Introducció a l'anàlisi visual mitjançant D3

Dades, *layouts*, visualitzacions

PID_00246770

Julià Minguillón Alfonso

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Índex

Introducció

En aquest material docent s'introdueixen els conceptes bàsics relacionats amb l'anàlisi visual de dades, així com l'ús d'una llibreria de software (D3) que permet visualitzar dades d'acord amb diversos formats preestablerts (*layouts*), proporcionant un entorn interactiu per a la seva exploració i anàlisi visual.

La visualització de dades és una eina molt eficaç per a realitzar-ne una anàlisi preliminar, aprofitant les capacitats del sistema visual humà per detectar i extreure coneixement en forma de patrons, tendències, *outliers*, etc. Per a això només cal representar les dades gràficament d'acord amb l'objectiu de l'anàlisi que es vulgui fer, utilitzant algun tipus predeterminat de visualització que es consideri adient (per exemple, per haver sigut satisfactòriament utilitzada amb anterioritat).

Tot i que la visualització de dades és una disciplina que requereix de coneixements diversos pertanyents a diferents àmbits (estadística, disseny gràfic, computació, psicologia...), en aquest material docent descriurem principalment models existents per crear visualitzacions d'una forma relativament ràpida i simplificada, que poden servir de base per crear visualitzacions interactives més complexes, com les usades per Mike Bostock* al *New York Times* per narrar històries mitjançant dades, per exemple.

*https://goo.gl/aRFFhU

Un estudi rigorós del procés de visualitzar dades requereix, entre altres aspectes, aprofundir i comprendre factors com la subjectivitat intrínseca a les representacions i el mapeig de dades, la importància de l'aspecte visual i les seves conseqüències cognitives, etc. Tanmateix, és possible intentar reutilitzar bones pràctiques en visualització de dades per explorar noves dades.

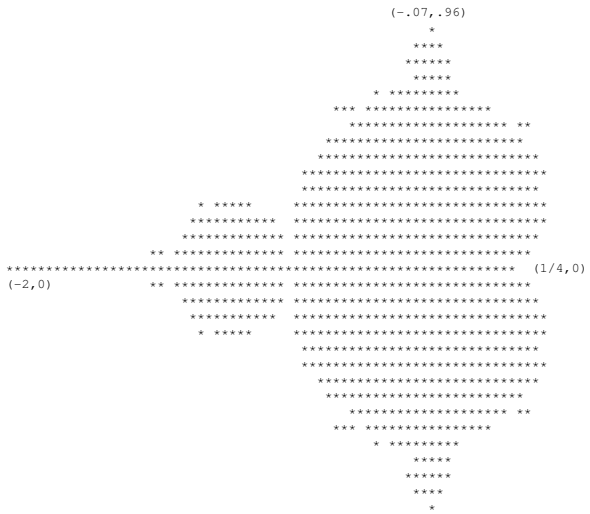
D'aquesta manera, sense haver de conèixer D3 en profunditat, és possible adaptar visualitzacions ja existents per a visualitzar noves dades, usant el *layout* o configuració més apropiat en cada ocasió, en funció de la naturalesa de les dades i dels objectius de cada visualització. Per a això es proporcionen un conjunt de visualitzacions interactives que inclouen plantilles (codi HTML + CSS + *scripts* D3) que poden ser modificades per visualitzar conjunts de dades en format CSV o JSON. Aquestes visualitzacions s'executen com pàgines web i són mostrades mitjançant un navegador.

L'objectiu és proporcionar un mecanisme senzill per transformar dades (en format tabular o jeràrquic) en visualitzacions interactives que permeten la seua manipulació gairebé sense haver de programar, usant els diferents *layouts* de D3 com un «motlle» que genera la visualització a partir dels «ingredients» adequats (les dades en el format preestablert).

