Univerza
v Ljubljani
Medicinska
fakulteta

GAJ VIDMAR

PRIKAZ VEČRAZSEŽNIH PODATKOV S KONVEKSNOLUPINSKIMI IN KONKORDANČNIMI DIAGRAMI

DOKTORSKO DELO

Imenovanje mentorja na seji senata dne 28. 6. 2005.
Komisija za oceno in zagovor imenovana na seji senata dne 18. 12. 2006.

Datum zagovora: 6. 3. 2007

Mentor: prof. dr. Janez Stare

Predsednik komisije: prof. dr. Borut Peterlin
Članica: prof. dr. Anuška Ferligoj
Član: prof. dr. Lovro Stanovnik

Izvleček

Poveden prikaz podatkov je bistven del sodobne statistične analize podatkov, s tem pa tudi raziskovalnega in strokovnega dela v biomedicini in na številnih drugih področjih. Cilj dela je izviren prispevek k tej problematiki na dveh področjih. Prvi sklop dela je namenjen izboljšanemu prikazu velike količine trorazsežnih podatkov, pri katerih je ena razsežnost opisna, drugi dve pa številski, z razširjenimi konveksnolupinskimi diagrami. Drugi sklop je posvečen možnim rešitvam doslej neobravnavanega problema prikaza konkordance s poudarkom na primerih, ko so ocenjevalci razdeljeni v skupine. Nove metode smo implementirali z odprtokodnim, prostim oziroma najširše dostopnim programjem, ki deluje v različnih okoljih, ter preizkusili na podatkih iz biostatistične prakse. Uporabnost razširjenih konveksnolupinskih diagramov in programja zanje smo dodatno ovrednotili z anketo pri potencialnih uporabnikih. Obravnavane diagrame smo opisali na podlagi Wilkinsonove grafične slovnice. Prikaz podatkov s predlaganimi metodami lahko pripomore k boljšemu vpogledu v preučevane probleme ter s tem izboljšal kakovost in razumljivost publikacij.

Abstract

Informative data visualisation is an essential part of modern statistical data analysis, and thus of research and professional practice in biomedicine and elsewhere. The thesis makes an original contribution in two fields of data visualisation. The first part is aimed at improving visualisation of large amounts of three-dimensional data consisting of one categorical and two numeric dimensions by means of augmented convex hull plots. The second part is dedicated to the previously unaddressed task of visualising concordance, with emphasis on comparison of concordance between groups of raters. The newly proposed methods were implemented in open-source public-domain or widely available software working on various platforms, and tested on real-life datasets from biostatistical practice. Feasibility of augmented convex hull plots and the software to produce them was additionally assessed with a survey of potential users. The proposed diagrams were described on the basis of Wilkinson's grammar of graphics. The proposed data visualisation methods can contribute to better insight into studied phenomena, and thus improve quality and clarity of publications.

I've seen things you people wouldn't believe.
(Roy Batty, replikant, Blade Runner, 1982)

Zahvala

Svojemu mentorju, predstojniku Inštituta za biomedicinsko informatiko prof. dr. Janezu Staretu, se zahvaljujem za podporo in nasvete. Sodelavki asist. Maji Pohar Perme, univ. dipl. mat., se zahvaljujem za sodelovanje pri izdelavi programja za razširjene konveksnolupinske diagrame, asist. dr. Ninu Rodetu s Fakultete za socialno delo pa za prispevek k razvoju konkordančnih diagramov. Prijatelju doc. dr. Primožu Ziherlu z Instituta Jožef Stefan se zahvaljujem za jezikovni pregled besedila. Hvaležen sem tudi doc. dr. Andreju Blejcu z Nacionalnega inštituta za biologijo za uvodne napotke glede prikaza statističnih podatkov s konveksnimi lupinami. Posebej se zahvaljujem podiplomskim študentom statistike, ki so sodelovali v anketi o razširjenih konveksnolupinskih diagramih.

Delo posvečam soprogi Meriti.

Opomba

Zaradi jedrnatosti in lažje razumljivosti se v celotnem besedilu uporabljajo zgolj moške jezikovne oblike, seveda pa se nanašajo na oba spola.

Vsebina

1. UVOD

Nazoren in poučen prikaz večrazsežnih (multidimenzionalnih) podatkov v dveh razsežnostih je sestavni del ali končni cilj številnih biostatističnih in bioinformacijskih metod, pa tudi najrazličnejših informatiziranih oziroma računalniško vodenih postopkov razvrščanja (klasifikacije), prepoznavanja (identifikacije), pregledovanja in vrednotenja podatkov v medicini in na drugih področjih (npr. Spence, 2000).

Pričujoče delo je usmerjeno v dva izseka iz te izjemno široke problematike, s katero se ukvarja na stotine monografij in na tisoče člankov iz različnih temeljnih in uporabnih znanosti, kot so statistika, računska geometrija in psihologija zaznavanja, ved, strok in tehnologij, povezanih z razvojem računalništva in informatike, kot so prikaz informacij (information visualisation), strojno učenje (machine learning) in podatkovno rudarjenje (data mining), pa tudi najrazličnejših drugih področij, od analize slik v medicini do poslovnega odločanja, od politologije do kemometrije. Teoretični uvod, ki sledi, zato ne more biti zaokrožen in podroben, vendar skuša biti celovit pri predstavitvi ključnih pojmov in sklicih na literaturo.

1.1. PRIKAZ DVORAZSEŽNIH IN TRORAZSEŽNIH PODATKOV

Sistematična in uravnotežena predstavitev vseh načinov prikaza dvorazsežnih podatkov in prikaza trorazsežnih podatkov v dveh razsežnostih ter uporabe teh načinov za prikaz večrazsežnih podatkov je naloga, ki presega obseg cele knjige. Pred tremi desetletji je za tovrsten poskus še zadoščala krajša monografija (Everitt, 1978), že pred petimi leti pa je celo najobsežnejša (Harris, 2000) komajda zaobjela celoten razvoj, pa še to bolj na nivoju tehnološkega kataloga kot v smislu znanstvenega razpravljanja. Preprost in zgoščen, a dovolj sodoben in širok pregled prikaza statističnih podatkov podaja Jacoby (1997, 1998). Osnovne metode in načela, ki jih predstavlja, so primerno izhodišče za razumevanje pričujočega dela.

Za podrobnejše vrednotenje obravnavane problematike je potrebno poznavanje specializiranih postopkov. Za prikaz trorazsežnih podatkov, pri katerih je ena razsežnost opisna, so to:

- daljinomerski zaboji z ročaji (rangefinder box plots, Becketti in Gould, 1987);
- elipsaste in pravokotne bivariatne razširitve zaboja z ročaji (bivariate extensions of the boxplot, Goldberg in Iglewicz, 1992);
- zaboji z ročaji na podlagi področja največje gostote (highest density region boxplots, Hyndman, 1996);
- robustni bivariatni zaboji z ročaji, dobljeni na podlagi lupljenja konveksnih lupin in glajenja z zlepki (robust bivariate boxplots, Zani, Riani in Corbellini, 1998);
- vrečasti diagram kot oblika bivariatnega zaboja z ročaji (bagplot – bivariate boxplot, Rousseeuw, Ruts in Tukey, 1999);
- dvorazsežni zaboj z ročaji (two-dimensional boxplot, Tongkumchum, 2005).

Za prikaz soglasja (med skupinami) ocenjevalcev pri rangiranju pa so posebnega pomena:

- naključno raztresenje kot eden od standardnih postopkov prikaza več podatkov z istima koordinatama (random jittering; Cleveland in McGill, 1984b);
- uporaba vzporednih osi za prikaz večrazsežnih statističnih podatkov in podatkovno rudarjenje (Wegman, 1990, 2003);
- prikaz večrazsežnih frekvenčnih porazdelitev urejenostnih spremenljivk s permutacijskimi politopi (mnogokotniki, telesi oziroma analognimi objekti v več razsežnostih – Baggerly, 1995; Thompson, 1994).

1.2. POJEM, MERE IN PRIKAZ KONKORDANCE

Preden pristopimo k problemu prikaza konkordance, moramo razumeti oziroma definirati pojem in mere konkordance. V statistiki se pojem konkordance nanaša na m rangiranj k objektov[1]. Tovrstni podatki so pogosti na področju kadrovanja, raziskav tržišča in merjenja

[1] Zaradi lažjega razločevanja *objekt* v nadaljevanju besedila označuje predmete, pojave oziroma pojme, ki jih rangirajo ocenjevalci, slovenska beseda *predmet* pa se uporablja kot strokovni izraz na področju računalniškega programja (prevod angleške besede *object* v samostalniški ali pridevniški rabi na področju programskih jezikov in podatkovnih zbirk).

javnega mnenja ter povsod drugod, kjer kandidate za sprejem na delovno mesto ali napredovanje, izdelke, politične stranke ali druge subjekte ali objekte rangirajo vodstveni delavci, strokovnjaki, ciljne skupine, opazovalci ali pa računalniški algoritmi. Stopnjo konkordance merijo različni koeficienti konkordance. Ker je grafični prikaz konkordance namenjen predvsem primerjavi konkordance med skupinami ocenjevalcev oziroma vzorci podatkov, v nadaljevanju podajamo tudi kratek pregled statističnih metod za primerjavo konkordance.

1.2.1. KOEFICIENTI KONKORDANCE

Pregled problematike konkordance podajajo številni viri (npr. Legendre in Lapointe, 2004; Palachek in Schucany, 1984; Siegel in Castellan, 1988). Izvzemši koeficient Imana in Conoverja (1987) je vsem meram konkordance skupno, da temeljijo na koeficientu urejenostne korelacije – bodisi Spearmanovem ρ bodisi Kendallovem τ. Najbolj znano in najpogosteje uporabljano mero sta uvedla Kendall in Babbington-Smith (1939). Njun (po slavnem prvem avtorju v praksi imenovan kar Kendallov) koeficient konkordance W je v bistvu povprečni ρ za vse možne dvojice rangiranj [1] in je neločljivo povezan s Friedmanovim testom (analizo variance za ponovljene meritve na podlagi rangov), tako da Friedmanova testna statistika χ^2 [2] služi za preizkušanje statistične značilnosti W glede na ničelno hipotezo o odsotnosti konkordance:

$$W = [(m - 1)\, \mathrm{E}\,(\rho) + 1] / m; \qquad [1]$$

$$\chi^2 = W\, m\, (k - 1), \quad df = k - 1. \qquad [2]$$

Kendallov koeficient soglasja je definiran na enak način kot W, le da namesto ρ uporabimo τ, kar lahko zapišemo kot $u_K = W_\tau$. Ehrenberg (1952) je za mero soglasja med več ocenjevalci predlagal preprosto povprečje vseh vrednosti τ, zato je njegov koeficient [3] linearno povezan z u_K:

$$u_E = \mathrm{E}\,(\tau_{\text{rangiranje-rangiranje}}) = [u_K\, m / (m - 1) - 1] / (m - 1). \qquad [3]$$

V zgornji definiciji lahko zamenjamo τ z ρ in u_K z W (Lyerly, 1952), s čimer ocenimo pričakovano vrednost ρ [$u_L = E(\rho_{\text{rangiranje-rangiranje}})$]. Če eno rangiranje izberemo kot ciljno oziroma kriterijsko, pa govorimo o korelaciji med skupino ocenjevalcev in kriterijem, ki je definirana kot povprečje korelacij med posameznim ocenjevalcem in kriterijem [$T_C = E(\tau_{\text{rangiranje-kriterij}})$] in je računsko povezana z nekaterimi metodami za primerjavo konkordance med skupinami.

1.2.2. PRIMERJAVA KONKORDANCE MED SKUPINAMI

Osnova za primerjavo konkordance med skupinami je statistika $\mathscr{L}$ za dve skupini ocenjevalcev [4], ki sta jo uvedla Schucany in Frawley (1973). Gre za posplošitev Pagevega testa oziroma statistike L (Page, 1963). Postopek je namenjen preizkusu konkordance znotraj vsake skupine in med obema skupinama (z m_1 in m_2 člani), ki rangirata k objektov. Če rangiranja predstavljajo m_1 oziroma m_2 vrstic, objekti pa k stolpcev v matriki rangov, in če s $\boldsymbol{s}$ in $\boldsymbol{t}$ označimo vektorja vsot stolpcev v teh dveh matrikah (z elementi S_i oziroma T_i), je statistika $\mathscr{L}$ definirana kot

$$\mathscr{L} = \Sigma_k S_i T_i = \boldsymbol{s}' \boldsymbol{t}. \qquad [4]$$

Asimptotično normalno porazdeljena testna statistika za preizkus ničelne hipoteze o odsotnosti konkordance je $\mathscr{L}^*$ [5], ki temelji na pričakovani vrednosti [6] in varianci $\mathscr{L}$ [7], ti pa izračunamo iz E (S_i), Var (S_i) in Cov (S_i, S_j):

$$\mathscr{L}^* = [\mathscr{L} - E(\mathscr{L})] / \text{Var}(\mathscr{L})^{1/2}; \qquad [5]$$

$$E(\mathscr{L}) = m_1 m_2 k (k+1)^2 / 4; \qquad [6]$$

$$\text{Var}(\mathscr{L}) = m_1 m_2 (k-1) k^2 (k+1)^2 / 144. \qquad [7]$$

Ker ima statistika $\mathscr{L}$ končen razpon [9, 10], sta Schucany in Frawley (1973) predlagala njeno standardizacijo. Dobljeni koeficient [8] je omejen na običajni razpon korelacijskih mer: $\mathscr{W} = 0$ kaže na odsotnost konkordance znotraj skupin in med skupinama, $\mathscr{W} = 1$ pomeni

popolno soglasje skupin glede rangiranja objektov, $\mathscr{W} = -1$ pa pomeni, da je skladnost znotraj vsake skupine popolna in da sta skupini rangirali objekte v obratnem vrstnem redu:

$$\mathscr{W} = [\mathscr{L} - \mathrm{E}(\mathscr{L})] / [\max(\mathscr{L}) - \mathrm{E}(\mathscr{L})] \quad \in [-1, 1]; \qquad [8]$$

$$\min(\mathscr{L}) = m_1 m_2 k (k + 1) (k + 2) / 6; \qquad [9]$$

$$\max(\mathscr{L}) = m_1 m_2 k (k + 1) (2k + 1) / 6. \qquad [10]$$

S posplošitvijo dvoskupinskega testa na več skupin sta Beckett in Schucany (1975, 1979) uvedla večskupinsko analizo konkordance (postopek sta na podlagi analogije z analizo variance poimenovala ANACONDA). Toda ta metoda se v praksi ni prijela, bržčas zaradi svoje nedorečenosti in zapletenosti, zaradi katere smo jo tudi poskusili dopolniti (Vidmar in Černigoj, 2004).

Pristop Schucanyja in sodelavcev je sicer že zgodaj dobil konkurenco. Hollander in Sethuraman (1978), ki sta dokazala, da je statistika $\mathscr{L}$ enaka povprečju vseh $m_1 m_2$ vrednosti ρ med enim ocenjevalcem iz prve in enim iz druge skupine, sta test konkordance med skupinama razvila z uporabo neparametričnega postopka Walda in Wolfowitza (1944) za primerjavo aritmetičnih sredin. Palachek in Kerin (1982) sta razvila dvoskupinski test konkordance na podlagi statistike U (npr. Randles in Wolfe, 1979) in ga primerjala s predhodnimi pristopi.

