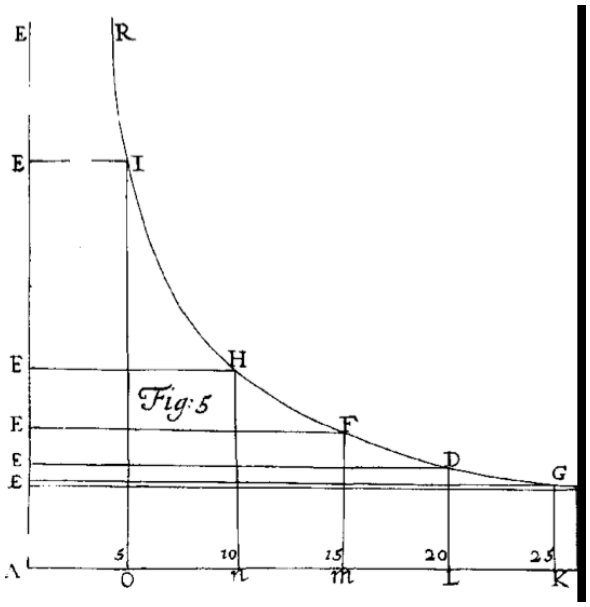
Representações Gráficas na Formação da Intuição, na Análise dos Dados e na Comunicação das Ideias

Dinis Pestana

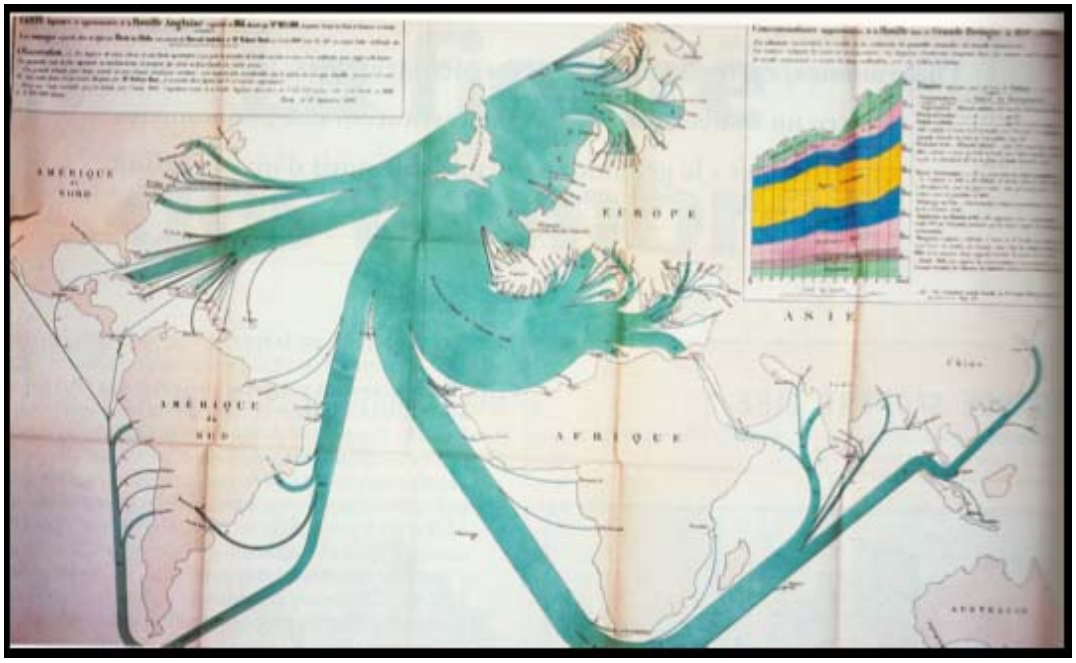
Universidade de Lisboa, FCUL, DEIO
CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa

Na capa das revistas publicadas pela *Royal Statistical Society* figura um molho de espigas e uma insígnia em latim que se pode traduzir "moer bem os dados". Analisar os dados é, afinal, o objectivo da Estatística. Tal como moer o grão o transforma em farinha, "moer" os dados transforma informação em conhecimento, se os dados forem bem moídos — e certamente aqui "bem" deve ser tomado na acepção puramente qualitativa, sem abusos, pois todos conhecemos o conselho "se torturares os dados, eles confessam" e as suas consequências.

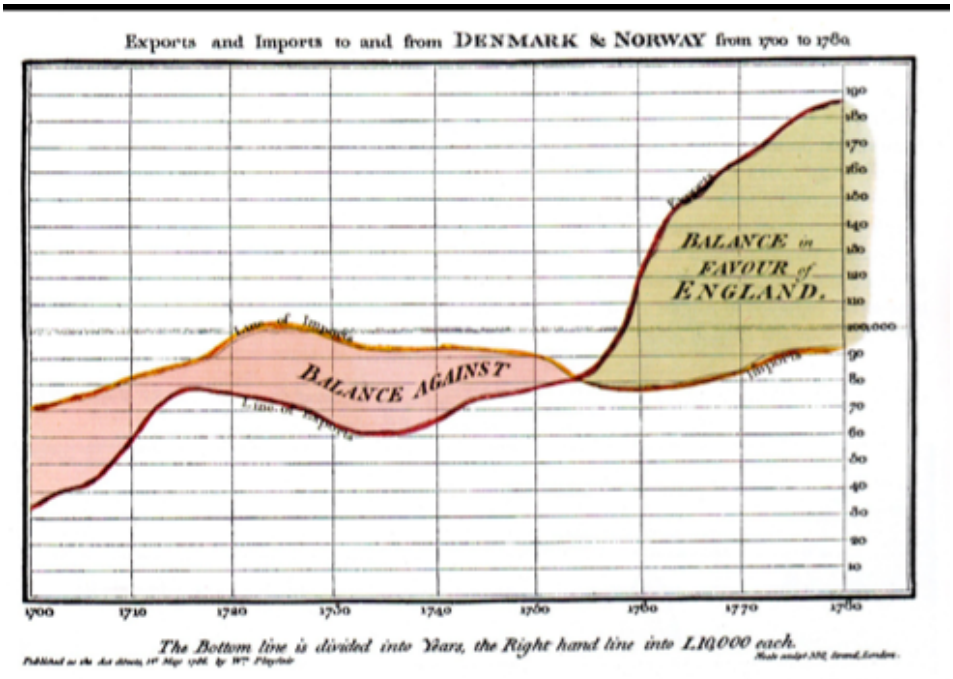
Bem cedo se tornou evidente que, neste moer dos dados, a análise gráfica tem um papel de relevo. Por exemplo, Halley (1686, 1693), quando se interessou por problemas demográficos (a *Royal Society* viria por isso a encomendar-lhe uma análise da demografia da população inglesa, que originou as modernas tabelas de mortalidade, uma outra forma de resumir os dados e evidenciar padrões "inventou" gráficos para se libertar dos acidentes da amostra e encontrar o modelo da população, como o que abaixo se reprroduz e inspirou muitos gráficos posteriormente usados em análise de sobrevivência.