1. La visualització com a exploració de dades

Aplicacions per a visualitzar dades han existit des de fa molt de temps, tot i que s'han popularitzat recentment donades les possibilitats que ofereix el maquinari disponible en l'actualitat, el qual ha anat doblant la seva capacitat, velocitat i resolució al llarg dels anys. Sense aquesta capacitat computacional, la generació de gràfics per ordinador era molt limitada, tant per la resolució gràfica com del temps de còmput necessari per generar els gràfics. De fet, les primeres visualitzacions havien d'utilitzar els recursos bàsics del joc de caràcters per representar gràfics. Llançades al mercat l'any 1981 amb l'IBM PC, les primeres targetes gràfiques CGA tenien una resolució de 320 x 200 píxels i una paleta de dos bits o quatre colors. Menys de quatre dècades després, el maquinari permet generar gràfics de fins a 7680 x 4320 píxels, amb vint-i-quatre bits de profunditat de color RGB i una freqüència de refresc de 60 Hz.

Figura 1. Primera visualització del conjunt de Mandelbrot, per Brooks i Matelski, 1978



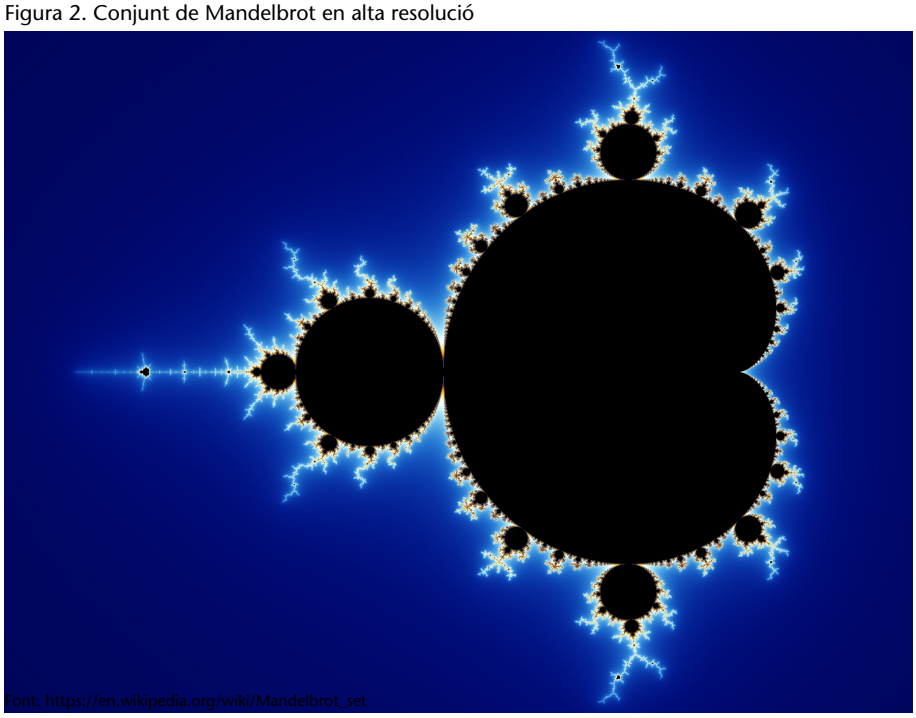
Font: http://mrob.com/pub/muency/brooksandmatelski.html

Un exemple de l'evolució de la capacitat i les possibilitats ofertes per les targetes gràfiques el proporciona la visualització d'objectes fractals, construïts a partir d'un cert algoritme i unes dades o paràmetres d'entrada. Es tracta de descripcions algorítmiques molt senzilles que generen imatges molt complexes, amb un grau de detall virtualment infinit. Quan Benoit Mandelbrot va presentar a finals dels setanta el conjunt que porta el seu nom, va haver d'utilitzar les limitades capacitats de representació de l'època, com en mostra la primera imatge creada per Robert W. Brooks i Peter Matelski, mostrada en la figura 1, usant només text. No obstant això, va ser amb la popularització de l'ordinador personal i les seves capacitats gràfiques quan es va poder visualit-

zar per primera vegada el conjunt de Mandelbrot en detall, a uns nivells de resolució insospitats anteriorment.

El conjunt de Mandelbrot és un bon exemple de com la visualització de dades transforma la comprensió d'un concepte, possibilitant el descobriment d'un coneixement fins al moment desconegut (la forma del conjunt en el pla complex i el nivell de detall, en aquest cas). Concretament, Benoit Mandelbrot va conjecturar que el conjunt que porta el seu nom no era connex, i va ser mitjançant-ne la visualització que es van descobrir els petits filaments que fan que el conjunt de Mandelbrot sigui connex. Sense l'ús d'ordinadors per a visualitzar-lo amb una resolució diferent, hagués estat impossible.