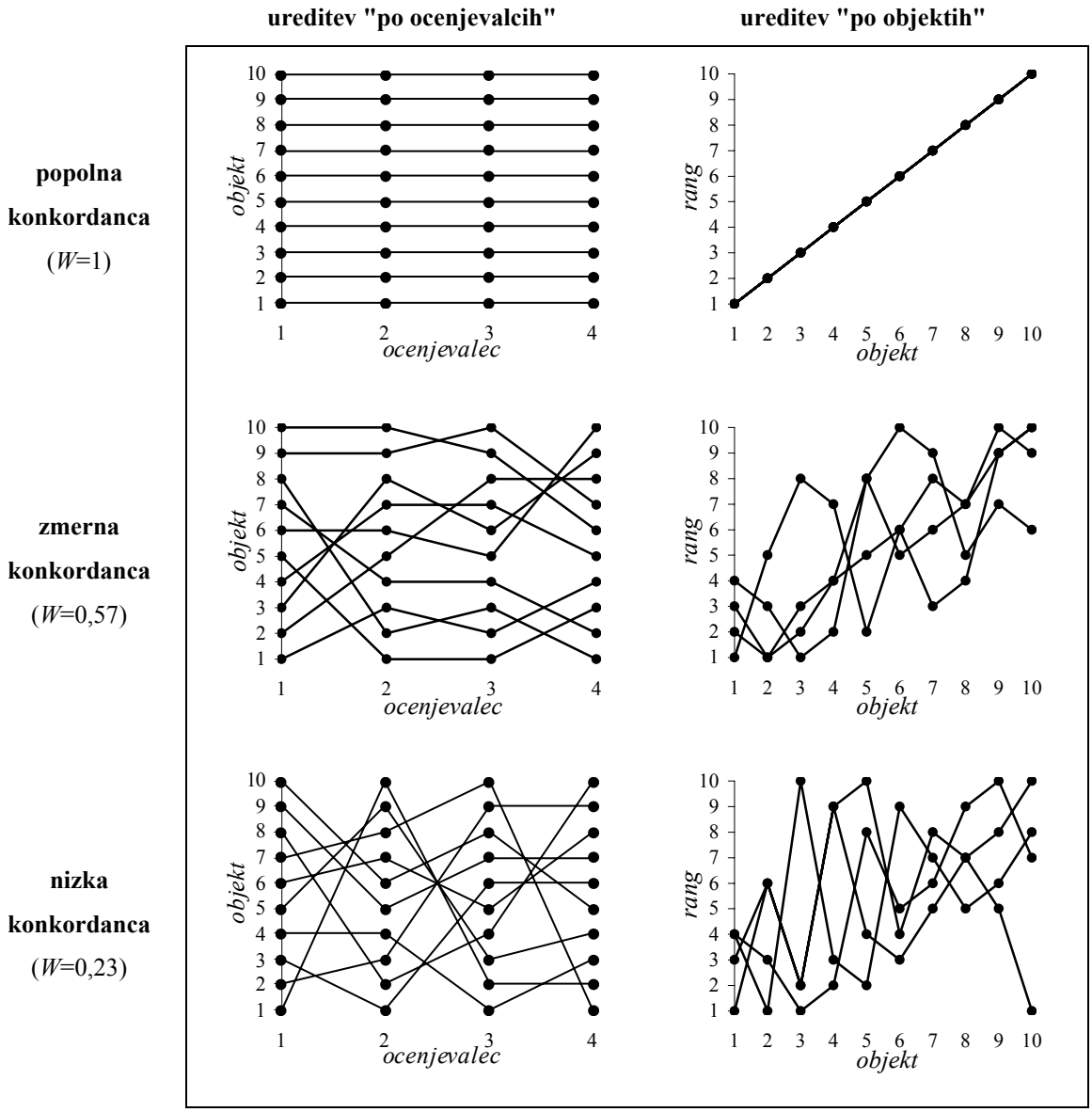
Najnovejši pristop k analizi konkordance je prispeval Legendre (2005), ki je izhajal iz problema povezanosti vrst v ekologiji. Zasnoval je permutacijski test na podlagi Kendallovega W z *a posteriori* testi za ugotavljanje, kateri ocenjevalci so skladni oziroma neskladni z drugimi ocenjevalci, pri katerih upoštevamo Holmov (1979) popravek. Permutacijski postopek je natančno opisal in predlagane metode uporabil na primeru terenske študije pršic v plasteh šotišča. Za permutacijsko testiranje konkordance je izdelal tudi prosto dostopno programje (http://www.bio.umontreal.ca/casgrain/en/labo/kendall_w.html).

1.2.3. OBSTOJEČE MOŽNOSTI ZA PRIKAZ KONKORDANCE

Ne da bi si izmislili novo vrsto statističnega diagrama, lahko range objektov, ki služijo za analizo konkordance, neposredno upodobimo z vzporednimi koordinatnimi osmi (Inselberg, 1985). Vzporedne koordinatne osi se sicer za prikaz podatkov še vedno uporablja razmeroma redko glede na to, kako preprosto je z njimi prikazati visokorazsežne objekte in nekatere njihove lastnosti (npr. Inselberg in Dimsdale, 1990). Šele odkar je razmah podatkovnega rudarjenja (data mining) povečal zanimanje za prikaz večrazsežnih podatkovij, tovrsten prikaz omogočajo tako specializirani programi za odkrivanje zakonitosti v podatkih (komercialni – Inselberg, 1998, ali javno dostopni – Fua, Ward in Rundensteiner, 1999) in vodilni komercialni statistični programski paketi, kot tudi javno dostopne dejavne spletne strani (npr. StatCrunch – West, Wu in Heydt, 2004) in brezplačni programski dodatki za raziskovalne namene (npr. VisuLab – http://www.inf.ethz.ch/personal/hinterbe/Visulab).

Kot je pokazal Wilkie (1980), lahko Kendallov koeficient urejenostne korelacije τ izračunamo in prikažemo na podlagi števila presečišč črt, ki povezujejo dvojice točk iz obravnavanih rangiranj, ki ju upodobimo na vzporednih oseh. Podobno velja, če prikažemo konkordančne podatke (s k črtami) v sistemu m vzporenih osi, ki predstavljajo ocenjevalci: več presečišč pomeni manjšo konkordanco, manj presečišč večjo.

Tovrstna ureditev "po ocenjevalcih" je ena od dveh možnosti prikaza vhodnih podatkov za analizo konkordance z vzporednimi koordinatnimi osmi. Druga možnost je ureditev "po objektih", pri kateri vsaka izmed m črt upodablja enega ocenjevalca, k osi pa predstavlja objekte. Na takem diagramu popolno konkordanco predstavlja popolno prekrivanje vseh črt. Pri ureditvi "po objektih" je osi primerno urediti po naraščajočem povprečnem rangu, kar pomeni, da je upodobitev popolne konkordance navzgor nagnjena daljica. Pri ureditvi "po ocenjevalcih" lahko osi uredimo glede na medsebojno podobnost ocenjevalcev, kot jo pokažejo *a posteriori* testi in postopki združevanja ocenjevalcev v skupine, ki jih je predlagal Legendre (2005). Različne stopnje konkordance z obema ureditvama prikazuje slika 1. Tovrsten prikaz konkordance ima tudi skupno podlago z enim izmed predlaganih konkordančnih diagramov: vzporedne osi smo uporabili pri metodi, predstavljeni v razdelku 3.3.2.1.2.



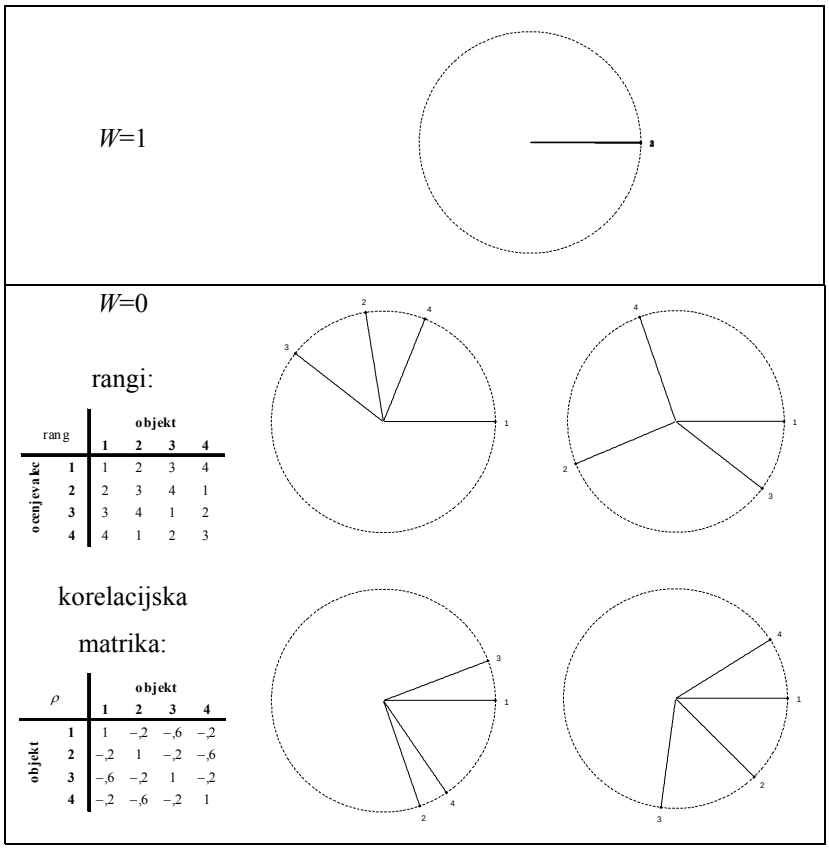
***Slika 1**: Prikaz konkordance (k=10 objektov; m=4 ocenjevalci; umetni podatki s tremi različnimi stopnjami konkordance) s prikazom rangov na vzporednih oseh: osi so lahko ocenjevalci (levo) ali objekti (desno).*

Med specializiranimi metodami je konkordančnim diagramom še najbližje grafični prikaz razpršenosti ocen ocenjevalcev, ki z urejenimi kategorijami ocenjujejo medicinske podatke (Nelson in Pepe, 2000). Temelj tega prikaza je, da soodvisnost med povprečjem in varianco pri urejenostnih podatkih zahteva, da moramo porazdelitev ocen opazovati ločeno glede na povprečno oceno ocenjevalca, če želimo dobiti popoln opis razlik med ocenjevalci. Rangi so urejenostni podatki, zato smo to načelo upoštevali pri konkordančnih mehurčnih diagramih, predstavljenih v razdelku 3.3.2.1.1. Vendar je rešitev, da preprosto dodamo povprečni rang kot tretjo razsežnost na prikaz celotne razpršenosti s histogramom, učinkovita šele pri

vzorcih, ki so neprimerno večji od tipičnih vzorcev, ki nastopajo v študijah konkordance. Poleg tega je osnovno načelo tovrstnega grafičnega prikaza bližje večjemu številu možnih ocen (torej približevanju zveznim spremenljivkam, po možnosti z zglajeno porazdelitvijo) kot pa manjšemu (torej rangom manjšega števila objektov, zlasti v primeru vezanih rangov).

Tretjo možnost prikaza konkordančnih podatkov z obstoječimi metodami nudi prikaz korelacijskih matrik. Postopki za neposreden prikaz večjih korelacijskih matrik, kot so matrike eliptičnih simbolov (elliptic glyphs – Murdoch in Chow, 1996) in kor(elo)grami (Friendly, 2002), so – vsaj brez preurejanja podatkov (npr. Friendly in Kwan, 2003) – manj primerni za prikaz konkordance, ker ne povedo ničesar o objektih in njihovih povprečnih rangih. To slabost ima tudi prikaz korelacijskih matrik, ki ga je predlagal Trosset (2005), a je za prikaz konkordance bolj obetaven. Postopek izhaja iz prikaza Pearsonove korelacije z diagrami *h* (Corsten in Gabriel, 1976; Seber, 1984), razširjenega na poljubne korelacijske matrike z uporabo večrazsežnega lestvičenja. Uporaba Trossetovega postopka na matrikah Spearmanovih koeficientov korelacije rangov se zdi naravna grafična dopolnitev Legendrovega postopka analize konkordance in združevanja ocenjevalcev v skupine, a ima znatne pomanjkljivosti.

Kot je razvidno iz slike 2, je Trossetova metoda pri prikazu popolne konkordance uspešna (in načeloma enolična, saj se črte vedno prekrivajo in je doseženi minimum ciljne funkcije vedno pod 10^{-8}), toda algoritem ne konvergira k ustrezni in enoznačni predstavitvi ničelne konkordance. Če uporabimo privzete vrednosti parametrov, programska koda (http://www.math.wm.edu/~trosset/Research/MDS/vc.s, ki smo jo priredili za okolje R) pri podatkovni matriki s $k=m=4$ ne konvergira niti pri statistično značilno neničelni konkordanci (W=0,675; p=0,033). Če je podatkovna matrika malo večja, resda konvergira že pri statistično neznačilni konkordanci (W=0,367; p=0,088; ciljna funkcija 2,692; k=4, m=6; podatki Schucanya in Frawleya, 1973, o francoskih ocenjevalcih vina, podrobneje predstavljeni v razdelku 3.3.2.1.1), a nizko konkordanco pri majhnih vrednostih k in n, pri kateri lahko algoritem odpove, je v praksi pričakovati. Vseeno ima tudi Trossetov pristop skupno osnovno z eno od predlaganih metod: polarni koordinatni sistem smo uporabili pri konkordančnih diagramih blazinice z bucikami (razdelek 3.3.2.2.2).



***Slika 2**: Prikaz popolne in ničelne konkordance s prikazom matrik Spearmanovih korelacijskih koeficientov po Trossetovem (2005) postopku (umetni podatki; k=m=4; pri ničelni konkordanci so poleg štirih primerov dobljenih diagramov prikazani tudi vhodni podatki in korelacijska matrika). Pri ničelni konkordanci Trossetov postopek ne najde ustrezne rešitve.*

1.3. WILKINSONOVA GRAFIČNA SLOVNICA

1.3.1. NASTANEK IN RAZVOJ

Wilkinson[2] je v devetdesetih letih dvajsetega stoletja zasnoval novo paradigmo statističnega prikaza podatkov, ki izhaja iz predmetno usmerjenega načrtovanja računalniške grafike. V kvantitativnih prikazih podatkov je odkril globoko slovnično strukturo in zasnoval dva temeljna kognitivna modela, ki sta hkrati izvedbena algoritma: enega za izdelavo grafik

[2] Leland Wiliknson (http://www.spss.com/research/wilkinson) je psiholog, statistik, računalničar in podjetnik. Je avtor statističnega programskega paketa SYSTAT, podpredsednik podjetja SPSS in profesor na Univerzi Northwestern v ZDA.

(Wilkinson, 1999, slika 2.1 na str. 22, tu reproducirana kot slika 4), ki ga podrobneje obravnavamo v nadaljevanju, in enega za "branje" grafik[3].

Wilkinsonova ustvarjalnost temelji na poznavanju in povezovanju spoznanj z različnih področij in združuje odlike ostalih najpomembnejših avtorjev s področja prikaza podatkov. Izhaja iz poglobljenega analitičnega pristopa k vizualizaciji kot celoviti znanstveni disciplini, katerega začetnik je Jacques Bertin (Bertin, 1981, 1983), hkrati pa lahko z vidika kognitivnih psiholoških modelov in človekove predelave informacij njegovo delo postavimo ob bok Stepehenu M. Kosslynu (Kosslyn, 1989, 1994). Njegov estetski navdih ter vpliv na raziskovalno in poslovno prakso je primerljiv z Edwardom R. Tuftejem (Tufte, 1983, 1990), obenem pa je vrhunski statistik kot William S. Cleveland (Cleveland, 1993, 1994).

Wilkinsonovo delo torej ne pomeni le teoretičnega mejnika v razvoju in razumevanju prikaza podatkov, temveč je tudi izrazito praktično. V okviru podjetja SPSS sta nastali dve delujoči implementaciji prikaza podatkov na podlagi grafične slovnice: knjižnica nViZn (http://www.spss.com/research/wilkinson/nViZn/nvizn.html), ki temelji na tehnologiji Java (http://java.sun.com/), je starejša in (zaradi visoke cene ter omejenega trženja) manj razširjena, jezik Graphics Production Library (GPL) pa so z verzijo 14 (http://www.spss.com/corpinfo/newsletter/0905_tip.htm) vključili v statistični programski paket SPSS za okolje Windows in v prihajajoči verziji 15 (http://www.spss.com/pdfs/S15CMPlr.pdf) s skoraj vsemi predvidenimi zmožnostmi spremlja drugo, razširjeno izdajo Wilkinsonove knjige (Wilkinson, Wills, Rope, Norton in Dubbs, 2006).

Ne glede na avtorsko zaščito in zasebno lastništvo implementacij je najbrž le vprašanje časa, kdaj bo Wilkinsonova grafična slovnica postala vodilni, če ne celo univerzalni standard za statistično vizualizacijo, saj uspešno poteka projekt njene implementacije v okolju R s paketom *ggplot* (http://had.co.nz/ggplot/). Njegov avtor Wickham je zanj in za paket *reshape*, ki prinaša v R zmožnosti sprotnega preurejanja analiziranih in prikazovanih podatkov oziroma tehnologije OLAP, prejel prestižno Nagrado Johna Chambersa za statistično

[3] Ta je s teoretičnega vidika morda še pomembnejši, a ga za razumevanje pričujočega dela ni nujno poznati, je razmeroma zapleten, hkrati pa ga avtor še ni povsem dodelal in preizkusil, zato ga ne obravnavamo.

programje (http://www.amstat-online.org/sections/graphics/newsletter/Volumes/v171.pdf) za leto 2006. V prihodnosti namerava Wickham tudi nadgraditi Wilkinsonovo delo s slovnico interaktivne grafike (http://had.co.nz/portfolio/cv.pdf).

