PV251 Vizualizace

Jaro 2017

Výukový materiál

---

6. přednáška: Vizualizace multivariate dat

Jako multivariate označujeme data, která se skládají z různých typů atributů. Jako příklad můžeme uvést datovou sadu, která vznikne tak, že sbíráme informace o váze w, výšce h a čísle bot s náhodného vzorku osob. Pak trojice (w_1, h_1, s_1), (w_2, h_2, s_2), … jsou příkladem sady multivariate dat (neboli dat o více proměnných).

V této přednášce se budeme zabývat technikami vizualizace seznamů a tabulek dat, které obecně nemají explicitní prostorové atributy. Přednáška bude organizována tak, že techniky probereme postupně podle typu grafických primitiv, který jsou používána při jejich renderování. Začneme tedy opět body, čarami a regiony a poté uvedeme techniky, které kombinují dvě či více z těchto základních technik. Na závěr uvedeme vlastnosti, které jsou pro všechny multivariate vizualizační techniky společné.

Techniky pro bodová data

Bodové grafy si můžeme v tomto kontextu představit jako typ vizualizace, který promítá záznamy z n-dimenzionálního datového prostoru do libovolného k-dimenzionálního prostoru výstupního zařízení (např. displeje), kdy jsou datové záznamy mapovány na k-dimenzionální body. Každý záznam je asociován s určitou grafickou reprezentací (značkou).

Bodové grafy mohou zobrazovat jednotlivé záznamy nebo souhrnné záznamy a mohou být strukturovány na základě využití různých projekčních technik. Nyní si ukážeme několik populárních metod pro vykreslování bodových dat.

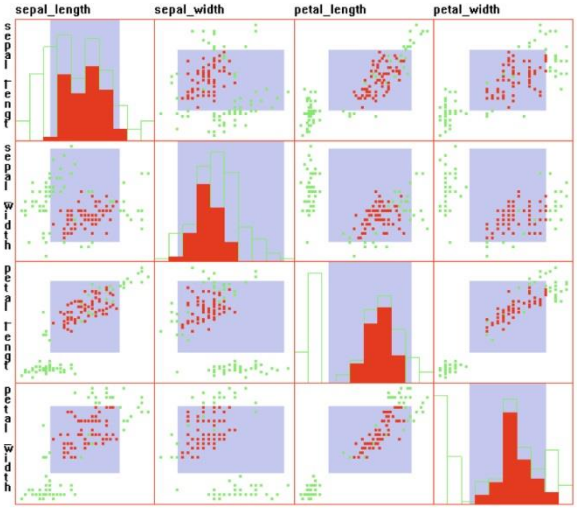
Bodové grafy (scatterplots)

S bodovými grafy jsme se již několikrát setkali a to hlavně proto, že bodové grafy jsou jedněmi z prvních a nejrozšířenějších vizualizačních technik používaných v analýzách dat. Většina nástrojů a balíků věnujících se analýze informace obsahuje jistou formu 2D a 3D bodových grafů. Jejich úspěch vychází z naší přirozené schopnosti odhadovat relativní pozici uvnitř omezeného prostoru. Se zvyšující se dimenzionalitou vstupních dat se vizuální analýza skládá z:

- **Hledání podmnožiny vstupních dimenzí** (dimension subsetting), kdy uživateli povolíme výběr pouze určité podmnožiny vstupních dimenzí, které budou zobrazeny. Při hledání vhodné podmnožiny dimenzí můžeme využít algoritmy pro hledání dimenzí obsahujících pro daný úkol nejužitečnější informace.
- **Redukce dimenze** – za využití technik jako jsou PCA (principal component analysis) nebo multidimensional scaling, které umožňují transformovat data o vyšších dimenzích do dat o dimenzích nižších, přičemž se snaží zachovat co nejvíce původních vztahů mezi datovými body.
- **Ukotvení dimenze** (dimension embedding) – mapování dimenzí na další grafické atributy kromě pozice, jako je například barva, velikost a tvar (samozřejmě počet dimenzí, které můžeme tímto způsobem zobrazit, je limitován).
- **Násobné zobrazení** (multiple displays) – zobrazení několika grafů, kdy každý zobrazuje některé z dimenzí (zobrazení pomocí superimposition nebo juxtaposition).

Násobné zobrazení

V případě zobrazení několika grafů zobrazujících různé dimenze zobrazovaných dat (násobné zobrazení) je nejčastější používanou technikou tzv. **matice bodových grafů (scatterplot matrix)**. Ta se skládá z mřížky obsahující bodové grafy, která má N^2 buněk, kde N je počet dimenzí. Tudíž každá dvojice dimenzí je vykreslena dvakrát – liší se pouze orotováním grafu o 90 stupňů. Uspořádání dimenzí je obvykle stejné v horizontální i vertikální ose, což vede k symetrii matice podél hlavní diagonály. Grafy na hlavní diagonále, které by měly zobrazovat proměnnou v dané dimenzi samu se sebou, se často používají pro sdělení informace o dimenzích v odpovídající řadě/sloupci nebo pro vykreslení histogramu dané dimenze.



Force-based metody

Existuje mnoho technik pro projekci bodů o velkých dimenzích do 2D nebo 3D prostoru zobrazení. Jejich hlavním cílem je pokusit se zachovat vlastnosti N-dimenzionálních dat při

projekci do jiné dimenze (např. vztahy existující mezi daty v původních dimenzích by měly být zachovány i po projekci). Projekce může zavést určité artefakty, které se mohou objevit ve výsledné vizualizaci a přitom v původních datech vůbec nebyla obsažena.

Nyní si popíšeme několik běžných projekčních metod, jako je Multidimensional scaling (MDS) či RadViz.

Multidimensional scaling

MDS představuje velkou sadu algoritmů pro redukci dimenze, které jsou běžně používány ve statistické analýze a information visualization. Typický MDS algoritmus má následující strukturu:

1. Mějme datovou množinu o M záznamech a N dimenzích. Vytvoříme MxM matici D_s obsahující výsledky měření podobnosti mezi jednotlivými páry vstupních dat. Toto měření lze provést různým způsobem, například použitím Euklidovské vzdálenostní metriky.

2. Předpokládejme, že vstupní data chceme promítnout do K dimenzí (pro účely zobrazení je K obvykle mezi 1 a 3). Sestrojíme matici L o rozměrech MxK, která obsahuje umístění promítnutých bodů. Těchto M umístění může být zvoleno náhodně nebo může být použita technika jako například Principle component analysis.

3. Spočteme matici L_s o rozměrech MxM, která obsahuje podobnost mezi všemi páry bodů z L.

4. Spočteme hodnotu tzv. *stress* – S, která je určena měřením rozdílů mezi D_s a L_s.

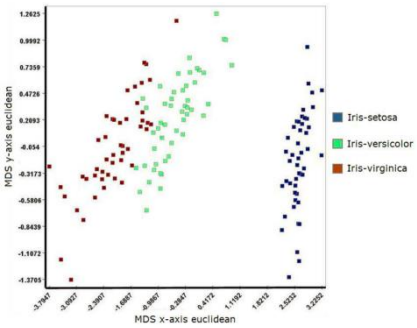
5. Pokud je S dostatečně malé nebo se v několika posledních iteracích významně nezměnilo, algoritmus končí.

6. Jinak posuneme pozice bodů v L ve směru, který zredukuje jejich jednotlivé hodnoty *stress*. To může být například vážený součet posunutí založeného na porovnání bodu se všemi ostatními body nebo pouze s nejbližšími sousedy.

7. Návrat na krok 3.

Je zřejmé, že existuje řada variant tohoto algoritmu. Jejich rozdíl spočívá zejména ve způsobu výpočtu podobnosti a hodnoty *stress*, v různé definici počátečních a koncových podmínek a v různé strategii updatování pozice bodů. Podobně jako u jiných optimalizačních algoritmů se nám může stát, že můžeme uvíznout v lokálním minimu, které ovšem má stále vysokou hodnotu *stress*. Běžné strategie, které se vypořádají s tímto problémem, příležitostně přidávají náhodný „skok“ v dané pozici bodu, kdy cílem je konvergovat k jinému umístění.