Figura 2. Conjunt de Mandelbrot en alta resolució



Per tant, la visualització d'un conjunt de dades és normalment una primera etapa indispensable per comprendre millor la seva naturalesa, detectant ràpidament aspectes tan intrínsecs a les dades com simetries, localitzacions, valors extrems, continuïtat, suavitat, agrupacions, etc. Mitjançant aquesta exploració és possible fer-se una idea preliminar que pot permetre afinar els objectius de l'anàlisi i la posterior visualització dels resultats, en un cicle continu.

2. La visualització com a anàlisi preliminar

La mateixa falta de coneixement sobre la naturalesa del conjunt de Mandelbrot pot succeir amb altres dades en general, els quals són de naturalesa complexa i poden combinar diferents aspectes al mateix temps, entre d'altres: ser multidimensionals, anar lligats a restriccions espaitemporals, longitudinals (que evolucionen en el temps), multimodals (combinant diferents fonts i orígens), així com provenir de l'execució de múltiples processos paral·lels o models. Visualitzar dades inclou gestionar tota aquesta complexitat per convertir-los en informació, és a dir, obtenir respostes a les preguntes o objectius de la visualització. L'anàlisi visual no substitueix a l'estadística clàssica o la construcció de models de mineria de dades, sinó que aporta una perspectiva diferent basada en les capacitats del sistema visual humà.

Així, l'objectiu de la visualització és mostrar la naturalesa de les dades, facilitant la seva comprensió i posterior exploració. Es tracta, doncs, de realitzar una anàlisi visual preliminar per detectar els aspectes clau presents en les dades: distribucions de cada variable, valors extrems, relacions entre variables, tendències, patrons, *outliers*, etc. Per a això és necessari poder disposar d'un entorn gràfic que permeti visualitzar dades usant diferents projeccions, combinant eines estadístiques amb models generats a partir de les dades, des de descriptors estadístics fins el resultat d'un algoritme de classificació no supervisat, per exemple, variant els paràmetres del mateix.

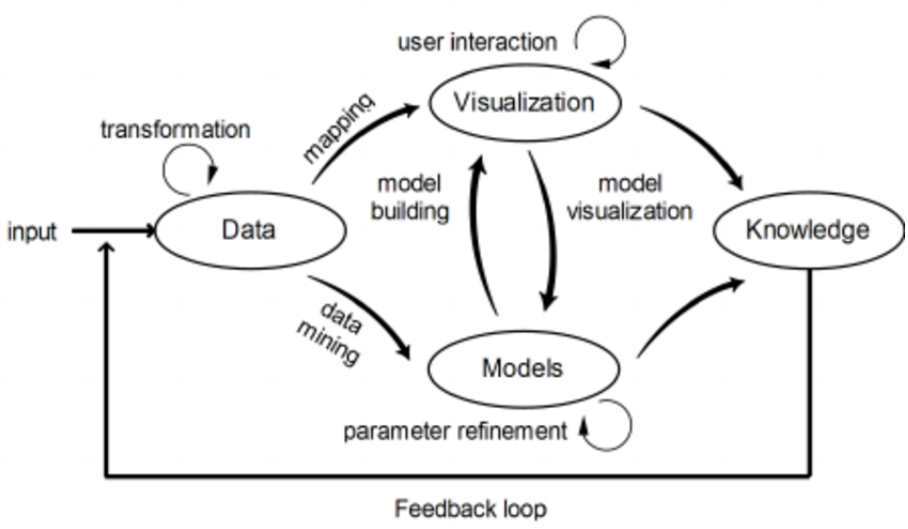
En aquest sentit, l'evolució de la visualització de dades no s'ha centrat només en la capacitat de generar gràfics complexos amb major resolució en un breu lapse de temps, sinó que ha anat incorporant elements interactius en la pròpia visualització, en forma d'operacions bàsiques (selecció, filtrat, etc.). D'acord amb el treball de Keim i altres publicat el 2008, l'anàlisi visual de dades es fonamenta en un mantra que és una versió modificada del proposat per Ben Shneiderman el 1996:

"Analyse First –
Show the Important –
Zoom, Filter and Analyse Further –
Details on Demand"

Així, el procés d'anàlisi visual consisteix en un cicle continu que s'inicia en les dades i les seves possibles transformacions, i que es bifurca en dues aproximacions complementàries, la visualització i la construcció de models, entre les quals hi ha un diàleg amb l'objectiu de extreure coneixement que pugui

ser usat per a iterar el procés d'anàlisi visual amb un major nivell de detall o complexitat, tal com mostra la figura 3. La capacitat d'interacció ha de permetre a l'usuari de la visualització realitzar, almenys, les operacions bàsiques definides per Ben Shneiderman (vista general, zoom, filtre i selecció).

Figura 3. Procés d'anàlisi visual definit per Keim *et al.* (2008)



Des d'una perspectiva d'anàlisi visual, les dues primeres etapes definides a la figura 3 són la transformació (o adaptació) de les dades i la seva visualització, incloent en aquesta la interacció. Per tant, un cop establert l'objectiu de l'anàlisi visual de les dades, es tracta de seleccionar un tipus de visualització interactiva que permeti realitzar aquesta exploració preliminar.

El conjunt d'eines estadístiques per analitzar, descriure i resumir dades és enorme, existint diferents versions en funció de la naturalesa de les dades i l'objectiu a assolir. Sense pretendre ser exhaustius, una anàlisi preliminar ha de plantejar-se, com a mínim, les següents preguntes:

- Es tracta d'analitzar una sola variable o bé una combinació de dues o més variables? En funció del nombre de variables que formen part de l'anàlisi (és a dir, de la visualització), les possibilitats són molt diferents, així com la resta de preguntes a plantejar-se.

- Es tracta de variables contínues o categòriques (incloent ordinals, nominals i binàries)? Per a cada tipus de variable o parella de variables hi ha un test estadístic diferent en funció de l'objectiu.

- S'estan analitzant les variables originals o bé hi ha hagut una etapa de transformació intermèdia, o són el resultat d'un model construït amb anterioritat? En tots dos casos, la visualització s'hauria de fer tenint en compte aquests processos.

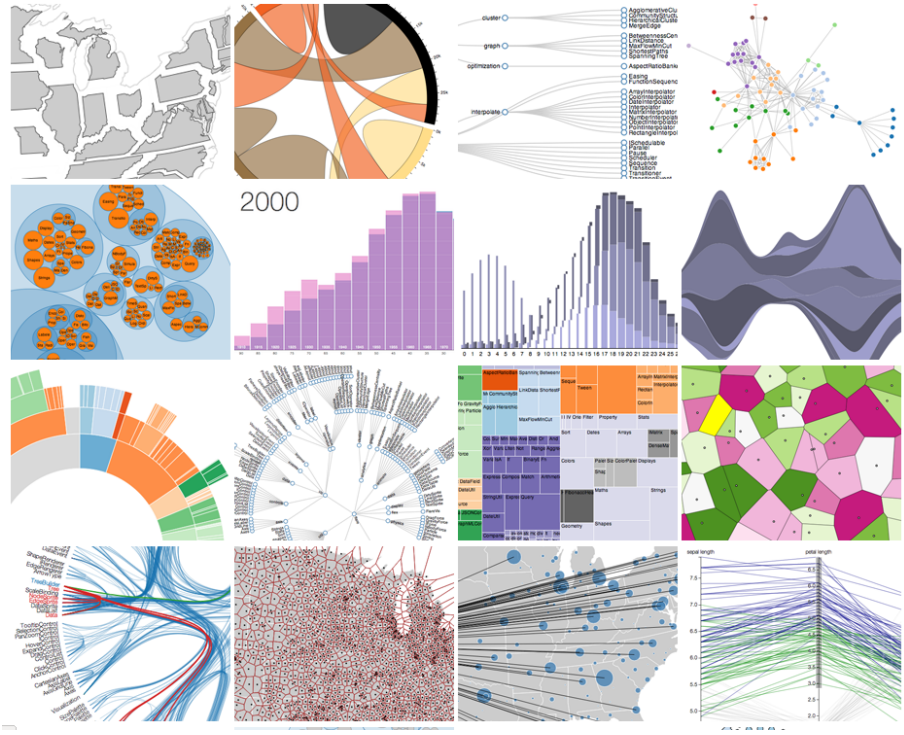
No obstant, en aquest material docent ens centrarem en dades multidimensionals categòriques, on les possibilitats per mostrar la naturalesa de les dades i les relacions entre una o més variables permeten diferents representacions més enllà dels típics histogrames i gràfics de barres, així com una certa manipulació mitjançant l'ús d'elements interactius.