1.3.2. TEORETIČNE OSNOVE

Na tem mestu ne moremo podati celovitega prikaza in poglobljene razlage Wilkinsonovega dela, zato se bomo osredotočili na predstavitev nekaterih ključnih pojmov in opis diagramov na podlagi grafične slovnice. Eno od izhodišč za to je pojem grafike (graphics), ki je nadrejen matematičnemu pojmu graf (graph) in presega laični pojem grafikon (chart)[4]. Z vidika izvedbe oziroma izdelave grafike lahko Wilkinsonov model zapišemo kot naslednje zaporedje (Wilkinson sicer namenoma uporablja številne neologizme, zato so slovenski izrazi tu in v nadaljevanju navedeni le v pomoč pri razumevanju in ne kot prevodi):

podatki/Data		algebra/Algebra		statistike/Statistics		koordinate/Coordinates		izdelovalec/Renderer
↘	↗	↘	↗	↘	↗	↘	↗	
spremenljivke/Variables		lestvice/Scales		geometrija/Geometry		estetike/Aesthetics		

Wilkinson tako definira grafiko kot preslikavo podatkov v estetske značilnosti grafičnih predmetov. Njegovo definicijo lahko povzamemo tudi z obrazcem

grafika (Graphics)	=	podatki (Data)	+	lestvice (Scales)	+	grafični predmeti (Graphical Objects = Grobs)	+	podobe (Facetting)

,

s katerim lahko opišemo katerikoli običajen statistični diagram. Grafični predmeti so lahko preprosti (črte, točke, stolpci, pravokotniki, mnogokotniki, besedilo, trakovi ipd.), srednjega

[4] V slovenščini nimamo povsem ustreznega prevoda za ta izjemno razširjeni pojem, s katerim se grafične prikaze podatkov označuje predvsem v svetu poslovanja, financ in trgovine. Čeprav je besedo *chart* (ki zhaja iz grške $\chi\alpha\rho\tau\eta\sigma$ oziroma latinske *charta*, ki pomenita list papirusa oziroma papirja) že pred dvesto leti uporabljal pionir prikaza statističnih podatkov, William Andrew Playfair (prim. Playfair, 2005), se je v veliki meri razširila "po zaslugi" pisarniških programskih paketov, ki uporabniku nudijo "katalog grafikonov" ter mu tako pogosto olajšujejo izdelavo risbosmetja (chartjunk – Tufte, 1983). Zato v pričujočem delu poleg besede grafika, pri kateri se držimo Wilkinsonovega pomena, uporabljamo še besedo diagram kot ožji in manj določen pojem.

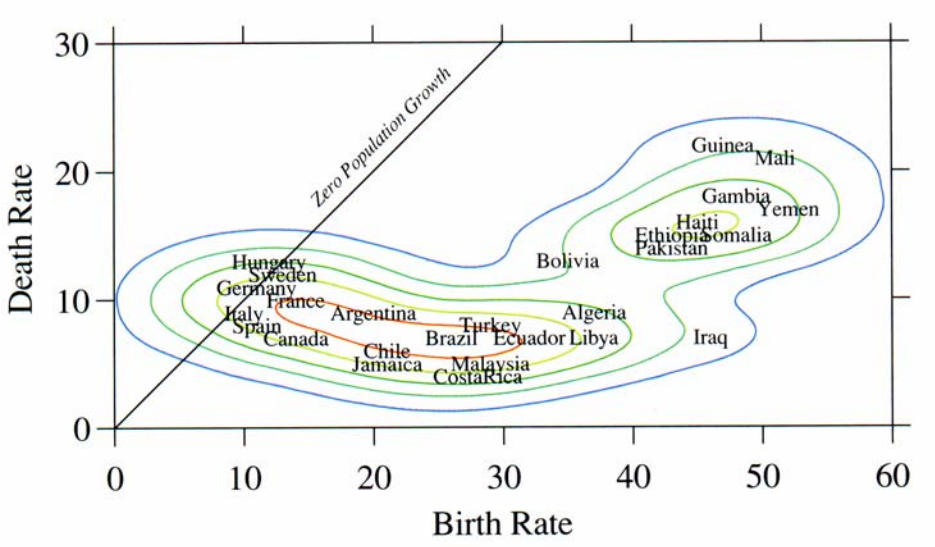
nivoja (raztresene točke, zaboj z ročaji) ali kompleksni (združujejo preproste grafične predmete s statistikami – vrečasti diagram, obrobe gostote verjetnosti, kvantilna regresijska črta, histogram ipd.). Lestvice oziroma preslikave estetik se razlikujejo glede na medsebojni položaj enot (opisne lestvice, zvezne lestvice, izenačene lestvice, projekcije pri zemljevidih, pretvorjene lestvice), zato moramo biti pri različnih lestvicah pozorni na različne estetike (pri opisnih npr. na obliko ali barvo, pri zveznih pa na velikost, kot vrtenja, debelino ali gradient). Vsaka vrsta lestvice ima tudi svojo obliko vodila (legende, prikazane osi).

Wilkinsonov opis (specifikacija) statistične grafike obsega sedem stavkov (ki se – z izjemo prvega – v okviru posameznega opisa lahko ponovijo):

1. DATA: niz operacij, ki iz podatkovij ustvarijo spremenljivke;
2. TRANS: pretvorbe spremenljivk (npr. rangiranje – *rank*);
3. FRAME: množica spremenljivk, povezanih z operatorji, ki določa prostor;
4. SCALE: pretvorbe lestvic (koordinatnih osi, npr. logaritmiranje – *log*);
5. COORD: koordinatni sistem (npr. polarni – *polar*);
6. GRAPH: grafi (npr. točkovni – *points*) in njihove estetske lastnosti (npr. barva – *color*);
7. GUIDE: eno ali več vodil (narisane osi – *axes*, legende – *legends* ipd.).

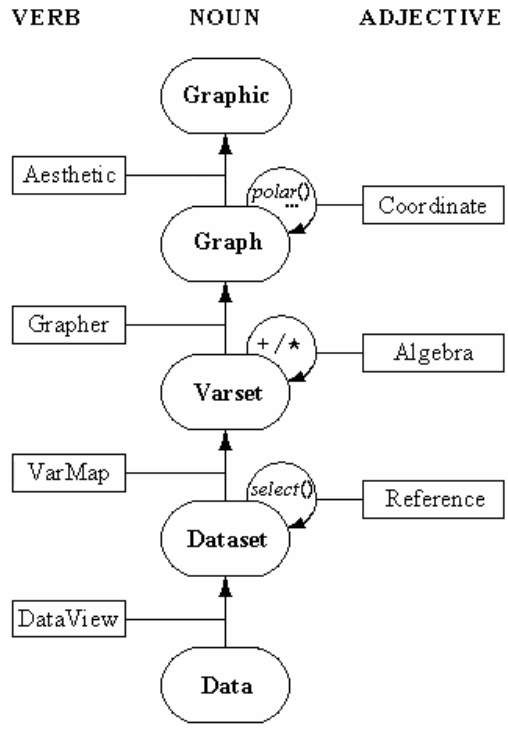
Sintakse opisa grafik tu ne moremo predstaviti v celoti, okvirno predstavo o njej pa najlaže dobimo na konkretnem primeru. Poučno in jasno grafiko, ki prikazuje stopnjo rodnosti in stopnjo umrljivosti za različne države z obrobami bivariatne gostote verjetnosti (slika 1.4 na str. 8 v izvirniku, tu reproducirana kot slika 3), opišemo oziroma izrišemo s šestimi stavki.

```
FRAME: birth*death
GRAPH: point(size(0),label(country))
GRAPH: contour.density.kernel.epanechnikov.joint(color.hue())
GUIDE: form.line(position((0,0),(30,30)),label("Zero Population Growth"))
GUIDE: axis1(label("BirthRate"))
GUIDE: axis2(label("DeathRate"))
```



***Slika 3**: Stopnja rodnosti in stopnja umrljivosti za različne države: opis na podlagi Wilkinsonove grafične slovnice in dobljena grafika (povzeto po Wilkinson, 1999, slika 1.4 na str. 8).*

Wilkinson svojo grafično slovnico prikazuje tudi kot proces (slika 4), v katerem samostalniške objekte (tipa Noun: podatki – Data, podatkovje – Dataset, spremenljivke – Varset, graf – Graph in grafika – Graphics) spreminjamo z glagolskimi objekti (tipa Verb: pogled na podatke – DataView, preslikava spremenljivk – VarMap, izdelovalec grafov – Grapher in estetika – Aesthetic) v skladu z pridevniškimi objekti (tipa Adjective: sklic – Reference, algebra – Algebra in koordinata – Coordinate).



***Slika 4**: Proces pretvorbe podatkov v grafiko v skladu z Wilkinsonovo grafično slovnico (povzeto po Wilkinson, 1999, slika 2.1 na str. 22).*

1.3.3. IMPLEMENTACIJA V JEZIKU GPL IN GRAFIČNA ALGEBRA

Implementacija grafične slovnice v jeziku GPL se od doslej navedenega razlikuje v nekaterih tehničnih vidikih. Specifikaciji je dodan stavek SOURCE, ki določa vir podatkov (to je lahko oznaka *userSource*, ki pomeni, da podatke priskrbi aplikacija, ki kliče GPL, ali ime datoteke). Namesto stavka GRAPH se uporablja stavek ELEMENT.

V grafični slovnici in še zlasti v implementaciji v jeziku GPL ima poseben pomen grafična algebra, ki določa, kako povezati podatke, da določimo položaj posameznih grafičnih prvin. Grafična algebra torej definira razsežnosti grafa oziroma podatkovni okvir, v katerem narišemo graf(e). Okvir običajnega razsevnega diagrama, na primer, določajo vrednosti ene spremenljivke, križane z vrednostmi druge spremenljivke.

Naloga grafične algebre je torej določiti spremenljivke, ki jih želimo analizirati z grafom. Če želimo v izraze grafične algebre v implementaciji GPL vključiti več kot eno spremenljivko, moramo uporabiti enega od treh operatorjev:

- Križanje (Cross, *) – Ta operator križa vse vrednosti ene spremenljivke z vsemi vrednostmi druge spremenljivke. Rezultat da za vsako statistično enoto (t.j. vrstico v matriki podatkov). Je najpogostejši operator, saj ga je potrebno uporabiti vsakič, ko ima graf več kot eno os (v dvorazsežnem koordinatnem sistemu je rezultat $A*B$ prikaz A na osi x in B na osi y). Križanje je uporabno tudi za izdelavo polj (podob, facetting), če je križanih spremenljivk več kot razsežnosti v koordinatnem sistemu. Če primer v dvorazsežnem pravokotnem koordinatnem sistemu križamo tri spremenljivke, zadnja spremenljivka določa polja (pri specifikaciji $A*B*C$ bi polja določala spremenljivka C).
- Gnezdenje (Nest, /) – Ta operator gnezdi vse vrednosti ene spremenljivke v vse vrednosti druge spremenljivke. Razlika med križanjem in gnezdenjem je, da pri slednjem rezultat obstaja le, če obstaja ustrezna vrenost spremenljivke, v katero je druga spremenljivka gnezdena. Gnezdenje vedno vodi do prikaza po poljih ne glede na koordinatni sistem.
- Mešanje (Blend, +) – Ta operator združi vse vrednosti ene spremenljivke z vsemi vrednostmi druge spremenljivke. Uporabimo ga, če želimo predstaviti dve spremenljivki na isti osi (npr. začetno in trenutno plačo zaposlenega). Pogosto ga uporabimo pri ponovljenih merjenjih (npr. v specifikaciji *salary2004+salary2005*).

Križanje in gnezdenje dodajata razsežnosti specifikaciji grafa, mešanje pa združi več nizov vrednosti v eno razsežnost. Kako medsebojno delujejo razsežnosti in algebra, je odvisno od koordinatnega sistema. Pomembno je tudi, da grafična algebra ni komutativna, a je asociativna in distributivna: $(X*Y)*Z = X*(Y*Z)$, $(X/Y)/Z = X/(Y/Z)$, $(X+Y)+Z = X+(Y+Z)$, $X*(Y+Z) = X*Y+X*Z$, $X/(Y+Z) = X/Y+X/Z$. Ob tem vidimo, da ima operator gnezdenja prednost pred ostalima dvema, križanje pa ima prednost pred mešanjem. Zato z operatorjem mešanja praviloma uporabimo oklepaje, saj spremenljivke navadno najprej zmešamo in nato rezultat križamo z drugimi spremenljivkami ali gnezdimo v druge spremenljivke.

Enotska spremenljivka (ki jo označujemo z 1) v algebri služi za nadomeščanje. Če ustvarimo os za enotsko spremenljivko, ima ta vrednost 1 na sredini in nobenih drugih vrednosti. Enotsko spremenljivko potrebujemo le, kadar v neki razsežnosti ni spremenljivke, a to razsežnost vseeno želimo vključiti v algebro. Če npr. želimo v dvorazsežnem pravokotnem

koordinatnem sistemu prikazati število zaposlenih glede na vrsto delovnega mesta (*jobcat*), vključimo v specifikacijo GPL *summary.count*(*jobcat*) in frekvenca se prikaže na navpični osi, a v tej razsežnosti ni nobene dejanske spremenljivke. Če želimo tovrsten graf uporabiti v več poljih (panelih), moramo torej pred spremenljivko, ki določa polja, določiti še spremenljivko za navpično razsežnost grafa v vsakem polju. Če želimo izdelati polja v stolpcih glede na spol zaposlenega (*gender*), moramo spremeniti specifikacijo v *summary.count*(*jobcat**1**gender*). Če želimo polja nanizati v vrsticah, pa moramo dodati enotsko spremenljivko še za določitev stolpcev, torej spremeniti specifikacijo v *summary.count*(*jobcat**1*1**gender*).

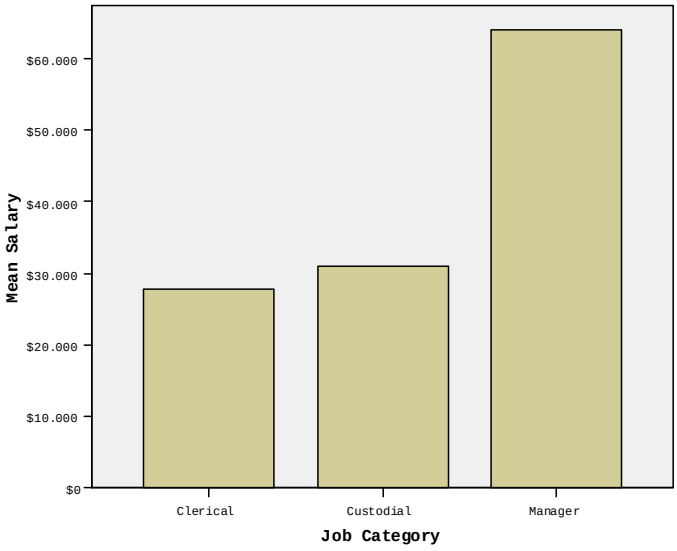
Grafična algebra lahko vsebuje tudi konstante, ki jih podamo kot nize znakov v dvojnih narekovajih (npr. "2005"). Konstanta se v algebri obnaša kot nova spremenljivka, ki ima pri vseh statističnih enotah enako vrednost. Njen učinek je torej odvisen od operatorjev in funkcij, s katerimi jo uporabimo. V funkciji *position* lahko konstante uporabimo za izdelavo ločenih osi. Če npr. želimo izdelati dve ločeni osi na diagramu s polji, gnezdimo vrednosti vsake od spremenljivk (v tem primeru *calls* in *orders*) v svojo konstanto in zmešamo rezultate, s čimer dobimo dve skupini statističnih enot, ki imata različna razpona osi (a). Konstante lahko v funkciji *position* uporabimo tudi zato, da ustvarimo dodatno kategorijo, v kateri so vse statistične enote, tako, da konstanto zmešamo z opisno spremenljivko (b). Konstante lahko koristno uporabimo tudi v funkcijah, s katerimi določamo estettike: z mešanjem dveh spremenljivk dobimo dve grafični prvini za vsako posamezno statistično enoto in če to želimo prikazati z ustreznimi barvami, estetsko funkcijo *color* uporabimo na "mešanici" dveh konstant (c).

```
(a) ELEMENT: line(position(date*(calls/"Calls"+orders/"Orders")))
(b) ELEMENT: interval(position(summary.mean((jobcat+"All")*salary)))
(c) ELEMENT: point(position(jobcat*(salbegin+salary), color("Beginning"+"Current")))
```

Uporabo jezika GPL najbolje razjasnjujejo primeri. Slika 5 prikazuje kodo GPL in stolpčni diagram, ki ga z njo izdelamo na podlagi podatkovja *Employee data*, ki je del standardne distribucije paketa SPSS. S stavkom DATA povežemo spremenljivki *jobcat* in *salary* s stolpcema v podatkovnem viru. Stavek prepozna ustrezna stolpca s pomočjo funkcije *name*, saj njen argument ustreza imenu spremenljivke v trenutnem podatkovnem viru (*userSource*).