A Alquimia da Estatística chamou-se Geografia Política, e na idade do ouro das represenntações gráficas muitos dos mais notáveis gráficos estavam, directa ou indirectamente, associados a mapas. Reproduzimos um mapa de Minard mostrando a importância da exportação da hulha no comércio externo do Reino Unido em meados do século XIX



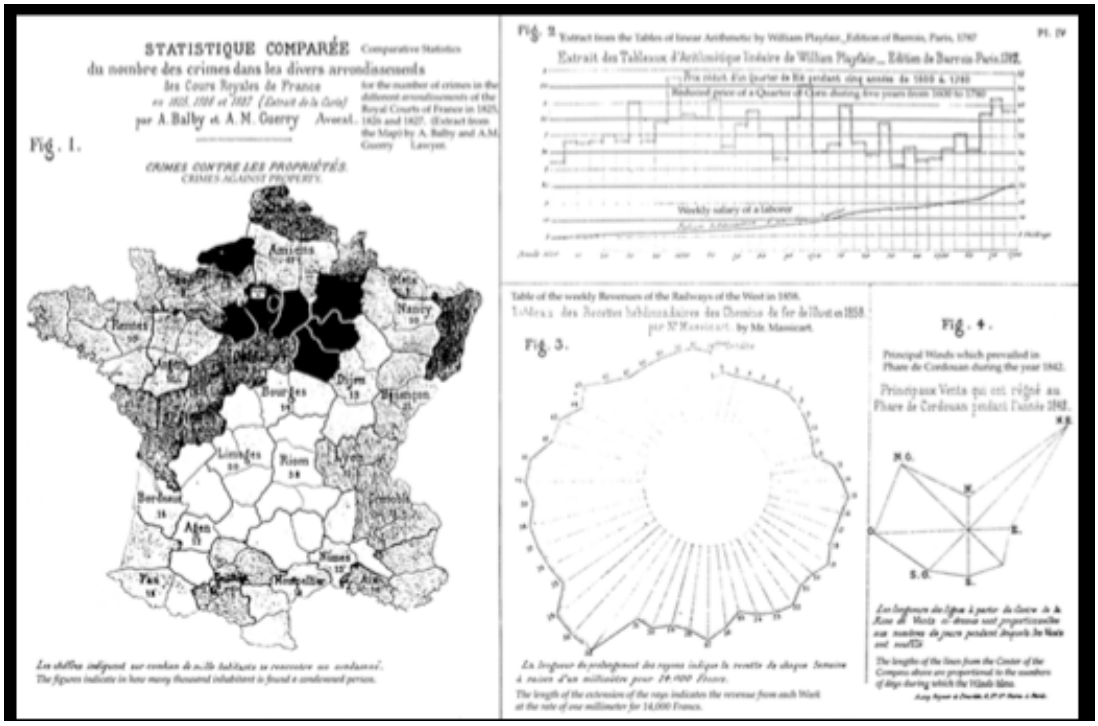
ou a representação ainda mais simples de Playfair,



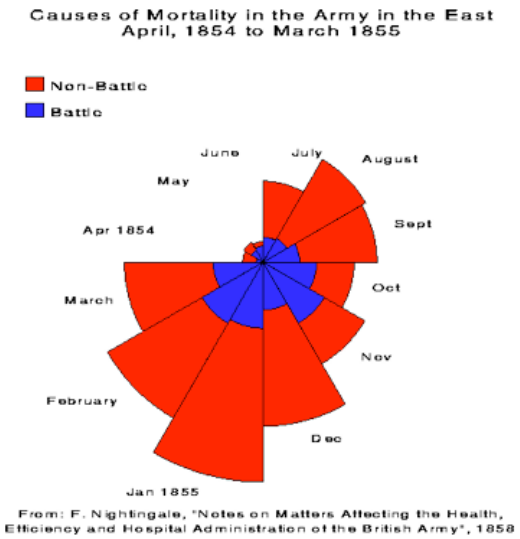
mostrando que, *circa* 1755, a balança comercial com a Noruega e a Dinamarca se torna favorável a Inglaterra, e isto sobretudo devido a um grande incremento nas exportações.

Uma página de Balbi — que também publicou uma notável monografia sobre Portugal, de que existe na Universidade dos Açores um exemplar de Mouzinho da

Silveira, atestando a influência que esse estudo estatístico teve sobre o nosso estadista (e consulte-se também a obra de Maria Leonor Machado de Sousa sobre Solano Constâncio, onde se analisa a importância da obra de Balbi na formação do pensamento do nosso constituinte que, no auto-in fligido exílio, se dedicou também ele a analisar os problemas portugueses com base em dados) — que abaixo se reproduz mostra uma excelente utilização de mapas, gráficos de radar, diagramas para análise de sucessões temporais, nomeadamente da sua tendência e sazonalidade.



Os gráficos de radar são aliás os antepassados dos *coxcomb* de Florence Nightingale (1857), que tanta importância tiveram na alteração das condições sanitárias do exército britânico.



Não faria sentido repetir aqui uma história da evolução das ideias sobre representações gráficas, uma vez que facilmente se acede a excelentes documentos[1] (recomenda-se também o excelentemente ilustrado artigo de Friendly, 2008).
No que mais propriamente se refere a representações gráficas em Estatística, com uma grande dose de simplificação podemos atribuir a Playfair, no século XIX, e a Tukey, no século XX, a invenção e popularização dos gráficos mais usados na descrição e exploração de dados, tornando-os um veículo previlegiado de comunicação da informação contida nos dados. Tukey preocupou-se ainda com a manutenção da integridade da informação, advogando métodos semi-gráficos, em que a informação numérica é mantida a par da (ou *na*) informação gráfica. Por exemplo, defendeu que, se cada dado devia ser representado por um símbolo, então se usasse um dígito, que contém decerto mais informação do que um ponto.
Por exemplo, num estudo sobre sobrevivência de indivíduos com cancro maligno da próstata, planeou-se uma experiência em que os de um grupo testemunha são tratados com um placebo, e os de um grupo experimental são tratados com um medicamento ainda em desenvolvimento, com efeitos secundários sobre o fígado que leva a que alguns dos doentes "morram da cura". Há, *post-mortem*, uma separação dos que morreram devido ao cancro e dos que morreram devido a falha da função hepática. Regista-se como variável resposta o tempo de sobrevivência (em semanas) após diagnóstico e admissão no protocolo experimental.
Os dados obtidos neste estudo experimental não fazem sentido imediato, pois têm o aspecto aparentemente caótico que é usual nos dados brutos.

