189
Atelier B	3711	3128	4413	4060	1515	4228	2316	4195	3215	3448

Rares sont les étudiants capables de trouver la bonne réponse. Mais lorsque ces données sont présentées sous une forme graphique appropriée, une grande majorité des étudiants perçoit que l'atelier B produit toujours plus que l'atelier A et que la différence de production est quasiment constante (voir Figure 1). A ce sujet, on peut citer Tukey : *The greatest value of a picture is when it forces us to notice what we never expected to see.*

1.2 Le DU DataViz

Notre société fait face à de plus en plus de données et souhaite communiquer autour de ces dernières de manière simple et accessible. On peut citer parmi ces acteurs le data-journalisme, les transports et le monde sportif. Partant de ce besoin, le département STID a ouvert en 2015 un Diplôme Universitaire (DU) sur la visualisation de données s'intitulant DU DataViz. Ce diplôme s'adressant à des personnes en formation continue porte sur la visualisation et l'aide à l'interprétation des données. D'une durée courte de 150 heures et conciliable avec une activité professionnelle, il vise des étudiants de niveau licence 3.

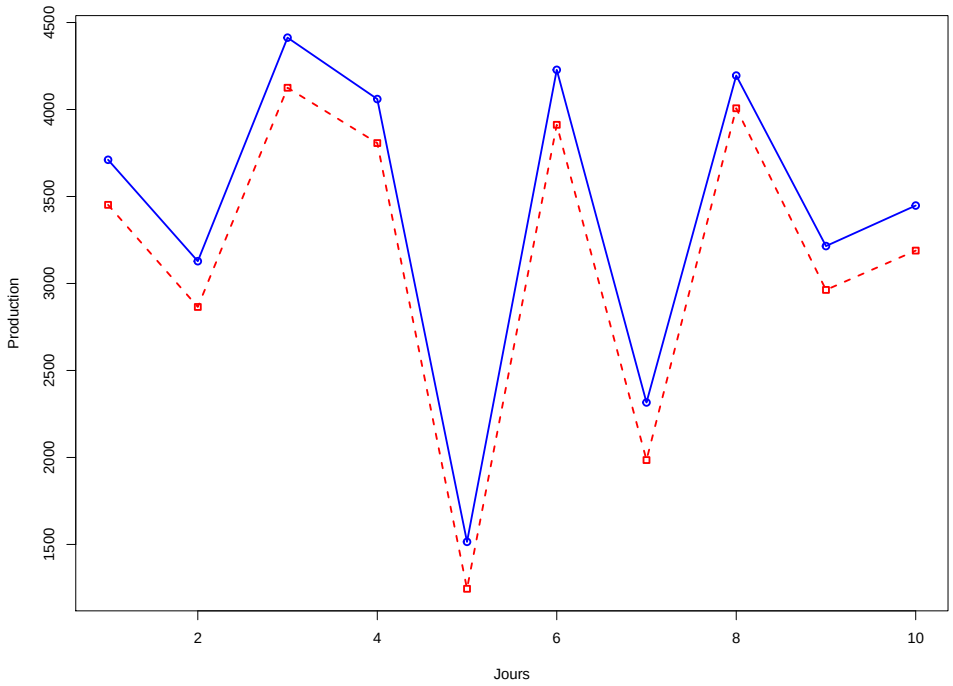


Figure 1 – *Représentation du tableau de production des deux ateliers A et B.*

L'équipe pédagogique est mixte, composée à la fois d'universitaires et de professionnels issus du monde socio-économique.

La formation a pour principal objectif d'explorer des données brutes et de les traduire en information interprétable à l'aide de représentations graphiques. A travers cette formation la data visualisation est conçue comme un outil d'analyse et de compréhension, qui offre la possibilité d'engager des stratégies, de faciliter la prise de décision, voire d'innover, mais également de communiquer et de transmettre. Cette formation, multidisciplinaire par nature, se situe à la croisée de la statistique, de l'informatique et de la communication :

- la statistique : choisir les bons indicateurs et les bonnes représentations, ainsi que connaître les pièges à éviter lors de l'analyse ;
- l'informatique : savoir élaborer des interfaces intuitives, interactives, afin de proposer une expérience utilisateur ergonomique et innovante ;
- la communication : savoir formaliser des visuels graphiques et esthétiques en cohérence avec le message et ses objectifs tout en étant en adéquation avec le public visé.

Ce public plus mature (que celui de la formation initiale de l'IUT) est très intéressé et très intéressant. Au delà des questions habituelles, les étudiants montrent une réelle curiosité concernant l'intérêt, la justification, la finalité ainsi que les enjeux sous-jacents aux techniques statistiques présentées. Leurs questionnements sont du type : Comment

applique-t-on « concrètement » une telle technique ? Quelle est l'origine d'une telle idée ? Comment justifier le choix d'une telle méthode ? Dans quel but cette technique a-t-elle été développée ? Quel était le contexte historique ? Ces interrogations et d'autres interpellent l'enseignant. En effet, comment justifier l'usage universel de la moyenne arithmétique sans décrire le processus qui lui confère ce statut ? Comment expliquer le rôle central acquis par la loi normale en seulement deux siècles sans en référer à ses origines et à son histoire ? Notre approche pédagogique se base sur l'élaboration d'un cadre pertinent basé sur une approche historique afin d'enseigner la data visualisation. Ce processus de contextualisation éveille l'attention et suscite une réflexion des étudiants favorisant des échanges pouvant se poursuivre au-delà des cours.

1.3 Le package HistData de R

Afin d'enrichir et d'illustrer des cours par des jeux de données et des exemples de data visualisation, d'aucun peut trouver une multitude de contenus sur le site[2] internet entièrement dédié à la visualisation de données historiques. Ce dernier est l'œuvre de Mickael Friendly, professeur de psychologie à York University au Canada. Il est à l'origine du Milestone Projet : un projet sur l'histoire de la visualisation de données qui a donné lieu à diverses publications (voir Friendly (2007), ainsi que Friendly *et al.* (2010, 2011)). 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 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, il est possible pour tout un chacun d'agrémenter ses cours en y ajoutant une dimension historique.

Dans la suite de cet article nous présentons six exemples de visualisation de données historiques. Le choix de ces exemples s'explique par leur pertinence à mettre en évidence à la fois l'impact historique qu'ont pu avoir les statistiques, mais également l'histoire de la statistique dans l'Histoire. Certains de ces exemples soulèvent des questions et apportent un éclairage sur les liens tissés avec d'autres disciplines, favorisant dans certains cas le développement de ces dernières. Pour plus de détails sur les jeux de données disponibles dans le package *HistData*, on pourra se reporter au Tableau 1 donné en Annexe 1. On y retrouve les principaux thèmes et figures historiques ainsi que les noms des jeux de données et leurs époques. Nous aimerions 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 toujours.

[2] `http://www.datavis.ca`

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

Il n'est guère évident de dater le commencement de l'histoire de la data visualisation, Friendly (2011) allant même jusqu'à considérer que les peintures rupestres font partie de la représentation de données. Dans le livre *Histoire de la statistique* (voir Droesbeke et Tassi, 1997), les auteurs situent l'origine de la représentation graphique « quantitative » dans la construction de cartes géographiques dont les plus anciennes datent d'environ 6000 ans. Ces dernières nous sont parvenues sur des tablettes d'argile gravées en Mésopotamie. L'histoire est ainsi jalonnée par de nouvelles représentations graphiques de données. On peut retenir, au V^{e} siècle, Macrobe dans son « *Commentaire au Songe de Scipion* » dans lequel il donne une description du mouvement des planètes au cours du temps. Il s'agit probablement du premier graphique de série temporelle de l'histoire. Par la suite, un jalon important est celui des travaux de Nicole Oresme (1320–1382), savant français de l'époque médiévale. En 1370, trois siècles avant René Descartes (1596–1650), ce dernier, dans son « *Tractatus de configuratione qualitatum et motuum* » jamais imprimé, représente sous la forme graphique le rapport entre deux variables, et préfigure ainsi les premières fonctions. Il introduit un système de coordonnées rectangulaires qu'il utilise pour donner des représentations graphiques de fonctions, signant ainsi les prémices de la géométrie analytique. Pour plus de détails sur l'histoire des graphiques statistiques on pourra se reporter à Droesbeke et Tassi (1997). C'est au XVII^{e} siècle qu'apparaissent les premiers graphiques statistiques avec les travaux de Michael van Langren qui feront l'objet de notre premier cas d'étude.

Dans la suite de cette partie nous présentons, dans leur ordre chronologique d'apparition, six exemples historiques mettant en avant l'utilisation et l'utilité de la statistique :

- la mesure de la différence de longitudes entre Tolède et Rome de Michael van Langren ;
- les graphiques de William Playfair ;
- les cartes choroplèthes de André-Michel Guerry ;
- les roses de Florence Nightingale ;
- la carte de John Snow ou les débuts de l'épidémiologie ;
- un des chefs d'oeuvre de Charles Minard.