Če bi podatke zajemali iz besedilne datoteke, v kateri so vrednosti ločene z vejicami (t.j. tipa CSV), bi moral argument funkcije *name* ustrezati eni od vrednosti v prvi vrstici datoteke. Vrsto delovnega mesta (*jobcat*) definiramo kot opisno spremenljivko (privzeto so spremenljivke zvezne). Stavek SCALE določi vrsto in razpon lestvic za razsežnosti grafa; v navedenem primeru določi linearno lestvico na drugi razsežnosti (osi *y*), ki mora vsebovati vrednost 0 (pri stolpčnih diagramih to namreč ni privzeta nastavitev). Stavek GUIDE je namenjen grafičnim prvinam, ki niso neposredno povezane s podatki, a pomagajo podatke razumeti; v navedenem primeru določa oznaki osi *x* in *y* (na osi se sklicujemo s funkcijo *dim*). Stavek ELEMENT določi vrsto grafa, spremenljivke in statistike, s čimer se graf dejansko izriše; v navedenem primeru določa *interval*, ki ustvari stolpce, tako da je položaj (*position*) stolpcev pri vsaki kategoriji spremenljivke *jobcat*, višina stolpcev pa je povprečna plača za dano kategorijo. V ta namen je uporabljena grafična algebra.

```
SOURCE: s=userSource(id("Employeedata"))
DATA: jobcat=col(source(s), name("jobcat"), unit.category())
DATA: salary=col(source(s), name("salary"))
SCALE: linear(dim(2), include(0))
GUIDE: axis(dim(2), label("Mean Salary"))
GUIDE: axis(dim(1), label("Job Category"))
ELEMENT: interval(position(summary.mean(jobcat*salary)))
```



***Slika 5**: Izdelava stolpčnega diagrama povprečne plače glede na vrsto delovnega mesta za podatkovje "Employee data" s programom SPSS z uporabo jezika GPL: specifikacija in dobljeni diagram (SPSS for Windows verzija 14.0.1, privzete nastavitvre).*

2. CILJI IN RAZISKOVALNA VPRAŠANJA

Cilj naloge je izviren prispevek k problematiki prikaza podatkov na dveh ožjih področjih. Prvi tematski sklop naloge je namenjen prikazu velike količine trorazsežnih podatkov, pri katerih je ena razsežnost opisna (kategorialna – predstavlja pripadnost določeni skupini oziroma populaciji), drugi dve pa sta številski (numerični). Drugi tematski sklop naloge se ukvarja s prikazom soglasja ocenjevalcev v primeru rangiranja (konkordanco), še zlasti v primeru, ko so ocenjevalci razdeljeni v skupine oziroma različne eksperimentalne pogoje.

V okviru prvega sklopa smo razvili in preizkusili postopke in programje za izdelavo razširjenih konveksnolupinskih diagramov (augmented convex-hull plots), v okviru drugega sklopa pa smo razvili in preizkusili postopke in programje za izdelavo diagramov za prikaz konkordance.

Eno ključnih vprašanj na področju prikaza podatkov je poenotenje problematike razumevanja, načrtovanja, izdelave in vrednotenja grafičnih prikazov. Najsodobnejši in najbolj celovit pristop k temu vprašanju predstavlja grafična slovnica, ki jo je zasnoval in razvil Wilkinson (1999). Zato smo grafično slovnico prvič predstavili v slovenskem prostoru in z njo opisali obravnavane diagrame.

V skladu z uveljavljenimi raziskovalnimi standardi na področju statistične vizualizacije (npr. Cleveland in McGill, 1984a) smo tudi empirično ovrednotili uporabnost predlaganih metod. Zaradi raznolike osnovne izobrazbe, širine strokovnih interesov ter poznavanja najsodobnejšega statističnega programja predstavljajo najprimernejšo populacijo potencialnih uporabnikov v Sloveniji študenti univerzitetnega podiplomskega študija Statistika, posebej tisti, ki so se odločili za izbirni predmet Statistika v medicini. Ovrednotili so razumljivost in uporabnost konveksnolupinskih diagramov ter uporabniško prijaznost in zmogljivost programja za njihovo izdelavo.

Raziskovalna vprašanja, iz katerih smo izhajali v prvem sklopu naloge, so:

- Ali razširjeni konveksnolupinski diagrami omogočajo hiter in pregleden prikaz velikih množic (desettisoče statističnih enot in več) podakov o dveh številskih lastnostih pri več skupinah enot?
- Ali konveksnolupinski diagrami in njihove izpeljanke z minimalno količino črt in oznak povzamejo obliko bivariatne porazdelitve ter medsebojno povezanost, srednjo vrednost, razpršenost in robno porazdelitev dveh številskih spremenljivk pri več skupinah enot?
- Je moč konveksnolupinske diagrame in njihove izpeljanke v celoti implementirati kot prilagodljiv in preprosto uporaben programski paket za okolje R?

V drugem sklopu naloge smo izhajali iz naslednjih raziskovalnih vprašanj:

- Lahko konkordanco uspešno prikažemo z diagrami na podlagi dodeljenih rangov (s konkordančnimi mehurčnimi diagrami in konkordančnimi diagrami z vzporednima osema) ter z diagrami na podlagi razlik v rangih (s stolpčnimi diagrami in diagrami blazinice z bucikami)?
- Imajo tovrstni diagrami ustrezno razmerje med podatki in črnilom in se podrejajo splošnim načelom kakovostnega prikaza podatkov?
- Ali tovrstni diagrami olajšajujejo razumevanje konkordančnih koeficientov ter omogočajo nazorno primerjavo konkordance med različnimi vzorci, skupinami oziroma eksperimentalnimi pogoji?

Obema sklopoma je skupno raziskovalno vprašanje, ali je nove tipe diagramov moč opisati na podlagi Wilkinsonove grafične slovnice.

Pri izdelavi programja nas je vodila želja, da bi raziskovalcem in drugim uporabnikom omogočalo enostavno in natančno izdelavo novih tipov diagramov ter s tem učinkovitejši oziroma ustreznejši prikaz obravnavanih vrst podatkov. Kakovostnejši prikaz podatkov namreč omogoča učinkovitejšo raziskovalno analizo podatkov (exploratory data analysis – EDA; Tukey, 1977), ustreznejšo izbiro statističnih modelov in s tem (vsaj posredno) tudi nov vpogled v preučevane probleme (tvorjenje hipotez). Kakovostni grafični prikazi tako

povečujejo razumljivost znanstvenih in strokovnih publikacij ter s tem olajšujejo prenos znanja.

Dolgoročni cilj dela je tudi, da bi pripomoglo k uveljavitvi prikaza podatkov kot samostojne znanstvene discipline v slovenskem prostoru, širilo zavest o pomenu in dobrobitih strokovnega in kakovostnega prikaza podatkov ter spodbujalo k poznavanju in uveljavljanju dobrih praks prikaza podatkov (Tufte, 1983; Wainer in Thissen, 1981). V tem pogledu ima vključitev študentov biostatističnega modula podiplomskega študija statistike poseben pomen za prihodnost raziskovalne dejavnosti v slovenski biomedicini.

3. MATERIALI IN METODE DELA

3.1. OPIS MATERIALOV

Nove metode smo implementirali z jezikom in okoljem R (R Development Core Team, 2004), elektronsko preglednico Microsoft® Excel in programskim paketom jsplot (http://ourworld.compuserve.com/homepages/jsieberer/).

3.1.1. JEZIK IN OKOLJE R ZA STATISTIČNO ANALIZO IN GRAFIKO

Jezik in okolje R za statistično analizo in grafiko (v nadaljevanju okolje R; http://www.r-project.org) je v slabem desetletju (prim. Ihaka in Gentleman, 1996) zraslo od zamisli do enega vodilnih statističnih programskih paketov, ki v akademskem okolju že prednjači kot najbolj razširjen in najzmoglivejši. Gre za brezplačno in odprtokodno implementacijo predmetnega jezika S, ki ga je zasnoval Chambers s sodelavci (Becker in Chambers, 1984), v skladu z licenco GNU (http://www.gnu.org). Okolje R je na voljo za različne izvedbe operacijskih sistemov UNIX/Linux, Windows in MacOS (mis smo uporabili verzijo za Windows). Okolje R je več kot zgolj programski paket za statistično analizo podatkov, saj združuje programsko podporo za delo s podatki, računske postopke in grafične prikaze, pri čemer ga odlikujejo:

- učinkovito shranjevanje, urejanje in spreminjanje podatkov;
- natančne in hitre računske operacije na skalarnih vrednostih, vektorjih in matrikah;
- obsežna in medsebojno povezana zbirka orodij za statistično analizo podatkov;
- zmogljiv grafični sistem za izpis na zaslon, v datoteke ali na tiskalnik;
- visoko razvit in pregleden programski jezik, ki vključuje pogojno izvajanje kode, zanke, rekurzivne funkcije in široko paleto podatkovnih struktur;
- natančno načrtovanje in podrobna usklajenost vseh sestavnih delov;
- praktično neomejena razširljivost, ki dopušča dodatke v drugih programskih jezikih (C, C++, Fortran).

® Microsoft je zaščitena blagovna znamka podjetja Microsoft Corporation.

Bistvo razširljivosti in prilagodljivosti okolja R so paketi (packages), pri čemer namestitvena distribucija vključuje le osnovni nabor paketov, ostale pa si uporabnik namešča po potrebi. Distribucija okolja R in vseh paketov poteka preko omrežja CRAN (Comprehensive R Achive Network), ki ga sestavljajo zrcaljeni datotečni strežniki po vsem svetu. Omrežje CRAN zagotavlja popolno usklajenost in osveženost vseh kopij izvorne kode in dokumentacije okolja R. Za vključitev v omrežje CRAN mora programski paket ustrezati zelo strogim in obsežnim zahtevam glede kakovosti programiranja, statistično-metodološke pomembnosti, združljivosti s celotnim okoljem R, razširljivosti in dokumentiranja zmogljivosti.

3.1.2. ELEKTRONSKA PREGLEDNICA MICROSOFT® EXCEL

Elektronske preglednice si navzlic predsodkom in nerazumnim odporom v zadnjih letih uspešno utirajo pot v akademski svet. Kljub omejitvam in pomanjkljivostim, predvsem na področju numeričnih algoritmov in ravnanja z manjkajočimi podatki, ki jih je včasih moč obiti oziroma premagati le z znatnim naporom in obsežnim znanjem (Heiser, 2006), so postale pomembno orodje na področju matematičnega in statističnega izobraževanja (Neuwirth in Arganbright, 2004), naravoslovja (de Levie, 2004) in tudi v statistični praksi[5].

Daleč najbolj razširjena in najbolj zmogljiva elektronska preglednica je Microsoft® Excel, ki deluje v operacijskih sistemih Windows in MacOS. Zanjo je na voljo izredno obsežna javna zbirka znanja v obliki uradnih novičnih skupin (*microsoft.public.excel.**) in spletnih strani vodilnih nosilcev naziva Microsoft® MVP[6], kot so Fernando Cinquegrani (http://www.prodomosua.it/ppage02.html), Jon Peltier (http://www.peltiertech.com) in John Walkenbach (http://www.j-walk.com/ss).

Na področju prikaza podatkov se elektronske preglednice odlikujejo glede možnosti za vključevanje dodatnih informacij v smislu nadgrajevanja grafikonov v ilustracije (Wilkinson,

[5] Primerno izhodišče za pregled področja je spletna stran združenja ASSUME (Association of Statistics Specialists Using Microsoft Excel) na naslovu http://www.jiscmail.ac.uk/files/ASSUME/welcome.html.

[6] Most Valuable Professional

1999). Zaradi svoje dinamičnosti in interaktivnosti so posebej primerne tudi za razvoj novih metod prikaza podatkov in hitro izdelavo prototipov diagramov.

3.1.3. PROGRAMSKI PAKET JSPLOT

Programski paket jsplot, ki je mnogo manj znan in razširjen kot okolje R, je delo enega samega avtorja in po obsegu celotne distribucije zelo majhen, a je kljub temu izjemno zmogljiv in prilagodljiv. Tudi jsplot je razvit v skladu z licenco GNU in deluje v obeh glavnih družinah operacijskih sistemov (UNIX/Linux in Windows). Njegova posebnost je, da združuje zmožnosti interaktivnega risarskega orodja, programa za risanje znanstvenih in tehničnih diagramov ter programskega jezika za razvoj grafičnih aplikacij. Posebna odlika jsplota je, da je njegov programski jezik hkrati tudi zapis (v obliki navadnega besedila) za vse risbe oziroma diagrame, s čimer so slednji samodejno tehnično dokumentirani, hkrati pa jim zlahka dodajamo nove elemente.

3.2. POTEK RAZISKAVE

Raziskavo so sestavljali trije deli: najprej smo razvili programje za implementacijo predlaganih diagramov iz obeh tematskih sklopov, sledila je uporaba novih metod za prikaz podatkov iz biostatistično-svetovalne prakse na Inštitutu za biomedicinsko informatiko Medicinske fakultete v Ljubljani, nato pa smo obravnavane diagrame opisali na podlagi Wilkinsonove grafične slovnice. Vrednotenje predlaganih metod smo dopolnili z anketo pri potencialnih uporabnikih razširjenih konveksnolupinskih diagramov na področju biostatistike.

V okviru drugega dela smo razširjene konveksnolupinske diagrame najprej uporabili za prikaz populacijskih podatkov s področja javnega zdravja (Artnik, Vidmar, Javornik in Laaser, 2006; Javornik in Korošec, 2003), ki se nanašajo na povezanost bioloških (v izbranemu primeru spola, ki določa skupino) in socio-ekonomskih dejavnikov (v izbranem primeru dohodka) z vzroki smrti in starostjo ob smrti (druga izbrana razsežnost). Na nov način smo prikazali tudi podatke iz obsežne retrospektivne študije ginekoloških kirurških posegov (Zupančič-Pridgar, 2003). Dodatno smo predstavili uporabnost konveksnolupinskih diagramov za manjša

podatkovja na znanem botaničnem primeru razvrščanja perunik na podlagi merjenja cvetov (Fisher, 1936).

Vzorec za študijo uporabnikov so sestavljali študenti podiplomskega študija statistike. V vzorec smo skušali zajeti 24 študentov, ki so v študijskih letih 2003/04, 2004/05 in 2005/06 obiskovali izbirni predmet Statistika v medicini oziroma modul Biostatistika. Sodelovanje je bilo prostovoljno. Študijo smo izvedli v obliki ankete, poslane in vrnjene po elektronski pošti. Anketa je obsegala predstavitev in preizkus razumevanja razširjenih konveksnolupinskih diagramov, primerjavo z razsevnimi diagrami, oceno zahtevnosti uporabe in uporabnosti paketa *chplot* za okolje R ter poizvedbo o poznavanju drugih metod za prikaz istovrstnih podatkov (priloga 3). Dopolnili smo jo z osnovnimi demografskimi značilnostmi anketirancev. Rezultate študije uporabnikov smo ustrezno povzeli in statistično analizirali.

Konkordančne diagrame smo razvili na podlagi preučitve možnosti prikaza konkordančnih podatkov z obstoječimi metodami ter teoretične in empirične analize prednosti, slabosti in omejitev teh možnosti. Nove konkordančne diagrame smo preizkušali na umetnih podatkih, znanih podatkih iz literature ter podatkih iz biostatistične in bibliometrične prakse.

3.3. UPORABLJENE METODE

3.3.1. RAZŠIRJENI KONVEKSNOLUPINSKI DIAGRAMI

Zasnovo in značilnosti konveksnolupinskih diagramov najbolje opisujejo značilnosti paketa *chplot*, ki ga smo ga razvili za okolje R. Paket sestavljajo tri komponente: funkcija *chplot* za risanje diagramov, funkcija *chadd* za dodajanje novih elementov na diagram ter podatkovna zbirka *hdr* (Javornik in Korošec, 2003), na katero se nanašajo primeri uporabe v navodilih za uporabo.

Funkcija *chplot* kot vhodne podatke zahteva le podatkovni okvir z vsaj dvema vektorjema (številskima spremenljivkama za vodoravno in navpično os, seveda pa podatki praviloma vsebujejo vsaj še vektor z opisno spremenljivko, ki določa skupino), a nudi številne možnosti v obliki neobveznih parametrov, za katere so privzete najbolj splošno uporabne vrednosti.