3. Eines per a l'anàlisi visual de dades

La creació de visualitzacions interactives pot abordar des de dues aproximacions diferents: la primera és utilitzar eines de programari dissenyades específicament per a això, sent QlikView i Tableau de les més populars en l'actualitat, principalment a causa de la seva potència i flexibilitat. La segona aproximació és utilitzar les capacitats gràfiques d'algun llenguatge de programació o llibreria per mostrar dades de forma més o menys interactiva, com D3 (mostrat a la figura 4), Processing o Caire, entre d'altres*, depenent en aquest cas d'un desenvolupament que pot ser costós i complex. Una tercera via que combina ambdues aproximacions és la utilització d'eines programari que generen visualitzacions en forma de codi que pot ser executat independentment de l'eina usada per a crear-les, com Quadrigram, per exemple.

https://goo.gl/PnOq4a

Figura 4. Exemples de diferents visualitzacions creades amb D3



Font: https://github.com/d3/d3/wiki/Gallery

En l'actualitat, la visualització de dades interactiva compta amb un nou aliat, el qual elimina la necessitat de crear aplicacions específiques i dota d'una interfície visual coherent per a la seva comesa. Es tracta dels navegadors web, que visualitzen pàgines que contenen un codi font que construeix (mitjançant la renderització) la visualització quan la pàgina és accedida i carregada. Una pàgina web és una combinació de CSS (fulls d'estil que determinen l'as-

pecte dels elements de la pàgina), contingut HTML pròpiament dit i codi JavaScript que permet manipular el DOM (de l'anglès Document Object Model, és a dir, l'estructura de la pròpia pàgina web vista com un document estructurat jeràrquicament), generant nous continguts que s'incrusten dinàmicament, incloent codi HTML i gràfics vectorials (SVG, o Scalable Vector Graphics). El gràfic (o millor dit, la forma de construir-lo) és part de la pàgina, i és visualitzat quan el navegador executa les ordres necessàries per mostrar el contingut de la pàgina.

D'aquesta manera, generar una visualització de dades interactiva es pot veure com la creació d'una pàgina web construïda dinàmicament i que visualitza aquestes dades d'acord amb una configuració preestablerta. En aquest sentit, D3 (o també D3.js) és una llibreria JavaScript que permet manipular dades en diferents formats (taules, CSV o JSON, entre d'altres) i generar gràfics vectorials de forma dinàmica que poden ser incrustats a la pàgina web per a la seva manipulació, incloent elements d'interactivitat, tant pel que fa a la interfície de l'usuari com a l'ús de transicions que aporten dinamisme a la visualització (Murray, 2013).

Donada la seva flexibilitat, D3 pot utilitzar-se per crear qualsevol tipus de visualització interactiva, generant els elements gràfics a partir de les dades que alimenten la visualització, des de gràfics de barres fins a complexes visualitzacions combinant diferents elements gràfics. Un dels aspectes més interessants de D3 és la incorporació de la interacció com a part de la pròpia visualització, de manera que esdevé la interfície d'accés a les dades, permetent la seva manipulació d'acord a les operacions bàsiques definides per Shneiderman (1996).

Les possibilitats que ofereix D3 per a visualitzar dades categòriques (principalment) són molt interessants, justificant la seva elecció per a una anàlisi exploratòria inicial i, fins i tot, ser l'esquelet inicial d'una visualització complexa. Per a això disposa d'un catàleg de configuracions (*layouts*) molt extens, com es veurà a continuació.