Obrázek ukazuje příklad datové množiny pro kosatce zobrazující čtyři numerické dimenze, kdy byla pro projekci použita technika MDS.



Problémy:

Pro většinu technik, které v této sekci představujeme, platí, že výsledky nejsou unikátní: drobné změny v počátečních podmínkách mohou vést ke zcela odlišným výsledkům. Dalším problémem je, že souřadný systém po projekci není pro uživatele zcela „smysluplný“ – vzhledem k dimenzím původních dat. Například je typické mapovat při běhu jednoho algoritmu datový bod na pozici v horní části displeje a při spuštění jiného algoritmu řešícího stejný problém může být stejný bod mapován na pozici v dolní části displeje. Proto je důležitá relativní pozice jednotlivých bodů, nikoliv absolutní.

RadViz

Další technikou je force-driven technika pro rozložení bodů nazývaná RadViz. Vychází z fyziky, přesněji z Hookova zákona a pro zobrazení využívá nalezení rovnovážné polohy bodu. Pro N-dimenzionální datovou množinu je na obvod kružnice umístěno N „kotevních“ bodů, které reprezentují fixní konce N strun přiřazených každému datovému bodu. Pro zjednodušení výpočtu a poskytnutí intuitivního pohledu na tento algoritmus umístíme kotvy na kružnici o poloměru 1.0, jejíž střed je v počátku souřadné soustavy.

Tedy pro daný normalizovaný vektor dat $D_i = (d_{i,0}, d_{i,1}, ..., d_{i,N-1})$ a sadu jednotkových vektorů A, kde A_j představuje j-tý kotevní bod, dostáváme následující výpočet rovnováhy:

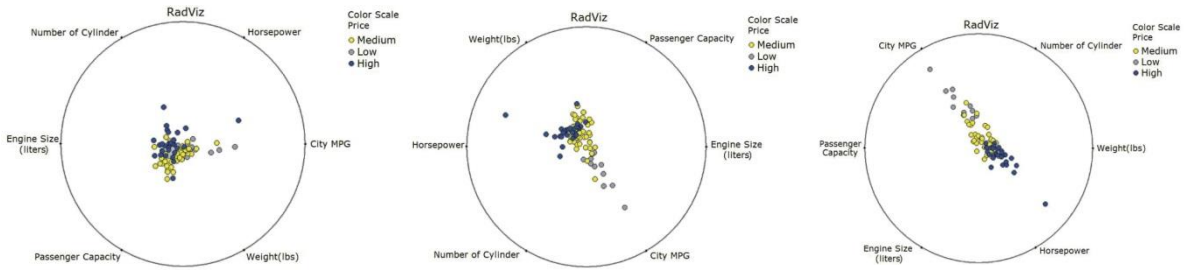
$$\sum_{j=0}^{N-1} (A_j - p) d_j = 0$$

Kde *p* je vektor pro bod v rovnovážné poloze. Výpočet *p* probíhá podle vzorce:

$$p = \frac{\sum_{j=0}^{N-1} (A_j d_j)}{\sum_{j=0}^{N-1} d_j}$$

Všimněme si, že různé rozmístění a uspořádání kotev vede k odlišným výsledkům a že body, které jsou v N dimenzích odlišné, mohou být mapovány na stejné místo ve 2D. Avšak toto je problém vyskytující se u všech projekčních technik a technik pro redukci dimenze. V případě

RadViz je jednoduchým řešením zpřístupnění interakce, jako například umožnění pohybu s kotvami a pozorování změn ve vizualizaci. Takto je často možné vysledovat vztahy v datech, jak ukazuje následující obrázek.

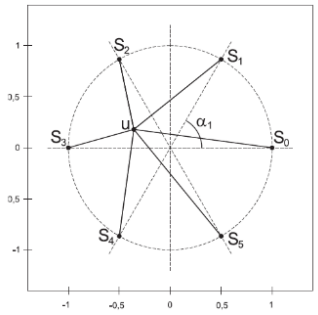


Obrázky ukazují různé pohledy na stejnou datovou množinu zobrazenou technikou RadViz. Obrázky znázorňují výsledky po ručním přeskládání jednotlivých dimenzí (posunem kotev). Automobily jsou navíc obarveny podle jejich ceny. Cílem je nalézt atributy aut, které nejlépe předpovídají, do které cenové kategorie auto spadne na základě daných atributů.

Je potřeba mít stále na paměti, že se jedná o ztrátovou transformaci.

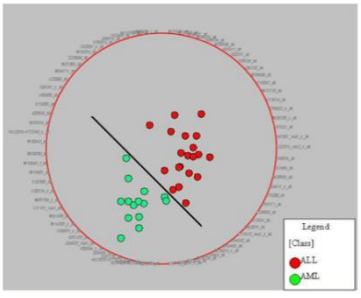
Definice ještě jednou - trochu jinak

Ještě trochu jiný pohled na definici RadViz (Radial Coordinate Visualization) je následující. Každý z parametrů výchozích dat je reprezentován jednou z pevných kotev umístěných na kružnici. Mějme bod $[y_1, y_2, \ldots, y_n]$ definovaný v n-dimenzionálním prostoru. Ke každé kotvě S_j je připevněna virtuální pružina, jejíž tuhost y_j se mění podle hodnoty daného parametru. Všechny pružiny jsou pevně spojeny v jednom bodě *u*. Požadovaným výsledkem je pak vyvážený systém pružin, tzn. suma je rovna nule.



Více zde: https://cyber.felk.cvut.cz/research/theses/papers/216.pdf

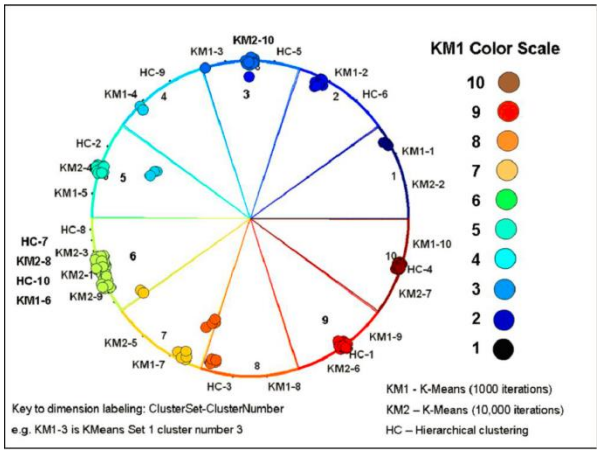
Dalším přístupem je implementace vyhledávacího algoritmu, který se snaží nalézt takové rozložení dimenzí po kružnici, které vede k maximálnímu rozptýlení dat, jako je vidět na obrázku.



Vectorized RadViz (VRV)

Vektorizovaná podoba RadViz (Vectorized RadViz - VRV) konstruuje násobné dimenze z jednotlivých dimenzí. Jako příklad si uveďme rozložení dimenze reprezentující počet válců motoru auta do pěti nových dimenzí: první obsahuje pouze 1 nebo 2 válce, druhá 3 nebo 4 válce, třetí 5 nebo 6 válců, čtvrtá 7 válců a pátá 8 válců. Počet nových dimenzí může být určen algoritmicky nebo ručně. Tento proces je velmi podobný metodě třídění dat do košů (např. podle nízké, střední a vysoké ceny automobilů).

Každá původní dimenze je tedy reprezentována vektorem nových dimenzí, kdy každá nová souřadnice v takovém vektoru nabývá hodnoty 0 nebo 1 podle toho, zda daný záznam obsahuje hodnotu odpovídající této dimenzi nebo ne. Proto pro každý takový záznam obsahuje každý nový vektor právě jednu dimenzi obsahující hodnotu 1 a všechny ostatní mají hodnotu 0. Příklad VRV je ukázán na následujícím obrázku.