Diagram določimo s formulo (ki je v jeziku S standardni način zapisa statističnih modelov) oblike $y \sim x \mid g$, ki pomeni diagram odvisnosti spremenljivke y (risana na navpični osi) v odvisnosti od spremenljivke x (risana na vodoravni osi) glede na vrednost spremenljivke g. Spremenljivko g, ki določa pogojno porazdelitev, lahko tudi izpustimo. Imena spremenljivk določajo imena osi (y in x) in naslov legende (g). Uporabljene spremenljivke morajo biti združene v podatkovni okvir (*data frame*).

Prva neobvezna odločitev uporabnika je med konveksnimi lupinami (privzeto) in bivariatnimi obrobami gostote verjetnosti. Robne porazdelitve je moč narisati z relativnimi frekvenčnimi mnogokotniki (privzeto) ali z oceno gostote verjetnosti. V prvem primeru črte povezujejo točke v sredini razrednih intervalov, pri čemer sta prvi in zadnji interval določena tako, da v njiju ni podatkov v nobeni od skupin. Tako pri frekvenčnih mnogokotnikih kot pri diagramih ocene robne gostote verjetnosti imata zaradi neposredne primerljivosti obe robni porazdelitvi enako lestvico na osi relativne frekvence.

Opisne statistike so privzeto prikazane z daljicama, ki se križata v točki (v prostorski statistiki imenovani povprečno središče), ki jo določata povprečji obeh številskih spremenljivk, ter prikazujeta 68% interval tolerance (torej segata en standardni odklon v obe smeri vodoravno in navpično). Namesto standardnega odklona je moč izbrati standardno napako povprečja (t.j. interval zaupanja za povprečje), spremeniti pa je moč tudi stopnjo zaupanja (vrednost 0 povzroči, da prikaza opisnih statistik ni). Možno je izbrati tudi nesimetričen prikaz z daljico od prvega do tretjega kvartila na obeh oseh in križanjem daljic pri medianski vrednosti (to točko v prostorski statistiki imenujejo mediansko središče). Namesto križev, ki so vedno vzporedni z osema, je moč narisati elipse s središčem v povprečju obeh spremenljivk, ki so nagnjene v skladu s korelacijo med spremenljivkama in prikazujejo bivariatni interval zaupanja za povprečje.

Diagram je moč izrisati v barvah (privzeta možnost; privzeti nabor barv je moč spremeniti) ali črno-belo (z različnim tipom črte za vsako skupino; tudi tu lahko namesto privzetega nabor določimo sami). Nastavljivo je še razmerje med osrednjim grafom in celotnim diagramom, ki določa tudi privzeti položaj legende: za razmerje manjše ali enako privzeti vrednosti 0,75 je legenda umeščena v desni zgornji kot, sicer pa znotraj glavnega grafa, pri čemer jo uporabnik

interaktivno namesti z mišjo. Izbor razmerja 1 povzroči, da se robni porazdelitvi ne izrišeta. Poleg oziroma preko razširjenega konveksnolupinskega diagrama uporabnik lahko izriše tudi običajni razsevni diagram z izvornimi podatki. Legenda, ki jo je seveda moč izpustiti, privzeto obsega le imena skupin, na zahtevo uporabnika pa za vsako skupino vsebuje tudi izvirno bivariatno mero razpršenosti – površino konveksne lupine oziroma obrobe gostote verjetnosti na točko. Privzeta pasovna širina, izračunana po standardnem obrazcu (Everitt in Rabe-Hesketh, 2003), seveda ne more ustrezati vsem podatkom in lahko povzroči nesklenjene obrobe gostote verjetnosti, zato se da pri obrobah gostote verjetnosti nastaviti stopnjo zaupanja in potenco v obrazcu za izračun pasovne širine. Eno ali obe osi je moč narisati v logaritemskem merilu, naslove osi je moč spremeniti iz privzetih imen spremenljivk.

Poseben argument funkcije *chplot*, ki je seznamskega tipa (*list*), podrobno določa lastnosti legende – poleg že omenjenih npr. velikost pisave in izris okvirja. Uporablja vse parametre v okolje R vgrajene funkcije *legend* (pri čemer imajo nekateri drugačne privzete vrednosti kot izven funkcije, npr. naslov legende) ter dva dodatna parametra (za izpis bivariatne mere razpršenosti in položaj legende). Omogoča tudi spreminjanje vseh ostalih grafičnih parametrov legende v skladu z vgrajenimi možnostmi okolja R.

Nekaj lastnosti funkcije je pomembnih predvsem s programerskega vidika. Enote z manjkajočim podatkom o katerikoli uporabljeni spremenljivki se izpusti iz izdelave diagrama (o čemer se izpiše opozorilo), po izdelavi diagrama pa funkcija vrnila vse parametre grafičnega podsistema v predhodno stanje. Poleg tega, da izriše diagram, funkcija *chplot* vrne kot argument poseben predmet, ki služi kot vhodni podatek za funkcijo *chadd*. Ta razširjene konveksnolupinske diagrame združuje s celotno paleto zmožnosti okolja R, saj omogoča uporabniku neomejeno dodajanje novih elementov na diagram (od navadnih črt do zapletenih statističnih grafov).

3.3.2. KONKORDANČNI DIAGRAMI

Ker mere konkordance temeljijo na merah urejenostne povezanosti (razdelek 1.2.1), je nove metode za prikaz konkordance smiselno razviti na podlagi izhodišč različnih koeficientov ordinalne korelacije (npr. Kendal in Dickinson Gibbons, 1990): na podlagi dejansko

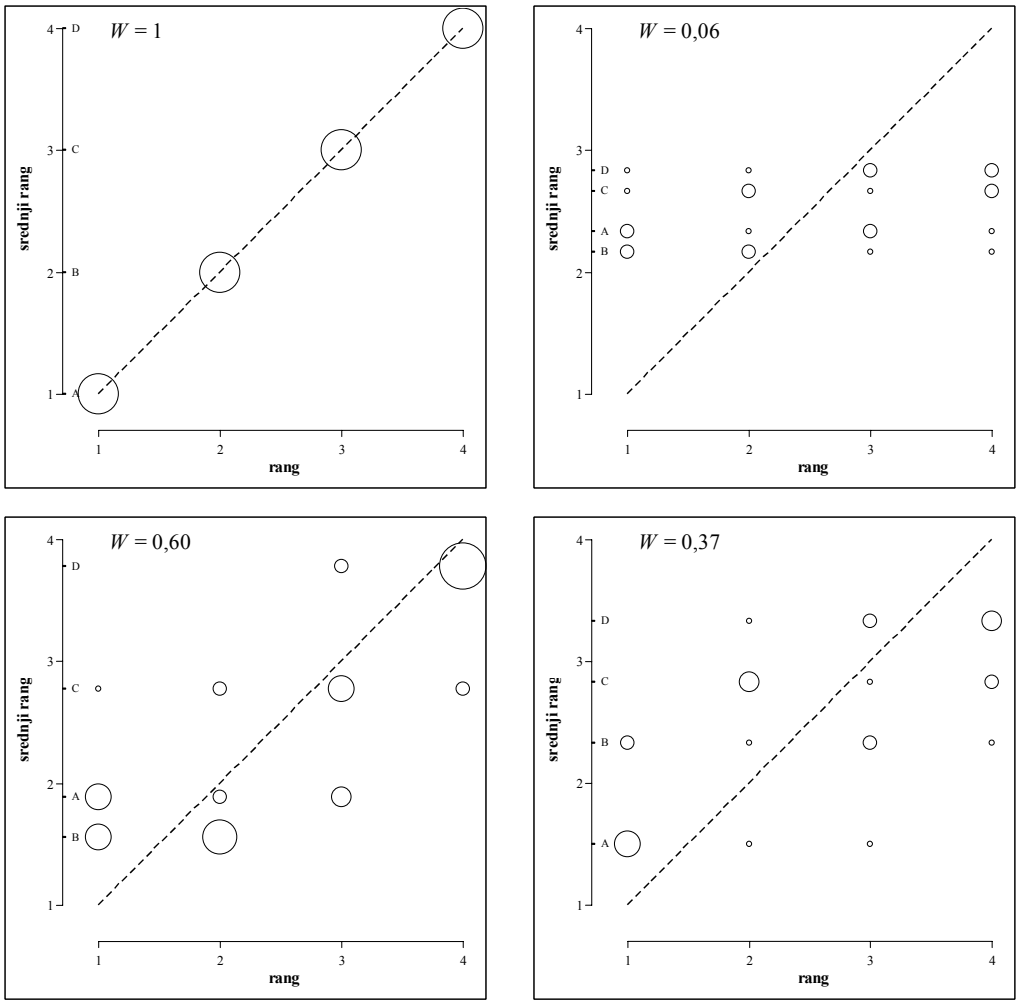
dodeljenih rangov (kar vključuje prikaz vseh opaženih dvojic rangov) ter na podlagi razlik med dvojicami rangov, dodeljenih istemu objektu. Prvi pristop, na katerem temeljita konkordančni mehurčni diagram in konkordančni diagram z vzporednima osema, je bolj neposreden; drugi pristop, na katerem temeljita konkordančni stolpčni diagram in diagram blazinice z bucikami, se bolj sklada s koeficienti konkordance, saj iz razlik med dvojicami rangov izhajata najpogosteje uporabljana koeficienta ordinalne korelacije, Spearmanov ρ in Kendallov τ.

3.3.2.1. DIAGRAMI NA PODLAGI DODELJENIH RANGOV

3.3.2.1.1. KONKORDANČNI MEHURČNI DIAGRAM

Konkordančni mehurčni diagrami imajo enak cilj kot diagrami razpršenosti med ocenjevalci, ki sta jih predlagala Nelson in Pepe (2000): prikazujejo frekvenco ocen (t.j. dodeljenih rangov) kot funkcijo povprečnega ranga (t.j. objekta). Zato predstavljajo posebno obliko razsevnega diagrama z dvema nizoma točk: v glavnem nizu velikost kroga prikazuje število istoležnih točk, v pomožnem nizu, ki leži na navpični osi in podaja ključno dodatno informacijo, pa imena objektov označujejo povprečne range.

Različne stopnje konkordance so na tak način prikazane na sliki 6. Popolno konkordanco predstavljajo krogi na glavni diagonali (diagram zgoraj levo), zato je glavna diagonala (s prekinjeno črto) prikazana na vseh tovrstnih diagramih, kar olajša presojanje soglasja oziroma nesoglasja med ocenjevalci. Brez vezanih rangov je pri številnih kombinacijah vrednosti m in k najmanjša možna stopnja konkordance večja od 0, torej povprečni rangi objektov niso enaki, a so si v vseh takih primerih zelo podobni (zgoraj desno) v primerjavi s popolno konkordanco, pri kateri se razlikujejo v največji možni meri. Diagrama v spodnjem delu slike 6 prikazujeta "klasične" podatke Schucanyja in Frawleya (1973) o dveh skupinah enologov, ki sta rangirali štiri vina, pri čemer je devet strokovnjakov iz ZDA (W=0,60; spodaj levo) med seboj soglašalo bolj kot štirje strokovnjaki iz Francije (W=0,37; spodaj desno).



***Slika 6**: Različne stopnje konkordance (popolna, najmanjša možna, srednja, nizka), prikazane s konkordančnimi mehurčnimi diagrami (k=4; m=9 pri W=0.60, sicer pa je m=6). Na zgornjih dveh diagramih so prikazani umetni podatki, na spodnjih pa podatki Schucanyja in Frawleya (1973) o dveh skupinah ocenjevalcev vin.*

Konkordančni mehurčni diagram ima visoko razmerje med podatki in črnilom (data-ink ratio – Tufte, 1983), še posebej, če krogov ne zapolnimo. Namesto mehurčnega diagrama bi lahko za prikaz konkordance uporabili tudi ustrezno prilagojen razsevni diagram z raztresenjem (jittered scatter-plot), a tovrsten diagram bi imel nižje razmerje med podatki in črnilom, pa še pri velikem številu ocenjevalcev bi lahko prišlo do prekrivanja točk med sosednjimi rangi.

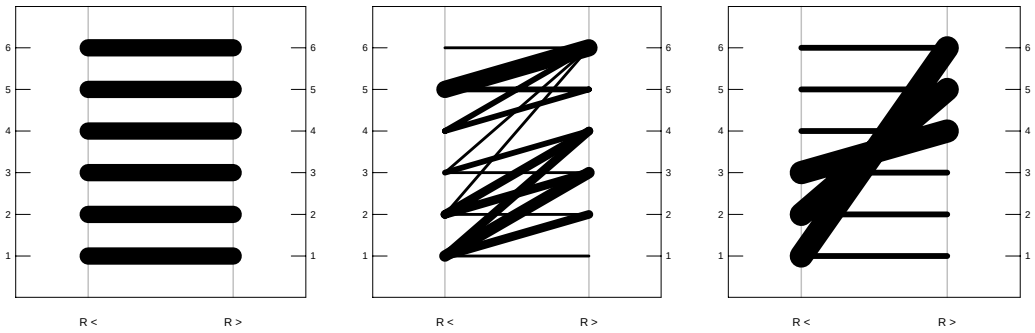
Če želimo doseči, da je na konkordančnem mehurčnem diagramu ploščina krogov premo sorazmerna s številom identičnih dvojic rangov, moramo za polmer za frekvenco *f* vzeti produkt kvadratnega korena *f* in polmera za frekvenco 1. Če uporabimo polmer, sorazmeren z *f*, bi lahko nastali graf označili za tipičen primer faktorja lažnivosti (lie factor –Tufte, 1983). Toda ogromno raziskav na področju psihofizike (npr. Gescheider, 1985) kaže, da je za

ploščino kot vidni dražljaj konkavna naraščajoča funkcija dober prvi približek splošnega odnosa med dejansko in zaznano količino (najsi to imenujemo Fechnerjev logaritemski zakon ali Stevensov potenčni zakon, pri katerem za ploščino velja potenca, manjša od 1), zato krog s polmerom 2 praviloma zaznamo le kot malo več kot dvakrat večjega od kroga s polmerom 1. Poleg tega je sodobna kognitivna znanost ovrgla smiselnost klasične psihofizične paradigme iskanja nepremenljivega odnosa med dražljajem in občutkom ter poudarila nujnost in pomen konteksta za zaznavanje (Lockhead, 1992, 1995). Zato smo se pri razvoju konkordančnih diagramov namesto za zapletene postopke psihofizičnega lestvičenja raje odločili za čim preprostejšo praktično izvedbo (s polmerom, sorazmernim z *f*) in ustvarjanje primernega konteksta za primerjave s premišljeno izbranimi grafičnimi elementi (mrežne črte, oznake). Hkrati lahko zaradi odprtokodne implementacije (priloga 5) uporabnik sam ustezno prilagodi sorazmernost polmerov.

3.3.2.1.2. KONKORDANČNI DIAGRAM Z VZPOREDNIMA OSEMA

Konkordančni diagram lahko izdelamo tudi tako, da za vsakega od objektov prikažemo vse dvojice rangov, ki so mu bile dodeljene. Pri tem se vplivu vrstnega reda ocenjevalcev v matriki podatkov na dobljeni diagram izognemo tako, da za eno koordinatno os vzamemo manjšega od rangov v dvojici (R<), za drugo (R>) pa večjega (oziroma enakega v primeru soglasja dveh ocenjevalcev glede ranga obravnavanega objekta).

Če postavimo koordinatni osi vzporedno, lahko število enakih dvojic {R<, R>} upodobimo z debelino črte. Popolno in ničelno konkordanco na tovrstnem diagramu predstavljata značilna in zlahka razpoznavna vzorca, različne stopnje zmerne in srednje konkordance pa je težje razločiti, še zlasti pri večjem številu objektov (primere podaja slika 7). Ne glede na to menimo, da si metoda že zaradi svoje nenavadnosti in s tem večje prepoznavnosti zasluži predstavitev.

