VIZUALIZACE VÍCEROZMĚRNÝCH DAT SYMBOLOVÝMI GRAFY

E + M

Jaroslav Myslivec, Hana Skalská

Úvod

Dostupnost technologií a množství dat, zaznamenávaných nebo uchovávaných v nejrůznějších datových zdrojích, vedou k intenzivním snahám nalézat nové možnosti, jak data efektivně využívat. Tento aspekt nabývá na důležitosti proto, že informace z dat jsou považovány za jeden z potenciálních zdrojů dalšího rozvoje. Je přitom známo, že větší množství dat neznamená samo o sobě větší informovanost v běžném životě, ani nepřináší konkurenční výhodu pro firmu, která data vlastní, ani automaticky nevytváří nové poznatky v jednotlivých oborech lidské činnosti, protože neexistují jednoznačné nebo automatizované postupy, které by využití dat zajišťovaly.

Předmětem studia řady oborů jsou proto otázky, jak využít data k získávání informací, které přinášejí nové poznatky a umožní lépe chápat vztahy, souvislosti, závislosti nebo struktury sledovaných jevů nebo událostí. Explorační metody, jejichž součástí je také obrazová reprezentace dat, patří do kategorie metod, souvisejících s touto problematikou.

Grafická prezentace (vizualizace) dat využívá přirozené schopnosti člověka analyzovat a vyhodnocovat zrakové vjemy při porovnávání objektů nebo při posouzení vztahů intuitivním způsobem, avšak efektivně. Další předností je rychlost úsudku. Významná může být též eliminace nebo minimalizace nutnosti numerického porovnávání, při kterém jsou často vyžadovány dodatečné znalosti, například znalost vlastností použitých kvantitativních ukazatelů.

Cílem vizualizace vícerozměrného datového souboru je vytvoření názorného grafického popisu jednotlivých objektů nebo souborů dat. Zde budou uvažována vícerozměrná číselná data uspořádaná do matice, ve které jsou v řádcích zaznamenány objekty a ve sloupcích vlastnosti (atributy, proměnné, rozměry), měřené nebo zaznamenávané na každém objektu. Pro *n* objektů a *m* atributů získáme matici řádu n x m.

Problém zobrazení vícerozměrných dat má několik aspektů. Jedním aspektem je návrh efektivních metod současného zobrazení většího počtu dimenzí na dvojrozměrném prostoru (obrazovka monitoru, stránka papíru). Další problémy jsou spojeny s technickou realizací, návrhem vhodného typu reprezentace, nebo přiřazením proměnných. Metody vizualizace navrhují různé možnosti, jak lze překonat tato omezení využitím barev, tvarů, nebo umístěním zobrazovaných objektů v závislosti na jejich vlastnostech. Metody jsou velmi často doplněny interaktivními nástroji, které umožní sledovat a analyzovat změny v zobrazení při změnách datových bodů.

Východiskem tohoto příspěvku je obsáhlá rešerše možností a modifikací vizuálních metod reprezentace vícerozměrných dat. Výsledkem tohoto studia je klasifikace vizualizačních metod. Klasifikace, která byla sestavena s ohledem na techniky zobrazení, obsahuje přehled a stručné charakteristiky jednotlivých kategorií.

Dále je práce zaměřena na symbolové grafy. Jsou studovány jejich možné varianty, jejich vlastnosti, aplikační výhody i rizika. Je navržena a popsána modifikace symbolového grafu v soustavě souřadnic stanovených pomocí hlavních komponent nebo přímo atributy objektů. Takové uspořádání umožní vizuální identifikaci shluků prvků s podobnými vlastnostmi. Metoda byla implementována do vlastní knihovny programů v jazyce R. Knihovna byla nazvána symbols [25]. Aplikace této knihovny a současně možnosti vybraných symbolových technik jsou ilustrovány pomocí reálných dat, která byla převzatá ze zdroje [29]. Data popisují hodnocení spokojenosti pacientů zdravotnických zařízení na základě 12 měřených dimenzí. Pomocí těchto ilustračních dat jsou prezentovány výhody, nevýhody a možnosti interpretace různých variant symbolových grafů.

1. Klasifikace vizualizačních metod

Vizualizační metody lze klasifikovat podle různých hledisek, například podle: použité zobrazo-

vací techniky, typu dat, řešené úlohy, způsobu interakce a dalších kritérií. Zde uvedená klasifikace je zaměřena na jediné hledisko, kterým jsou používané techniky. Vizualizační metody lze potom třídit do těchto kategorií:

- Jednorozměrné, dvojrozměrné nebo trojrozměrné grafy, například histogramy (sloupcové grafy), výsečové grafy, spojnicové grafy, bodové grafy (x - y nebo x - y - z), krabicové grafy, stromkové a další.
- Symbolové grafy, které každý objekt reprezentují jedním symbolem (ikonou), každý atribut je představován vymezenou částí symbolu. Patří sem polygony (hvězdicové a sluníčkové grafy) [26], obličejové grafy [4], nebo barevné ikony (Color Icons) [22].
- Transformace původního ortogonálního rozměrného souřadnicového systému do systému neortogonálního, zobrazitelného ve dvou dimenzích. Sem patří matice bodových grafů (vícerozměrný souřadnicový systém je rozdělen na dílčí dvojrozměrné grafy), technika rovnoběžných souřadnic [15] (všechny souřadnicové osy jsou rovnoběžné), metoda zobrazení určitého segmentu dat v matici bodových grafů (Prosection Views [11]) a další.
- Pixelové grafy [18] reprezentují každou hodnotu atributu barevným pixelem (bodem) na obrazovce. Pixely odpovídající stejnému atributu jsou seskupeny do jedné oblasti. Každý atribut objektu je zobrazen pouze pomocí jediného pixelu, proto lze takto zobrazit velké počty objektů. Tyto grafy se využívají pro posuzování trendů a vztahů mezi dimenzemi, nejsou vhodné pro hledání podobných objektů, shluků, nebo pro identifikací odlehlých objektů. Typickými představiteli jsou Recursive Pattern [20] a Circle Segments [2].
- Hierarchické metody, které spočívají v postupném „vnořování" jednotlivých rozměrů (atributů) do sebe. Například metoda Dimensional Stacking [21] vnořuje do souřadnicového systému dvou atributů postupně další souřadnicové systémy ostatních atributů. Na podobném principu jsou založeny metody Worlds-within-Worlds [8] nebo Treemap [16].
- Metody založené na snížení dimenze dat například faktorovou analýzu, analýzou hlavních komponent, metodou vícerozměrného škálování, korespondenční analýzou, nebo metodou optimální projekce (Projection Pursuit [14]) a další.
- Hybridní techniky, využívající vhodné kombinace výše zmíněných metod.

Jiné způsoby klasifikace vizualizačních metod lze najít například v [19], [32].

2. Symbolové grafy

Symbolové grafy (ikony, „Glyphs" [31]), reprezentují každý objekt jedním grafickým symbolem. Jednotlivé atributy jsou přiřazeny na určitou část symbolu tvarem, velikostí, barvou, nebo jinak. Předností je srozumitelnost a možnost současného zobrazení velkého počtu atributů. Mají význam při porovnávání objektů mezi sebou, pro identifikaci podobných nebo výjimečných objektů, nebo pro identifikaci shluků.

Příkladem symbolového grafu je šipka. Její různou délku, tloušťku, orientaci a barvu člověk snadno rozlišuje, proto může šipka zobrazovat čtyři různé atributy vícerozměrných dat.

Návrh grafu nelze zcela automatizovat, protože výsledný vzhled grafu je ovlivněn přiřazením atributů a rozmístěním symbolů. Přiřazení atributu určité části symbolu („mapování") by nemělo být náhodné. Pokud uživatel vychází při mapování z logické vazby zobrazovaných atributů, zvyšuje se vypovídací schopnost grafů. Například u obličejových grafů je mapování jedním z nejdůležitějších kroků tvorby symbolu. Vztahy mezi atributy jsou lépe postřehnutelné u atributů mapovaných na ty části symbolu, které jsou umístěny blízko sebe.