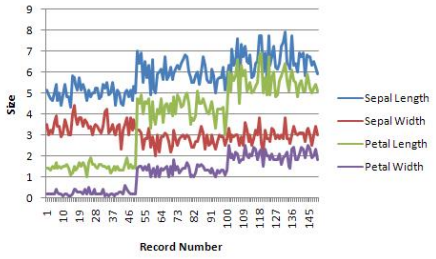


Techniky pro čárová data

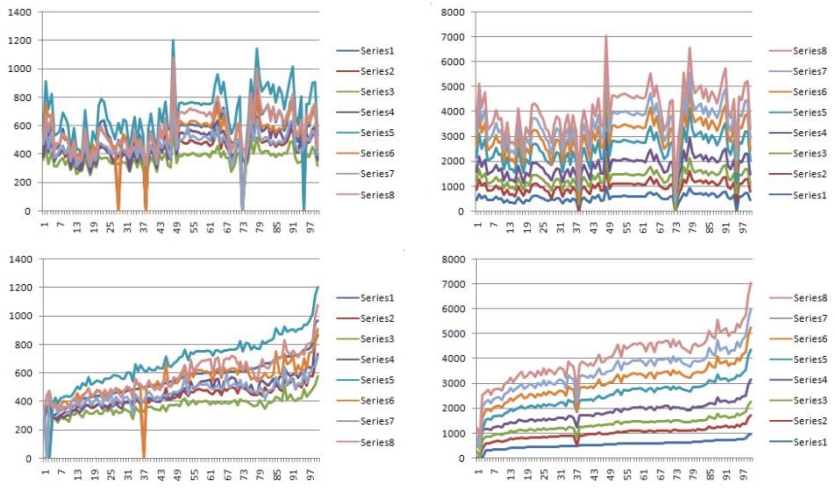
Metody pro vizualizaci bodových dat reprezentovaly každý záznam pomocí značky. Techniky pro čárová nebo úsečková data zobrazují záznamy tak, že spojují odpovídající body přímou nebo zakřivenou čarou. Tyto čáry nejen zdůrazňují vztahy mezi datovými hodnotami, ale zároveň předávají další vnímatelné vlastnosti pomocí různých zkosení, zakřivení, křížení a dalších charakteristik úsečkových vzorů. Opět popíšeme některé metody věnující se tomuto typu zobrazení.

Čárový graf je vizualizační technika o jedné proměnné, při které vertikální osa reprezentuje možný rozsah hodnot proměnných a horizontální osa reprezentuje jisté uspořádání záznamů v dané datové množině.

Většina technik pro zobrazení jedné proměnné může být rozšířena pro více proměnných (multivariate data) – pomocí již známých technik superimposition nebo juxtaposition. Nejpoužívanější technikou v tomto případě jsou právě čárové grafy, které jsou schopny pro rozumné množství dimenzí vykreslit data za použití běžné sady os. Další dimenze se rozlišují pomocí barvy, typu vykreslované čáry, její šířky nebo dalších grafických atributů (viz obrázek).



Při zvyšování počtu dimenzí, nebo pokud již dochází k velkému překryvu dat, stává se použití techniky skládání (superimpositioning) problematické. Obrázek vlevo nahoře znázorňuje 8-dimenzionální datovou množinu (platy na fakultách pro různé funkce na 100 různých univerzitách). Je zřejmé, že při tomto klasickém zobrazení pomocí superimpositiningu je obtížné se v datech vyznat. Avšak pro zlepšení interpretace je možné použít některé strategie. Obrázek vpravo nahoře tzv. vrstvený čárový graf (stacked line chart), kdy namísto použití společné základny pro vykreslení je pro každou další dimenzi použit jako základ graf předchozí dimenze.



Obrázky vlevo a vpravo dole ukazují použití jiné strategie – třídění záznamů podle jedné dimenze.

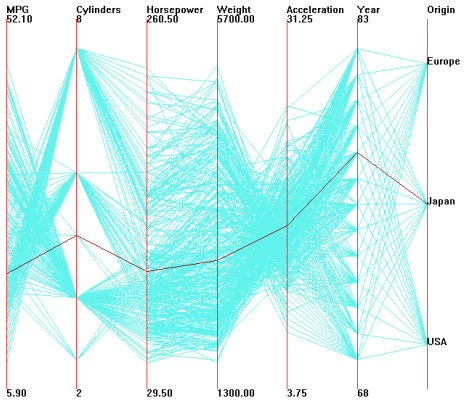
Efektivita výše uvedených příkladů závisí z velké části na skutečnosti, že dimenze mají společné jednotky v osách. Pokud mají jednotlivé proměnné (odpovídající dimenzím) různé jednotky, stává se situace výrazně složitější. Jeden z běžných přístupů je využití násobných vertikálních os, kdy každá je označena zvlášť. Dalším možným přístupem je vytvoření sady grafů, pro každou dimenzi jeden. Tyto grafy pak vertikálně naskládáme (obvykle po aplikaci škálování ve vertikální dimenzi, aby bylo možné většinu grafů zobrazit současně).

Paralelní souřadnice

Paralelní souřadnice (graf paralelních souřadnic je označován jako PCP – parallel coordinates plot) byly poprvé zavedeny Inselbergem v roce 1985 jako mechanismus pro studium geometrie o vyšších dimenzích. Od té doby se řada dalších vědců, včetně samotného Inselberga, zabývala rozšířeními PCP pro jejich použití pro analýzu multivariate dat.

Základní myšlenkou paralelních souřadnic je, že osy jsou místo ortogonálního umístění rozmístěny paralelně za sebou. Osy jsou reprezentovány rovnoměrně rozloženými vertikálními nebo horizontálními čarami, které reprezentují příslušné uspořádání jednotlivých dimenzí.

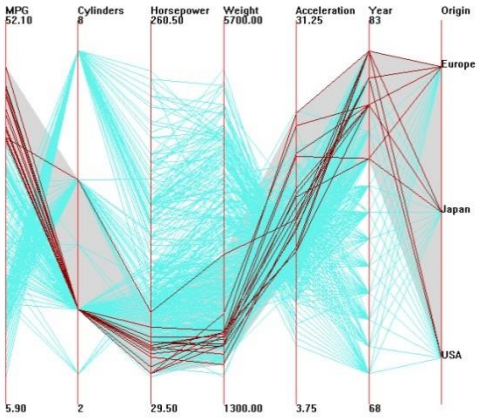
Datový bod je vykreslen jako polyčára, která protíná každou osu na pozici úměrné své hodnotě v odpovídající dimenzi.



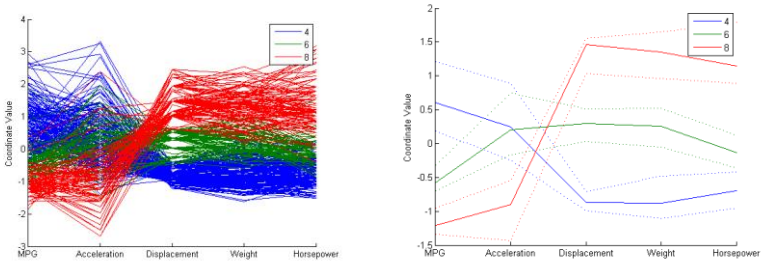
Obrázek ukazuje příklad 7-dimenzionální datové množiny zobrazené pomocí paralelních souřadnic. V obrázku je pro ilustraci zvýrazněn jeden datový bod – v podobě polyčáry.

Pro interpretaci grafu se díváme na podobné čáry (indikující korelaci mezi páry dimenzí), podobné průsečíky a čáry, které jsou buď izolované, nebo mají výrazně odlišný sklon od svých sousedů. Problém paralelních souřadnic spočívá v tom, že dokáží zobrazit vztahy pouze mezi dvojicemi dimenzí. Pro překonání tohoto omezení je možné využít interaktivní výběr a zvýrazňování záznamů, které umožní uživateli vidět vztahy, které pokrývají všechny dimenze.

Jako příklad si uveďme následující obrázek. Čáry zobrazené tmavě červenou barvou byly izolovány pomocí tažení myši přes vysoké hodnoty souřadnice MPG (spotřeba), čímž vybereme záznamy spadající do tohoto rozsahu v dané dimenzi. Světle šedé regiony identifikují obsah N-dimenzionální oblasti, který obsahuje vybrané body.



Při velkém množství dat se stávají paralelní souřadnice nepřehledné. V takovém případě se může zobrazit pouze střední hodnota daných klastrů. Nebezpečím však je, že takto eliminujeme mezní hodnoty, které jsou často velmi zajímavé.