Grupo de controle: (3.0, 13.9, 1.6, 8.7, 2.0, 8.0, 13.7, 2.7, 1.2, 10.4, 2.1, 27.6, 2.4, 41.4, 16.1, 5.4, 15.5, 3.6, 3.7, 25.2, 6.6, 13.6, 4.8, 20.1, 11.0, 4.4, 20.3, 1.5, 22.0, 4.6, 5.3, 7.2, 16.2, 1.3, 8.6, 0.6, 1.8, 0.7, 5.6, 4.0, 9.6, 13.8, 5.2, 3.2, 6.5, 5.3, 6.3, 12.8, 21.2, 0.7, 15.9, 29.6, 8.8, 24.2, 4.2, 6.1, 10.6, 0.3, 6.6, 2.3);
Grupo experimental, morte devida ao cancro: (14.2, 51.9, 33.3, 27.1, 4.4, 1.2, 19.5, 15.5, 27.0, 3.0, 14.6, 1.4, 29.2, 4.3, 17.6, 13.0, 3.3, 9.0, 5.6, 3.3, 3.6, 19.3, 1.7, 20.5, 8.7, 3.4, 5.4, 1.9, 47.2, 1.6, 8.9, 15.1, 18.5, 7.9, 9.4, 15.4, 10.8, 7.1, 9.2, 8.6, 20.5, 6.0, 3.1, 71.3, 14.5, 44.1, 9.7, 18.0, 5.5, 10.8, 1.8, 11.6, 12.6, 30.0, 21.5, 31.0, 12.7, 10.2, 13.5, 15.4, 8.9, 6.8, 3.1, 10.1, 0.2, 6.8, 7.6, 12.5, 9.6, 22.9, 2.2, 5.5, 4.6, 2.4, 5.6, 20.2, 0.4, 15.5, 19.3, 41.2, 6.5, 15.0, 9.2, 2.2, 3.4, 59.2, 9.2, 33.9, 16.0, 1.5, 6.2, 3.3, 7.6, 11.0, 0.8, 6.9);
Grupo experimental, morte por falha da função hepática: (1.4, 4.8, 53.7, 69.1, 41.1, 26.7, 0.9, 10.6, 17.7, 18.3, 29.7, 36.5, 3.3, 22.4, 18.6, 4.3, 0.4, 9.4, 7.2, 6.7, 10.5, 19.2, 18.3, 4.3, 3.7, 0.4, 3.5, 7.6, 36.4, 4.8, 4.5, 5.7, 6.0, 21.1, 6.5, 2.2, 3.2, 20.6, 13.0, 15.5, 14.9, 9.7, 73.2, 28.7, 36.3)

Em certo sentido, pode dizer-se que a Estatística é a ciência que nos ensina a "ler" os dados, a decifrar a informação que não está ainda evidente na forma embrionária em que é obtida. Uma tabulação ordenada dos dados é decerto mais legível, ou um histograma, como os que Playfair popularizou. Mas Tukey propôs uma algoritmo de ordenação dos dados que em simultâneo os representa semigraficamente, e que não só revela os padrões subjacentes como mantém a integridade da informação.

[1] Por exemplo file:///Users/macbookpro/Desktop/Gallery%20of%20Data%20Visualization %20-%20Historical%20Milestones.webarchive

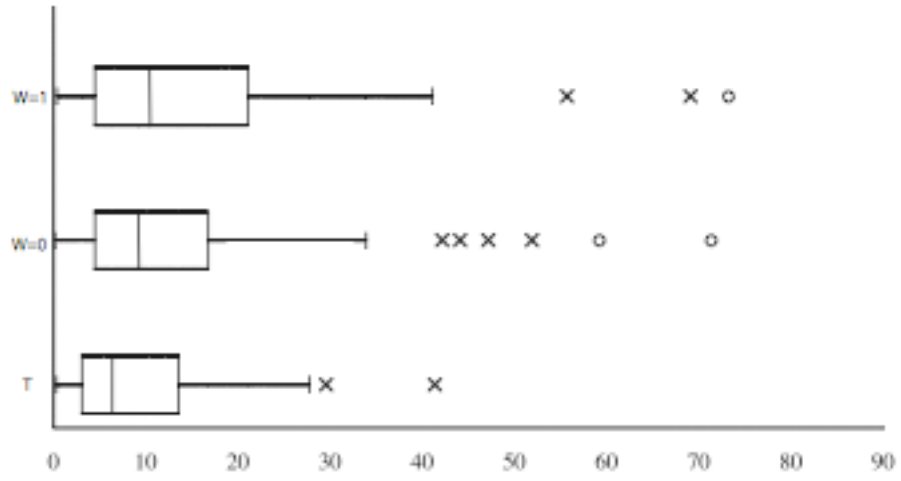
Sem mais comentários, que decerto seriam desnecessários, apresenta-se o diagrama de caule-e-folhas de Tukey, comparando de forma imediata a sobrevivência de doentes não tratados e tratados.