Pro interpretaci grafů jsou podle Warda [31] podstatné dvě vlastnosti jednotlivých částí symbolů, vlastnosti geometrické (tvar, velikost, umístění, orientace) a vzhledové (barva, textura, průhlednost). Jejich kombinováním vzniká prostor pro zobrazení velkého počtu atributů. Symbol může obsahovat také komponenty, které nejsou vázány se žádným atributem. Zcela nezávislé na datech mohou být například barva a umístění hvězdicových grafů, nebo velikost tváře u obličejových diagramů.

Kromě symbolů samotných může být novou informací také jejich umístění v grafu. Umístění symbolů lze řídit buď samotnými daty (hodnotami atributů), nebo strukturou dat. Překrývání symbolů ovlivní jejich velikost a následně počet zobrazených objektů, tím se také změní vypovídací schopnost grafu. Podobně rozložení grafu na obrazovce ovlivňuje výsledný vizuální efekt. Mezi

symboly nemusí být žádný volný prostor, nebo naopak lze symboly oddělit volným místem mezi nimi.

Při umístění symbolů řízeném daty je pozice symbolu zpravidla určena hodnotami dvou atributů, které jsou přiřazeny na osy souřadnicového systému. Výběr atributů zobrazovaných na osách grafu by neměl být náhodný. Vhodné je zvolit buď atributy, které jsou co nejvíce korelovány s ostatními, nebo atributy, které jsou důležité z pohledu uživatele. Pozici symbolu lze potom lépe interpretovat a navíc je možné posoudit korelace mezi atributy na osách. Nevýhodou je subjektivita volby atributů a možnost volby pouze dvou, respektive tří (pro trojrozměrný graf) atributů.

Jinou možností je zobrazení transformovaných proměnných na osách. Transformované souřadnice vyjadřují určitou kombinaci všech atributů. Váhy atributů se stanoví například faktorovou analýzou, metodou hlavních komponent, metodou vícerozměrného škálování, samoorganizujících map, apod. Výhodou je zapojení všech atributů do určení pozice symbolu a výsledné umístění „podobných" objektů blízko sebe, nevýhodou je větší výpočetní náročnost a nezřídka také obtížná interpretace zobrazených os.

Pokud existuje vztah mezi objekty (například uspořádání, časová řada, hierarchická struktura apod.), doporučuje se řídit umístění symbolů na základě struktury dat. Řazení prvků lze provést zleva doprava, shora dolů, spirálovitě od středu ke kraji [32], nebo rekurzivně (dle pixelové techniky „Recursive Pattern" [20]), výhodou je snadnější interpretace a efektivní využití plochy obrazovky.

Způsob řazení může také napomoci k odhalení zákonitostí nebo změn v datech. Při hierarchickém uspořádání symbolů se úsečkami spojují objekty vyšší úrovně s objekty nižší úrovně (klasické stromové zobrazení nebo trojrozměrné stromy). Hierarchické uspořádání je proto vhodnější pro zobrazení vztahů mezi objekty, nikoliv pro zobrazení samotných objektů.

Síťová struktura je zobecněním hierarchické struktury. Jednotlivé uzly jsou nahrazeny symboly, které představují dané objekty. Příkladem je zobrazení webové struktury (systém Narcissius [13]).

Volba umístění symbolů na obrazovku je ovlivněna typem řešené úlohy a charakterem zobrazovaných dat. Přístup řízený daty poskytuje umístění objektů, které je intuitivní a podporuje porovnání jednotlivých objektů. Přístup řízený strukturou dat podporuje zobrazení strukturálních vazeb mezi objekty.

Většina symbolových grafů umožňuje zobrazit značný počet atributů, přesto při rostoucím *m* (zhruba od 20 atributů), může docházet k omezení rozlišitelnosti, nebo obtížné interpretaci atributů.

Z hlediska počtu zobrazených objektů jsou klasické symbolové techniky omezeny velikostí a rozlišovacími schopnostmi zobrazovací plochy. Protože každý symbol zabírá určitý prostor, je při velkém počtu současně zobrazovaných symbolů problém buď s velikostí symbolů, nebo jejich překrýváním. Tento problém se netýká grafů Color Icons nebo Stick Figures, jejichž předností je možnost zobrazení velkého počtu objektů a lze je proto využít pro odhalení struktur v datech.

3. Prezentace vybraných symbolových vizualizačních technik

V této části jsou ze symbolových grafů podrobněji popsány profily, polygony a obličejové grafy (Chernoffovy diagramy). Jednotlivé typy grafů jsou popsány a také prezentovány pomocí procedur knihovny nazvané *symbols* [25], vytvořené pomocí statistického jazyka R (knihovna dostupná u prvního autora). Různé typy symbolových grafů a současně možnosti této knihovny, včetně navržené modifikace umístění symbolů pomocí souřadnic nebo ve zvoleném uspořádání, jsou prezentovány pomocí reálných dat. Knihovna zahrnuje profily, sloupce, polygony (hvězdičkové, sluníčkové a polygonové grafy), Color Icons a Stick Figures.

Zvolená data (získána z veřejného zdroje [29]) popisují výsledky průzkumu spokojenosti pacientů v devatenácti zdravotnických zařízeních. Výsledky, vyjádřené procentem spokojených zákazníků pro jednotlivé kategorie, jsou v tabulce 1. Spokojenost je měřena pomocí 12 atributů, jejichž značení a význam jsou následující: **Přijetí** = Přijetí do nemocnice, **Respekt** = Respekt, ohled a úcta, **Koordinace** = Koordinace a integrace péče, **Informace** = Informace, komunikace a vzdělávání, **Pohodlí** = Tělesné pohodlí, **Cit. opora** = Citová opora a zmírnění strachu a úzkosti, **Rodina** = Zapojení rodiny a přátel, **Propuštění** = Propuštění a pokračování péče, **Sestry** = Spokojenost se sestrami, **Lékaři** = Spokojenost s lékaři, **Služby** = Spokojenost se všeobecnými službami. **Souhrn** vyjadřuje, jak je pacient spokojen s nemocniční péčí souhrnně ve všech sledova-

Tab. 1: Spokojenost pacientů v procentech ve čtrnácti zdravotnických zařízeních

Nemocnice	Přijetí	Respekt	Koordinace	Informace	Pohodlí	Cit. opora	Rodina	Propuštění	Sestry	Lékaři	Služby	Souhrn
Plzeň I.	84,9	84,6	85,4	84,7	78,0	81,2	92,9	90,6	83,9	83,6	68,4	84,0
Plzeň II.	83,2	81,6	80,9	80,7	76,9	77,1	91,0	88,7	79,3	80,6	63,9	81,3
Ostrava	82,4	78,6	81,0	82,3	79,0	75,7	90,0	88,1	81,1	78,2	66,9	81,2
Brno	83,0	82,1	81,1	81,3	73,1	75,0	77,7	68,5	77,2	80,6	65,2	77,8
Brno (d)	78,7	76,1	75,7	75,0	71,8	68,1	70,5	59,2	72,1	74,3	60,6	72,5
HK	85,5	84,4	85,8	84,8	81,0	75,9	85,1	69,7	83,8	83,4	68,8	82,3
Vinohrady	83,0	85,6	85,9	83,0	74,1	78,8	86,0	74,0	81,5	83,7	64,1	80,6
Bulovka	75,0	79,6	79,7	76,4	69,3	69,5	72,7	64,3	74,6	77,0	52,8	73,4
Olomouc	83,7	81,6	84,7	83,3	79,0	74,4	82,2	73,0	82,1	80,2	70,6	80,6
Thomayerova	81,6	79,5	79,4	78,9	72,9	69,7	76,7	62,9	75,6	78,0	55,9	75,6
Sv. Anna	80,5	80,6	82,1	79,0	76,4	71,6	79,7	69,1	81,8	77,4	63,3	77,7
Motol	79,1	83,0	85,1	81,0	74,6	70,1	79,7	66,1	77,5	81,7	63,8	77,7
IKEM	79,9	80,1	83,5	81,1	80,0	73,7	87,0	74,6	82,2	81,4	70,3	80,0
MOU	86,4	88,3	87,4	88,4	83,8	79,9	88,4	77,3	84,7	87,0	76,0	85,2
ÚN Brno	78,2	84,9	82,4	78,4	78,3	72,5	77,6	72,4	82,5	79,2	63,7	78,5
UPMD	87,7	78,1	70,0	73,8	64,9	66,6	70,9	65,7	69,3	74,9	62,2	72,3
Plzeň II. (d)	61,2	60,7	59,3	65,3	56,4	52,4	83,3	75,5	55,8	61,0	40,2	62,5
Thomayerova (d)	79,4	77,3	72,0	74,2	59,6	65,2	63,5	50,3	64,5	75,0	46,9	67,4
Motol (d)	65,6	68,5	52,2	57,8	51,0	47,9	67,1	41,6	38,1	64,4	40,8	54,7

Zdroj: [29].

ných atributech. Vyšší hodnoty atributů odrážejí příznivější hodnocení.