Vědci v posledních desetiletích, kdy se zabývali paralelními souřadnicemi, významně rozšířili možnosti paralelních souřadnic. Některé z těchto technik jsou:

- Hierarchické paralelní souřadnice – zobrazují datové klastry namísto původních dat
- Použití poloprůhledných čar pro odhalení klastrů v rozsáhlých datových množinách
- Klastrování, přeskupování a rozmístění os na základě korelace
- Přeskupení os za účelem zlepšení vizuálního uspořádání
- Shlukování dat do pásů klastrů
- Zahrnutí histogramů do os
- Napasování křivek na průsečíky pro zlepšení interpretace spojitosti přes osy

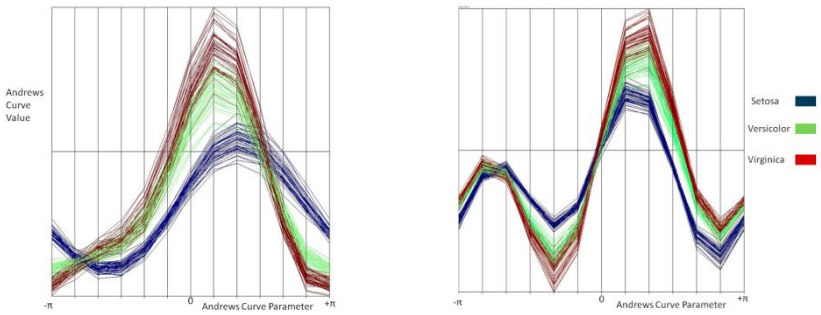
Andrewsovy křivky

Další technikou pro vizualizace multivariate dat pomocí čar jsou tzv. Andrewsovy křivky. Byly vyvinuty v roce 1972 Davidem F. Andrewsem. Každý multivariate datový bod $D = (d_1, d_2, ..., d_N)$ je použit pro vytvoření křivky ve tvaru:

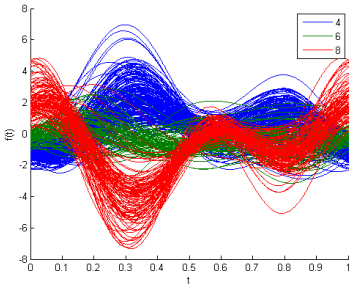
$$f(t) = \frac{d_1}{\sqrt{2}} + d_2 \sin(t) + d_3 \cos(t) + d_4 \sin(2t) + d_5 \cos(2t) + \ldots$$

Pro lichý počet dimenzí je poslední člen d_N ve tvaru $\cos\left(\frac{N-1}{2}t\right)$, zatímco pro sudý počet dimenzí je to $\cos\left(\frac{N}{2}t\right)$.

Podobně jako u ostatních vizualizačních technik pro multivariate data, pořadí jednotlivých dimenzí může mít významný vliv na výslednou Andrewsovu křivku. Obrázky ukazují stejná data, pouze je jiné pořadí jednotlivých dimenzí.

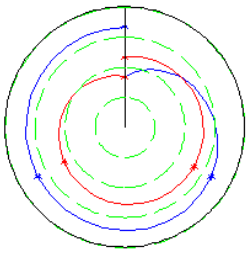


V následujícím Andrewsově grafu je každé pozorování vyjádřeno hladkou funkcí z intervalu [0,1].

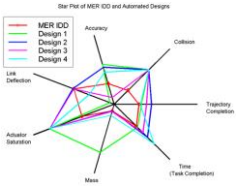


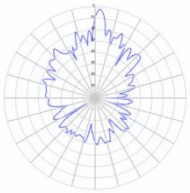
Techniky radiální osy

Pro každou techniku, která má souřadný systém orientován horizontálně a/nebo vertikálně, existuje ekvivalentní technika využívající radiální orientaci. Například **kruhový čárový graf** je takový graf, kde vykreslené čáry představují offset z kruhové základny (viz obrázek).

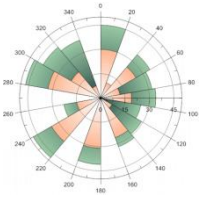


Velký graf může být transformován rozdělením na stejně velké segmenty a mapováním každého segmentu na základnu o různém poloměru. Toto je užitečné zejména pro studování cyklických událostí. Jednotlivé varianty kruhových čárových grafů zahrnují rovněž **radar** a **hvězdicové grafy**. Kromě těchto populárních technik byla vyvinuta řada dalších kruhových diagramů, jako například:

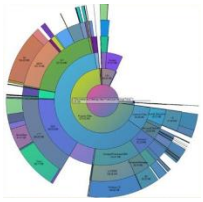




- **Polární grafy** – grafy zobrazující polární souřadnice
- **Kruhové sloupcové diagramy** – podobné jako kruhové čárové grafy, pouze namísto čar jsou zobrazeny sloupce

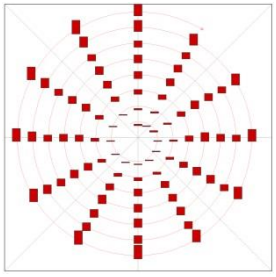


- **Kruhové plošné grafy** – podobně jako čárové grafy, plocha pod čarami je navíc vyplněna barvou nebo texturou
- **Kruhové sloupcové grafy** – sloupce jsou reprezentovány kruhovými oblouky se společným středem. Rozdíl mezi kruhovým sloupcovým diagramem a grafem je ten, že v jednom je sloupec rovný a základna je zakřivená, zatímco ve druhém je to přesně naopak.



Typy technik pro radiální osy

Veškeré techniky využívající radiální osy a zahrnující více než jednu kružnici využívají buď soustředné kružnice, nebo spojitou spirálu. Jako příklad sloupcového grafu se spirální základnou uvádíme následující obrázek.



Tato metoda na rozdíl od soustředných kružnic nevykazuje nespojitosti na konci každého cyklu. Porovnání uvnitř a mezi jednotlivými cykly je poměrně jednoduché, obzvláště v případě, kdy jsou sloupce orientovány podél vertikální osy (jako na obrázku) místo kolmé orientace na spirálu.

Díky znalosti lidského vnímání víme, že v tomto případě je měření rozdílu mezi sousedními prvky obtížnější než při použití společné základny (tradiční sloupcový graf). Avšak tradiční

sloupcové vyjádření nám nedovoluje jednoduše sledovat vzory mezi jednotlivými prvky na stejné pozici v různých cyklech.

Techniky pro plošná data

U technik pro plošná data jsou pro zobrazení hodnot využity vyplněné polygony o dané velikosti, tvaru, barvě a dalších atributech. Ačkoliv díky chybám lidského vnímání víme, že naše schopnost přesně interpretovat plochu je horší než schopnost měřit jiné atributy, jako například délku, přesto byla vyvinuta řada velmi efektivních technik pro zobrazení dat této kategorie. Cílem některých z těchto technik není ukázat samotná hrubá data, ale jejich shluky nebo rozložení hodnot.

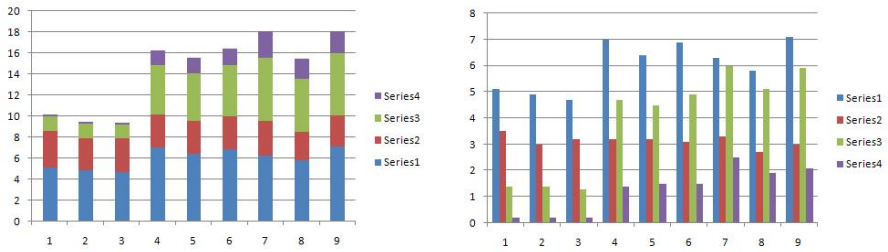
Mnoho těchto technik pro vizualizaci plošných dat bylo původně navrženo pro univariate data (o jedné proměnné), jako jsou například koláčové nebo sloupcové grafy. Některé z nich byly rozšířeny o více dimenzí. Nyní si uvedeme některé z těchto metod.