	$60; \frac{1}{2} \times 10^1$		$141; \frac{1}{2} \times 10^1$	
23	44444333322222111110000	0	000000111111112222333333333333344444444	39
(16)	9888876666655555	0∗	55555566666666677777788889999999999	(34)
21	33332100	1	000000112223334444	68
13	6655	1∗	5555555677888889999	50
9	42100	2	00001122	31
4	975	2∗	677899	23
		3	0133	17
		3∗	666	13
1	(1	4	114)	10
		4∗	7)	7
		5	13)	6
		5∗	9)	4
		6		
		6∗	9)	3
		7	13)	2

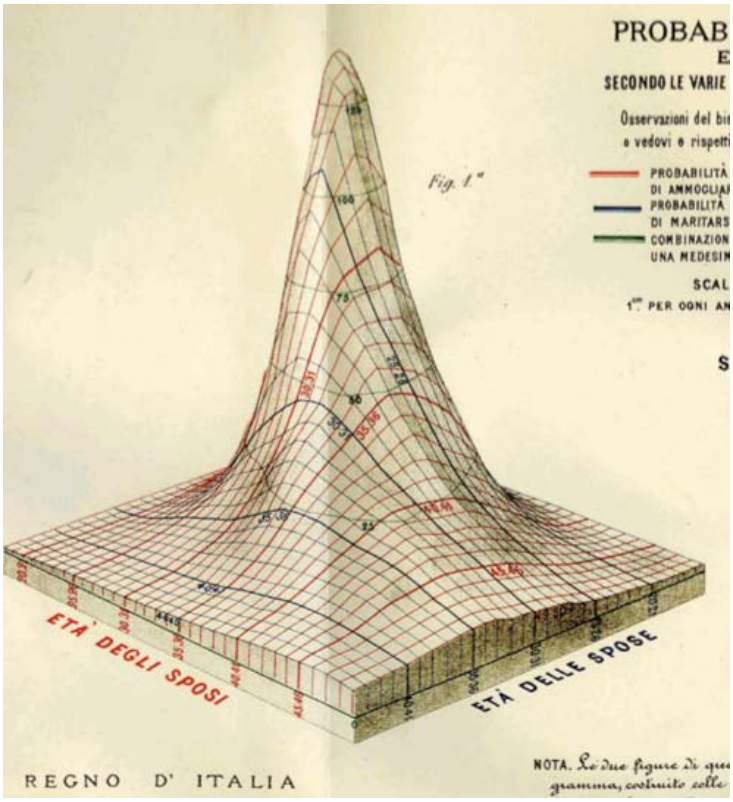
Este diagrama não só torna evidente que há um aumento do tempo de sobrevivência dos tratados, como, para qualquer utilizador de Estatística com alguma experiência, sugere uma modelação exponencial (que evidentemente terá que ser formalmente investigada com um teste de ajustamento adequado) que, a verificar-se, permitirá a estimação de parâmetros importantes (por exemplo médias populacionais) por métodos fidedignos como o da verosimilhança máxima.

Os diagramas de caule-e-folhas prestam-se ao tratamento de uma ou de duas amostras. No entanto, com mais do que duas amostras torna-se fastidioso compará-las duas a duas.

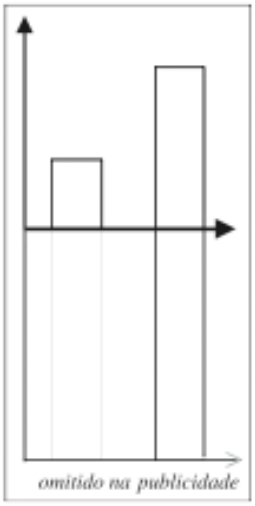
Um outro método de apresentação gráfica inventado por Tukey — as caixas-com-bigodes, a que Murteira (1993) deu o nome mais preciso de diagramas de extremos e quartis — permite uma apresentação simultânea de vários grupos. Por exemplo, neste estudo, as caixas-com-bigodes paralelas revelam de forma muito imediata que os doentes tratados têm maior sobrevivência, e que os que morrem de efeitos secundários do medicamento experimental sobre o fígado têm ainda, tendencialmente, uma maior sobrevivência.



Anote-se que, na época de ouro (Friendly, 2008) das representações gráficas, surgiram excepcionais representações multivariadas. A título de exemplo reproduzimos um dos “estereogramas” de Perozzo (1880).



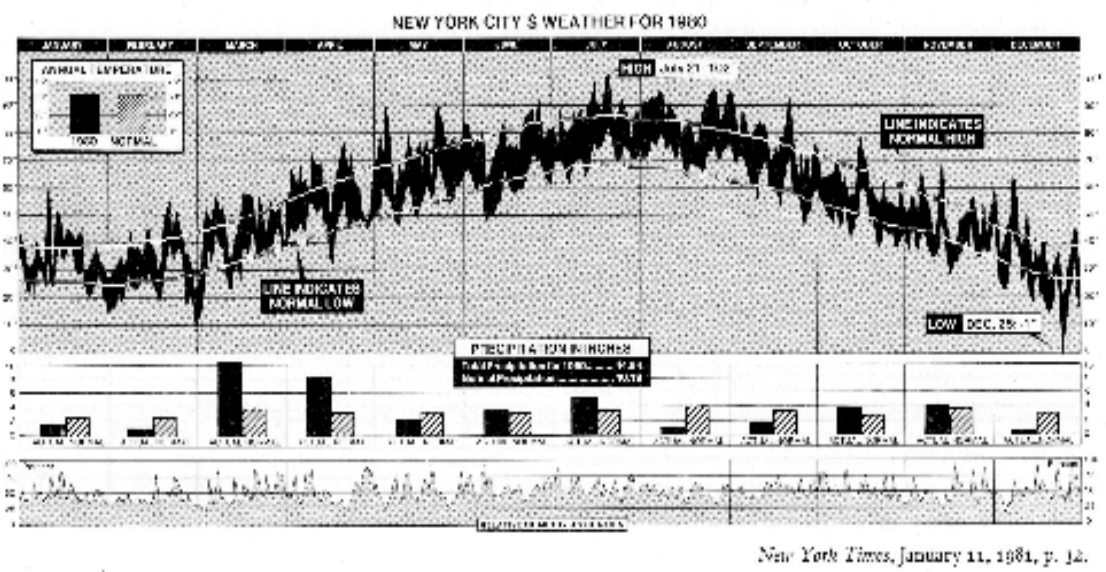
O exemplo da sobrevivência de doentes tratados e não tratados procura mostrar como os gráficos podem ser bem usados, quer para dar origem a novas ideias (por exemplo, que um modelo exponencial pode ser adequado), quer para divulgar informação.
Mas os gráficos servem também para deturpar informação. Os primeiros gráficos "errados" provavelmente foram apenas acidentes, mas depressa os profissionais de *marketing*, publicidade e criação de imagem se aperceberam do extraordinário potencial dos gráficos na manipulação da informação. Por exemplo, o famoso "hexaclorofene duplica a sua protecção contra a cárie" nada quer dizer se se apoiar num gráfico como o abaixo reproduzido, em que a ausência de escala permite todos os enganos.