Tous ces exemples seront illustrés par des graphiques réalisés avec R et mis en correspondance avec leurs homologues historiques. Lorsque nous présentons ces données et leurs graphiques historiques aux étudiants, nous insistons sur le contexte historique. Pour chaque exemple nous présenterons donc succinctement le contexte historique, nous donnerons le (ou les) graphique(s) et une version revisitée possible à faire en R, ainsi que le jeu de données correspondant du package *HistData*.

2.1 La mesure de la différence de longitudes entre Tolède et Rome de Michael van Langren

Michael van Langren (1598–1675) était un cartographe, ingénieur, mathématicien, cosmographe et astronome à la cour du roi Philippe IV d'Espagne (1605–1665). A cette époque les Pays Bas Méridionaux (la Belgique actuelle) faisaient partie du royaume d'Espagne. Le XVII^{e} siècle a connu une grande activité de mesure des grandeurs physiques (temps, distances et localisations spatiales). Un des problèmes au centre des recherches et préoccupations de l'époque était celui du calcul des longitudes. Ce problème difficile était d'un intêret primordial pour les navigateurs. Michael van Langren y consacra une grande partie de sa vie sans y apporter une réponse complète. Cependant, cela le mènera à être l'auteur en 1644 d'une des premières représentations de données statistiques (voir Tufte, 1997) donnée en Figure 2. Historiquement, ce graphique est considéré comme le premier graphique statistique mais Friendly *et al.* (2010) ont établi qu'une version antérieure a vu le jour en 1628.

Cette représentation graphique avait pour but d'illustrer et « d'estimer » la différence de longitudes entre les villes de Tolède (Espagne) et de Rome (Italie). Elle est considérée comme statistique car elle représente une série de données. Il s'agit en effet de mesures réalisées à différentes époques par des scientifiques dont l'autorité était reconnue. Parmi les plus célèbres on peut citer Claude Ptolémée (90–168), Gérard Mercator (1512–1594) ou encore Tycho Brahé (1546–1601). C'est d'ailleurs ce dernier qui fut le premier à introduire la notion statistique fondamentale d'indice de tendance centrale. En effet, il eut recours à de « nombreuses » observations d'une même quantité afin d'en estimer la valeur. La moyenne arithmétique apparaît alors pour la première fois dans l'œuvre de l'astronome Danois afin d'éliminer des erreurs d'observations. Ceci permit à Johannes Kepler (1571–1630) de formuler les lois régissant les mouvements des planètes autour de leur orbite (voir Droesbeke et Tassi, 1997).

Michael van Langren a, quant à lui, mis au point un graphique en une dimension faisant apparaître la grande dispersion des mesures et a situé la ville de Rome dans une position centrale, réalisant de ce fait la première représentation graphique d'un indice de tendance centrale. Il est important de remarquer que van Langren ne donne pas de valeur numérique à la longitude de la ville de Rome mais situe cette dernière explicitement par son nom orthographié en grands caractères s'étalant sur une large amplitude. Malgré la place occupée par le mot ROMA, la longitude suggérée est loin de la valeur réelle de 16.5° indiquée par la flèche ajoutée à ce graphique. On retrouve également dans la Figure 2, d'une part, la représentation de van Langren et, d'autre part, une version revisitée en R (voir le jeu de données *Langren*), toutes deux superposées à une carte actuelle de l'Europe. Il est à noter la surestimation et la grande dispersion des mesures.

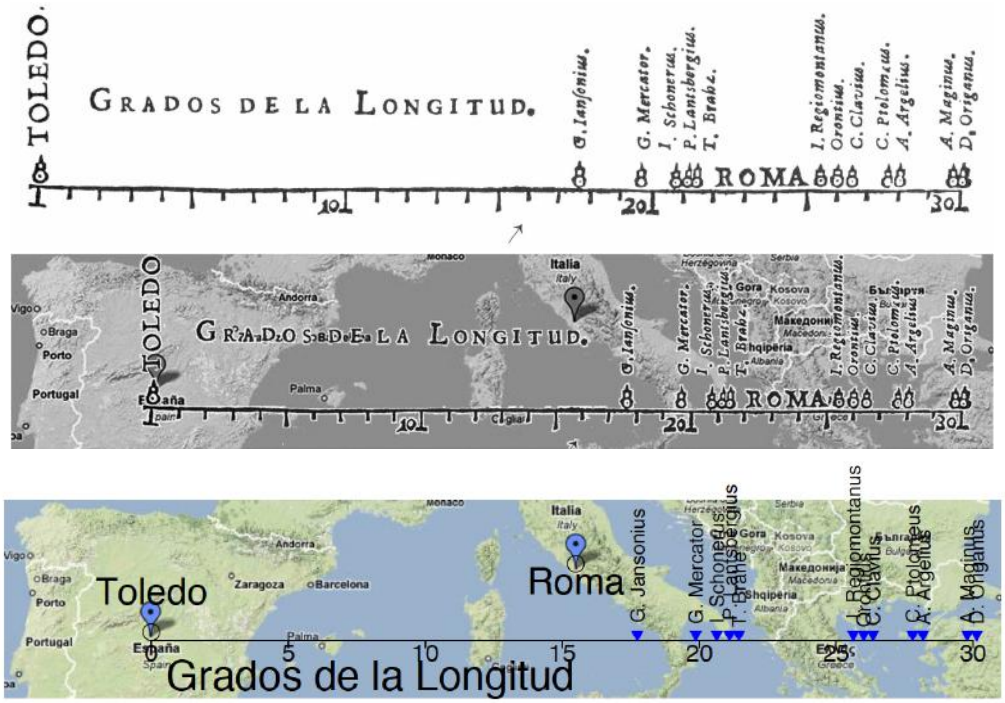


Figure 2 – *En haut : graphique original de van Langren. Source : Friendly et al. (2010). Au milieu : le même superposé sur une carte actuelle. En bas : graphique de van Langren revisité en R.*

2.2 Les graphiques de William Playfair

Le second cas qui nous intéresse est celui de William Playfair (1759–1823), ingénieur et économiste écossais. On attribue à ce pionnier l'invention des graphiques statistiques de base largement utilisés aujourd'hui. Les constructions graphiques de William Playfair sont considérées comme des modèles du genre. En effet, on lui attribue aujourd'hui l'anticipation intuitive de certaines idées fondamentales exploitées en psychologie expérimentale. Plus précisément les idées à l'origine des travaux portant sur les interactions entre la perception naturelle et les capacités cognitives modélisant l'appréhension et l'interprétation des représentations graphiques (voir Spence, 2006).

Il a notamment conçu les diagrammes circulaires et les diagrammes en barres. Le diagramme circulaire de la Figure 3, publié en 1801 dans *The Statistical Breviary* (voir Playfair, 2005), exploite visuellement la proportionnalité entre angles traduisant la proportionnalité entre les tailles des différents états des Etats-Unis d'Amérique. On peut remarquer un manque de lisibilité pour les plus petits états. Ces derniers auraient pu être intercalés entre d'autres ayant de plus grands secteurs angulaires.

Le diagramme en barres de la Figure 3, publié en 1786 dans *The Commercial and Political Atlas* (voir Playfair, 2005), représente les échanges commerciaux de l'Ecosse avec différents pays. On y retrouve les exportations et importations juxtaposées pour chacun des pays. La longueur des barres est dans une échelle absolue figurant dans l'entête. On peut ainsi synthétiser visuellement des informations comparatives. On peut signaler

la difficulté de lecture pour les pays à faibles échanges commerciaux. Ne disposant que de données sur une année (de Noël 1780 à Noël 1781), William Playfair a été amené à concevoir le diagramme en barre par manque de données temporelles. C'est presque en s'excusant qu'il écrivit : « *This chart does not comprehend any portion of time, and it is much inferior in utility to those that do* ». Playfair est surtout connu pour ses nombreuses représentations chronologiques. En effet, sur les 44 graphiques que compte *The Commercial and Political Atlas*, 43 concernent des séries chronologiques, le dernier étant le diagramme en barres précédent.

On retrouve en Figure 4 son graphique le plus célèbre, paru dans Playfair (1821), sous le titre *Chart shewing at one view the price of the quarter of wheat and wages of labor by the week from the year 1565 to 1821*. Il y représente et étudie l'évolution du salaire hebdomadaire de ce qu'il désigne comme étant un « bon mécanicien » en le comparant à l'évolution du prix du blé et ceci sur une période s'étalant de 1565 à 1821. Ainsi il représente sur un même graphique deux séries chronologiques et une frise historique se situant en haut du graphique où l'on retrouve les règnes des différents monarques anglais. L'axe horizontal représente le temps avec une unité de 5 ans, l'axe vertical représente la monnaie avec une unité de 5 shillings. Les barres représentent le prix du blé pour un quarter (1 quarter = 12.7 kg) et la courbe le salaire hebdomadaire d'un « bon mécanicien ». Playfair a voulu montrer que le pouvoir d'achat d'un « bon mécanicien » a connu une tendance à la hausse durant la période considérée pour atteindre son maximum en 1821. Dans la Figure 4, on trouve également une version revisitée en R de ce fameux graphique (voir le jeu de données *Wheat*).