Sloupcové diagramy/histogramy

Jednou z nejpoužívanějších vizualizačních technik, kromě čárových grafů, bodových grafů a map, je **sloupcový diagram**, kde jsou pro zobrazení numerických hodnot využity obdélníkové sloupce. Jejich efektivita vyplývá z toho, že lidské vnímání je dobře přizpůsobeno na rozpoznávání délky a obecně lineárních vlastností. Proto jsou sloupcové grafy běžně využívány pro zobrazení různých typů dat. Běžně jsou používány jak horizontální, tak vertikální sloupce. Pokud má být danému sloupci přiřazen textový popisek, je z důvodu možné délky popisku jednodušší zobrazení pomocí horizontálních sloupců. Avšak otočením popisků o 90 stupňů můžeme dosáhnout jednoduchého použití i pro vertikální sloupce.

Jedním ze zásadních rozhodnutí, které je třeba při návrhu sloupcových diagramů učinit, je určení, kolik sloupců je zapotřebí pro co nejlepší reprezentaci dat. Pokud sloupce reprezentují stav N proměnných a pokud N není příliš velké, je možné použít mapování 1:1 mezi proměnnými a sloupci. Jestliže je cílem zobrazit souhrn nebo rozložení datové množiny, můžeme využít **histogram** pro zaznamenání počtu výskytů datových hodnot. Pokud data obsahují nominální hodnoty, pak je rozhodnutí jednoduché. Máme tolik sloupců, kolik je různých hodnot. Pro spojitá data nebo hodnoty typu integer o velkém rozsahu je nutné data rozdělit do intervalů hodnot a každému intervalu přiřadit jeden sloupec.

Pokud zobrazujeme multivariate data, máme několik možností, jak použít sloupcové diagramy. Běžnou technikou je vrstvený sloupcový graf, kde se každý sloupec skládá z několika kratších sloupců reprezentujících hodnoty v každé dimenzi. Pro jejich rozlišení se běžně používá barva, textura a další (viz obrázek vlevo).

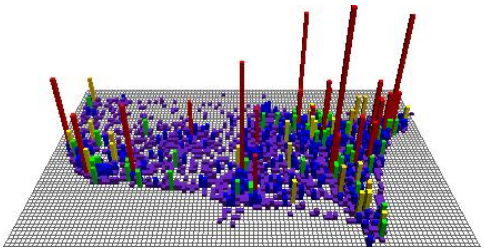


Podobný přístup zobrazuje sloupce pro jednotlivé proměnné těsně vedle sebe (viz obrázek vpravo). Sloupce tedy mají společnou základnu, čímž se zjednodušuje jejich interpretace.

Výběr mezi těmito dvěma přístupy často záleží na počtu proměnných a na počtu sloupců. Skládané sloupce nevyžadují dodatečné nároky na horizontální prostor, zatímco druhá technika „sousedních“ sloupců může vyžadovat mnohem větší prostor v tomto směru.

Cityscapes

Jednou z verzí 3D sloupcových grafů jsou tzv. cityscapes, se kterými jsme se setkali již dříve. V nich jsou místo 2D obdélníků využity 3D kvádry. Sloupce jsou rozloženy na mřížce, kdy dvě dimenze dat jsou využity na umístění příslušného sloupce na povrch mřížky. Další dimenze jsou využity na ovládání velikosti a barvy skládaných kvádrů. Cityscapes získaly svůj název díky tomu, že výsledná vizualizace často vypadá jako budovy ve městě. Pokud jsou obsazeny všechny buňky mřížky, je takový graf někdy nazýván **3D histogram**.



Problémy 3D sloupcových grafů

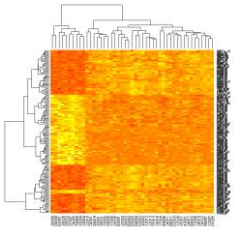
Problém se zobrazením sloupců ve 3D spočívá v jejich častém překrytí (viz obrázek). Pro eliminaci tohoto překryvu existuje řada různých technik. Jedna z možných poskytuje uživateli možnost rotovat se scénou, čímž je možné učinit zakryté sloupce viditelnými. Další možností je zmenšit tloušťku sloupců, čímž snížíme plochu, kterou sloupec zabírá, což vede rovněž ke snížení počtu sloupců, které může daný sloupec zakrývat. Třetí přístup spočívá ve změně průhlednosti jednotlivých sloupců.

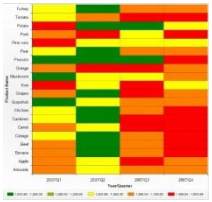
Všechny tyto metody mají své nedostatky, nicméně cityscapes metoda je i přes to velmi oblíbenou vizualizační technikou, zejména pro geografická data.

Tabulková zobrazení

Multivariate data jsou často uložena v tabulkách, proto byla vyvinuta řada vizualizačních technik, které pracují s těmito strukturami. Tyto techniky se většinou liší typem interakcí, které podporují.

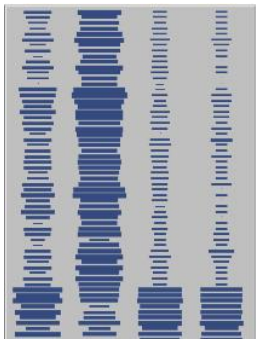
Jedním z příkladů jsou tzv. **heatmapy**. Vznikají zobrazením tabulky záznamů za použití barvy namísto textu. U této vizualizační techniky jsou všechny datové hodnoty mapovány na stejný normalizovaný barevný prostor a každá hodnota je renderována jako barevný čtverec či obdélník. Použití různých barevných map společně s povolením uživateli roztahovat či zmenšovat barvy pro zdůraznění nebo naopak potlačení některých rozsahů hodnot významně zvyšuje použitelnost této techniky.





Permutace či přeskládatelné mřížky jsou v podstatě heatmapy, které umožňují reorganizovat řádky a sloupce za účelem odhalení určitých vlastností dat. Sloupce a řádky mohou být reorganizovány, aby maximalizovaly diagonalizaci – vytvoření matice s buňkami zarovnanými podél hlavní diagonály. Jiné varianty přeskládají data za účelem izolování klastrů s podobnými hodnotami nebo vzorců v datových hodnotách.

Další technikou jsou tzv. **survey plots** (přehledové grafy). Jsou variantou permutační matice, kdy namísto obarvování buněk pracujeme s jejich velikostí. Navíc středy buněk zarovnáváme na jednotlivé atributy. To zmírňuje chyby ve vnímání barvy způsobené různými nežádoucími efekty sousedních barev.

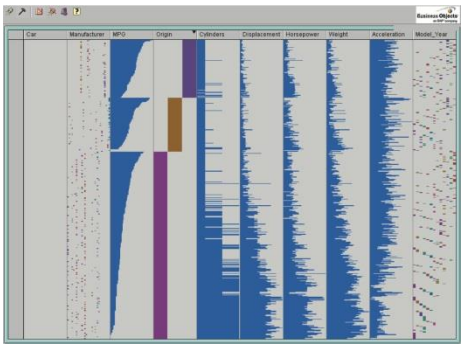


Avšak protože měření plochy je mnohem více náchylné k chybě než měření délky, má tato metoda rovněž své chyby.

Obrázek ukazuje survey plot spočtený pomocí nástroje DataLab. Každý sloupec je vizuální reprezentací jedné z čtyř dimenzí datové sady pro kosatce.

Konečně si ukážeme techniku, která kombinuje dosavadní přístupy a poskytuje level-of-detail mechanismus poskytující zoomování za účelem zobrazení celé tabulky v podobě několika různých pohledů.

Data mohou být tímto způsobem zobrazena různě, podle toho, jak velký prostor obrazovky uživatel alokuje pro zobrazení daného řádku nebo sloupce. Třídění sloupců pomáhá rychle identifikovat trendy a korelace v datech (viz obrázek).



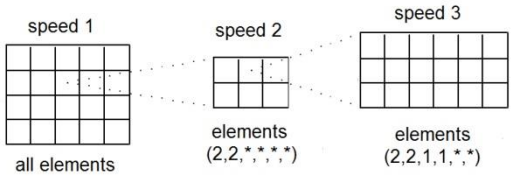
Skládání (stacking) dimenzí

Technika skládání dimenzí byla vyvinuta LeBlancem a spol. a je zaměřena na mapování dat z diskrétního N-dimenzionálního prostoru do 2D obrázku takovým způsobem, že se minimalizuje zakrytí (okluze) dat za současného zachování většiny prostorové informace.