Se os métodos gráficos têm uma longa tradição de popularidade, a facilidade de criação de gráficos em computador tornou-os uma praga. Já há uma vintena de anos um manual da especialidade estimava o número de gráficos criados anualmente em 1 a 3 milhões. Actualmente esse número decerto decuplicou — e provavelmente com um concomitante abaixamento da sua qualidade. A galeria de gráficos que o Excel permite facilita o trabalho de qualquer utilizador da Estatística, mas é também uma fonte generosa de disparates por parte dos abusadores que não sabem o que fazem, e não se preocupam com as judiciosas frases "antes de ligar o computador ligue o cérebro" e "quando o *input* é lixo, o *output* só pode ser lixo". Recordo por exemplo uma análise de dados económicos, todos eles rondando 50 milhões de euros, mas nitidamente com um padrão crescente até 2002, e decrescente depois dessa data. O analista representou esses dados mantendo, no eixo das ordenadas, os limites 0 e 100. Por isso lhe pareceram ter um padrão constante, sem perceber que estava apenas a usar uma escala imprópria para revelar o verdadeiro padrão (a preocupação, aliás, era o facto de a recta de regressão ser horizontal, o que o levou a tecer o mimoso comentário "esta recta não é linear"). Recordo também uma análise, de bastante melhor qualidade, mas em que se tinha usado sistematicamente uma regressão parabólica para estudar a influência da educação formal na progressão das carreiras, usando como *proxi* a taxa de aumentos salariais. Como a parábola tem um ápice, mesmo nos casos de estudos superiores depois dos 50 anos parecia não haver viagra capaz de endireitar aquelas curvas!

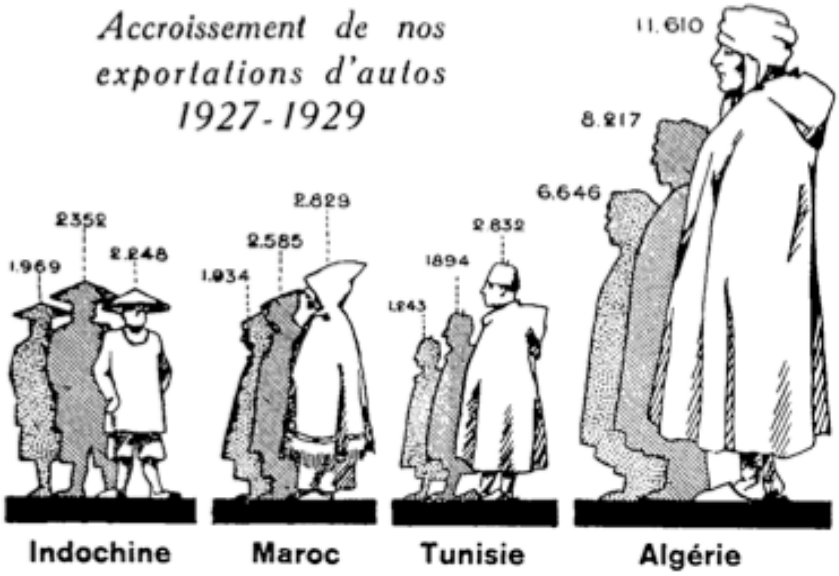
Software estatístico mais especializado permite a criação de gráficos ainda mais sofisticados e animação gráfica, nomeadamente de dados multidimensionais, que podem revelar facetas novas sobre a estrutura dados. Já há uma vintena de anos surgiu um programa *MacSpin*, que foi pioneiro em rodar os pontos no espaço tridimensional, e na colecção de livros da *Springer Verlag* sobre o cada vez mais popular ***R*** surgiu recentemente um *Lattice* totalmente dedicado às potencialidades gráficas daquele programa, que aparentemente se vai tornar o padrão, pelo menos nos meios universitários.

Tufte (1983) inspirou muito trabalho de excelente qualidade sobre gráficos, mas em livros subsequentes veio a dedicar muitas páginas ao abuso da Estatística com maus gráficos. Reproduz-se, de Tufte (1983), mais um excepcional exemplo de um gráficvo notável na veiculação de informação, neste caso sobre o clima em Nova Iorque.

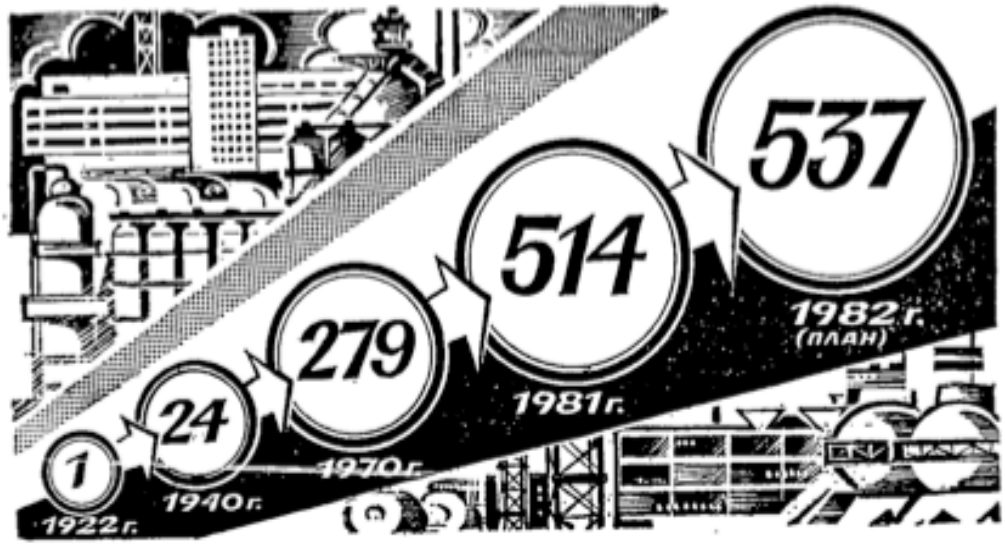


New York Times, January 11, 1981, p. 32.

Mas, por outro lado, recorda-se o seu discurso consistente contra os gráficos que não preservam a integridade da informação, ou a deturpam, e nomedamente os seus avisos sobre os muitos enganos decorrentes do uso de pictogramas. Por exemplo, o pictograma seguinte é enganador, porque a representação das quantidades é a altura, mas a percepção que transmite é que é o volume (que cresce por um factor 8 quando a altura duplica!).