4. *Layouts* D3 per l'anàlisi visual de dades

Com ja s'ha comentat, D3 ofereix una bona col·lecció de *layouts* ja predefinits que poden ser usats per crear visualitzacions interactives amb relativament poc esforç, tal com mostra la figura 4. La idea bàsica és reutilitzar les visualitzacions ja existents per adaptar-les a les dades que es desitja visualitzar (i viceversa, adaptar les dades als requeriments de la visualització), de manera que s'utilitza cada *layout* com un «motlle» o «caixa negra» que transforma dades en una visualització interactiva.

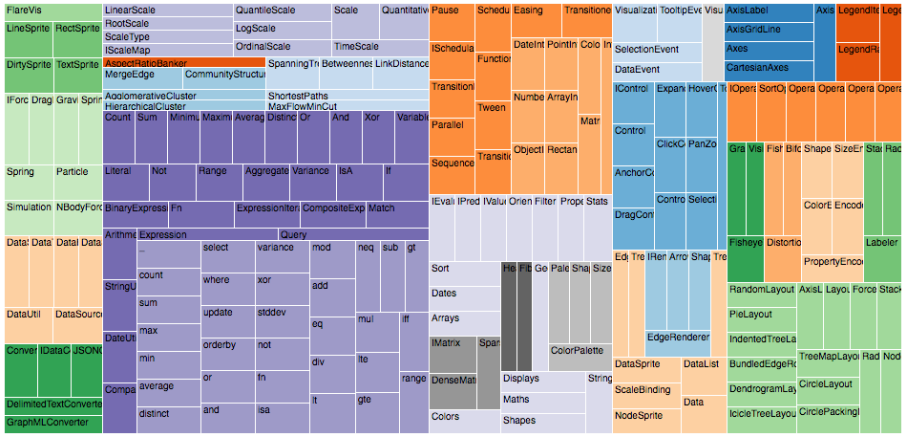
A continuació es mostra una selecció de *layouts* de D3 obtinguts de la pròpia galeria (`https://github.com/d3/d3/wiki/Gallery`), que permeten representar-ne dades i realitzar-ne una exploració senzilla mitjançant l'ús d'elements interactius integrats en la pròpia visualització. Com a complement a aquest material docent, s'ha preparat un exemple per a cada un d'aquests *layouts*, que incorpora un major nivell de detall i una visualització interactiva sobre unes dades reals que es poden descarregar per reutilitzar la visualització.

Enlace de interés

Els exemples interactius es troben a http://oer.uoc.edu/VIS/D3/

4.1. *Treemap*

Aquest tipus de visualització permet observar com es distribueixen els valors d'una o més variables categòriques en una àrea, normalment rectangular, permetent realitzar comparacions senzilles per mida o proporció, així com detectar l'existència de patrons en el cas de combinar dues o més variables. Per exemple, es pot visualitzar com es distribueixen els pressupostos d'un país en les diferents partides que el componen, observant clarament quins apartats o regions s'emporten la major part.



4.2. *Bubble*

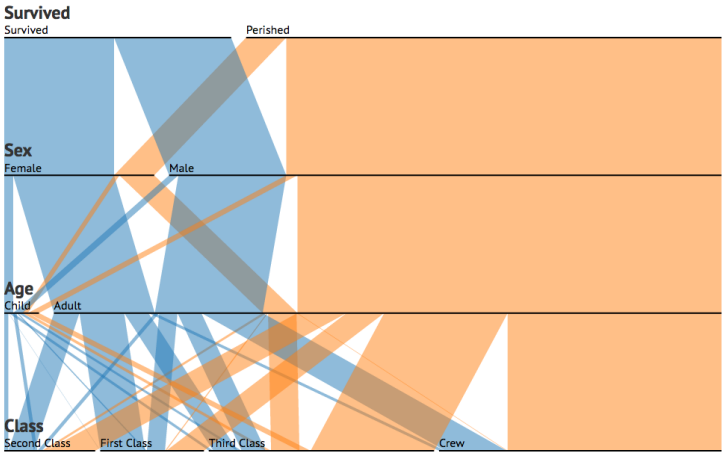
Es tracta d'una visualització jeràrquica que permet visualitzar les relacions entre una o més variables categòriques, de manera similar a un *treemap* però amb una representació basada en cercles. L'aprofitament de l'espai és menor que en el cas del *treemap*, però les estructures jeràrquiques presents en les dades poden detectar-se millor, especialment quan l'estructura jeràrquica no està equilibrada.