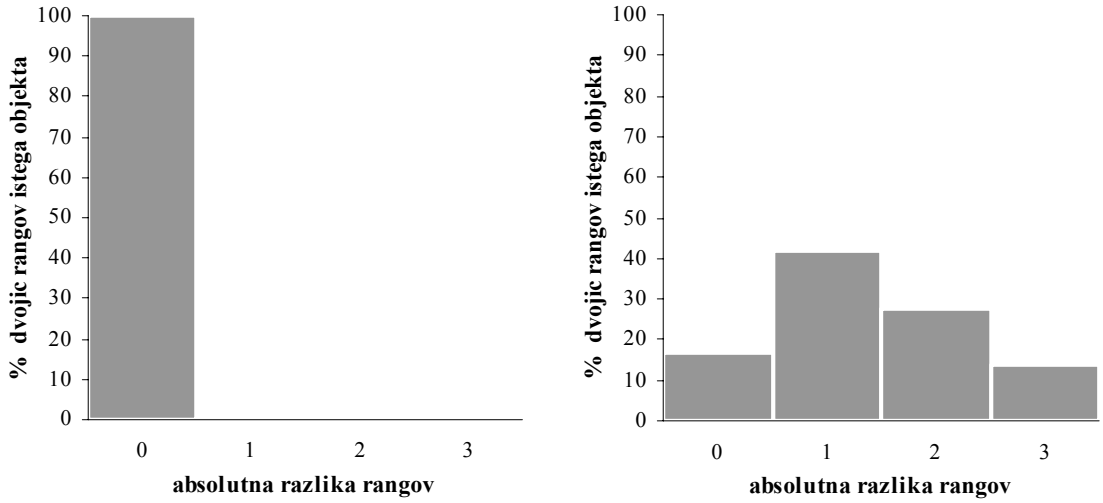


***Slika** 7: Konkordančnimi diagrami z vzporednima osema za umetne podatke (k=6; m=4) s popolno, srednjo in nizko konkordanco.*

3.3.2.2. DIAGRAMI NA PODLAGI RAZLIK RANGOV

3.3.2.2.1. KONKORDANČNI STOLPČNI DIAGRAM

Za vsako dvojico rangov, dodeljeno istemu objektu, velja, da je najmanjša možna razlika med rangoma 0, največja možna absolutna razlika pa k–1. Množica vseh opaženih razlik, ki ima $m\,(m-1)\,/\,2$ elementov, je vzorec, ki predstavlja populacijsko porazdelitev razlik. Ker je razlika rangov urejenostna spremenljivka, jo je najpreprosteje in najprimerneje prikazati s stolpčnim diagramom. Čim manjša je konkordanca, tem bolj se srednja vrednost porazdelitve razlik odmakne od 0 in tem več je stolpcev na desni strani diagrama. Primera konkordančnih diagramov, dobljena na tak način, prikazuje slika 8.



***Slika 8**: Konkordančnimi stolpčni diagram za popolno (W=1; levo) in najmanjšo možno (W=0,01; desno) konkordanco (k=4; m=9).*

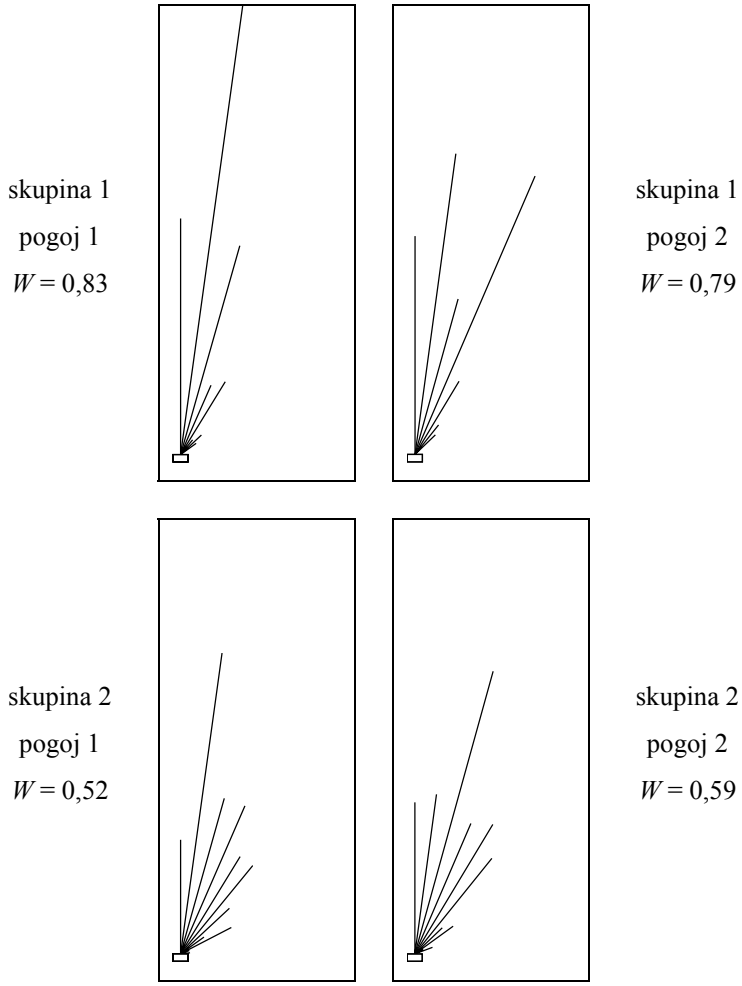
3.3.2.2.2. DIAGRAM BLAZINICE Z BUCIKAMI

Razlike med dvojicami rangov je moč na ravnini učinkovito prikazati tudi v polarnih koordinatah. Kot osnovo tovrstnega konkordančnega diagrama vzamemo navpično črto, ki predstavlja odsotnost razlike, torej dvojice enakih rangov dodeljene istemu objektu, ter merjenje kota, sorazmernega z razliko rangov, v smeri urnega kazalca. Odvisnost diagrama od zaporedja rangiranj v podatkovni matriki odpravimo tako, da upoštevamo absolutne vrednosti razlik. Največja možna razlika (k–1) predstavlja vodoravno os. Kot odmika daljice od vodoravne osi (ϕ) za dano absolutno vrednost razlike rangov ($|d|$) torej izračunamo po obrazcu

$$\phi = 90° \, (1 - |d| \, / \, (k - 1)). \qquad [11]$$

Če neke razlike v podatkih ne opazimo, pripadajoče daljice ne izrišemo. Število pojavljanj absolutne razlike določa dolžino daljice. V izhodišče diagrama narišemo pravokotnik višine, ki ustreza frekvenci 1, in širine, ki ustreza frekvenci 2. Dobljeni diagram spominja na šop bucik, zapičenih v blazinico (slika 9), pri čemer večji konkordanci ustrezajo daljše pokončne bucike in majhen delež kratkih poleglih bucik (zgornja diagrama), manjši konkordanci pa obratno (spodnja diagrama).

Slika 9 prikazuje podatke iz socialnopsihološkega eksperimenta (Vidmar in Černigoj, 2004), v katerem je sodelovalo devet skupin štirih do šestih študentov. Avtorja sta zanimali razvrstitev in skladnost ocen udeležencev glede pomena dvanajstih vedenjskih pravil pod dvema pogojema: kako bi reagirali, če bi od pravil odstopali drugi ("zunanji vidik norm", pogoj 1), in kako bi se počutili, če bi sami bili kršitelji pravil ("notranji vidik norm", pogoj 2). Ker je bil poudarek študije na metodologiji in ne na socialnopsiholoških spoznanjih, smo za prikaz izbrali dve skupini, ki sta se najbolj razlikovali glede znotrajskupinske konkordance, hkrati pa se v obeh ocene med eksperimentalnima pogojema niso znatno razlikovale. Učinkovitost prikaza konkordance lahko ocenimo na podlagi vidnega vtisa upoštevaje, da je vrednost $\mathscr{W}$ za primerjavo pogojev 1 in 2 znotraj 1. skupine 0,75 in da $\mathscr{W}$ za primerjavo skupin 1 in 2 znotraj pogoja 1 znaša 0,56.



***Slika 9**: Diagrami blazinice z bucikami za dve skupini študentov, ki sta pod dvema eksperimentalnima pogojema ocenjevali vedenjska pravila z vidika socialnih norm (Vidmar in Černigoj, 2004; k=12, m=5).*

4. REZULTATI

4.1. PROGRAMSKI PAKET CHPLOT

4.1.1. POTEK IZDELAVE IN OBJAV

Programski paket *chplot* za izdelavo razširjenih konveksnolupinskih diagramov je bil 18.9.2004 sprejet v omrežje CRAN (verzija 1.0 za R 1.9). Sledila je verzija 1.1 (sprejeta v CRAN 21.1.2005, združljiva z R 2.1), 2.6.2005 je bila v CRAN sprejeta verzija 1.1-1 (z manjšimi popravki), 12.7.2006 pa verzija 1.2 (ki je uvedla definicijo diagrama v obliki formule za opis modelov ter omogočila uporabo standardne R-ove funkcije *legend* za določanje lastnosti legende). Na ta način je paket *chplot* v širšem smislu postal del standardnega okolja R, s čimer je brezplačno na voljo statistikom, podatkovnim analitikom in vsem drugim uporabnikom po svetu (preko domače strani *CRAN* → *Packages* → *chplot* oziroma na spletnem naslovu http://cran.r-project.org/src/contrib/Descriptions/chplot.html).

O pomenu in uporabnosti konveksnolupinskih diagramov priča objava članka v mednarodnem znanstvenem časopisu *Computer Methods and Programs in Biomedicine* (Vidmar in Pohar, 2005; priloga 1), vključitev v omrežje CRAN pa potrjuje, da je programski paket zmogljiv in dodelan, podrobno dokumentiran in standardno nadgradljiv. Preprostost primerov, predstavljenih v uporabniškem priročniku (priloga 2), kaže, da je paket zaradi premišljenih prednastavitev uporaben tudi za manj vešče uporabnike okolja R.

4.1.2. REZULTATI ANKETE

Anketna vprašanja oziroma naloge so bile (za podrobnosti glej prilogo 3):

1. Namestitev paketa *chplot*, prikaz podatkovja *hdr* z razširjenim konveksnolupinskim diagramom in izbor ustreznega sklepa:
   a) ženske v splošnem živijo dlje in več zaslužijo;
   b) moški v splošnem živijo dlje, a zaslužijo manj;
   c) moški v splošnem prej umrejo, a zaslužijo več;
   d) ženske v splošnem prej umrejo in manj zaslužijo.

2. Prikaz podatkovja *sim* (4 skupine po 10000 točk) in ugotovitev, katera skupina je bila (najverjetneje) nažrebana iz porazdelitve z največjim povprečjem x in y.

3. Prikaz podatkovja *iris* (Fisherjevih perunik) z diagramom, ki ga je predlagal Cleveland (1993), in z razširjenim konveksnolupinskim diagramom (glej sliko16 v razdelku 4.3.1).

4. Prikaz podatkovja *porodi* in ustrezna interpretacija.

5. V primerjavi z ostalimi paketi za okolje R je *chplot*
   a) izrazito preprost;
   b) razmeroma preprost;
   c) podobno zahteven kot večina drugih paketov v okolju R;
   d) razmeroma zahteven;
   e) izrazito zahteven.

6. Celotna zamisel o razširjenih konveksnolupinskih diagramih je
   a) izjemno koristna;
   b) zmerno koristna;
   c) nekaj srednjega (ni povsem nepomembna, kaj posebnega pa tudi ni);
   d) v glavnem nekoristna;
   e) popolnoma nekoristna.

7. Morebitno poznavanje drugih grafičnih metod, ki se jih da uporabiti namesto razširjenih konveksnolupinskih diagramov za prikaz vrednosti dveh številskih spremenljivk glede na vrednost ene opisne spremenljivke (skupine, kategorije).

8. Leta starosti.

9. Dodiplomska izobrazba.

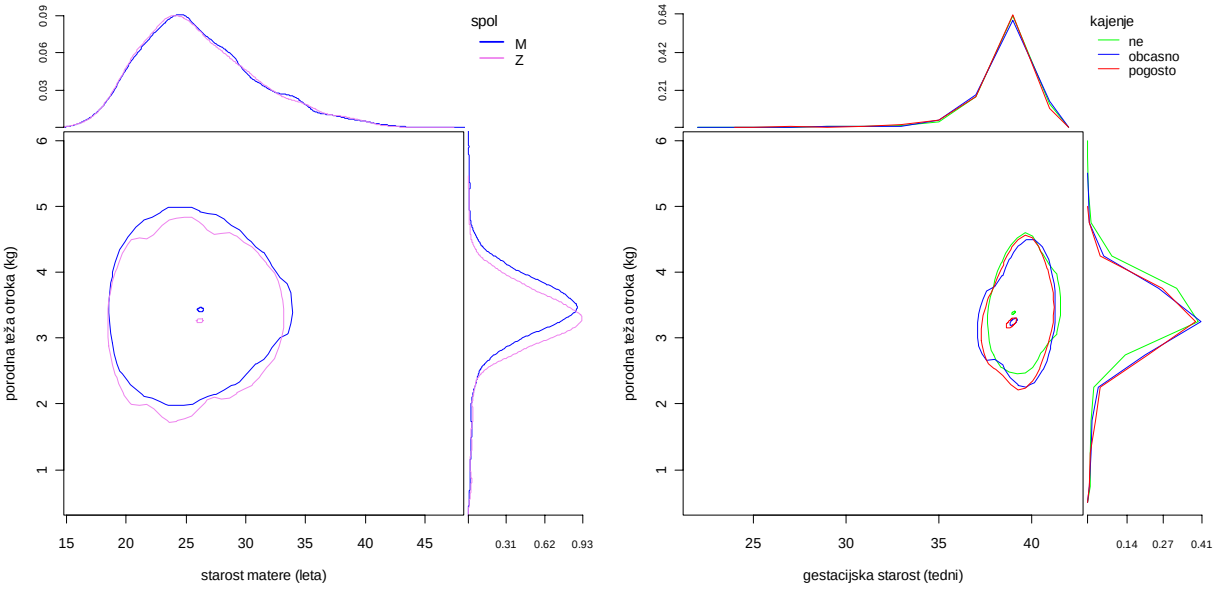
Pri prvem vprašanju je pravilni odgovor *c* (moški v splošnem umrejo prej, a zaslužijo več – slika 14 v razdelku 4.3.1), pri drugem vprašanju pa je pravilni odgovor *skupina 1* (zelene barve – slika 13 v razdelku 4.3.1). Primera smiselnega prikaza podatkov pri četrti nalogi skupaj s kodo za okolje R podaja slika 10, ustrezne interpretacije pa so

- na podlagi sinteze vseh elementov diagrama:
  - dečki se v povprečju rojevajo nekoliko težji od deklic;
  - spol otroka ni povezan s starostjo matere;
  - kajenje neugodno vpliva na porodno težo otroka, ne pa na trajanje nosečnosti;
- na podlagi robnih porazdelitev:
  - porodna teža je približno normalno porazdeljena;
  - gestacijska starost je levo asimetrično porazdeljena (na desni je omejena zaradi umetno sproženega poroda v 42. tednu nosečnosti);
  - starost matere je desno asimetrično porazdeljena;
- na podlagi elips zaupanja:
  - ocene povprečij so zaradi velikega vzorca (6356 enot) zelo natančne;
  - med porodno težo otroka in starostjo matere ni korelacije;
  - porodna teža otroka in gestacijska starost sta visoko pozitivno povezani.

Na anketo se je odzvalo 12 podiplomskih študentov statistike. Njihova starost je bila od 27 do 48 let, povprečje 35,7 let, mediana pa 33,5 let. Kljub raznoliki dodiplomski izobrazbi (ekonomija, elektrotehnika, fizika, matematika, medicina, meteorologija, računalništvo in informatika, športna vzgoja, sociologija) so bili odgovori zelo podobni.