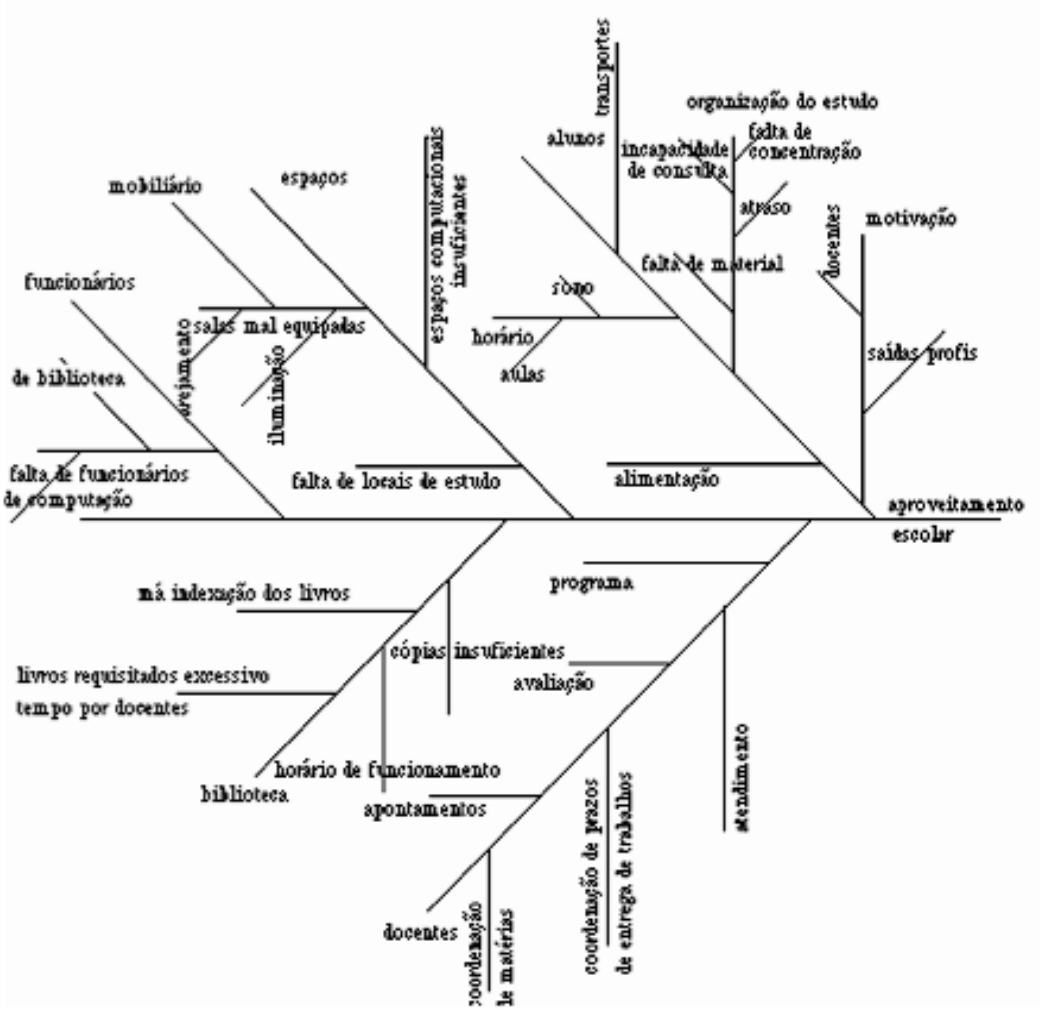
O controle (ou melhor, a perversão) da informação é muitas vezes feita usando pictogramas com escala arbitrária, como por exemplo no gráfico seguinte que aparenta um crescimento entre 1981 e 1982 em total desconformidade com a escala usada para mostrar o crescimento económico entre 1970 e 1981.



Рост продукции промышленности (1922 г. = 1).

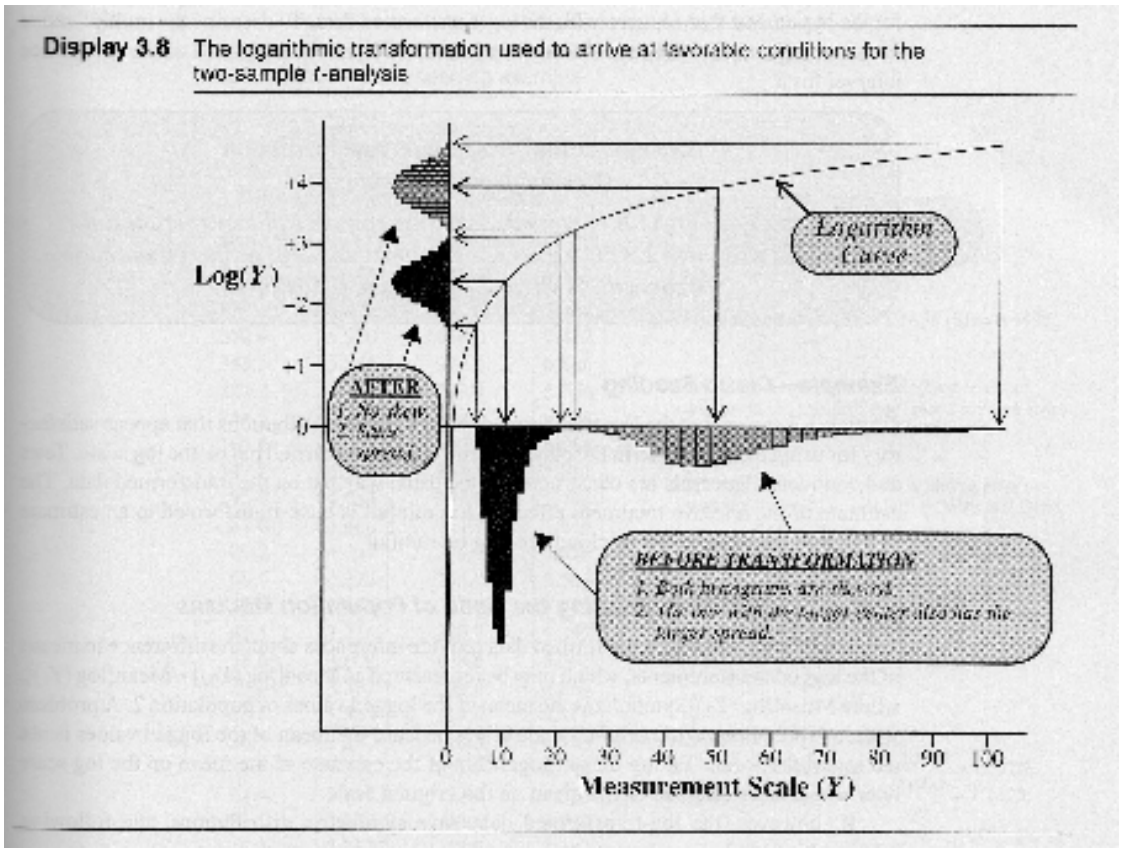
Wilkinson (1999), que tinha criado o saudoso package ***SYSTAT*** (absorvido pelo ***SPSS***, mas com perda de algumas características que mereceriam ter sido preservadas) contém muita informação sobre psicologia da cognição, relevantes para a produção de melhores gráficos.