Ve stručnosti je mapování prováděno následovně:

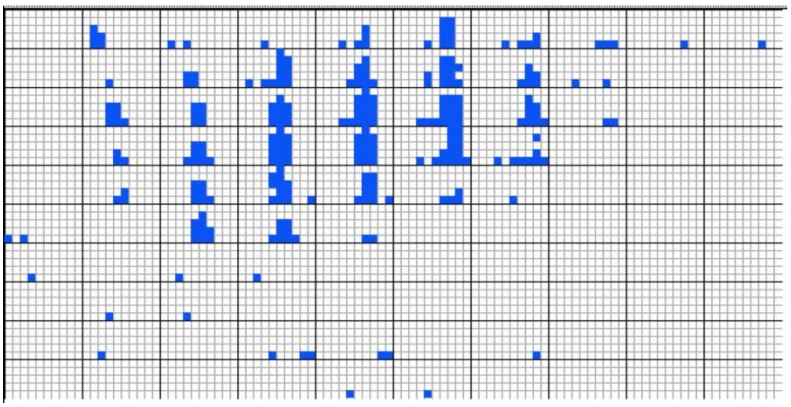
- Začínáme s daty dimenze 2N + 1 (pro sudý počet dimenzí je nutné dodat dodatečnou implicitní dimenzi kardinality 1).
- Vybereme konečnou kardinalitu pro každou dimenzi.
- Jednu z dimenzí vybereme jakou závislou proměnnou. Zbytek je považován za nezávislé proměnné.
- Nyní vytvoříme uspořádané dvojice nezávislých proměnných (N párů) a každému páru přiřadíme jeho jedinečnou hodnotu (označovanou jako *rychlost*) od 1 do N. Dvojice odpovídající rychlosti 1 vytvoří virtuální obraz, jehož velikost odpovídá kardinalitě dimenzí (první dimenze z dvojice je orientována horizontálně, druhá vertikálně). V každé pozici tohoto virtuálního obrazu je vytvořen další virtuální obraz, který odpovídá dimenzím o rychlosti 2. Opět, velikost tohoto obrazu je závislá na kardinalitě odpovídajících dimenzí. Tento proces je opakován, dokud nejsou zahrnuty všechny dimenze. Tímto způsobem dosáhneme toho, že každé umístění v prostoru o mnoha dimenzích má svoje unikátní umístění ve 2D obrázku, který je výsledkem mapování.

Hodnota závislé proměnné v daném umístění v prostoru o mnoha dimenzích je poté mapována na barvu/intenzitu tohoto místa ve 2D obrázku. Celý proces ilustruje obrázek. Je zobrazena 6-ti dimenzionální datová sada, kde dimenze d_1 až d_6 mají kardinality 4, 5, 2, 3, 3, a 6.



Jinými slovy, tato technika funguje následovně. Začíná se diskretizací rozsahů v každé dimenzi. Každé dimenzi je pak přiřazena orientace a uspořádání. Dimenze se dvěma nejnižšími uspořádáními se použijí pro rozdělení virtuální obrazovky na sekce, přičemž kardinalita dimenzí určuje, kolik sekcí v horizontální a vertikální ose je generováno. Další takto vytvořená sekce je použita pro rekurzivní definici virtuální obrazovky v dalších dvou dimenzích stejným způsobem. Tento proces se opakuje, dokud nejsou zpracovány všechny dimenze a data nejsou umístěna na jejich pozici v obrazovce.

Jako příklad si uveďme 4D data vizualizována pomocí skládání dimenzí.



Skládání dimenzí může být zobrazeno pomocí N-dimenzionálního histogramu, jestliže je barva buňky nastavena úměrně datovým hodnotám, které jsou na ni mapovány.

Na podobném principu jsou založeny metody Worlds-within-worlds nebo treemaps, kterými se budeme zabývat v dalších přednáškách.

Kombinace technik

Kromě základních technik založených na bodech, čarách a plochách existuje řada hybridních technik kombinujících znaky těchto předchozích technik. Ukážeme si dvě nejznámější techniky tohoto typu: **glyfy** (piktogramy) a tzv. **dense pixel displays**.

Glyfy a ikony

V kontextu vizualizace dat a informace je za glyf považována vizuální reprezentace části dat nebo informace, kde je grafická entita a její atributy řízeny jedním nebo více atributy vstupních dat. Například šířka a výška kvádru může být řízena výsledky studenta v pololetí a na konci roku, zatímco barva může být asociována s pohlavím studenta.

Definice je hodně obecná, protože zahrnuje značky bodových grafů, sloupce histogramu nebo dokonce celé čáry v grafu.

Mnoho autorů vyvinulo seznamy grafických atributů, na které mohou být datové hodnoty mapovány. Ty zahrnují pozici (1D, 2D, 3D), velikost (délka, plocha, objem), tvar, orientaci, materiál (jas, sytost, intenzita, textura, průhlednost), styl čáry (šířka, pomlčky, sbíhání) a dynamiku (rychlost pohybu, směr pohybu, rychlost blikání).

Nyní si uvedeme řadu možných mapování na různé typy glyfů, včetně:

- Mapování 1:1, kde je každý datový atribut mapován na jednoznačné a odlišné grafické atributy
- Mapování 1:mnoho, kde je pro dosažení větší přesnosti a jednoduchosti interpretace použita sada redundantních mapování
- Mapování mnoho:mnoho, kde je několik nebo všechny datové atributy mapovány na společný typ grafického atributu, odděleného v prostoru, orientací nebo jiným typem transformace

Mapování 1:1 je často navrženo tak, aby využívalo znalostí uživatele – využívá se intuitivní párování dat na grafické atributy za účelem zjednodušení procesu porozumění. Příklady zahrnují mapování barvy na teplotu nebo mapování směru toku na orientaci čar.

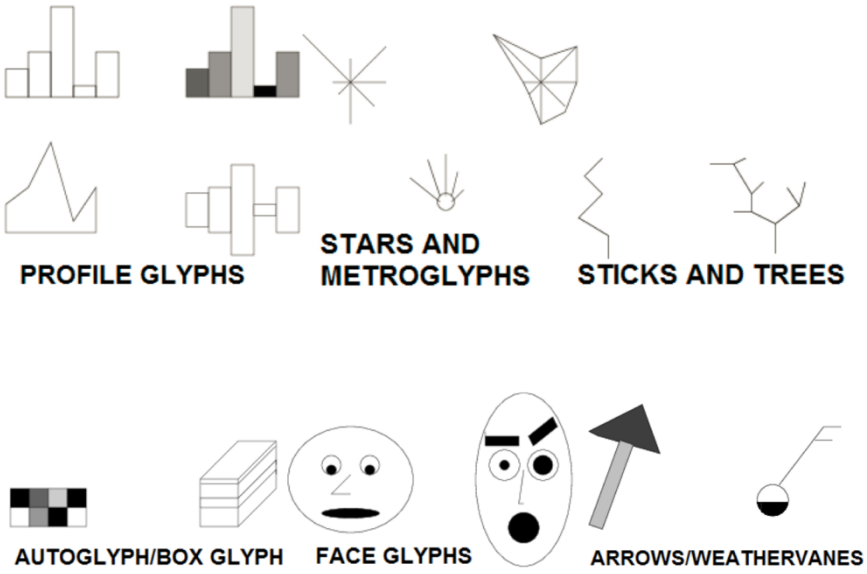
Redundantní mapování může být užitečné v situacích, kdy je počet dimenzí vstupních dat nízký a cílem je snížit možnost špatné interpretace na minimum. Příkladem je mapování populace zároveň na velikost a barvu, čímž zpřístupníme analýzu i lidem poruchou vnímání barvy a navíc zpřesníme porovnání dvou populací s podobnými hodnotami.

Mapování mnoho:1 je nejvýhodnější v situacích, kde je důležité nejen porovnat hodnoty rozdílných záznamů ve stejné dimenzi, ale rovněž porovnat různé dimenze stejného záznamu. Příkladem je mapování každé dimenze na výšku vertikálního sloupce umožní jak porovnání uvnitř záznamů, tak mezi nimi.