Symbol „(d)“ u některých oddělení v tabulce a grafech značí dětská oddělení, která jsou hodnocena odděleně vzhledem ke specifičnosti pacientů (odpovědi dětských a dospělých pacientů nelze směšovat).

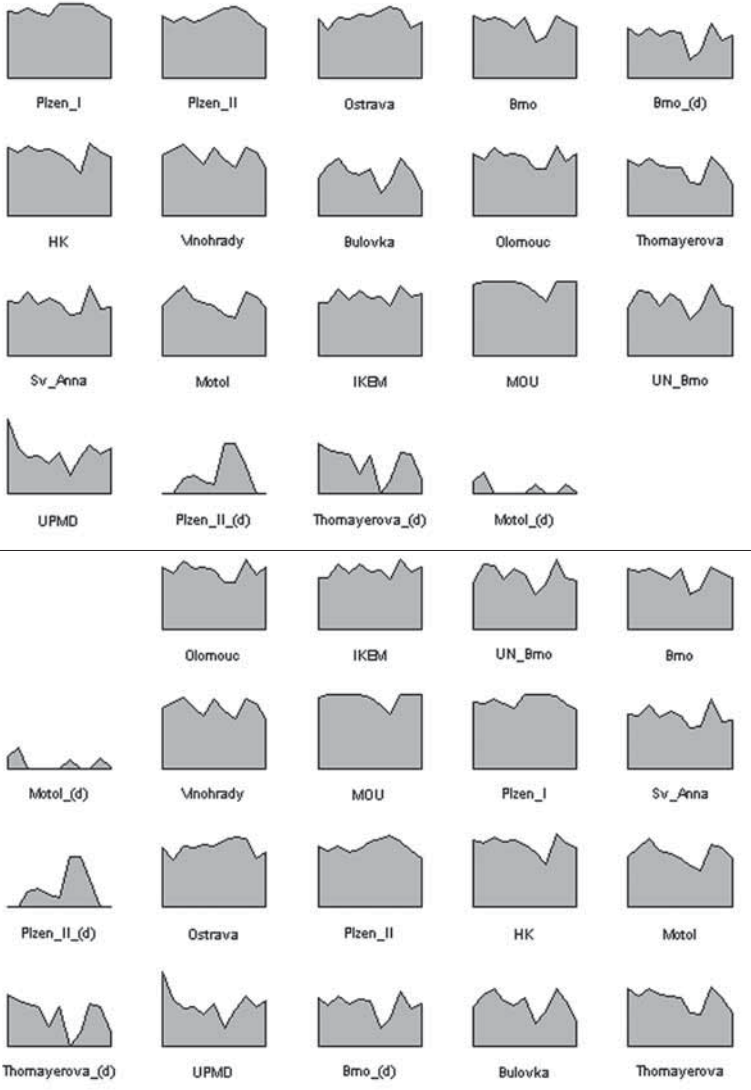
Pro všechny grafy byla data před zobrazením transformována na interval (0 – 1).

Profily

Profily patří mezi nejrozšířenější symbolové techniky. Každý objekt popsaný atributy je reprezentován sloupci, jejichž výška představuje hodnotu daného atributu. Místo sloupců se někdy vykresluje lomená čára, vytvořená postupným spojením vrcholů jednotlivých sloupců. Tím vznikne výsledný „profil“ daného objektu. Pokud se profily jednotlivých objektů zakreslí do jediného grafu, přechází tato technika do metody zvané „systém rovnoběžných souřadnic“, která se řadí mezi geometricky transformované techniky. Profily lze porovnávat mezi sebou a nacházet podobné objekty, vybočující měření a na základě podobnosti identifikovat vícerozměrné shluky. S rostoucím počtem objektů se toto porovnání stává obtížnější.

Na obrázku 1 vlevo jsou zobrazena data tabulky 1 pomocí profilů, atributy jsou řazeny v pořadí sloupců tabulky 1. Snadno lze identifikovat zdravotnická zařízení s vysokou spokojeností pacientů

Obr. 1: Profily nemocnic. Původní a spirálovité řazení podle spokojenosti



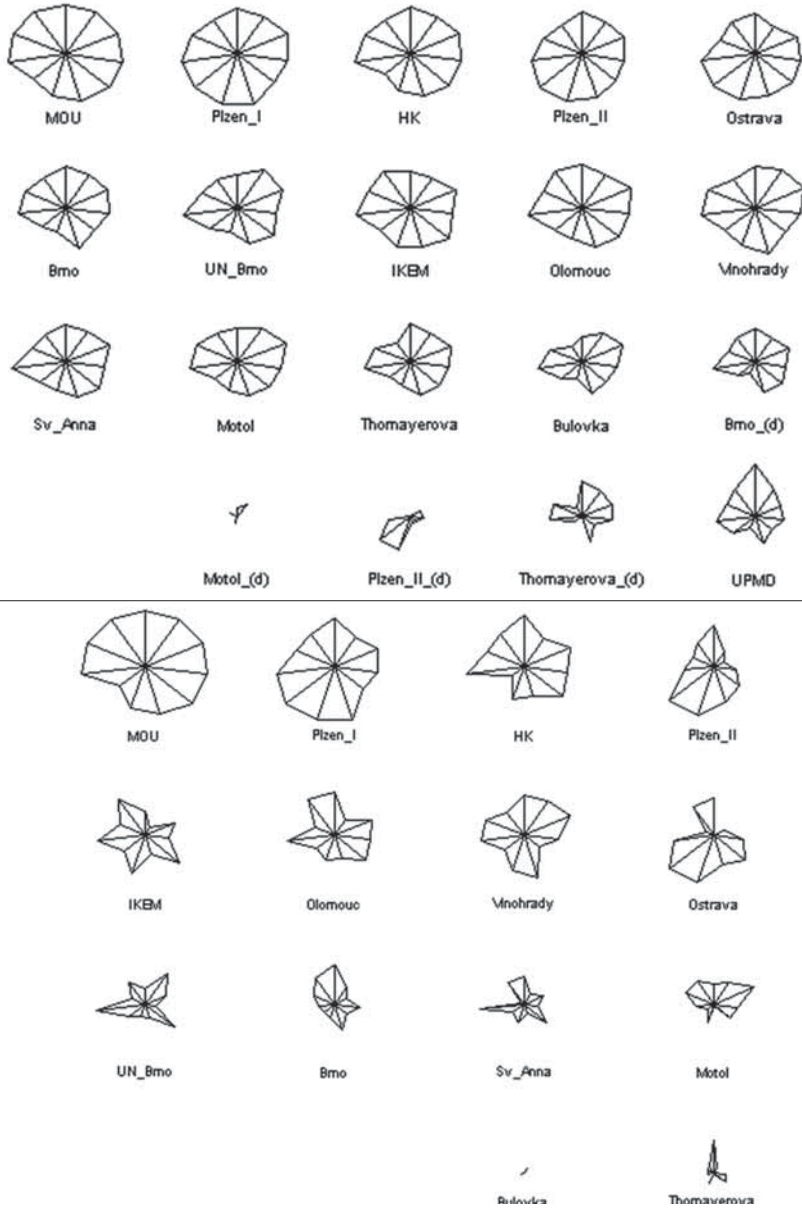
Zdroj: [25].

ve většině atributů (MOU, Plzeň, Ostrava, HK), podobně naopak zařízení s nízkou spokojeností (většinou dětská oddělení). Jednotlivé profily lze detailněji porovnávat. Například UPMD vykazuje oproti ostatním velmi vysokou spokojenost pacientů v prvním atributu (Přijetí do nemocnice), nízkou pak u atributu sedmého (Zapojení rodiny a přátel). V ostatních dimenzích je spokojenost přibližně průměrná. Na obrázku 1 vpravo jsou stejné profily seřazeny podle Souhrnné spokojenosti pacientů spirálovitě od středu. Objekty s podobnou hodnotou Souhrnné spokojeností se tak dostaly blízko sebe, podobné objekty lze lépe identifikovat a porovnat.