Quando se estuda o passado, o que sobrenada são os casos excepcionais. Mas naturalmente a contrapartida das 7 maravilhas do Mundo devem ser umas 777 biliões de construções toscas, reles, para esquecer. Os exemplos excepcionais contidos na bibliografia que indicámos são contrabalançados pela chateza de milhões de péssimos gráficos, muitos dos quais apenas procuram esconder a mediocridade de quem os faz, e ainda por cima tem o mau gosto e insensatez de os exibir!
Em contrapartida à acidez destes comentários, anote-se que continuam a ser inventados gráficos que merecem ser conhecidos, por exemplo, os gráficos em espinha de peixe (também denominados, devido ao seu uso, diagramas de causa e efeito), inventados no fim do século XX. Ilustramos com um possível estudos das razões do insucesso escolar:

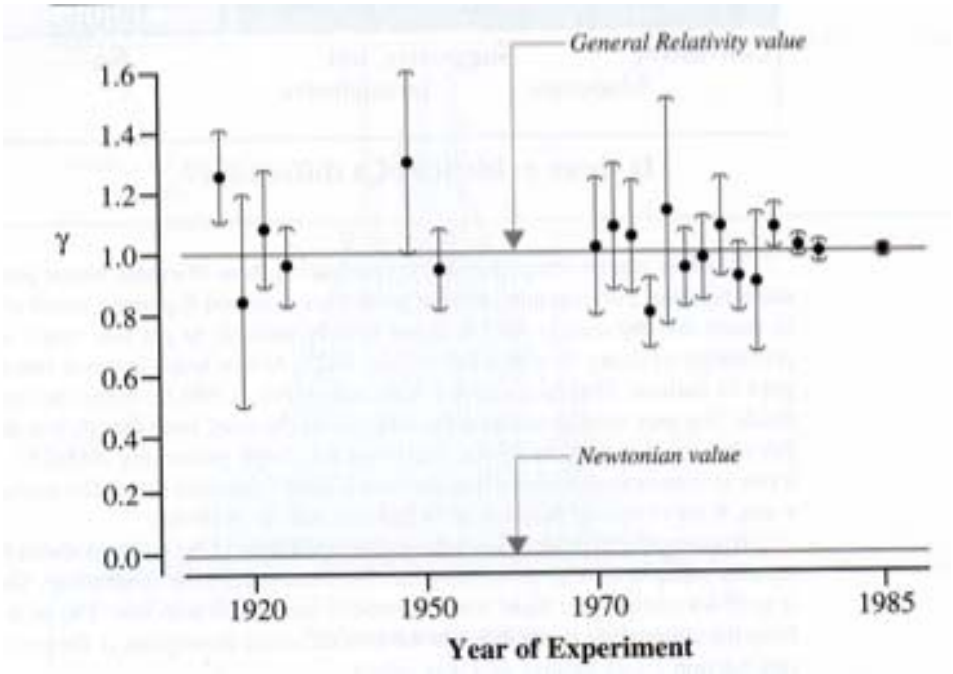


Como testemunho mais pessoal: na minha investigação uso os gráficos com uma parcimónia que ronda a avareza, e creio que apenas publiquei gráficos em trabalhos de índole didáctica. Gráficos de construção simples podem ser inspiradores, mas de modo nenhum substituem uma análise assente em métodos decerto menos apelativos,

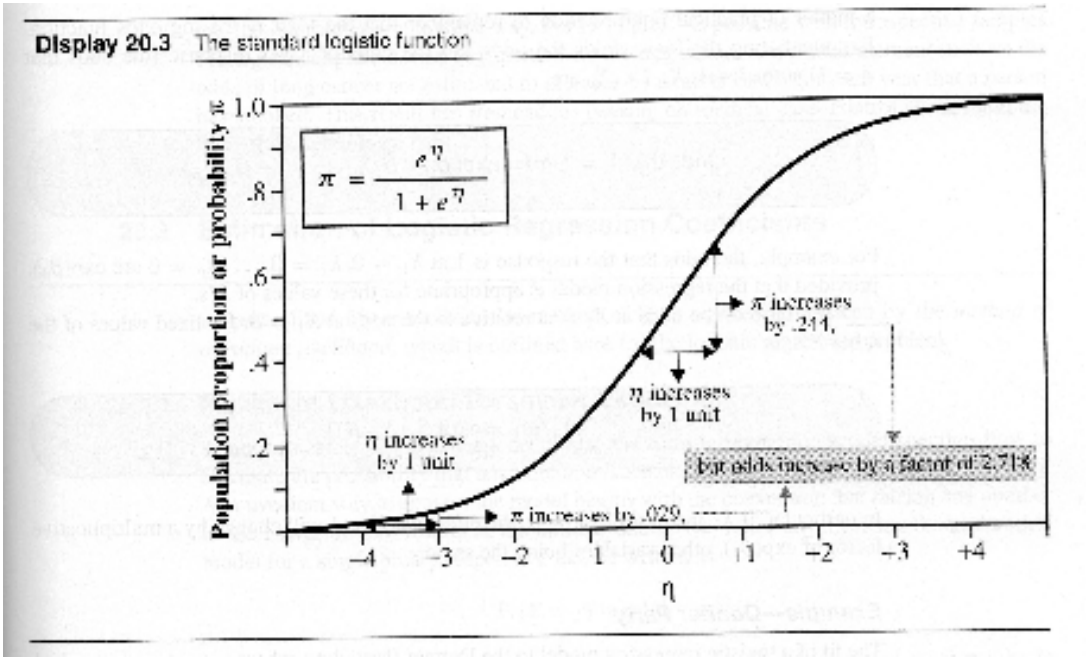
mas mais seguros. Os gráficos são mais próprios da análise exploratória de dados do que da análise estatística confirmatória, que decerto merece um estatuto de maior relevo.
Já no plano da divulgação, e em obras didácticas, considero os gráficos são mensageiros de eleição. Mesmo esboços muito rudimentares no quadro podem iluminar uma aula. Um livro que muito aprecio (Ramsey and Shafer, 2002) tem gráficos notáveis que em muito contribuem para a sua qualidade. Por exemplo, o seguinte gráfico mostra de forma notável que a transformação logaritmica, ao encolher a cauda direita e esticar a cauda esquerda, contribui para simetrizar os dados com grande assimetria direita, e lhes dá uma forma mais próxima da do modelo "normal".