Pédagogiquement cette représentation est intéressante à présenter aux étudiants. Bien que le prix du blé et le salaire s'expriment dans la même unité, le shilling, ce graphique soulève le problème de la cohérence des échelles. En effet, on peut se demander ce qui a motivé le choix du quarter pour le prix du blé et de la semaine pour le salaire. Playfair aurait pu tout aussi bien choisir un salaire mensuel ou le prix d'une autre mesure du poids du blé. Est-ce uniquement l'usage de l'époque ou le souci de superposer les deux graphiques tout en les dissociant ? On peut également se poser la question de la représentation différente de deux séries statistiques de même nature : courbe unidimensionnelle pour le salaire et diagramme en barres pour le prix du blé. Enfin, on peut remarquer que Playfair n'a pas envisagé de représentation graphique bidimensionnelle mettant en jeu le rapport du salaire par le prix du blé en fonction du temps. Il passa de ce fait à côté d'un concept fondamental de la statistique : la notion de lien entre deux variables. Ce sera l'objet de notre troisième exemple.

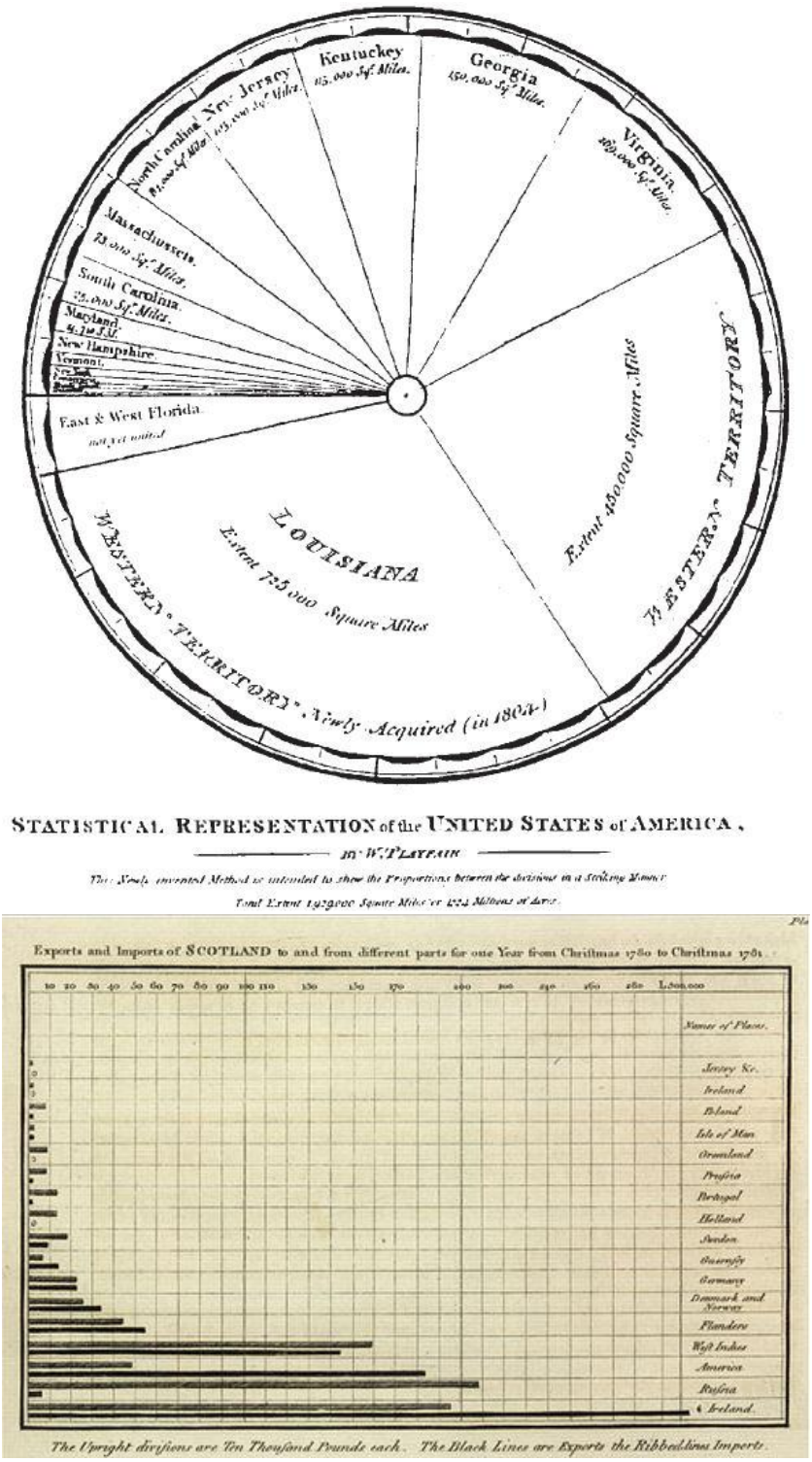


Figure 3 – *Graphiques originaux de Playfair. En haut : Statistical Reprensation of the United States of America. En bas : Exports and Imports of Scotland to and from different parts for one Year from Christmas 1780 to Christmas 1781. Source : Playfair (2005).*

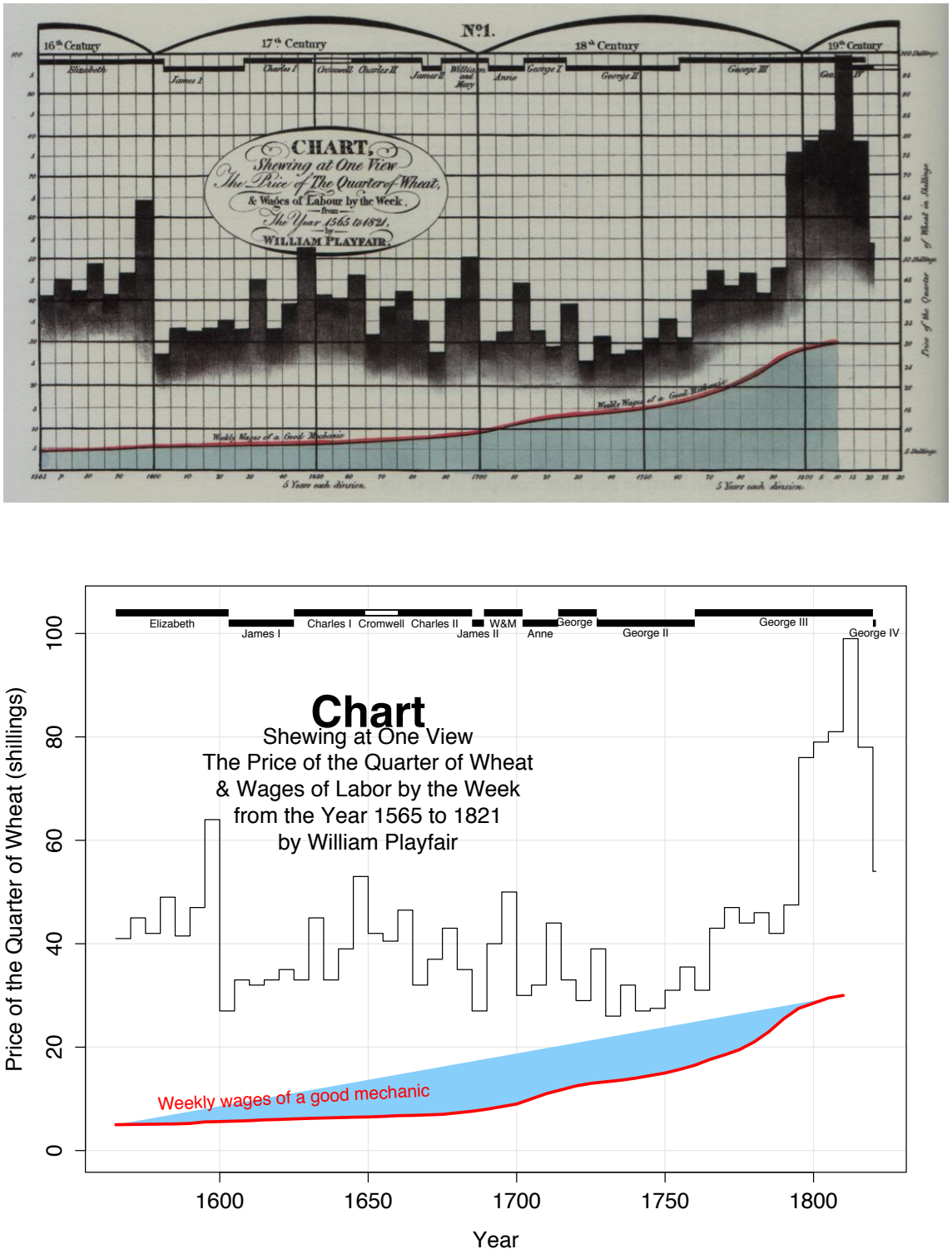


FIGURE 4 – *En haut : Chart shewing at one view the price of the quarter of wheat and wages of labor by the week from the year 1565 to 1821. Source : Playfair (1821). En bas : le même revisité en R.*

2.3 Les cartes choroplèthes de André-Michel Guerry

Une carte choroplèthe est une carte partitionnée en régions selon une thématique. Les aires géographiques (khorê = zone/région) ainsi délimitées sont coloriées selon une échelle de tons représentant une information quantitative (plethos = multiple). C'est au polytechnicien français Charles Dupin (1784–1873), l'un des précurseurs de la cartographie moderne, que l'on doit la première carte choroplèthe de l'histoire donnée en Figure 5 (voir Dupin, 1826) qu'il nomma alors « carte teintée ». En 1826, sa carte figurative de l'instruction populaire de la France, avec différentes colorations des départements, met l'accent sur l'opposition entre la France du Nord et la France Méridionale en termes de scolarisation.