Polygony

Polygon [26] je geometrický obrazec vytvořený pro každý z objektů. Ze středu souřadnicového

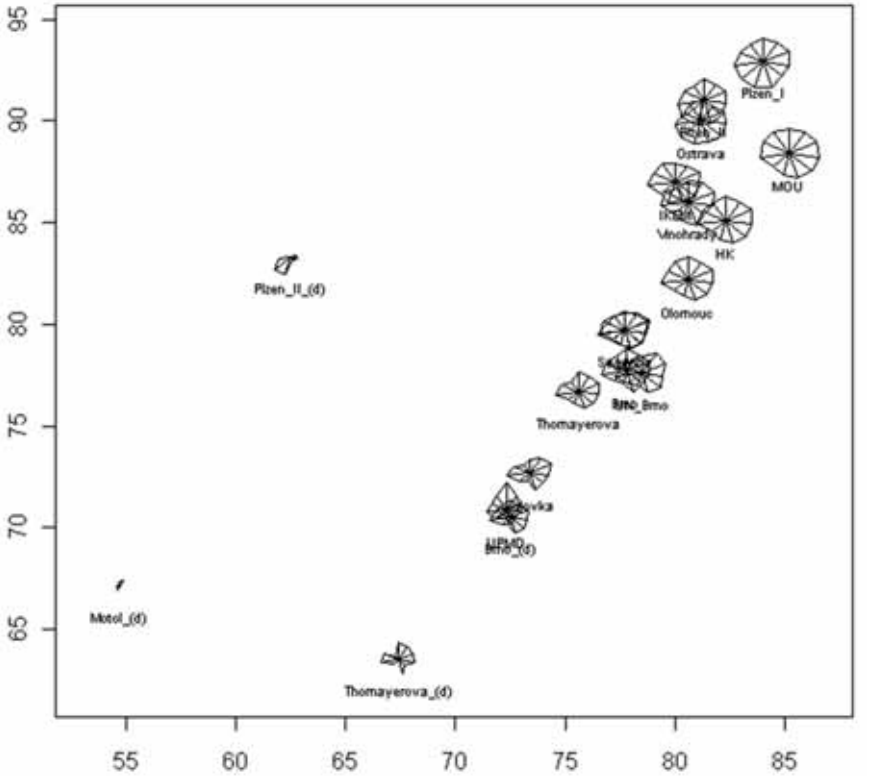
Obr. 2: Hvězdicové grafy spokojenosti. Napravo bez dětských oddělení



Zdroj: [25].

systému se pro každý atribut vede úsečka („paprsek"), jejíž délka odpovídá relativizované hodnotě daného atributu. Úsečky mezi sebou svírají shodný úhel. Koncové body úseček se postupně spojí a vytvoří uzavřený obrazec. Graf se označuje jako hvězdicový graf. Obdobou je sluníčkový graf (nebo graf slunečních paprsků – sun ray plot), který má všechny paprsky protažené do stejné délky. Hodnoty, kterých atribut nabývá, jsou vyznačeny na každém paprsku a tyto body spojeny

Obr. 3: Hvězdicové grafy, souřadnice Souhrnná spokojenost (x) a Rodina (y)



Zdroj: [25].

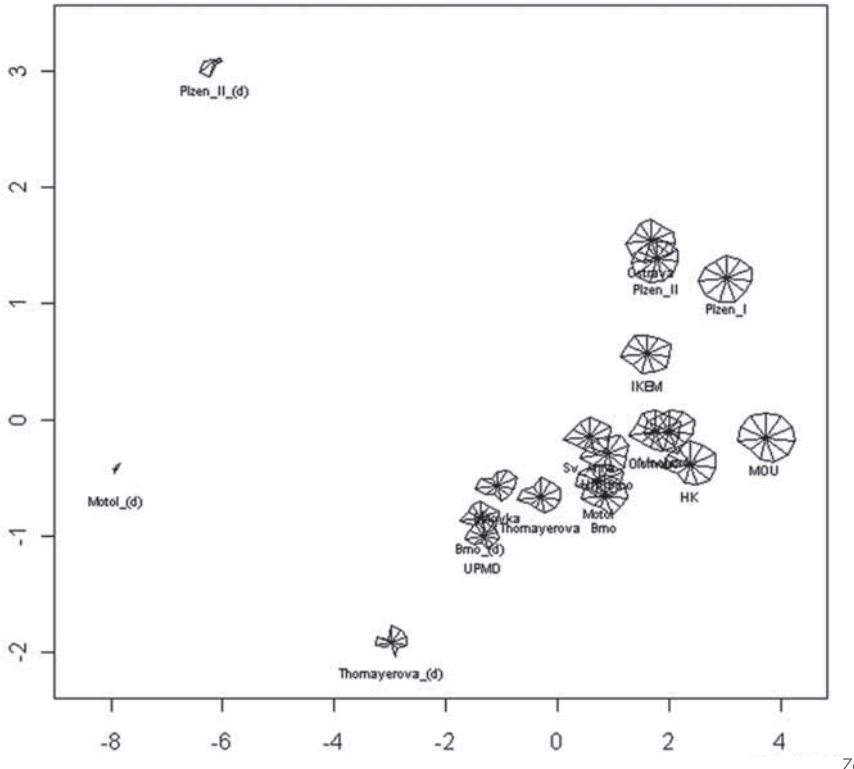
úsečkami. Podobně je vytvořen polygonový graf, který vyplňuje plochu vzniklého geometrického obrazce. Vzájemným porovnáním polygonů lze vyhledat podobné objekty, vybočující měření, případně identifikovat shluky.

Obrázek 2 vlevo zobrazuje hvězdicové grafy spokojenosti pacientů. První atribut (Přijetí) je zobrazen úsečkou vedenou kolmo vzhůru ze středu symbolu, další následují ve směru hodinových ručiček. Objekty jsou seřazené dle Souhrnné spokojenosti zleva doprava, na dalším řádku zprava doleva atd. Dobře odlišitelná jsou zařízení s vysokou či nízkou spokojeností, podobné i odlehlé objekty. Na hvězdicových grafech je patrné, že většina atributů spolu silně koreluje, neboť v jednotlivých hvězdicích jsou paprsky přibližně stejné délky. Pouze atributy Rodina a Propuštění (7. a 8. paprsek) příliš nekorelují s ostatními, zato silně korelují navzájem.

Protože dětská pracoviště jsou specifická, byla vypuštěna z obrázku 2 vpravo. Snížení počtu objektů umožnilo detailněji porovnávat atributy jednotlivých objektů mezi sebou. Například srovnání „Ostrava“ a „Vinohrady“ ukazuje, že v Ostravě jsou pacienti spokojenější s Tělesným pohodlím, Zapojením rodiny, Propuštěním a Službami, kdežto na Vinohradech s Respektem, Koordinací, Citovou oporou a Lékaři (což nebylo z obrázku vlevo ihned patrné). Nepravidelnost tvaru polygonu „Ostrava“ naznačuje nevyrovnané výsledky v jednotlivých atributech.

Obrázek 3 prezentuje umístění symbolů, které je řízeno původními daty. Na ose je Souhrnná spokojenost, na ose Zapojení rodiny a přátel (tento atribut má se Souhrnnou spokojeností nejnižší korelaci). V grafu můžeme identifikovat shluky objektů, například Plzeň II-Ostrava, IKEM-Vinohrady-HK, Sv. Anna-Motol-Brno-UN Brno (nemocnice u Sv. Anny a Motole dokonce mají identické umístění) nebo Bulovka-UPMD-Brno (d). Nalézáme také odlehlé objekty jako dětská

Obr. 4: Hvězdicové grafy, souřadnicový systém 1. a 2. hlavní komponenty



Zdroj: [25].