Um dos exemplos que os referidos autores exploram no texto é a capacidade de transmitir conhecimento com gráficos, comparando gráficos com fórmulas e comentários incluídos, como o que acima se reproduz, com gráficos sem essas informações incluídas, isto é relegando as fórmulas e comentários para o texto que acompanha o gráfico; deixamos ao leitor o trabalho de adivinhar a conclusão a que chegam.
É também merecedor de reprodução o gráfico que usa o exemplo da confirmação da teoria da relatividade mostrando que os intervalos de confiança de um parâmetro (que na mecânica newtoniana seria 0, e na mecânica relativista seria 1) não deixa dúvidas sobre qual das teorias deve ser usada no caso de estarem em jogo velocidades próximas da velocidade da luz.



Um outro gráfico de Ramsey and Shafer (2002) é notável pela informação rigorosa e condensada das principais características da curva logística, tornando evidente porque se deve usar regressão logística quando o objectivo é predizer probabilidades de acontecimentos em situações dicotómicas.



Na tradição de associar mapas e Estatística, desenvolveu-se toda uma área de "mapas estatísticos", associando bases de dados a mapas, que permitem a construção de imagens surpreendentes. Termino, aliás, com um mapa "explodido" de Portugal, produzido há alguns anos por um dos meus alunos pós-graduados, mostrando que Beja é o menor distrito do País, quando se pensa em número de casos de loucura.

Tão certeira como a famosa conclusão de que todos os tipos de crime, nas cidades inglesas, tinham aumentado ao longo do século XIX acompanhando o aumento do número de pastores da igreja anglicana!

Agradecimentos

Agradeço à Professora Doutora Olga Pombo o excepcional trabalho editorial, que melhorou substancialmente a apresentação. O trabalho foi parcialmente subsidiado por FCT/OE.

Bibliografia

• Balbi, A. (1822). *Essai statistique sur le royaume de Portugal et d'Algarve, comparé aux autres Etats de l'Europe et suivi d'un coup d'œil sur l'état actuel des sciences, des lettres et des beaux-arts parmi les Portugais des deux hémisphères*, Rey et Gravier, Paris. 2 vol. BNF: FB-16466/7.

• Bertillon, J. (1889). *Atlas de statistique graphique de la ville de Paris, année 1888*, tome I. Masson, Paris. Préfecture du département de la Seine, Secrétariat général, Service de la statistique municipale.
• Bertillon, J. (1891). *Atlas de statistique graphique de la ville de Paris, année 1889*, tome II. Masson, Paris. Préfecture du département de la Seine, Secrétariat général, Service de la statistique municipale.
• Cobb, G. W. (1998). *Introduction to Design and Analysis of Experiments*, Springer, New York.
• Friendly, M. (2008). The Golden Age of Statistical Graphics, *Statistical Science* **23**, 502-535.
• Halley, Edmund (1686). On the height of the mercury in the barometer at different elevations above the surface of the earth, and on the rising and falling of the mercury on the change of weather. *Philosophical Transactions*, *16*:104-115.
• Halley, Edmund (1693). An estimate of the degrees of mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw, with an attempt to ascertain the price of annuities on lives. *Philosophical Transactions*, *17*:596-610
• Machado de Sousa, M. L. (1979). *Solano Constâncio, Portugal e o Mundo nos Primeiros Decénios do sec. XIX*, Arcádia, Lisboa.
• Minard, C. J. (1862). *Des Tableaux Graphiques et des Cartes Figuratives*. E. Thunot et Cie, Paris. ENPC: 3386/C161, BNF: Tolbiac, V-16168; 8 p. and plate(s).
• Murteira, B. (1993), *Análise Exploratória de Dados: Estatística Descritiva*, McGraw Hill, Lisboa.
• Nightingale, F. (1857). *Mortality of the British Army*. Harrison and Sons, London.
• Nightingale, F. (1858). *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army*. Harrison and Sons, London.
• Perozzo, Luigi (1880). Della rappresentazione graphica di una collettività di individui nella successione del tempo. *Annali di Statistica*, *12*:1-16. BL: S.22.
• Pestana, D., e Velosa, S. (2008). *Introdução à Probabilidade e à Estatística*, Vol. I, 3ª ed., Fundação Gulbenkian, Lisboa.
• Playfair, W. (1786). *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century*. Corry, London. Republished in *The Commercial and Political Atlas and Statistical Breviary* (H. Wainer and I. Spence, eds.) 2005. Cambridge Univ. Press, Cambridge.
• Playfair, W. (1801). *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*. Wallis, London. Republished in *The Commercial and Political Atlas and Statistical Breviary* (H. Wainer and I. Spence, eds.) 2005. Cambridge Univ. Press, Cambridge.
• Ramsey, F. L., and Schafer, D. W. (2002).*The Statistical Sleuth — A Course in Methods of Data Analysis*, 2nd ed., Duxbury, Belmont.
• Statistischen Bureau (1897). *Graphisch-statistischer Atlas der Schweiz* (Atlas Graphique et Statistique de la Suisse). Buchdruckeri Stämpfli & Cie, Departments des Innern, Bern.
• Tufte, E. R. (1983). *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Conn.
• Tukey, J. W. (1972). Some graphic and semigraphic displays. In *Statistical Papers in Honor of George W. Snedecor*, T. A. Bancroft, ed.. 293–316. Iowa State University Press, Ames, IA.
• Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.
• Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statist. Sci*. **5**, 327-339.
• Wainer, H. (2003). A graphical legacy of Charles Joseph Minard: Two jewels from the past. *Chance* **16**, 56-60.
• Wilkinson, L. (1999). *The Grammar of Graphics*, Springer, New York.