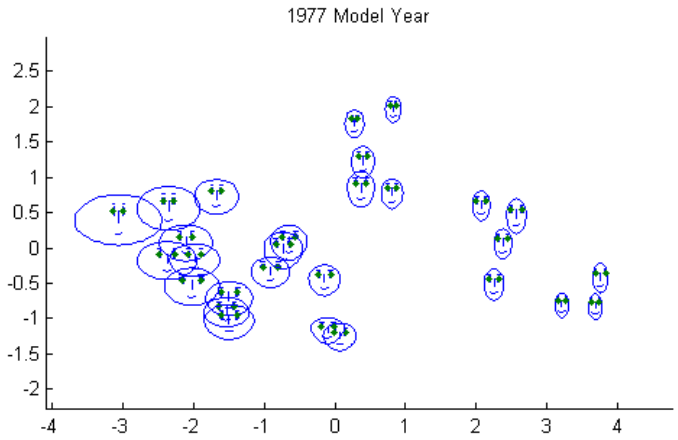
Následující seznam obsahuje podmnožinu glyfů, které byly navrženy v různých studiích a běžně se používají. Některé z nich jsou navrženy speciálně pro určité aplikace, jako například vizualizace toku, zatímco ostatní jsou obecné.

- Profily – výška a barva sloupců
- Hvězdicovité glyfy – délka stejnoměrně rozmístěných paprsků vycházejících ze středu
- Metroglyfy – délka paprsků
- „Tyčinkové“ obrázky – délka, úhel a barva větví
- Stromy – délka, tloušťka, úhly a větve; struktura větví je odvozena z analýzy vztahů mezi dimenzemi
- Autoglyfy – barva kostek
- Kvádry – výška, šířka, hloubka prvního kvádru + výška následujících kvádrů
- Ježci – „trny“ vektorového pole s různou orientací, tloušťkou a zúžením
- Tváře – velikost a pozice očí, nosu, úst, zakřivení úst, úhel obočí
- Šipky – délka, šířka, zúžení, barva základní čáry a samotné šipky
- Polygony – zvýrazňující lokální deformace ve vektorovém poli pomocí změn v orientaci a tvaru
- Dashtubes (přerušovaná trubka) – textura a průhlednost pro zobrazení data vektorových polí
- Weathervanes (větrná korouhev)
- Kruhové profily – vzdálenost od středu k vrcholům pod stejnými úhly
- Barevné glyfy – barevné čáry napříč krychle
- Bugs (brouci) – tvar křídel řízen průběhem v čase, délka tykadel, velikost a barva těla, velikost značek na těle
- Wheels (kola) – 3D kolo mapuje čas na výšku, hodnotu proměnné na poloměr
- Boids (hejna ptáků) – tvar a orientace primitiv pohybujících se v časově proměnném poli
- Procedurální tvary – „blobby“ objekty řízené až 14-ti dimenzemi
- Glyphmaker – mapování řízené uživatelem

- Icon Modeling Language – atributy 2D kontury a parametry, které ji extrahují do 3D a dále ji transformují či deformují



Chernoffovy obličeje



Při používání glyfů v oblasti information visualization si musíme být vědomi řady nepřesností a omezení této techniky. Nejdůležitější jsou nepřesnosti ve vnímání, které závisí na tom, jaké grafické atributy jsme použili. Některé atributy, jako například délka čáry, mohou být posuzovány mnohem přesněji než ostatní, jako například orientace nebo barva.

Další zdroje nepřesností plynou například ze skutečnosti, že vztahy mezi sousedními grafickými atributy jsou mnohem lépe interpretovatelné než ty, které jsou ve větší vzdálenosti. Podobným způsobem funguje porovnání dvou glyfů. Pokud jsou umístěny blízko sebe na obrazovce, je jejich porovnání jednodušší, než když jsou ve větší vzdálenosti.

Konečně počet dimenzí dat a záznamů, které je možné efektivně zobrazit pomocí glyfů, je omezen.

Pokud již máme vybrán typ glyfu, který chceme použít, existuje N! různých uspořádání dimenzí, které mohou být při mapování použity. Existuje několik strategií pro volbu vhodného uspořádání:

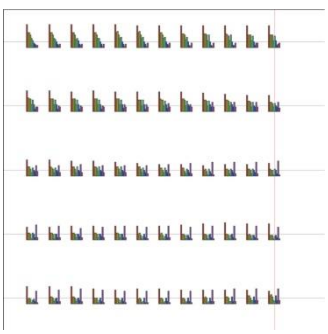
- Dimenze mohou být tříděny na základě jejich korelace – podobné dimenze jsou mapovány na sousední hodnoty. To pomáhá odhalit obecné trendy v datech.
- Dimenze mohou být mapovány takovým způsobem, že zvýšíme vliv glyfů se symetrickým tvarem, které jsou jednodušší pro vnímání a zapamatování. Tvary, které jsou méně symetrické než jejich sousedé, rovněž dominují.
- Dimenze mohou být roztříděny podle jejich hodnot v jednom záznamu. Například pokud data obsahují multivariate časové úseky, pak třídění na základě prvního záznamu může zvýraznit trendy v čase, kdy je vidět, které vztahy mezi dimenzemi jsou trvalé nebo se naopak významně mění v čase.
- Dimenze mohou být tříděny ručně na základě znalostí uživatele o dané doméně. Sémanticky podobné dimenze mohou být shlukovány pro zjednodušení interpretace.

Poslední důležitá úvaha při návrhu vizualizace pomocí glyfů je rozmístění glyfů na obrazovce. Existují tři základní typy strategií pro rozmístění:

1. Uniformní
2. Řízené daty
3. Řízené strukturou

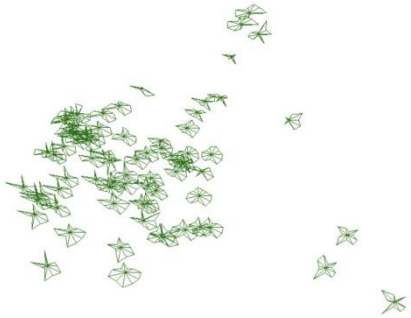
Uniformní rozmístění

Glyfy jsou škálovány a rozmístěny rovnoměrně po obrazovce (stejné mezery mezi glyfy). Tato strategie eliminuje překryvy a zároveň je efektivně využit prostor obrazovky. Různá třídění záznamů odhalují různé vlastnosti dat (viz obrázek).



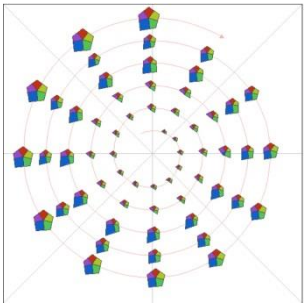
Rozmístění řízené daty

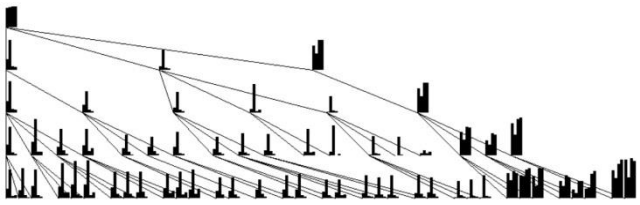
Datové hodnoty jsou využity pro řízení rozmístění glyfů. Zde jsou možné dva přístupy. V prvním jsou vybrány dvě (pro 3D zobrazení tři) dimenze, které řídí rozmístění. Ve druhém přístupu jsou pozice odvozeny s využitím algoritmů, jako například PCA a MDS.



Rozmístění řízené strukturou

Pokud mají data implicitní nebo explicitní strukturu, jako například cyklickou nebo hierarchickou, může být tato informace využita pro řízení rozmístění dat. Například glyfy mohou být rozmístěny do spirály nebo mřížky.





Dense Pixel Displays

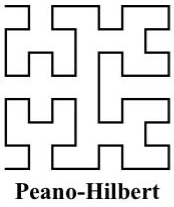
Dense pixel displays, známé také jako pixelově orientované techniky, jsou hybridní metodou na pomezí bodových a regionálních (plošných) metod. Technika byla vyvinuta Keimem a jeho kolegy a mapuje každou hodnotu na jednotlivé pixely a pro každou dimenzi vytváří vyplněný polygon. Tyto typy zobrazení maximálně využívají prostor obrazovky, přičemž umožňuje zobrazení milionů hodnot na jediné obrazovce. Každá datová hodnota řídí barvu jednoho pixelu – změnou použité barevné mapy můžeme potenciálně odhalit nové vlastnosti dat. Pokud máme dánu vstupní datovou množinu a barevnou mapu, je nutné ještě vyřešit rozmístění datových záznamů a jejich uspořádání.