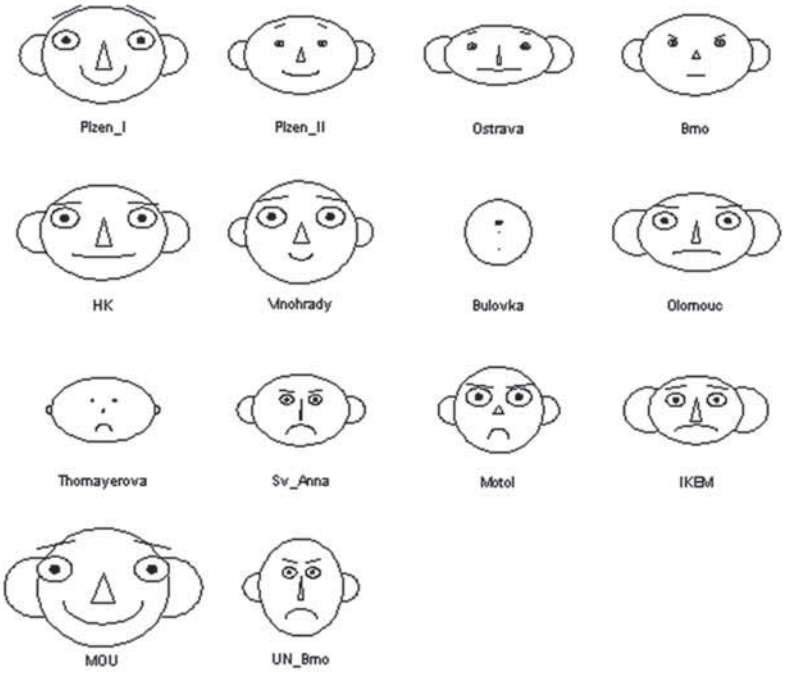
oddělení Plzeň, Motol a Thomayerova. Jelikož jsou v grafu zobrazeny symboly, je možné sledovat podobnosti a rozdílnosti objektů i v jiných atributech, než které jsou vyneseny na osách (např. blízko sebe umístěné UPMD a Brno (d) se výrazně liší v Přijetí do nemocnice – podobné je to i na obrázku 4). Tento styl zobrazení zvýrazní rozdíly v hodnotách atributů na osách souřadných. Nevhodná volba atributů na osách může vést k zavádějícím výsledkům.

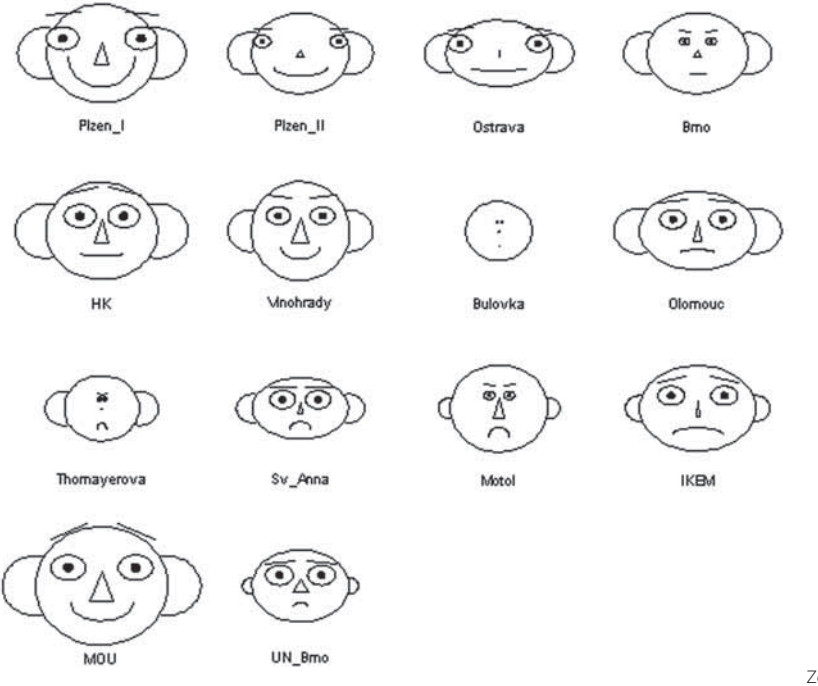
Pro eliminaci subjektivity při volbě atributů na osách lze hvězdicové grafy umístit v souřadnicovém systému první a druhé hlavní komponenty (obrázek 4). První hlavní komponenta (detailní výsledky této analýzy zde nejsou uvedeny) vysvětlila 81 % variability původních dat a bylo možné ji interpretovat jako Souhrnnou spokojenost (koresponduje s obrázkem 3). Druhá hlavní komponenta, která vysvětlila 11 % variability, byla nejvíce pozitivně sycena atributy Rodina a Propuštění z nemocnice, negativně atributy Respekt a Přijetí do nemocnice. Grafy 3 a 4 jsou podobné, obsahují podobné shluky a odlehlé objekty, protože volba atributů pro zobrazení na osách obrázku 3 koresponduje s faktory, které vysvětlily největší podíl variability spokojenosti mezi objekty. Oba grafy mají objekty uloženy na původně vypočítanou pozici, proto dochází k jejich překrývání.

Obličejové grafy

Obličejové grafy (Chernoff faces [4]) využívají schopnost člověka rozpoznávat a hodnotit rozdíly mezi lidskými tvářemi. Každý objekt je reprezentován schematickým obličejem, ve kterém tvar či velikost jednotlivých rysů (délka nosu, tvar úst, sklon obočí, šířka tváře) představují hodnotu odpovídajícího atributu (obrázek 5). Původní Chernoffův návrh mohl zobrazit až 18 atributů (pět jich definovalo tvar horní a dolní části obličeje, jeden délku nosu, tři umístění, zakřivení a šířku úst, pět umístění, vzhled a velikost očí, jeden umístění zornic a poslední tři umístění, tvar a velikost obočí).

Obr. 5: Obličejové grafy spokojenosti při opačné volbě mapování atributů





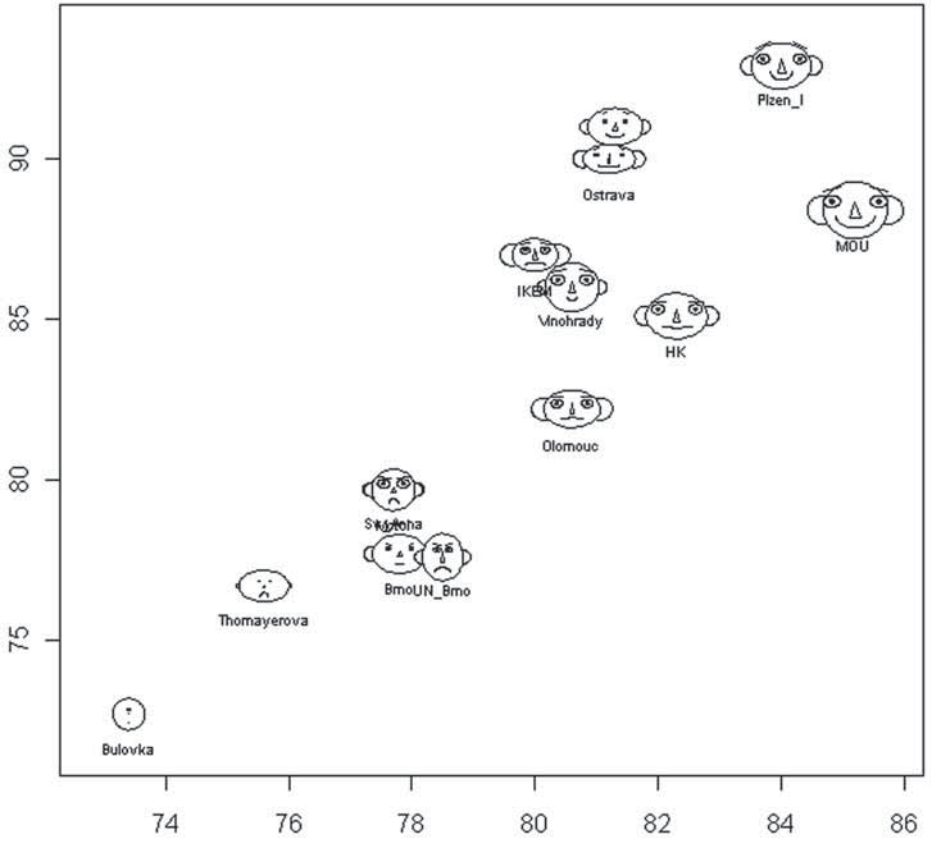
Zdroj: [25].