4.3. *Chord*

En aquest cas es tracta d'una visualització radial que utilitza una corona circular on es distribueixen els valors d'una variable categòrica, d'acord amb la seva probabilitat o freqüència. Entre cada parell de valors ha un arc (*chord*) que permet conèixer la relació entre els dos valors, també proporcional al nombre d'elements que participen en aquesta combinació. Per exemple, és possible visualitzar les combinacions més comunes que apareixen en les matrícules dels estudiants, analitzant les parelles d'assignatures més freqüents.

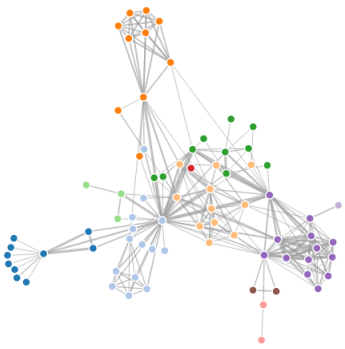
4.4. *Parallel sets*

Són útils per a representar dades categòriques, permetent visualitzar tant les freqüències d'aparició de cada valor per a cada variable com les combinacions de valors entre variables. En aquest segon cas, es poden utilitzar com una anàlisi visual complementària a l'ús de taules de contingència. Un exemple típic és la segmentació d'estudiants respecte els seus atributs: gènere, grup d'edat, estudis previs, carrera escollida, etc.



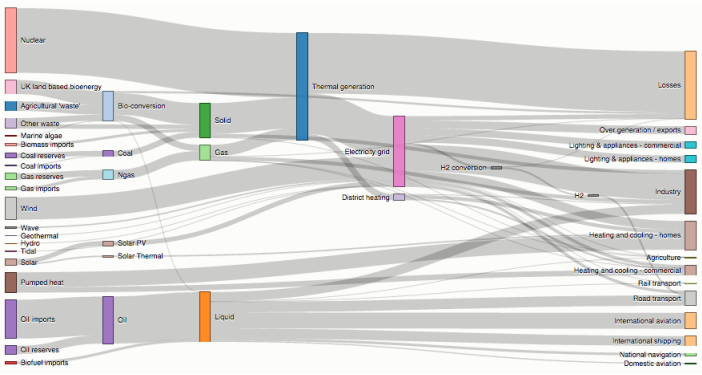
4.5. *Force-directed graphs*

A vegades és més interessant visualitzar les relacions entre elements que els atributs dels elements en si mateixos. Els grafs permeten visualitzar aquestes relacions, utilitzant una disposició que reflecteix, de certa manera, la distància (entesa com similitud o dissimilitud) entre els diferents elements, però de forma relativa, no absoluta. La visualització mostra els diferents grups o clústers que es formen, permetent establir categories de forma senzilla. Per exemple, és possible visualitzar un graf que visualitza les coautories entre autors que publiquen conjuntament, on cada autor és un node del graf i cada arc entre nodes representa el nombre de vegades que un autor ha compartit autoria amb un altre.



4.6. *Sankey*

En aquest cas es tracta de visualitzar un flux de dades entre diferents conceptes d'acord amb certs criteris, normalment espacials o temporals, encara que també pot usar-se per mostrar canvis del valor d'una o més variables en funció d'una altra, normalment el temps. En aquest cas, permet comparar valors que pot prendre una variable respecte al total, mostrant les proporcions de cada possible valor.