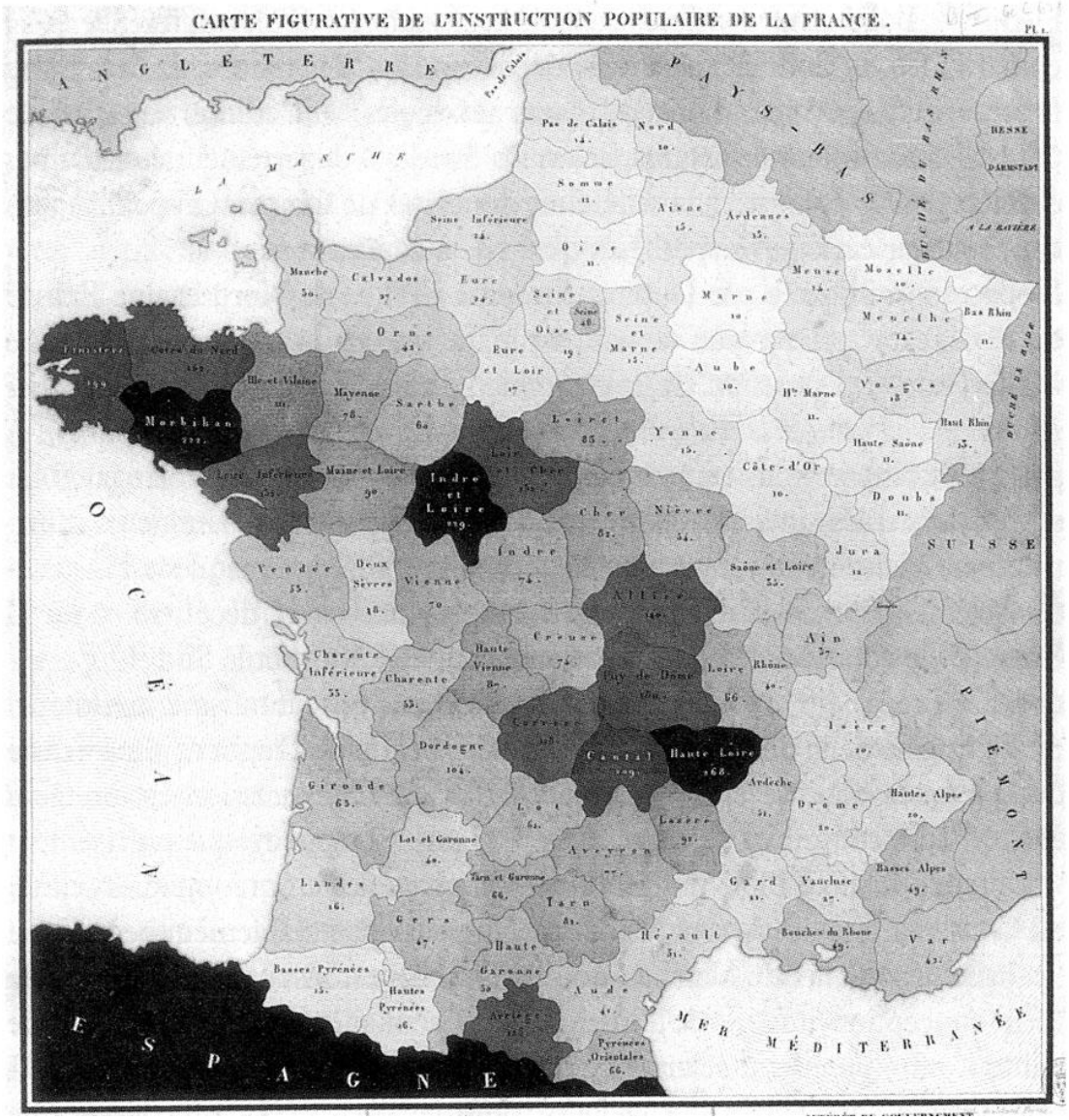


Figure 5 – *Graphique original de Dupin : carte figurative de l'instruction populaire de la France (1826). Source : Dupin (1826).*

Néanmoins, c'est au statisticien et juriste français André-Michel Guerry (1802–1866) que l'on doit le développement de ces cartes (voir Guerry, 1833) et leur usage en tant qu'outils d'étude du lien entre deux variables. C'est en 1829 que Guerry, avec l'aide du géographe vénitien Adriano Balbi (1782–1848), produisit les trois cartes choroplèthes de France intitulées *Statistique comparée de l'état de l'instruction et du nombre des crimes dans les divers Arrondissements des Académies et des Cours Royales de France*.

Ces cartes données en Figure 6 représentent les départements de France coloriés selon le nombre de crimes contre les personnes pour la première, selon les atteintes à la propriété pour la seconde et enfin selon l'instruction pour la dernière. Dans cette figure, on retrouve une version revisitée en R qui enrichit ces cartes d'une échelle de tons de couleurs où figure une information numérique manquante dans les cartes originales. Guerry produisit au cours de sa vie de nombreuses cartes de France dont, entre autre, celles des taux de suicides, des vols et des taux d'alphabétisation. Toutes ces cartes sont reproduisibles à l'aide du jeu de données *Guerry*.

Le 2 juillet 1832, l'avocat André-Michel Guerry, âgé de 29 ans, présenta à l'Académie Française des Sciences un mémoire (voir Guerry, 1933) intitulé « *Essai sur la statistique morale de la France* ». Historiquement, des travaux analogues ont été menés parallèlement par le mathématicien et astronome belge Adolphe Quetelet (1796–1874). Ce dernier est considéré comme initiateur des nouvelles méthodes quantitatives en sciences humaines et sociales. On pourra citer sa célèbre théorie de « l'homme moyen » (voir Quételet, 1835). Les données *ChestSizes* à ce sujet sont disponibles dans le package *HistData* (voir Annexe 1).

Les travaux de Guerry ont permis d'apporter une contribution majeure à ce que l'on appellera la « statistique morale ». Ce travail mènera à de nouvelles perspectives pour la criminologie, la sociologie et plus largement pour les sciences sociales. Il tentait d'apporter des réponses cartographiques aux questions sociales de l'époque telle que : *Est-ce que le niveau d'instruction et de criminalité sont liés ?* En cherchant des relations entre variables sociales et morales telles que le taux de criminalité et la richesse, Guerry souhaitait montrer l'existence de lois semblables aux lois physiques régissant la vie sociale. Il en conclura par exemple que c'est plutôt la tentation que le besoin qui régit le vol. Guerry se consacra tant à cette recherche de relation qu'il inventa l'ordonnateur statistique, une machine établissant des rapports de concordance entre des variables statistiques. Ce sont les balbutiements de la notion de corrélation.

Certains historiens des sciences mettent en relief la modestie de Guerry ainsi que son désintérêt pour la publication de ses travaux. Parmi les nombreux travaux peu connus, l'un des plus importants est sans doute celui concernant les diagrammes circulaires de Playfair enrichi d'une dimension temporelle (donnés à droite en Figure 7). Cette idée sera reprise et améliorée par Florence Nightingale. Ceci fera l'objet de notre quatrième exemple.