Přidáním dalších komponent (např. uši, vlasy apod.) lze jejich počet zvětšit. Vzhledem k některým grafickým omezením se jich ale většinou používá 10 až 12. Pro *k* částí obličeje lze *m* atributů, $m \leq k$, mapovat $z = \frac{k!}{(k-m)!}$ způsoby.

Obličejové grafy napomáhají při hledání podobných nebo odlišných objektů a při identifikování shluků. Posloupnost Chernoffových tváří a změny v jejich vzhledu mohou být využity při odhalování změn v časových řadách (např. [4] [17] [30]). Nevýhodou jsou některá grafická omezení (například obtížně rozpoznatelná poloha zornic v příliš malých očích), možnost vlivu jedné komponenty tváře na vzhled jiné komponenty [9], nebo nutnost vhodného přiřazení jednotlivých atributů rysům tváře (nebezpečí, že nepříliš významný atribut se stane příliš výrazným rysem tváře a bude odvádět pozornost od atributů s užitečnější informací).Některé studie se zabývaly také otázkami využití obličejových grafů ke klasifikaci a problémem, zda některé části obličeje jsou lépe rozlišovány než jiné ([5] [6] [24]).

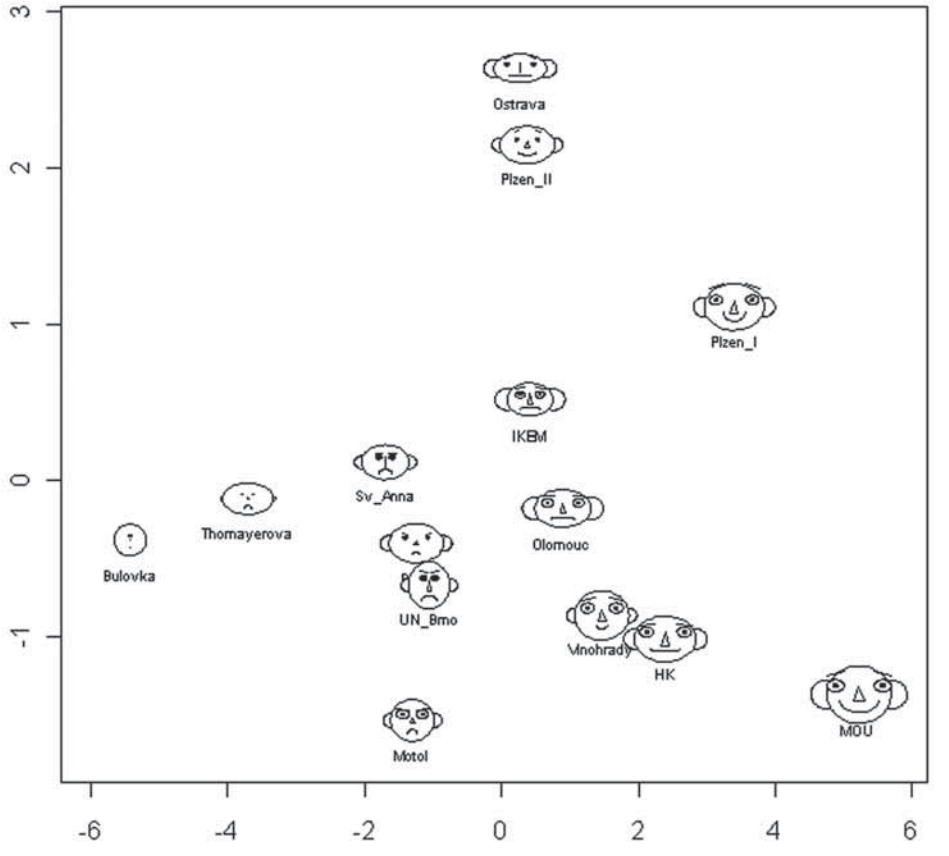
Obrázek 5 znázorňuje Chernoffovy diagramy pro data tabulky 1 (byla vypuštěna dětská pracoviště). Mapování atributů na obrázku 5 vlevo: Přijetí=šířka tváře, Respekt=výška tváře, Koordinace=velikost očí, Informace=vzdálenost očí od sebe, Pohodlí=šířka úst, Cit. opora=tvar úst, Rodina=úhel obočí, Propuštění=umístění obočí, Sestry=délka nosu, Lékaři=šířka nosu a Služby=velikost uší. Pro srovnání jsou na obrázku 5 vpravo atributy mapovány v opačném pořadí. Je zřejmé, že obličeje obou částí grafu se poněkud liší, ale vzhledem ke zmíněné vysoké korelaci mezi atributy jsou závěry podobné.

Obr. 6: Obličejové grafy, souřadnicový systém Souhrnné spokojenosti a Rodiny



Zdroj: [25].

Obr. 7: Obličejové grafy. Souřadnicový systém 1. a 2. hlavní komponenty



Zdroj: [25].

Na první pohled je možné identifikovat, v čem jsou jednotlivé objekty podobné a v čem se liší. Například Plzeň II., Ostrava a Brno jsou na obrázku 5 vlevo navzájem podobné velikostí očí (Koordinace), přičemž Brno se od ostatních dvou liší ve sklonu obočí (Rodina). Podobné objekty Sv. Anna a UN Brno se nejvíce liší v atributu Přijetí (šířka tváře v obrázku 5 vlevo a velikost uší vpravo). Vzhledem k tomu, že člověk dokáže detekovat odlišnosti v lidských rysech velice snadno, je porovnávání Chernoffových tváří pro uživatele přirozené.

Obrázky 6 a 7 se liší v souřadnicovém systému. Na obrázku 6 jsou obličejové grafy umístěny v souřadnicovém systému Souhrnné spokojenosti a Rodiny, na obrázku 7 v souřadnicovém systému první a druhé hlavní komponenty. Opět můžeme identifikovat shluky i vybočující objekty, navíc lze sledovat vzhled jednotlivých rysů tváře v závislosti na pozici tváře. Obličeje zleva doprava se více usmívají, mění se sklon obočí, velikost očí, nosu i celé tváře.

Dalšími variantami obličejových grafů jsou asymetrické tváře [9], které 18 atributů přiřazují na rysy levé poloviny tváře a dalších 18 na rysy pravé poloviny. Kromě zdvojnásobení počtu zobrazených atributů dochází v takto modifikovaných grafech k menšímu ovlivňování jednoho rysu tváře druhým. Výsledná asymetrická tvář vyžaduje ještě větší pozornost při volbě vhodného mapování atributů.

Rozšířením Chernoffových grafů je metoda EVA (Empathic Visualization Algorithm [23]), která pro návrh mapování atributů využívá genetický algoritmus tak, aby výsledný obličejový graf co nejlépe

odpovídal daným datům. Metoda FAV (Figural Animation Visualization [28]) mapuje atributy nejen na lidskou tvář, ale na celou postavu a navíc přidává animaci. Zde jsou prezentovány pouze Chernoffovy diagramy.

Další typy symbolových grafů

Pro doplnění přehledu symbolových grafů zmíníme též představitele dalších technik. Barevné ikony (Color Icons [22]), spojují výhody vnímání barev a textur (někdy také tvarů). Objekt je reprezentován plochou na obrazovce (ve většině případů čtvercem), která je úsečkami rozdělena na několik částí. Úsečky jsou barveny podle hodnoty atributu, který je úsečce přiřazen. Barvy políček, na která je ikona těmito úsečkami rozdělena, jsou interpolovány z barev okolních úseček. Při jiném přístupu se atributy mapují na barvy políček. Zvýrazněním některých úseček je možno dát větší důraz na zvolené atributy. Barevné ikony, jako jedny z mála symbolových grafů, jsou vhodné pro zobrazení velkého počtu objektů, pokud jsou popsány maximálně šesti až osmi atributy (počet znázorněných atributů je možné zvýšit různými modifikacemi, uvedenými v [22] a [7]).