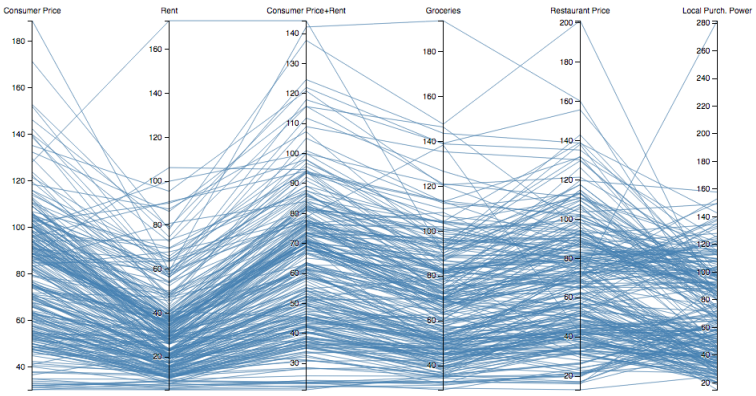


4.7. *Sunburst*

Aquest tipus de visualització és equivalent al *treemap* però utilitzant una disposició radial en lloc de rectangular. Igual que *bubble*, l'aprofitament de l'espai és menor, però la disposició radial comporta un ordre que pot aprofitar-se per mostrar els elements d'acord amb algun paràmetre. D'altra banda, el fet d'allunyar-se del centre també implica una certa ordenació que permet establir prioritats entre els diferents elements visualitzats. Finalment, es poden mostrar diferents nivells de profunditat per a cada segment, per la qual cosa és una representació més apta que *treemap* per a estructures jeràrquiques molt desbalancejades.

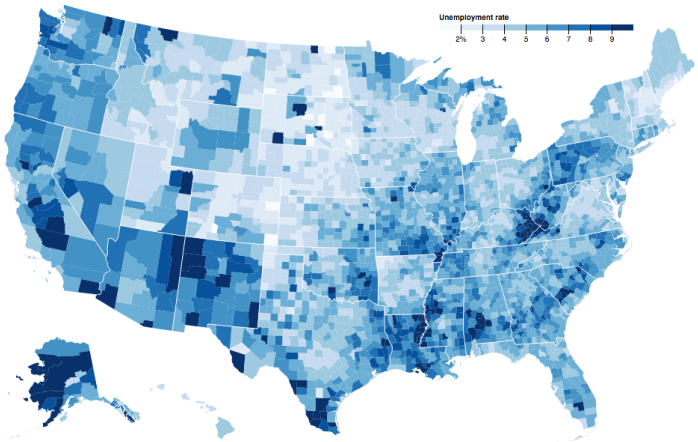
4.8. *Parallel coordinates*

Encara que a primera vista pugui semblar confusa, aquesta visualització és molt utilitzada, ja que aprofita les capacitats del sistema visual humà per detectar patrons i tendències usant línies i posicions relatives, facilitant la seva interpretació en clau de la correlació entre variables. Permet representar un espai multidimensional més enllà de la típica representació en dues dimensions ortogonals, d'aquí el seu nom de *coordenades paral·leles*.



4.9. *Choropleth*

En molts casos les dades que s'han de visualitzar tenen un origen relacionat amb una àrea o regió geogràfica. Aquest tipus de visualització permet superposar una gradació de color a un mapa, mostrant per a cada segment un valor possible en forma de la intensitat del color. D'aquesta manera és possible detectar fàcilment agrupacions on es concentren els valors extrems de la variable utilitzada per a la visualització.



Bibliografia

Brooks, R.; Matelski, J. P. (1981). «The dynamics of 2-generator subgroups of PSL (2, C)». A: *Riemann surfaces and related topics: Proceedings of the 1978 Stony Brook Conference* (vol. 97, pàgs. 65-71). Princeton Press University.

Keim, D.; Andrienko, G.; Fekete, J. D.; Görg, C.; Kohlhammer, J.; Melançon, G. (2008). «Visual analytics: Definition, process, and challenges». A: *Information visualization* (pàgs. 154-175). Springer Berlin Heidelberg.

Murray, S. (2013). *Interactive data visualization for the Web*. O'Reilly Media, Inc.

Shneiderman, B. (1996). «The eyes have it: A task by data type taxonomy for information visualizations». A: *Proceedings IEEE Symposium on Visual Languages* (pàgs. 336-343). IEEE.

Pàgines web

- D3: https://d3js.org
- QlikView: http://www.qlik.com/es
- Tableau: http://www.tableau.com/
- Processing: https://processing.org/
- Cairo: https://cairographics.org/
- Quadrigram: http://www.quadrigram.com/
- The Data Visualization Catalogue: http://www.datavizcatalogue.com/