Na prvi dve vprašanji so pravilno odgovorili vsi. Tudi podatke o perunikah je uspelo prikazati vsem, prikaz podatkov o porodih pa je uspel desetim od dvanajstih respondentov. Med uspešnimi in neuspešnimi pri tej nalogi ni bilo statistično značilne razlike v povprečni starosti (test Manna in Whitneya: eksaktni p=0,349). Težavnost uporabe paketa chplot so vsi označili kot srednjo (zdi se jim podobno zahteven kot večina drugih paketov v okolju R), celotno zamisel o razširjenih konveksnolupinskih diagramih pa je 10 respondentov označilo kot zmerno koristno, 1 kot nekaj srednjega in 1 kot izjemno koristno (negativnih mnenj ni bilo). Polovica respondentov je navedla, da poznajo drugačne načine prikaza istovrstnih podatkov, a

vsi so kot alternativo razširjenim konveksnolupinskim diagramom navedli zgolj razsevne diagrame z različno obarvanimi točkami skupin.




```
chplot(pto~starost|spol, chull=F, clevel=0.99,
  band.power=0.03, descriptives="ellipse",
  dlevel=0.99, mar.den=T, col=c("blue","violet"),
  xlab="starost matere (leta)",
  ylab="porodna teža otroka (kg)")
```

```
chplot(pto~gestac|kajenje, chull=F, clevel=0.95,
  band.power=0.1, descriptives="ellipse",
  dlevel=0.99, col=c("green","blue","red"),
  xlab="gestacijska starost (tedni)",
  ylab="porodna teža otroka (kg)")
```

***Slika 10**: Primera smiselnega prikaza podatkovja porodi z razširjenim konveksnolupinskim diagramom (dodana je pripadajoča koda za okolje R).*

4.2. KONKORDANČNI DIAGRAMI

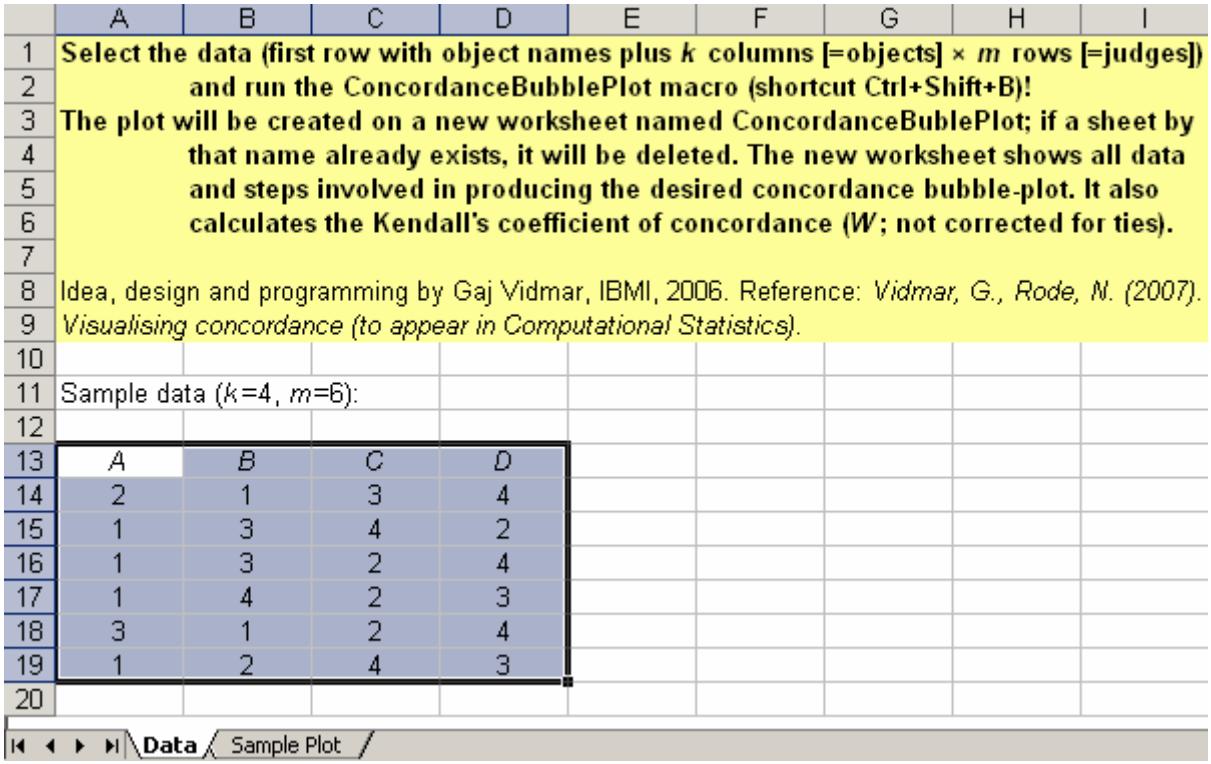
Po predstavitvi zamisli na mednarodnih znanstvenih srečanjih (*Methodology and Statistics*, Ljubljana, 2001 – http://vlado.fmf.uni-lj.si/trubar/preddvor/2001/default.htm; *Compstat Satellite Workshop on Data and Information Visualization*, Berlin, 2006 – http://z5.cms.hu-berlin.de/Zope/ise_stat/wiwi/ise/stat/forschung/veranstaltungen/div2006) in zaključenem razvoju programja je bil članek o konkordančnih diagramih sprejet v objavo v mednarodnem znanstvenem časopisu *Computational Statistics* (Vidmar in Rode, 2007; priloga 4).

Konkordančne mehurčne diagrame smo implementirali z elektronsko preglednico Microsoft® Excel, konkordančne diagrame blazinice z bucikami pa s programskim paketom jsplot. Konkordančne diagrame z vzporednima osema, ki imajo predvsem ilustrativni pomen in

hkrati zahtevajo največjo natančnost izrisa, smo izdelali ročno s programom SigmaPlot 2000 (verzija 6.0 za okolje Windows). Za prikaz porazdelitve razlik rangov s stolpčnim diagramom lahko uporabimo katerokoli statistično programje ali elektronsko preglednico.

4.2.1. DELOVNI ZVEZEK ZA IZDELAVO KONKORDANČNIH MEHURČNIH DIAGRAMOV

Delovni zvezek ima delovna lista (*Data*, na katerem je uporabniški vmesnik, in *Sample Plot*, na katerem je primer izdelanega diagrama) in modul s programsko kodo (priloga 5). Za uspešno delovanje torej potrebuje največ srednjo stopnjo varnosti Excelovih makrov. Matriko s podatki, ki ima v prvi vrstici imena objektov, označimo, nato pa poženemo makro *ConcordanceBubblePlot* (preko menija ali z bližnjico *Ctrl+Shift+B*). Makro odpre nov delovni list *ConcordanceBubblePlot* (če list s tem imenom že obstaja, ga izbriše), na katerem je razviden celoten postopek izdelave diagrama. Del izpisa je tudi Kendallov koeficient konkordance *W* (izračunan brez popravka za vezane range). Izgled delovnega zvezka prikazuje slika 11. Delovni zvezek je javno dostopen na spletnem naslovu http://www.mf.uni-lj.si/ibmi-english/biostat-center/programje/ConcordanceBubblePlot.xls.

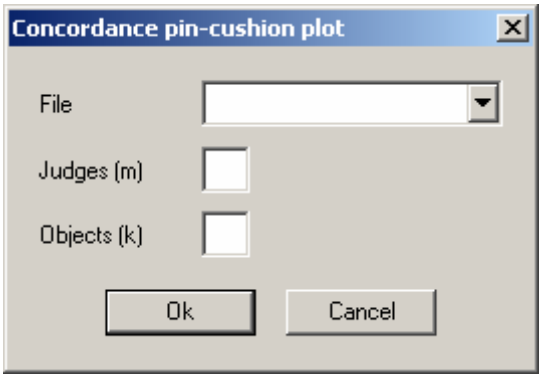


	A	B	C	D	E	F	G	H	I
1	**Select the data (first row with object names plus *k* columns [=objects] × *m* rows [=judges])**								
2	**and run the ConcordanceBubblePlot macro (shortcut Ctrl+Shift+B)!**								
3	**The plot will be created on a new worksheet named ConcordanceBublePlot; if a sheet by**								
4	**that name already exists, it will be deleted. The new worksheet shows all data**								
5	**and steps involved in producing the desired concordance bubble-plot. It also**								
6	**calculates the Kendall's coefficient of concordance (*W*; not corrected for ties).**								
7									
8	Idea, design and programming by Gaj Vidmar, IBMI, 2006. Reference: *Vidmar, G., Rode, N. (2007).*								
9	*Visualising concordance (to appear in Computational Statistics).*								
10									
11	Sample data (*k*=4, *m*=6):								
12									
13	*A*	*B*	*C*	*D*					
14	2	1	3	4					
15	1	3	4	2					
16	1	3	2	4					
17	1	4	2	3					
18	3	1	2	4					
19	1	2	4	3					
20									

***Slika 11**: Delovni zvezek za izdelavo konkordančnih mehurčnih diagramov z elektronsko preglednico Excel.*

4.2.2. PROGRAMSKA KODA ZA IZDELAVO KONKORDANČNIH DIAGRAMOV BLAZINICE Z BUCIKAMI

Programska koda v prilogi 6 omogoča izdelavo konkordančnega diagrama blazinice z bucikami. Interaktivna verzija, ki jo uporabnik zažene v okviru programa jsplot, odpre pogovorno okno (slika 12), v katerem uporabnik izbere datoteko s podatki ter vpiše število ocenjevalcev in število. Podatki morajo biti zapisani v besedilni datoteki s toliko vrsticami, kot je ocenjevalcev, znotraj vsake vrstice pa morajo biti rangi (cela števila) ločeni s preslednicami (ali tabulatorji). Pogovorno okno ima vgrajeno preverjanje ustreznosti vnosa.



***Slika 12**: Pogovorno okno za izdelavo konkordančnega diagrama blazinice z bucikami s programom jsplot.*

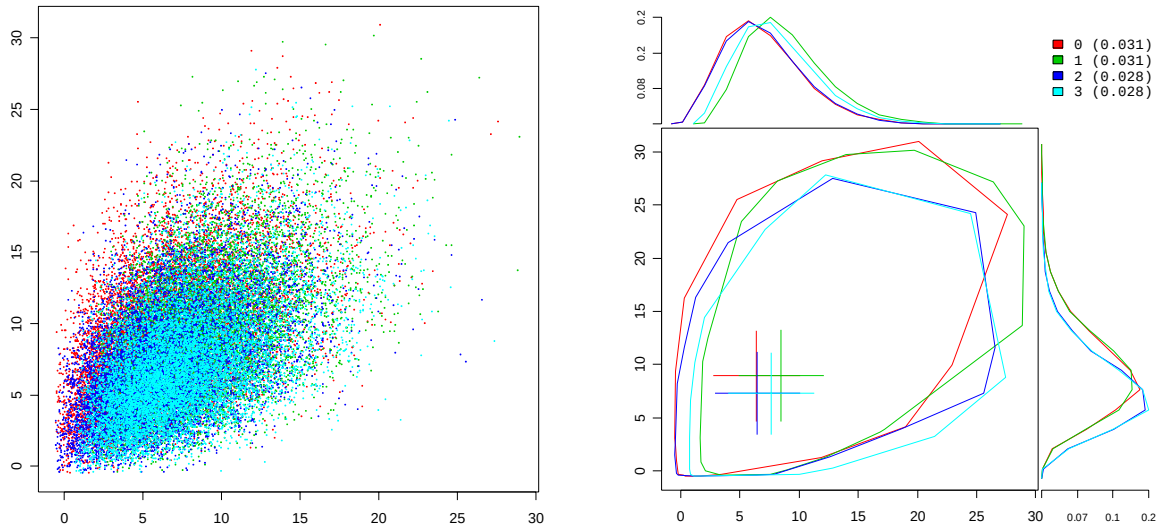
Samodejna verzija zahteva, da so podatki v datoteki *cbat.dat* v isti mapi kot datoteka s programsko kodo, in nima vgrajenega preverjanja ustreznosti podatkov. Podatkovna datoteka mora imeti v prvi vrstici dve celi števili: število ocenjevalcev in število objektov. Poženemo jo z ukazno vrstico v skladu z možnostmi, ki jih nudi jsplot – če želimo izdelati sliko v obliki Encapsulated PostScript, npr. uporabimo ukaz *jsplot -b -e -f test.eps cpcpbat.plt* (v okolju Windows ga vnesemo v ukazni vrstici ali pa vključimo v bližnjico).

Obe verziji (interaktivna je v datoteki *cpcp.plt*, samodejna pa v datoteki *cpcpbat.plt*) sta skupaj s primeri podatkovnih datotek stisnjeni v arhiv, javno dostopen na spletnem naslovu http://www.mf.uni-lj.si/ibmi-english/biostat-center/programje/ConcordancePincushionPlot.zip. Obe verziji zahtevata podatke brez vezanih rangov oziroma celoštevilske range.

4.3. PRIMERI UPORABE PREDLAGANIH DIAGRAMOV

4.3.1. PRIMERI RAZŠIRJENIH KONVEKSNOLUPINSKIH DIAGRAMOV

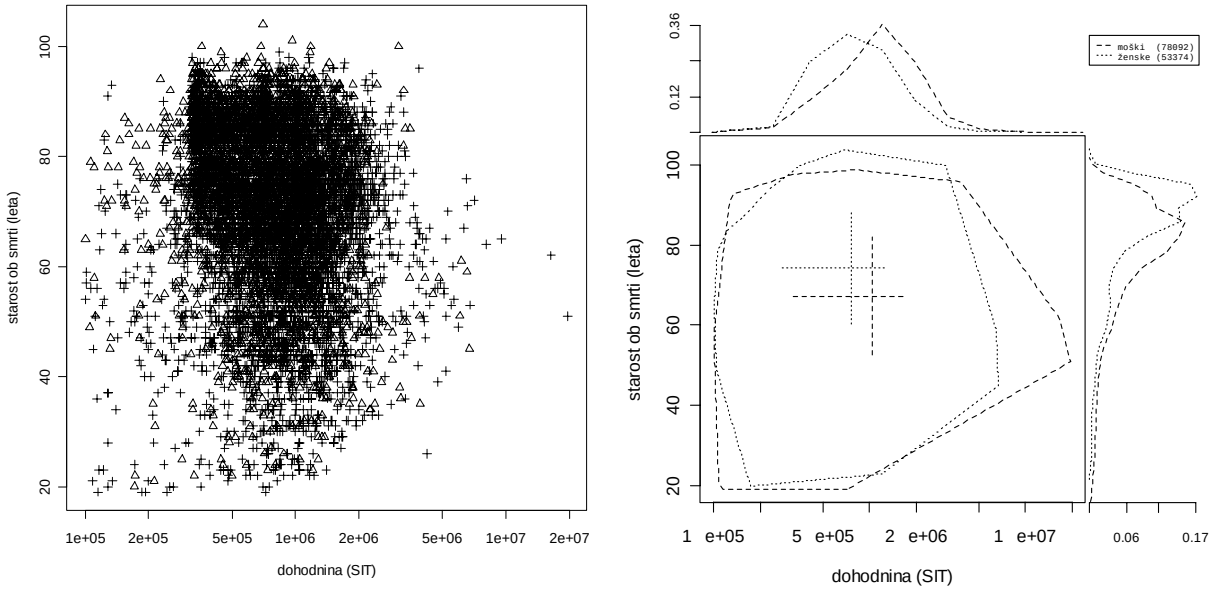
Za začetek predstavljamo razširjene konveksnolupinske diagrame na umetnih podatkih, ki kar najbolje prikazujejo njihove odlike. V ta namen smo štiri skupine po 10000 dvojic vrednosti nažrebali iz porazdelitev enake desno asimetrične oblike, a različne srednje vrednosti, ter jih prikazali z razsevnim diagramom in razširjenim konveksnolupinskim diagramom (slika 13). Na obeh digramih je pripadnost skupini zaznamovana z barvo, pri čemer je dodatna slabost razsevnega diagrama vpliv vrstnega reda risanja na končni izgled (svetlo modra barva prevladuje, ker je bila skupina 3 narisana zadnja).



***Slika 13**: Prikaz štirih skupin po 10000 dvojic vrednosti, vzorčenih iz porazdelitev enake (desno asimetrične) oblike, a različne srednje vrednosti, z razsevnim diagramom (levo) in razširjenim konveksnolupinskim diagramom (desno).*

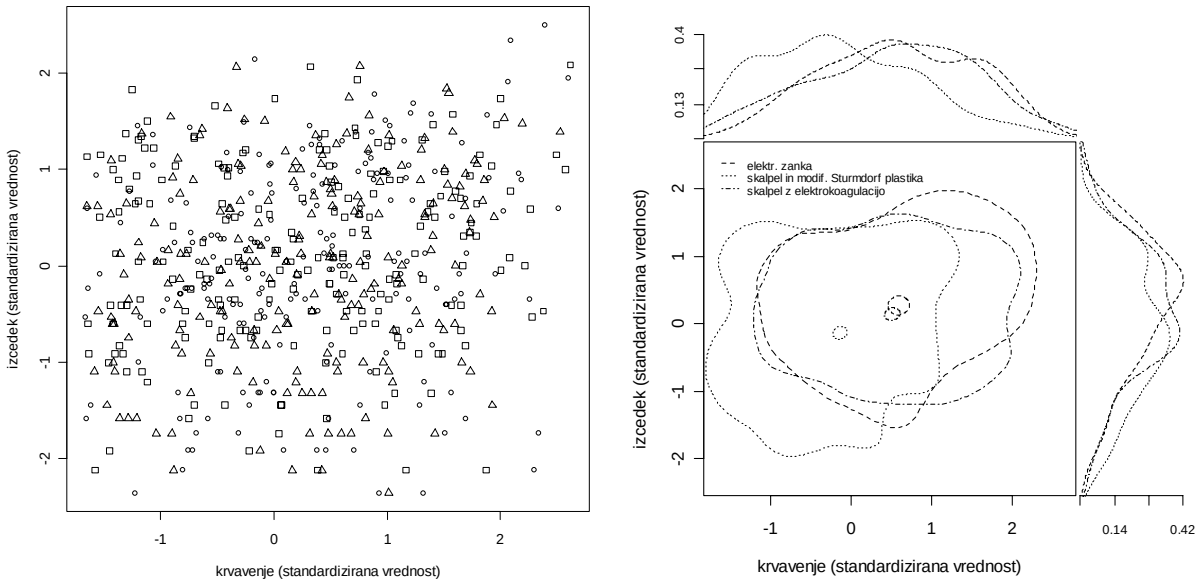
Naslednja dva primera izhajata iz naše redne biostatistične svetovalne prakse. Podatki za prvega so bili zbrani za potrebe raziskave o scioekonomskih dejavnikih umrljivosti v Sloveniji, ki je potekala v okviru Poročila o človekovem razvoju (Javornik in Korošec, 2003) in se je zaključila z objavo v mednarodnem znanstvenem časopisu (Artnik, Vidmar, Javornik in Laaser, 2005). Podatkovje *hdr*, ki je vključeno v paket *chplot* (glej razdelek 3.3.1), obsega podatke o višini odmerjene dohodnine, starosti ob smrti in spolu za 9051 oseb, umrlih v Sloveniji v letu 1998.