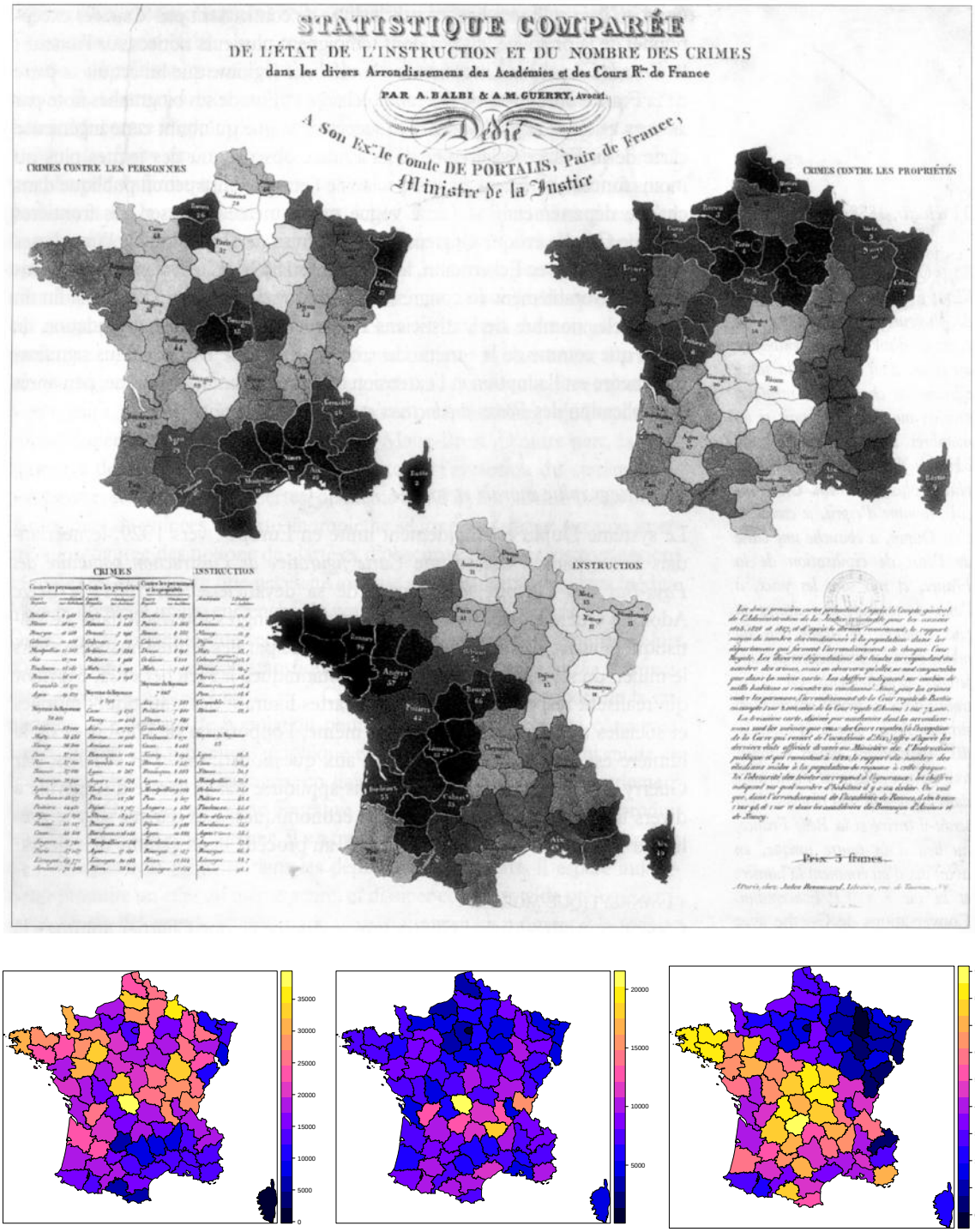


FIGURE 6 – *En haut : Statistique comparée de l'état de l'instruction et du nombre des crimes dans les divers Arrondissements des Académies et des Cours Royales de France de A. Balbi et A.-M. Guerry. Source : Guerry (1833). En bas : les cartes de France (à gauche crimes contre les personnes, au milieu crimes contre la propriété, à droite l'instruction) revisitées en R.*

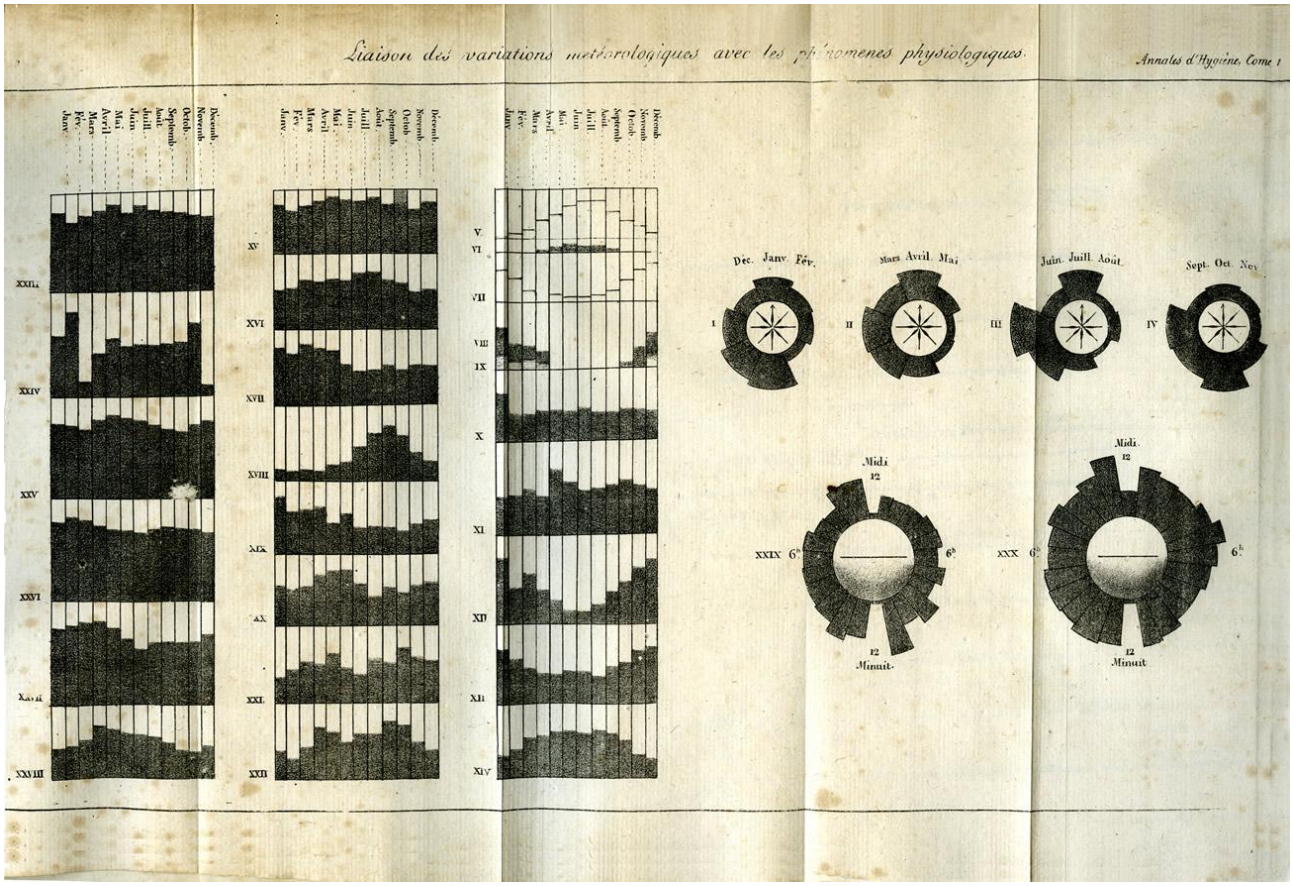


FIGURE 7 – *Graphique original de Guerry : Liaison des variations météorologiques avec les phénomènes physiologiques. Source : Guerry (1829).*

2.4 Les roses de Florence Nightingale

Dans ce quatrième cas, nous nous intéressons à la célèbre infirmière britannique Florence Nightingale (1820–1910), une des pionnières de l'utilisation des statistiques dans le domaine de la santé et plus particulièrement dans la représentation visuelle de l'information. Suite à la guerre de Crimée (1853–1856), elle décida d'illustrer les causes saisonnières de mortalité des patients de l'hôpital militaire dont elle s'occupait. Pour cela, elle utilisa une version améliorée des diagrammes circulaires développés par Playfair en 1801 (voir Figure 3) et enrichis par Guerry en 1833 (voir Figure 7) où sa remarquable contribution fut d'exploiter la dimension temporelle.

Dans la Figure 8, on retrouve les deux graphiques de Nightingale, publiés en 1857, également appelés *Diagram of the causes of mortality in the army in the East* (voir Nightingale, 1857). Chronologiquement, le premier étant celui de droite, il s'étale d'avril 1854 à mars 1855. En gris on retrouve la proportion de morts par maladie, en rose la proportion de morts par blessures de guerre et enfin, en noir, la proportion de morts d'autres causes. A gauche, le second graphique est dans la continuation du premier et décrit les proportions de morts d'avril 1855 à mars 1856. Il manque dans les diagrammes de Nigh-

tingale une échelle absolue afin d'avoir une idée du nombre de morts pour chaque cause. La version revisitée des diagrammes donnée en Figure 8 pallie ce manque (voir jeu de données *Nightingale*).