Metoda Stick Figures [27] zobrazuje objekt pomocí několika různých čar, hlavního těla symbolu a jednotlivých větví (limbs). Dva atributy jsou mapovány na osy souřadné, další určují úhel pro nasměrování těla a úhly, pod kterými z něj vyrůstají větve. Původní návrh Picketta a Grinsteina [27] definoval dvanáct způsobů, kterými mohou být na hlavní tělo napojeny čtyři větve. Celkem je zobrazeno sedm atributů. Zvýšení počtu zobrazovaných dimenzí je možné tím, že další atributy se mapují jako délka, šířka nebo barva větví, nebo lze přidat samotné větve. S rostoucím počtem atributů se stick figures stávají složité a nepřehledné a většinou se používají pro zobrazení maximálně dvaceti atributů. Jsou vhodné pro zobrazení velkého počtu objektů, protože mohou vytvářet přehledné textury, na základě kterých je možné identifikovat struktury a vzory v datech.

Symbolová technika autoglyfy [3] (Shape Coding) zobrazuje každý objekt obdélníkem, který je rovnoběžný s osami souřadnicového systému. Obdélník je rozdělen do čtvercových polí, z nich každé je přiřazeno jednomu atributu. Hodnoty atributu jsou nejprve normalizovány a poté rozděleny do dvou kategorií (podle zvolené hraniční hodnoty či směrodatné odchylky). Čtvercová pole jsou obarvena šedivě a černě pro každou z kategorií a bíle pro chybějící hodnoty daného atributu. Místo systému černá-šedá-bílá lze použít jiné barevné schéma. Samotné symboly se dají různým způsobem seřadit, ale nejčastější je třídění po řádcích z leva doprava nebo shora dolů. Autoglyfy jsou vhodné pro nominální atributy s malým počtem variant, zejména pro binární atributy.

Uvedený výčet symbolových vizualizačních technik není vyčerpávající, do této kategorie spadají dále TileBars [12], Dashtubes [10], šipky [33], metroglyfy [1] a další.

Závěr

Tento příspěvek vychází z rozsáhlé rešerše zaměřené na vizualizaci dat. Podrobně jsou studovány symbolové grafy a doporučena modifikace umístění symbolů, která je prezentována pomocí vytvořeného softwaru.

Základní filozofií symbolových grafů je zobrazení každého objektu pomocí určitého symbolu. Jednotlivé atributy, kterými je objekt popsán, jsou přiřazeny daným částem (komponentám) symbolu. Hodnota atributu je reprezentována vzhledem komponenty symbolu (tvarem, velikostí, barvou apod.). Některé typy symbolových grafů jsou široce rozšířeny (např. profily nebo polygony) a staly se základním prvkem exploratorní analýzy dat. Statistické softwarové systémy, jako SAS, SPSS, Stata nebo Statistica, většinou obsahují nějakou verzi profilů (někdy jako sloupcové grafy nebo ve formě rovnoběžných souřadnic), polygony se vyskytují řidčeji (ve Statistice různé varianty, hvězdičkové grafy v SAS), obličejové grafy výjimečně (např. Statistica nebo Unistat). Jejich umístění na obrazovce však v těchto programech nebývá nijak řešeno. Řada dalších symbolových technik prozatím chybí v nejpoužívanějších statistických programech a proto není jejich potenciálu příliš využíváno.

Symbolové grafy jsou vhodné zejména pro svoji jednoduchost a přehlednost. Jediným pohledem na graf lze získat poměrně jasnou představu o jednotlivých objektech zkoumaných dat. Lze je využít pro porovnávání objektů mezi sebou, pro vyhledání podobných objektů, pro identifikaci nepodobných či odlehlých objektů a pro rozpoznání shluků. Při vhodném použití mohou napomoci k nalezení korelace mezi atributy nebo pro odhalení trendů v datech.

Protože ze všech kategorií vizualizačních technik jsou nejvíce náročné na prostor potřebný pro

znázornění objektu, jsou počty současně zobrazených objektů limitovány velikostí zobrazovací plochy (výjimkou jsou Stick Figures nebo Color Icons). Při velkém počtu objektů (řádově stovky nebo dokonce tisíce) je vhodné je kombinovat s jinými vizualizačními technikami, nebo zvolit zcela jinou zobrazovací metodu. Dalším podstatným faktorem, který je třeba brát v úvahu, je u většiny symbolových grafů potřeba vhodného mapování atributů k jednotlivým částem symbolu.

Málo známou vlastností symbolových grafů, která může výrazně ovlivnit vypovídající schopnost výsledného grafického zobrazení, je volba umístění zobrazovaných objektů. Při výběru umísťovací strategie je třeba uvážit typ úlohy, kterou uživatel řeší a typ dat, která má k dispozici.

Vybrané symbolové vizualizační techniky (profily, polygony a obličejové grafy) zde byly podrobně popsány a prezentovány pomocí dat, která popisují hodnocení spokojenosti pacientů různých zdravotnických zařízení. Pomocí reálných dat bylo ukázáno, že symbolové grafy umožní nejen porovnání jednotlivých objektů (zde nemocnic), ale také rozpoznání skupin podobně hodnocených objektů a označení atributů (složek spokojenosti), kterými se jednotlivé objekty liší, nebo naopak podobají. Tyto výhody symbolových grafů je možné podpořit a zvýraznit volbou vhodné strategie umístění symbolů.

Byly navrženy a implementovány metody založené na umístění symbolů, které mohou zvýšit vypovídací schopnost symbolových grafů a napomoci jejich analýze. Vizualizace dat byla prezentována jako nástroj pro zobrazení kvantitativního obsahu dat, jehož cílem je přehledná informace. Atraktivita grafu však nesmí být na úkor přesnosti a spolehlivosti informace, která je v grafu obsažená. Při sestavování grafu pomocí grafického softwaru je nezbytné pečlivě kontrolovat automatická nastavení, která by mohla informaci v grafu zkreslit, proto je nutné rozumět principům a možnostem daného typu grafu ve vztahu k zobrazeným atributům.

Literatura

[1] ANDERSON, E. A Semigraphical Method for the Analysis of Complex Problems. *Proceedigs of the National Academy of Sciences of the U.S.A.,* 1957, vol. 43, no. 10, pp. 923-927. ISSN 1091-6490.

[2] ANKERST, M., KEIM, D. A., KRIEGEL, H-P. 'Circle Segments': A Technique for Visually Exploring Large Multidimensional Data Sets. In *Proceedings of Visualization '96*, Hot Topic Session. San Francisco, 1996. ISBN 0-89791-864-9.

[3] BEDDOW, J. Shape Coding of Multidimensional Data on a Microcomputer Display. In *Proceedings of the First IEEE Conference on Visualization.* San Francisco: IEEE Computer Society Press, 1990, pp. 238-246. ISBN 0-8186-2083-8.

[4] CHERNOFF, H. The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association,* 1973, vol. 68, no. 342, pp. 361-368. ISSN 0162-1459.

[5] CHERNOFF, H., RIZVI, M. H. Effect on Classification Error of Random Permutations of Features in Representing Multivariate Data by Faces. *Journal of the American Statistical Association,* 1975, vol. 70, no. 351, pp. 548-554. ISSN 0162-1459.

[6] De SOETE, G., De CORTE, W. On the Perceptual Salience of Features of Chernoff Faces for Representing Multivariate Data. *Applied Psychological Measurement*, 1985, vol. 9, no. 3, pp. 275-280. ISSN 0146-6216.

[7] ERBACHER, R., GONTHIER, D. The Color Icon: A New Design and a Parallel Implementation. In *Proceedings of SPIE '95 Conference on Visual Data Exploration and Analysis II.* San Jose, 1995, pp. 302-312. ISBN 0-8194-1767-X.