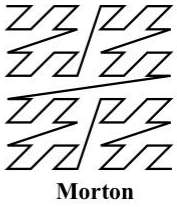
Ve své nejjednodušší formě každá dimenze datové množiny generuje oddělený „podobrázek“ na obrazovce. Takto můžeme každou dimenzi považovat za nezávislou sadu čísel, kdy každá řídí barvu odpovídajících pixelů. Poté je nutné rozmístit prvky v těchto sadách takovým způsobem, že zdůrazníme vztahy mezi body, které jsou v sadě blízko sebe.

Například vytvoříme podobrázek, kde střídáme průchod zleva doprava a zprava doleva, přičemž pokud dosáhneme kraje podobrázku, posuneme se o jednu řadu níže.

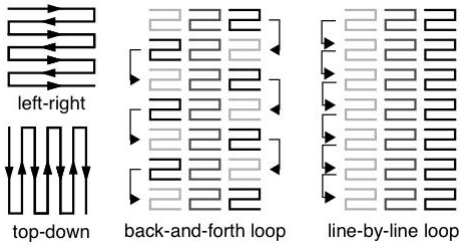
Další možností je využití spirálového rozmístění, kdy je první datový bod umístěn do středu podobrázku a následné body jsou rozmístěny soustředných čtvercích.

Existuje celá řada různých způsobů rozmístění, některé z nich jsou uvedeny na obrázcích.





vyplnění obrazovky

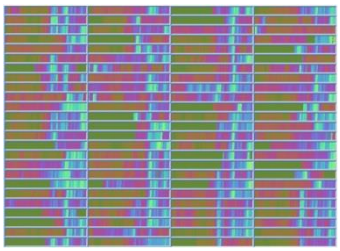


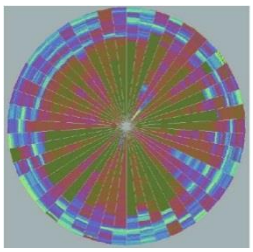
rekurzivní vzory

Podobrázky odpovídající datům pro každou dimenzi mohou být umístěny na obrazovku různými způsoby. Nejjednodušší je vytvoření mřížky podobrázků, které maximalizuje využití obrazovky. Mřížky mohou mít různá uspořádání na základě uspořádání dimenzí, což umožňuje odhalit korelace mezi dimenzemi.

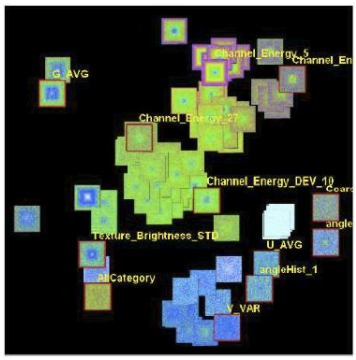
Technika zvaná **rekurzivní vzory** (recursive pattern) využívá právě mřížkové rozmístění podobrázků.

Jinou variantou jsou tzv. **kruhové segmenty** (circle segments), kde místo rozmístění pixelů do obdélníkových podobrázků umisťujeme pixely do kruhových „klínů“. Začínáme ve středu kruhu a proplétáme se tam a zpět směrem od středu. Každá dimenze zabírá N-tinu kruhu, kde N je počet dimenzí.





Další přístupy pro rozmístění podobrázků na obrazovku již povolují překrytí. Příkladem je technika „Value and Relation“ Yanga a spol, která využívá multidimensional scaling pro umístění podobných dimenzí na obrazovku společně. To umožňuje zobrazit klastry dimenzí a význačná jednotlivá data.

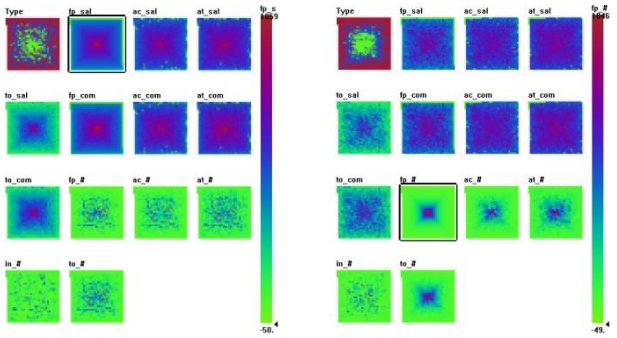


Posledním důležitým tématem při návrhu pixel-oriented displays je uspořádání dat. Pro některé typy dat, jako například časové sekvence, je uspořádání předurčeno a je fixní. Avšak v ostatních případech může přeskládání záznamů odhalit mnoho zajímavých vlastností.

Pokud jsou například data uspořádána na základě jedné z dimenzí, objeví se klastry hodnot v této dimenzi. Stejně tak to platí pro ostatní dimenze.

Dalším možným přístupem je uspořádání záznamů na základě jejich N-dimenzionální vzdálenosti od vybraného bodu.

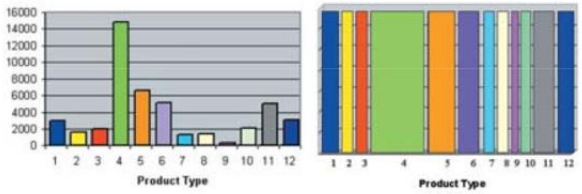
Obrázek ukazuje stejná data, která se liší uspořádáním záznamů.



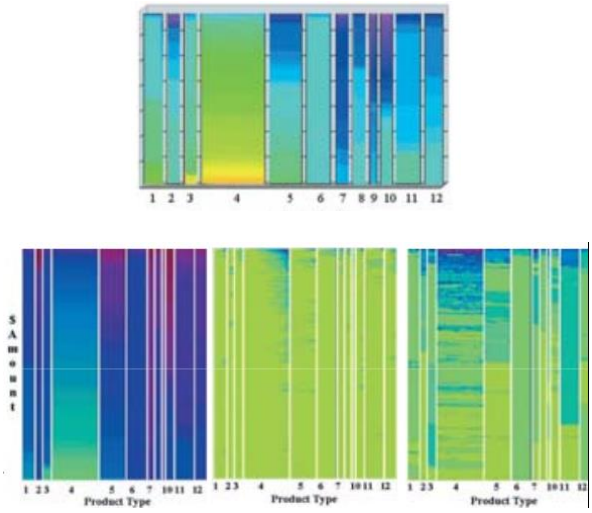
Pixelové sloupcové diagramy

Dense pixely mohou být rovněž umístěny do standardního sloupcového diagramu. Abychom efektivně využili prostor obrazovky, pixelové sloupcové diagramy často využívají šířku sloupce namísto jeho výšky pro reprezentaci agregovaných parametrů dat. Navíc jsou sloupce obarveny pixel po pixelu, abychom mohli zobrazit detailní informaci o jednotlivých hodnotách dat agregovaných ve sloupcích.

- Přetížení klasického sloupcového diagramu – zahrnutí více informací o jednotlivých prvcích.



- Každý pixel sloupce odpovídá datovému bodu patřícího do skupiny reprezentované tímto sloupcem



Příklad uvádí vztah typu produktu vůči ceně. Barva je mapována na:

a) Utracenou částku
b) Počet návštěv
c) Velikost prodejů

Další obrázek ukazuje několik pixelových sloupcových diagramů, které využívají stejné rozmístění pixelů uvnitř sloupců (rozdělení podle měsíce, v ose y setříděno podle počtu nákupů a v ose x podle počtu návštěv). Vizualizace umožňuje uživateli pozorovat zajímavá fakta o transakcích, jako například:

- V prosinci byl největší počet zákazníků, zatímco v únoru, březnu a květnu jich bylo nejméně
- Od února do května byl největší počet nákupů
- Počet nákupů v prosinci je průměrný
- Od března do června se zákazníci vraceli častěji než v jiných měsících. Prosincoví zákazníci byli většinou jednorázoví.
- Zákazníci kupující nejvíce se vracejí častěji a kupují více věcí.