Ces diagrammes 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 anglais de réaliser l'ampleur du désastre et menèrent à une réforme médicale. Les dirigeants concernés n'auraient probablement pas pu comprendre des rapports statistiques traditionnels.

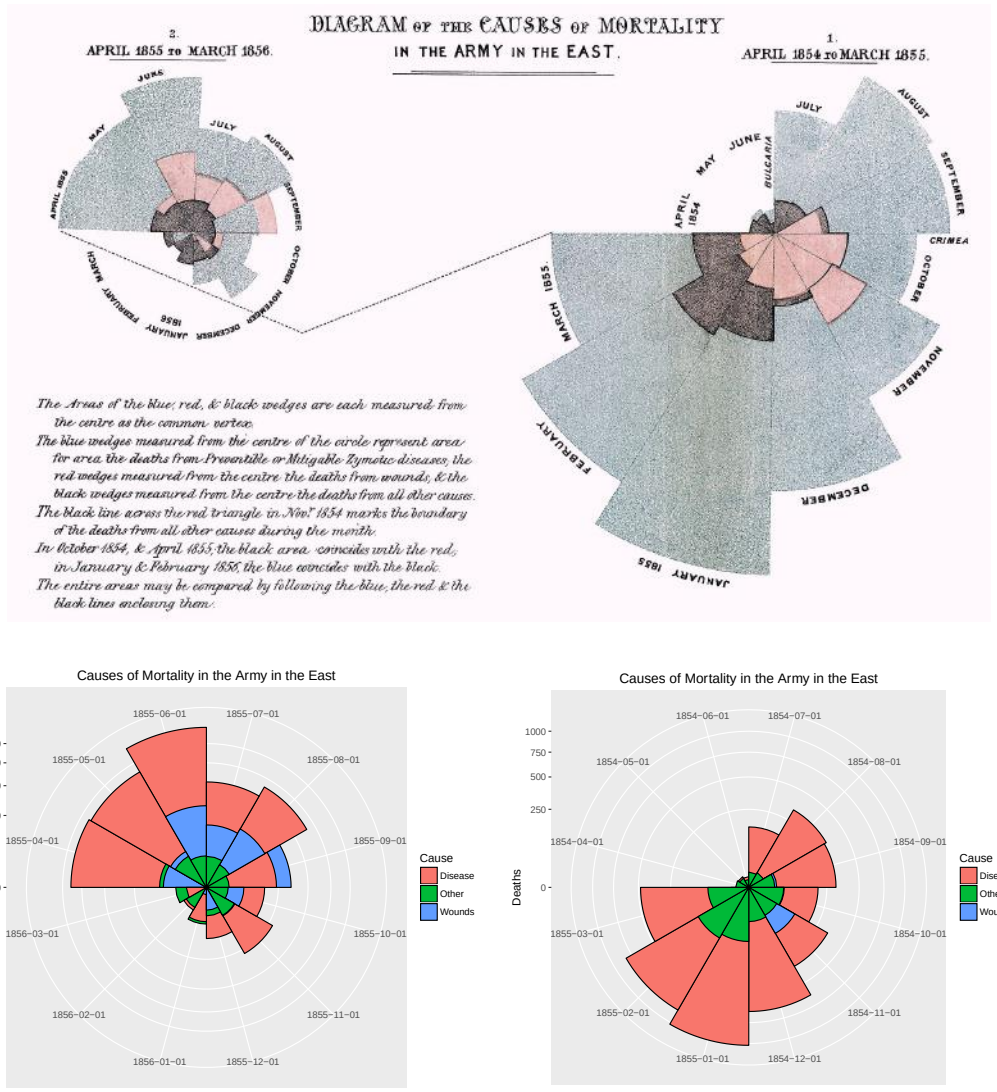


FIGURE 8 – *En haut : Diagram of the causes of mortality in the army in the East. Source : Nightingale (1857). En bas : les diagrammes revisités en R.*

2.5 La carte de John Snow ou les débuts de l'épidémiologie

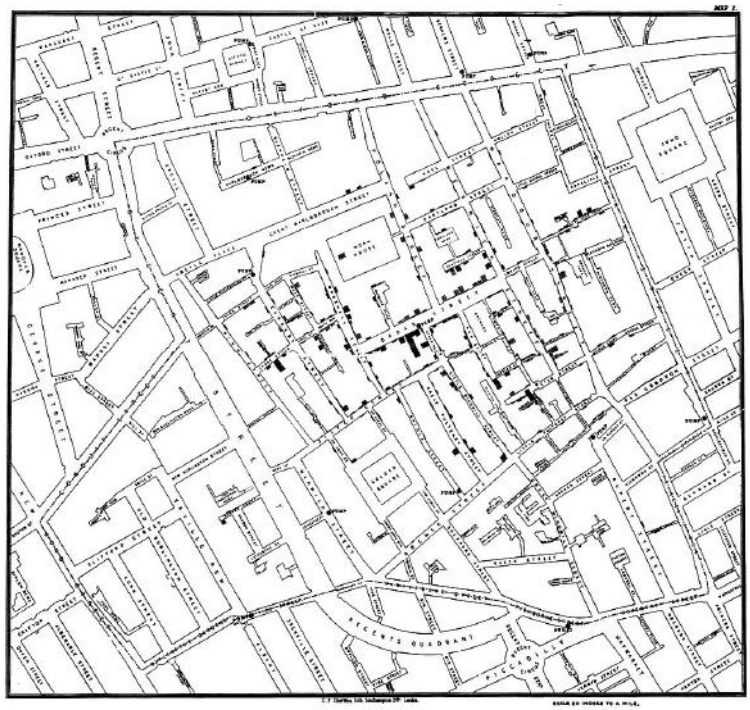
Au XIX^{e} siècle, la théorie dominante était que le choléra se propageait dans l'air. John Snow (1813–1858), un médecin britannique, pionnier dans les domaines de l'anesthésie, de l'hygiène et de la santé publique, était sceptique quant à la théorie des miasmes pour expliquer les épidémies de choléra. Suite à des expériences cliniques, il émit l'hypothèse que le choléra devait se développer à la suite de l'ingestion et non plus de l'inhalation, suspectant alors que l'eau jouait un rôle dans sa propagation. En 1849, il publia son essai : « *On the mode of communication of cholera* » (voir Snow, 1855) dans lequel il défend son hypothèse. Ces premiers écrits rencontrent le scepticisme de ses contemporains car, à l'époque, la théorie microbienne n'était pas bien établie.

Dans les années 1850, Soho, un quartier de Londres, a de sérieux problèmes d'insalubrité dus à une augmentation massive de sa population ainsi qu'à l'absence d'installations sanitaires appropriées. Les habitations ne bénéficiaient pas de distribution d'eau potable, les résidents étaient alors obligés de s'approvisionner à des pompes manuelles, localisées dans différentes rues. Suite à une saturation des installations de traitement des ordures, les dirigeants de la ville prirent la décision de jeter ces dernières dans la Tamise. Cela contamina les réserves d'eau et déclencha une épidémie de choléra. Cette épidémie majeure frappa le quartier Soho le 31 août 1854. Les trois jours suivants, 127 personnes habitant Broad Street ou les environs moururent. Les trois quarts des habitants fuirent le quartier dans les semaines suivantes. Le 10 septembre, en l'espace de 10 jours, 500 personnes étaient mortes. En tout, l'épidémie fit 616 morts.

C'est lors de cet événement que Snow eut l'idée révolutionnaire de noter sur un plan les lieux de résidence et de travail de 578 victimes grâce à des barres noires perpendiculaires aux rues (voir Figure 9). Pour cela, il interrogea les résidents du quartier, établit la répartition géographique des cas de choléra et représenta sur sa carte la localisation des points d'eau de la ville. En regardant son plan, il s'aperçut que le nombre de décès augmentait au fur et à mesure que l'on se rapprochait d'une pompe à eau publique située dans Broad Street, identifiant ainsi la source de l'épidémie. Il deviendra célèbre, en 1855, en publiant un compte rendu détaillé de l'épidémie de 1854 (voir Snow, 1855) dont sa fameuse carte donnée en Figure 9. Si on regarde celle-ci avec attention, on peut y déceler une anomalie ! En effet, juste à côté de la célèbre pompe, on trouve une brasserie où l'on ne dénombre aucune victime. Cela est probablement dû au fait que les travailleurs avaient droit tous les jours à une ration de bière gratuite. Par conséquent ils ne buvaient pas l'eau de la pompe ; de plus, le processus de fermentation tue les bactéries de choléra.

On peut recréer en R la carte de John Snow à l'aide des jeux de données *Snow* et *SnowMap* (voir Figure 9). Si John Snow n'a pas découvert le germe causal du choléra, il a su montrer sa transmission par l'eau. Cela constitue un événement majeur de l'histoire de la santé publique, qui peut être considéré comme l'acte fondateur de l'épidémiologie. Aujourd'hui encore, cette technique de cartographie statistique est utilisée afin d'identifier le point d'origine de certaines épidémies comme celle d'ebola en Afrique ou encore du choléra à Haïti.