[8] FEINER, S., BESHERS, C. Worlds within Worlds – Metaphors for Exploring n-Dimensional Virtual Worlds. In *Proceedings of Symposium on User Interface Software and Technology '90.* New York: ACM, 1990, pp. 76-83. ISBN 0-89791-410-4.

[9] FLURY, B., RIEDWYL, H. Graphical Representation of Multivariate Data by Means of Asymmetric Faces. *Journal of the American Statistical Association*, 1981, vol. 76, no. 376, pp. 757-765. ISSN 0162-1459.

[10] FUHRMANN, A., GRÖLLER, E. Real-Time Techniques for 3D Flow Visualization. In *Proceedings of the Conference on Visualization '98.* Los Alamitos: IEEE Computer Society Press, 1998, pp. 305-312. ISBN 1-58113-106-2.

[11] FURNAS, G. W., BUJA, A. Prosection Views: Dimensional Interface through Section and Projections. *Journal of Computation and Graphical Statistics*, 1994, vol. 3, no. 4, pp. 323-353. ISSN 1061-8600.

[12] HEARST, M. A. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*. Denver: ACM Press, 1995, pp. 59-66. ISBN 0-201-84705-1.

[13] HENDLEY, R. J., DREW, N. S., WOOD, A. M., BEALE, R. Narcissus: Visualising Information. Case Study. In *Proceedings of the 6th Conference on Visualization '95*. Los Alamitos: IEEE Computer Society Press, 1995, pp. 90-96. ISBN 0-8186-7201-3.

[14] HUBER, P. J. Projection Pursuit. *The Annals of Statistics,* 1985, vol. 13, no. 2, pp. 435-475. ISSN 0090-5364.

[15] INSELBERG, A., DIMSDALE, B. Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. In *Proceedings of the First IEEE Conference on Visualization*. San Francisco: IEEE Computer Society Press, 1990, pp. 361-370. ISBN 0-8186-2083-8.

[16] JOHNSON, B., SHNEIDERMAN, B. Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. In *Proceedings of the 2nd Conference on Visualization '91*. San Diego: IEEE Computer Society Press, 1991, pp. 284-291. ISBN 0-8186-2245-8.

[17] JOHNSON, R. A., WICHERN, D. W. *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall, 2002. 767 p. ISBN 0-13-092553-5.

[18] KEIM, D. A. Designing Pixel-Oriented Visualization Techniques: Theory and Applications. *IEEE Transactions on Visualization and Computer Graphics*, 2000, vol. 6, no. 1, pp. 59-78. ISSN 1077-2626.

[19] KEIM, D. A. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 2002, vol. 8, no. 1, pp. 1-8. ISSN 1077-2626.

[20] KEIM, D. A., KRIEGEL, H-P., ANKERST, M. Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data. In *Proceedings of the 6th Conference on Visualization '95*. Atlanta: IEEE Computer Society Press, 1995, pp. 279-286. ISBN 0-8186-7187-4.

[21] LeBLANC, J., WARD, M. O., WITTELS, N. Exploring N-Dimensional Databases. In *Proceedings of the First IEEE Conference on Visualization*. San Francisco, 1990, pp. 230-239. ISBN 0-8186-2083-8.

[22] LEVKOWITZ, H. Color Icons: Merging Color and Texture Perception for Integrated Visualization of Multiple Parameters. In *Proceedings of the 2nd Conference on Visualization '91*. San Diego: IEEE Computer Society Press, 1991, pp. 22-25. ISBN 0-8186-2245-8.

[23] LOIZIDES, A., SLATER, M. The Empathic Visualization Algorithm: Chernoff Faces Revisited. In *Technical Sketch, ACM Siggraph 2001 Conference Abstracts and Applications*. Los Angeles: ACM Press, 2001. pp. 175-175.

[24] MORRIS, CH. J., EBERT, D. S., RHEINGANS, P. An Experimental Analysis of the Effectiveness of Features in Chernoff Faces. In *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making, Proceedings of SPIE Vol. 3905.* Washington: ETATS-UNIS, 2000, pp. 12-17. ISBN 0-8194-3517-1.

[25] MYSLIVEC, J. Symbols. R Package. *The Comprehensive R Archive Network* [online]. 2009, listopad [cit. 2010-03-30]. Dostupné z: <http://cran.r-project.org/mirrors.html>.

[26] MYSLIVEC, J. Využití grafických metod pro analýzu vícerozměrných dat. In *IMEA 2006, Sborník příspěvků z 6. ročníku doktorandské konference.* Hradec Králové: Gaudeamus, 2006, pp. 102-109. ISBN 80-7041-164-3.

[27] PICKETT, R. M., GRINSTEIN, G. G. Iconographic Displays for Visualizing Multidimensional Data. In *Proceedings of IEEE Conference on Systems, Man and Cybernetics*. New Jersey: IEEE Press, 1988, pp. 514-519. ISBN 7-8003-039-3.

[28] PFLUGHOEFT, K. A., ZAHEDI, M., SOOFI, E. Figural Animation Visualization: The System and Its Application. In *Third Annual SIGDSS Pre-ICIS Workshop: Designing Complex Decision Support: Discovery and Presentation of Information and Knowledge*. Las Vegas, 2005.

[29] RAITER, T. *Měření kvality zdravotní péče prostřednictvím spokojenosti pacientů* [online]. c2008 [cit. 2008-10-31]. Dostupné z: <http://portalkvality.mzcr.cz/Pages/5-Mereni-kvality-zdravotni-pece-prostrednictvim-spokojenosti-pacientu.html>.

[30] SCOTT, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley Publishing, 2006. 384 p. ISBN 0-47-169755-9.

[31] WARD, M. O. A Taxonomy of Glyph Placement Strategies for Multidimensional Visualiza-

tion. *Information Visualization*. 2002, vol. 1, no. 3/4, pp. 194-210. ISSN 1473-8716.
[32] WARD, M. O., LIPCHAK, B. N. A Visualization Tool for Exploratory Analysis of Cyclic Multivariate Data. *Metrika*, 2000, vol. 51, no. 1, pp. 27-37. ISSN 0026-1335.
[33] WITTENBRINK, C. M., SAXON, E., FURMAN, J. J., PANG, A., LODHA, S. Glyphs for Visualizing Uncertainty in Environmental Vector Fields. *IEEE Transactions on Visualization and Computer Graphics*, 1996, vol. 2, no. 3, pp. 266-279. ISSN 1077-2626.

Ing. Jaroslav Myslivec
Univerzita Pardubice
Filozofická fakulta
Katedra věd o výchově
jaroslav.myslivec@upce.cz

doc. RNDr. Hana Skalská, CSc.
Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informatiky a kvantitativních metod
hana.skalska@uhk.cz

Doručeno redakci: 13. 10. 2009.
Recenzováno: 16. 11. 2009, 11. 3. 2010
Schváleno k publikování: 23. 6. 2010

ABSTRACT

VISUALIZATION OF MULTIDIMENSIONAL DATA USING SYMBOL PLOTS

Jaroslav Myslivec, Hana Skalská

Visualization of multidimensional data is presented here as a tool for displaying quantitative content of data. The aim of visualization methods is to arrange information about the data file in such a way that new visual information is descriptive and informative enough and able to communicate ideas easily and intuitively. With the using of graphical information it is possible to reach more profound understanding of the data and derive relevant information and knowledge from the data. This article is concerned with the categorization of visualization methods and brings extensive references related to this field. It focuses on selected symbol plots (profiles, polygons, face diagrams) in more details. The article describes fundamentals of these selected visualization techniques and also their advantages and disadvantages. It studies the characteristics of symbol plots when different mapping methods of attributes or different methods of placement of the symbols are used. Package "symbols" was designed and developed in the range of R language, which can change placement of symbols in the graph. This way it allows such an arrangement of symbols that is useful for visual identification of clusters with similar subjects. Data from public database is used here for the presentation of methods and possibilities of the software as well. This paper presents new implementation which allows displaying symbols either in coordinate system of suitably chosen attributes or in coordinate system of two principal components. The new arrangement in placement of objects helps a user to discover associations and/or mutual relations in attributes, as well as to find clusters or structures of similar objects in data file.

Key Words: *Visualization, symbol plots, attributes mapping, symbol placement.*

JEL Classification: *Y10, C44, D83, C81, C82.*