Dans notre dernier cas d'étude, on s'intéressera à la célèbre carte figurative de Charles Joseph Minard.



Snow's Cholera Map of London (sp)

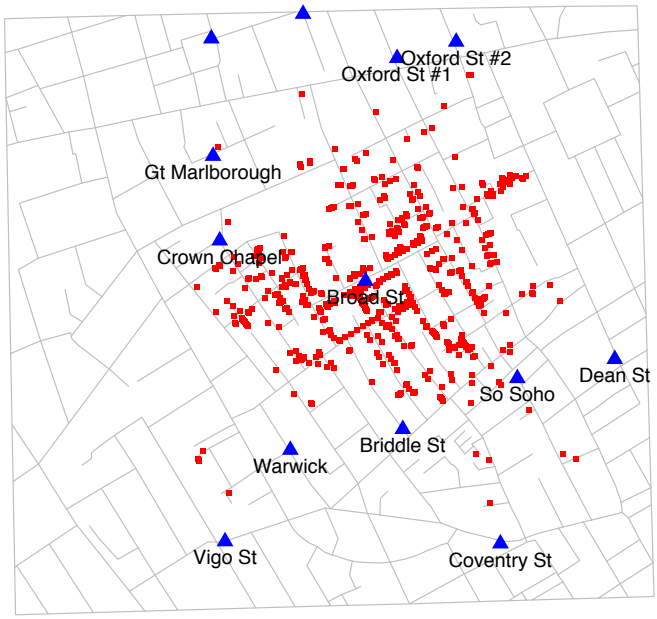


Figure 9 – *En haut : carte originale de John Snow sur l'épidémie de 1854 dans le quartier londonien de Soho. Source : Snow (1855). En bas : la carte revisitée en R.*

2.6 Un des chefs d'œuvre de Charles Minard

Charles Joseph Minard (1781–1870) était un polytechnicien et inspecteur des ponts et chaussées français qui a consacré une bonne partie de ses activités aux représentations graphiques appliquées au génie civil et aux statistiques. Son intérêt pour la synthétisation graphique des données s'est poursuivie même après son départ à la retraite et c'est en 1869, à l'âge de 88 ans, qu'il publia sa fameuse *Carte figurative des pertes successives en hommes de l'Armée française dans la campagne de Russie en 1812–1813*. Incontestablement celle-ci est considérée comme le chef d'œuvre de la représentation graphique. Tufte la qualifie de « meilleur graphique statistique jamais tracé » (voir Tufte, 1997).

Cette carte en deux dimensions donnée en Figure 10 intègre et synthétise pas moins de six niveaux d'informations. Elle relate la chronologie et met en exergue les événements de la campagne de Russie de Napoléon Bonaparte (1769–1821). Elle donne la localisation (principales villes et rivières), l'itinéraire de l'armée et indique les points de séparation et de regroupement des troupes. Au bas de ce graphique on peut lire parallèlement au mouvement des troupes une série chronologique des températures mesurées dans l'échelle de Réaumur faisant figurer des dates importantes. On peut ainsi suivre l'évolution des effectifs de l'armée et des pertes humaines représentées par l'épaisseur de la bande (beige ou noire) qui donne l'évolution des effectifs et leur répartition selon les lieux géographiques (1 mm pour 6000 hommes).

En résumé, de gauche à droite on peut suivre sur la bande beige l'évolution des effectifs de l'armée française du Premier Empire. Des 422.000 hommes partis de Kowno seuls 100.000 hommes environ arriveront à Moscou sans compter les 100.000 autres hommes qui se séparèrent du reste de l'armée au niveau de Wilna (60.000 partis en direction de Polotrk et 40.000 au Nord). Par la suite, on peut lire de droite à gauche sur la bande noire la retraite des troupes. On voit au niveau de Botr un regroupement des troupes de l'armée en retraite (20.000 hommes) avec ceux qui les quittèrent à Wilna (30.000 hommes). Il s'ensuit la traversée tragique de la rivière Bérézina par cette armée de 50.000 hommes faisant 22.000 morts. Seulement 4.000 hommes atteindront le point de départ et seront rejoint par 6.000 autres revenant du Nord. L'empereur regagnera la France avec 10.000 soldats uniquement. On peut comprendre aisément, rien qu'en regardant l'épaisseur du trait de Minard, pourquoi la rivière Bérézina est entrée dans l'histoire.

Une version revisitée en R est donnée au bas de la Figure 10 (voir le site[3] ainsi que le jeu de données *Minard*). Cette carte inspirera de nombreuses versions revisitées et même un concours de data visualisation : « Re-Visioning Minard Contest »[4]. En 1869, Minard illustrera de la même manière la deuxième guerre punique d'Hannibal Barca (247 av. J.-C. – 183 av. J.-C.) et plus particulièrement son voyage vers l'Italie ainsi que sa fameuse traversée des Alpes (voir Figure 10). Il produira au cours de sa vie plus d'une cinquantaine de cartes sur divers sujets (voir le site[5] et Minard, 1862).

[3]http ://www.yvesago.net/pourquoi/2014/03/r-la-campagne-de-russie-par-minard.html
[4]http ://www.datavis.ca/gallery/re-minard.php
[5]http ://visionscarto.net/charles-joseph-minard-cinquante-cartes

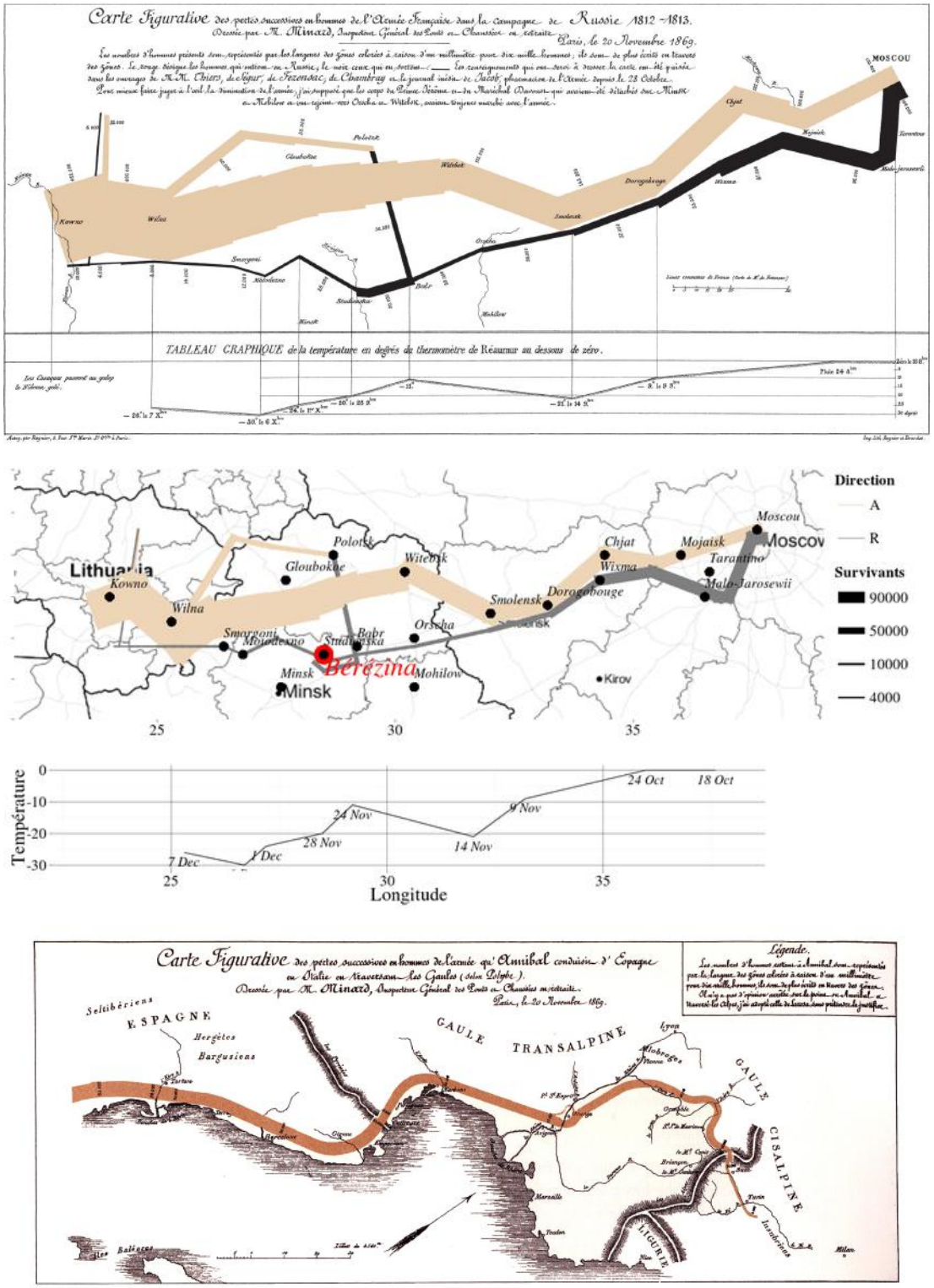


FIGURE 10 – *En haut : Carte figurative des pertes successives en hommes de l'armée française dans la campagne de Russie en 1812–1813. Au milieu : la carte revisitée en R. En bas : Carte figurative des pertes successives en hommes de l'armée qu'Hannibal conduisit d'Espagne en Italie en traversant les Gaules (selon Polybe). Source : Tufte (1997).*

3 Conclusions et perspectives

De nos jours, la data visualisation ne cesse de se développer et de se renouveler. L'apport d'une grande puissance de traitement numérique permet d'ajouter une dimension dynamique et interactive aux nouvelles représentations graphiques et enrichit la sémiologie graphique. On assiste même à l'émergence de métiers spécifiques tels que le journalisme de données alliant l'infographie à la visualisation des données (voir à ce sujet l'excellent article du journal : *The New York Times*[6]). Parmi les grands noms de la data visualisation actuelle on peut citer Hans Rosling (1948–2017), médecin, statisticien et conférencier suédois. Ce dernier est connu pour son travail sur la visualisation de données interactive ; il a notamment développé le logiciel Trendalyzer qui convertit les statistiques internationales en graphiques animés et interactifs (voir à ce sujet sa célèbre conférence TED[7]).

En conclusion, il nous semble primordial de contextualiser la naissance et l'usage des méthodes statistiques. Le package *HistData* de R nous offre cette possibilité. Les données disponibles abordent une très grande diversité de thèmes et de disciplines : sociologie, commerce, psychologie, médecine, militaire, épidémiologie, cartographie, biologie, physique, sciences sociales et criminologie. Il est à nos yeux essentiel de faire ce travail historique. Le choix des six exemples présentés dans cet article a été essentiellement motivé par leur importance historique, fondatrice et pédagogique. Cependant, de nombreux autres cas tout aussi intéressants ont vu le jour et ont contribué au développement de la statistique par des innovations graphiques et même parfois conceptuelles. Les principaux jeux de données du package *HistData* sont répertoriés dans l'Annexe 1.

Dans cette perspective, le département STID a décidé d'ouvrir deux cours d'histoire de la visualisation de données s'adressant aux étudiants en formation initiale et aux étudiants du DU DataViz. Ces cours consistent en une introduction d'une durée d'une heure et demie afin de leur présenter les motivations historiques de la discipline. Par la suite, ils auront des travaux pratiques et participeront à un concours de data visualisation[8]. Avec les différents intervenants, nous sommes en train de penser les choses de manière globale, afin que les étudiants puissent aborder la statistique sous un nouvel angle pédagogique. Ils pourraient par exemple dans un premier temps travailler sur des données historiques et cela sans connaitre le graphique historique correspondant. Cela leur permettrait de mettre en œuvre des techniques statistiques vues en cours et par la suite développer leurs propres outils de visualisation de données, pour enfin comparer leurs résultats avec les graphiques historiques.

Nous nous donnons comme perspective de continuer à insérer un maximum d'exemples historiques dans nos cours ou tout du moins de chercher à les contextualiser le plus possible. Nous invitons nos confrères à en faire de même en espérant que cet article en inspirera certains.

[6]https ://www.nytimes.com/interactive/2018/03/19/upshot/race-class-white-and-black-men.html
[7]https ://www.ted.com/talks/
[8]http ://www.stid-paris.fr/concours-dataviz-2016

Références

[1] Dupin, C. (1826), *Carte figurative de l'instruction populaire de la France*, Jobard : BNF : Ge C 6588.

[2] Droesbeke, J.-J. et Ph. Tassi (1997), *Histoire de la statistique*, Presses Universitaires de France - PUF.

[3] Friendly, M., P. Valero-Mora, and J. Ibáñez Ulargui (2010), The first (known) statistical graph : Michael Florent van Langren and the "Secret" of Longitude, *The American Statistician*, **64**(2), 174–184.

[4] Friendly, M. and J. D. Daniel (2011), Milestones in the history of thematic cartography, statistical graphics, and data visualization, *URL http ://www. datavis. ca/milestones*, 32.

[5] Friendly, M. (2007), A Brief History of Data Visualization, *Handbook of Computational Statistics : Data Visualization*, Springer–Verlag, **Vol. III**, Ch. 1, 1–34.

[6] Galton, F. (1886), Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute*, 15, 246–263.

[7] Guerry, A.-M. (1829), Mémoire sur les variations météorologiques comparées aux phénomènes physiologiques, *Annales d'Hygiène Publique et de Médecine Légale*, **1**, 228.

[8] Guerry, A.-M. (1833), *Essai Sur La Statistique Morale de la France*, Paris : Crochard.

[9] Minard, C. J. (1862), *Tableaux graphiques et cartes figuratives*, Bibliothèque numérique patrimoniale des ponts et chaussées.

[10] Nightingale, F. (1857), *Mortality of the British Army*, London : Harrison and Sons.

[11] Pearson, K. and A. Lee (1903), On the laws of inheritance in man, *Biometrika*, 2, 357–463, Table 31.

[12] Playfair, W. (1821), *Letter on our agricultural distresses, their causes and remedies ; accompanied with tables and copperplate charts shewing and comparing the prices of wheat, bread and labour, from 1565–1821*, London : W. Sams.

[13] Playfair, W. (2005), *Playfair's commercial and political atlas and statistical breviary*, Cambridge University Press.

[14] Quetelet, A. (1835), *Sur l'homme et le développement de ses facultés, essai d'une physique sociale*, Paris : Bachelier.

[15] Snow, J. (1855), *On the Mode of Communication of Cholera.*, London : (n.p), 2nd edition.

[16] Spence, I. (2006), *William Playfair and the psychology of graphs.*, In : 2006 JSM proceedings, American Statistical Association, Alexandria, pp 2426–2436.

[17] Tufte, E. R. (1997), *The Visual Display of Quantitative Information*, Chesshire, CT : Graphics Press.

Annexe 1

Le tableau suivant donne les principaux jeux de données du package *HistData* en précisant également le contexte relatif au jeu de données ainsi que les figures historiques mises en jeu. Pour plus de détails, se référer au document[9] R spécifique.

Tableau 1 – *Tableau résumant les principaux jeux de données du package HistData*

Package HistData			
Figures historiques	Thème(s) et/ou Contexte	Jeu de données	Siècle
John Arbuthnot	Premier test statistique de l'histoire	*Arbuthnot*	18e
Arthur Bowley	Exportations anglaises et irlandaises	*Bowley*	19e
Henry Cavendish	Estimation de la densité moyenne de la Terre et de la constante de gravitation	*Cavendish*	18e
Adolphe Quetelet	Théorie de l'homme moyen, loi normale	*ChestSizes*	19e
William Farr	Epidémie de choléra	*Cholera*	19e
Cushny and Peebles	Tests statistiques et médicaments	*CushnyPeebles*	20e
Edgeworth	Première anova à 2 facteurs de l'histoire	*Dactyl*	19e
Karl Pearson	Alcoolisme et hérédité	*DrinksWages*	20e
Waite et Karl Pearson	Tests d'indépendance et empreintes	*Fingerprints*	20e
Sir Francis Galton	Les débuts de l'eugénisme, corrélation, régression, loi normale bivariée	*Galton* *GaltonFamilies*	19e 19e
André-Michel Guerry	Cartes de France choroplèthes, sociologie, variables « morales »	*Guerry*	19e
Edmond Halley	Tables d'espérance de vie et assurance	*HalleyLifeTable*	17e
William Jevons	Psychologie cognitive et philosophie	*Jevons*	19e
Michael van Langren	Premier graphique statistique sur la longitude entre Tolède et Rome	*Langren*	17e
W. R. Macdonell	Anthropométrie de criminels	*Macdonell*	20e
Albert Michelson	Estimation de la vitesse de la lumière	*Michelson*	19e
Charles Minard	Carte figurative de la campagne de Russie de Napoléon	*Minard*	19e
Forence Nightingale	Diagrammes relatifs à la guerre de Crimée	*Nightingale*	19e
John Snow	Naissance de l'épidémiologie et carte de l'épidémie de choléra à Londres	*Snow* *SnowMap*	19e 19e
J. F. W. Herschel	Premier nuage de points de l'histoire	*Virginis*	19e
William Playfair	Graphique temporel	*Wheat*	19e
William Gosset « Student »	Cellules et erreurs d'échantillonnage	*Yeast*	20e
Charles Darwin Ronald Aylmer Fisher	Données sur des plantes reprises pour de nombreux tests statistiques	*ZeaMays*	19e

[9]https ://cran.r-project.org/web/packages/HistData/HistData.pdf