Primerjava socioekonomskega statusa (ki ga približno in zgolj ilustrativno povzema dohodnina) in starosti ob smrti med spoloma je prikazana na sliki sliki 14 (vodoravna os je narisana v logaritemskem merilu). Zaradi velikega števila prekrivajočih se podatkov in velike zgoščenosti bivariatne porazdelitve v manjšem obsegu je razsevni diagram povsem neuporaben, razširjeni konveksnolupinski diagram pa nazorno prikaže v splošnem nekoliko višje dohodke in nekoliko krajšo življenjsko dobo moških v primerjavi z ženskami. Hkrati nas bivariatna mera razpršenosti opozarja na večjo raznolikost moške populacije glede obravnavanih spremenljivk.



***Slika 14**: Prikaz odvisnosti starosti ob smrti od dohodnine glede na spol za podakovje hdr z razsevnim diagramom (levo; križi=moški, trokotniki=ženske) in razširjenim konveksnolupinskim diagramom (desno).*

Drugi primer prikaza podatkov iz biostatstične svetovalne prakse se nanaša na študijo dejavnikov, ki vplivajo na zaplete po konizaciji (Zupančič-Pridgar, 2003). Na sliki 15 je z razsevnim diagramom in z razširjenim konveksnolupinskim diagramom prikazan vpliv operativne tehnike na trajanje nožničnega izcedka in krvavitev po operaciji. Zaradi primerljivosti sta obe številski spremenjlivki standardizirani (izraženi v vrednostih *z*). Za razliko od razsevnega diagrama razširjeni konveksnolupinski diagram jasno pokaže prednost konizacije s skalpelom in modificirano Sturmdorf plastiko, ki vodi do manjše pogostnosti zapletov obeh vrst, kar se je v analizah tudi potrdilo kot statistično značilno. Hkrati nas nagnjenost elips zaupanja že ob prvem pregledu podatkov opozori na korelacijo med pogostnostjo obeh vrst zapletov, ki je prav tako statistično značilna in klinično pomembna.



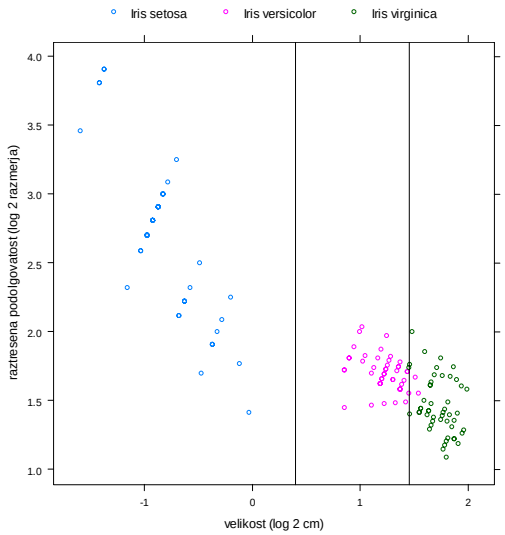
***Slika 15**: Prikaz podatkov o razlikah med operativnimi tehnikami glede zapletov po konizaciji (Zupančič-Pridgar, 2003) z razsevnim diagramom (levo; kvadrati=električna zanka, trokotniki=skalpel in modificirana Sturmdorf plastika, krog=skalpel z elektrokoagualcijo) in razširjenim konveksnolupinskim diagramom (desno).*

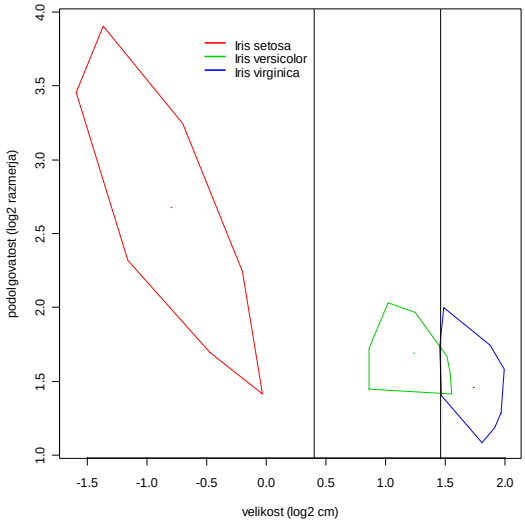
Četrti primer uporabe razširjenih konveksnolupinskih diagramov se nanaša na podatke s področja botanike, ima pa posebno težo in posebej širok pomen. "Fisherjeve perunike" (Fisher, 1936) so namreč po vsej verjetnosti največkrat analizirano in prikazano podatkovje v statistiki in na sorodnih področjih, zato jih je izredno težko analizirati ali prikazati na izviren način oziroma učinkoviteje kot doslej. Naloga je še toliko zahtevnejša, ker se jim je posebej posvetil eden od najvplivnejših raziskovalcev prikaza podatkov (Cleveland, 1993, str. 299-301). Podatki, ki jih je zbral Anderson (1935), za vsakega od 150 cvetov (po 50 za vsako od treh vrst) obsegajo štiri značilnosti: dolžino in širino čašnih in venčnih listov (sepal lengh, sepal width, petal length, petal width).

Slika 16 primerja Clevelandov diagram (izvorna koda funkcije *book.5.21* v jeziku S z naslova http://cm.bell-labs.com/cm/ms/departments/sia/wsc/visualizing.scripts je prirejena za okolje R) s konveksnolupinskim diagramom, pri katerem prikaz robnih porazdelitev ni potreben za ilustracijo razvrščanja. Oba prikaza izhajata iz pretvorbe izvornih podatkov v duhu Tukejevega diagrama *m*, saj so razlike med vrstami najbolj jasne, če v logaritemskem merilu opazujemo odvisnost razmerja dolžine in širine venčnih listov ("podolgovatosti") od njunega zmnožka ("velikosti"). Če simbole, ki se na razsevnem diagramu delno prekrivajo, nadomestimo s konveksnimi lupinami, je slika še jasnejša in bolj očitno je, da so večji venčni

listi manj podolgovati (tako znotraj posamezne vrste kot pri vseh vrstah skupaj), ter da lahko na podlagi velikosti popolnoma ločimo vrsto *setosa* od ostalih dveh (vrsto *versicolor* od vrste *virginica* pa skoraj popolnoma).

Naknadno dodane navpične črte tudi na konveksnolupinskem diagramu prikazujejo Clevelandovi ločevalni pravili (kar v okolju R dosežemo z dvema preprostima ukazoma *abline*). Opisne statistike niso prikazane (*dlevel*=0) in parameter *ratio* je postavljen na 1, da je prikaz robnih porazdelitev izpuščen, kar omogoča tudi neposredno primerljivost s Clevelandovim izvirnikom. Za razliko od Clevelandovega prikaza konveksnolupinski diagram ne bi bil nič manj jasen, če bi bil izdelan brez uporabe barv (t.j. s parametrom *bw*=T). Dodatno prednost daje našemu diagramu preglednost in enostavnost programske kode oziroma uporabe potrebnega paketa za okolje R (*chplot* namesto sicer zmogljivejšega in širše uporabnega, a tudi bistveno bolj zapletenega paketa *lattice*).






```
data(iris)
library(lattice)
set.seed(19)
petal.length <- iris[,3,]
petal.width <- iris[,4,]
variety <- factor(iris[,5,])
n <- length(levels(variety))
mea <- (log(petal.length,2)+log(petal.width,2))/2
dif <- jitter(log(petal.length,2)-
 log(petal.width,2), 2)
xyplot(dif ~ mea,
 panel = function(...){
 panel.superpose(...)
 panel.abline(v = c(0.4, 1.46))
 },
 groups = variety,
 aspect = 1,
 xlab = "velikost (log 2 cm)", ylab =
 "raztresena podolgovatost (log 2 razmerja)",
 key = list(points = Rows(trellis.par.get
 ("superpose.symbol"), 1:n),
 text = list(paste("Iris", levels(variety))),
 columns = n))
```

```
data(iris)
 library(chplot)
 x<-log(sqrt(iris[,3]*iris[,4]))/log(2)
 y<-log(iris[,3]/iris[,4])/log(2)
 variety <- factor(paste("Iris",iris[,5,]))
 param<-chplot(y ~ x | variety,
  legend=list(area.in=F),
  xlab="velikost (log2 cm)",
  ylab="podolgovatost (log2 razmerja)",
  dlevel=0,ratio=1)
 chadd(param,1,abline,v=1.46)
 chadd(param,1,abline,v=.4)
```

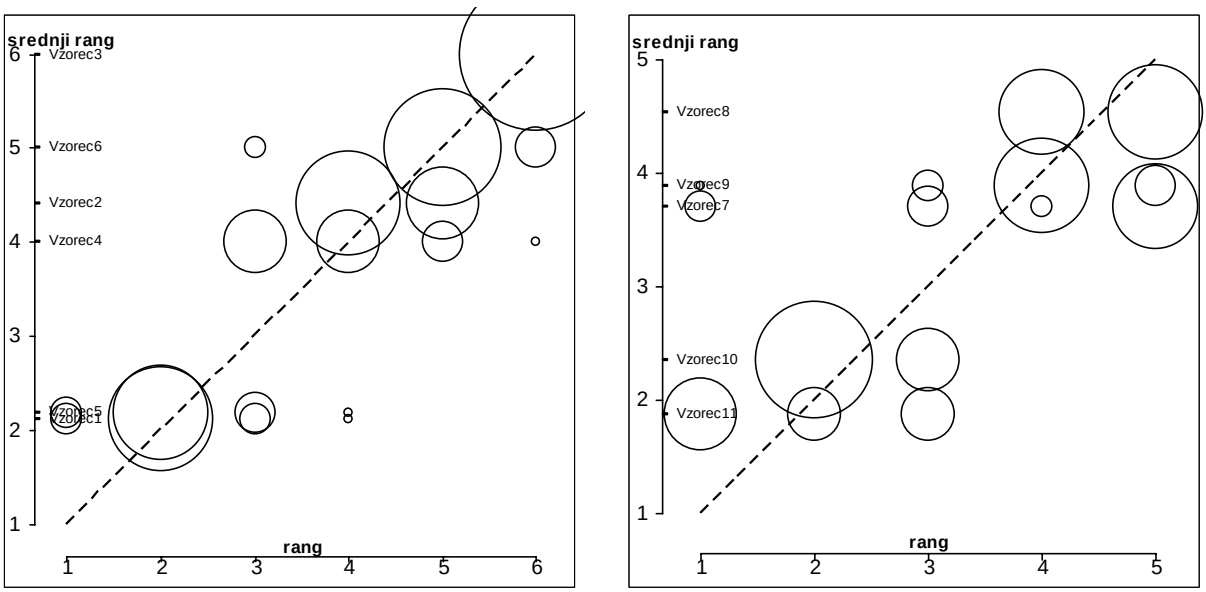
***Slika 16**: Razločevanje med vrstami perunik (Fisher, 1936) na podlagi velikosi in oblike venčnih listov, kot ga je predlagal in prikazal Cleveland (1993; levo), in z konveksnolupinskimi diagrami (desno). Pod diagramoma je pripadajoča koda za okolje R; namenoma so uporabljene privzete barve, seveda pa se da kateregakoli od diagramov tudi barvno izenačiti z drugim (z opcijo color).*

4.3.2. DODATNI PRIMERI KONKORDANČNIH DIAGRAMOV

Konkordančne diagrame lahko koristno uporabimo v medicini, zlasti na področjih, kjer je pogosto ocenjevanje v obliki rangiranja. Študije soglasja med ocenjevalci ali diagnostičnimi postopki, ki zaradi urejenostne narave podatke uporabljajo Kendallov *W* kot statistično mero soglasja, se sicer uveljavljajo šele v zadnjem času: iskanje po bibliografski podatkovni zbirki PubMed s presekom izrazov *Kendall* in *concordance* da 49 zadetkov, od katerih jih je 28 iz let od 1996 dalje, od tega kar 11 iz let 2005 in 2006.

Za prvi primer smo izbrali kar najbolj svež članek iz uveljavljene mednarodne revije, iz katerega se je dalo dobiti surove podatke, na katerih so bili izračunani koeficienti konkordance (Netto in sod., 2006). Gre za del prospektivne randomizirane večcenterske študije s 312 bolniki, razdeljenimi v tri skupine glede na vrsto zdravljenja. V študiji soglasja je sodelovalo 17 patologov, ki so na 6 biopsijskih vzorcih (*Vzorec1*-*Vzorec6*) ocenjevali stopnjo in stadij kroničnega hepatitisa C (HCV) v skladu s shemo Battsa in Ludwiga, na 5 biopsijskih vzorcih (*Vzorec7*-*Vzorec11*) pa akutno celično zavrnitev (acute cellular rejection, ACR) v skladu z Banffovo ocenjevalno shemo.

Za prikaz s konkordančnim mehurčnim diagramom (slika 17) smo izbrali podatke z razmeroma visoko (stadij HCV) in razmeroma nizko stopnjo konkordance (globalna stopnja ACR), pri čemer smo izvorne ocene rangirali z dodelitvijo najvišjega možnega ranga v primeru vezanih rangov. To je sicer manj običajno od dodelitve srednjega ranga, a odpravi necele range in s tem poenostavi prikaz, na vrednost koeficienta konkordance pa tako ne vpliva, katero vrednost v razponu med spodnjo in zgornjo mejo dodelimo vezanim rangom. Na prvem diagramu se večje soglasje jasno kaže v večjem deležu ploščine krogov v bližini glavne diagonale, hkrati pa na obeh diagramih povprečni rangi izpostavljajo vzorca, ki so ju patologi ocenjevali zelo podobno (1 in 5 pri stadiju HCV ter 7 in 9 pri globalni stopnji ACR).



***Slika 17**: Konkordančna mehurčna diagrama soglasja patologov (m=17) glede ocene stadija kroničnega hepatitisa (levo; k=6 ocenjevanih vzorcev; W=0,85) in ocene globalnega stadija akutne celične zavrnitve (desno; k=5 ocenjevanih vzorcev; W=0,57) v okviru obsežne večcenterske študije (Netto in sod., 2006).*

Drugi primer se nanaša na področje scientometrije[7]. Podatke o faktorju vpliva (IF, iz zbirke JCR – *Journal Citations Reports* –proizvajalca Thomson Scientific, dostopne preko spletnega vmesnika COBISS/OPAC Insituta informacijski znanosti) za leta 1994 do 2005 smo primerjali med sedmimi znanostmi oziroma področji, pri čemer smo analizirali skladnost rangiranja področij med štirimi možnimi merami "pomembnosti" področja ter med obravnavanimi dvnajstimi leti znotraj vsake mere. Področja so bila (navedena po abecedi) agronomija, fizika, kemija, medicina, psihologija, statistika in računalništvo (tabela 1), mere pa najvišji IF (max), povprečje 20 najvišjih IF (M_{20}), mediana 10 najvišjih IF (Me_{10}) in število časopisov z IF na danem področju (No). Najvišja vrednost mere je dobila rang 1, najnižja 7.

Dobljeni rangi so zbrani v tabeli 2 in konkordanca med njimi je prikazana na sliki 18. Za izračun oziroma prikaz konkordance med leti smo podatke iz tabele 2 ustrezno preuredili. Ugotovili smo presenetljivo visoko soglasje med merami, čeprav se število časopisov z IF na danem področju ne zdi smiselna mera oziroma nima razvidne veljavnosti, poleg tega pa so s statističnega vidika ocene kvantilov izrazito nestabilne, še zlasti ekstremnih kvantilov, kot je maksimum.

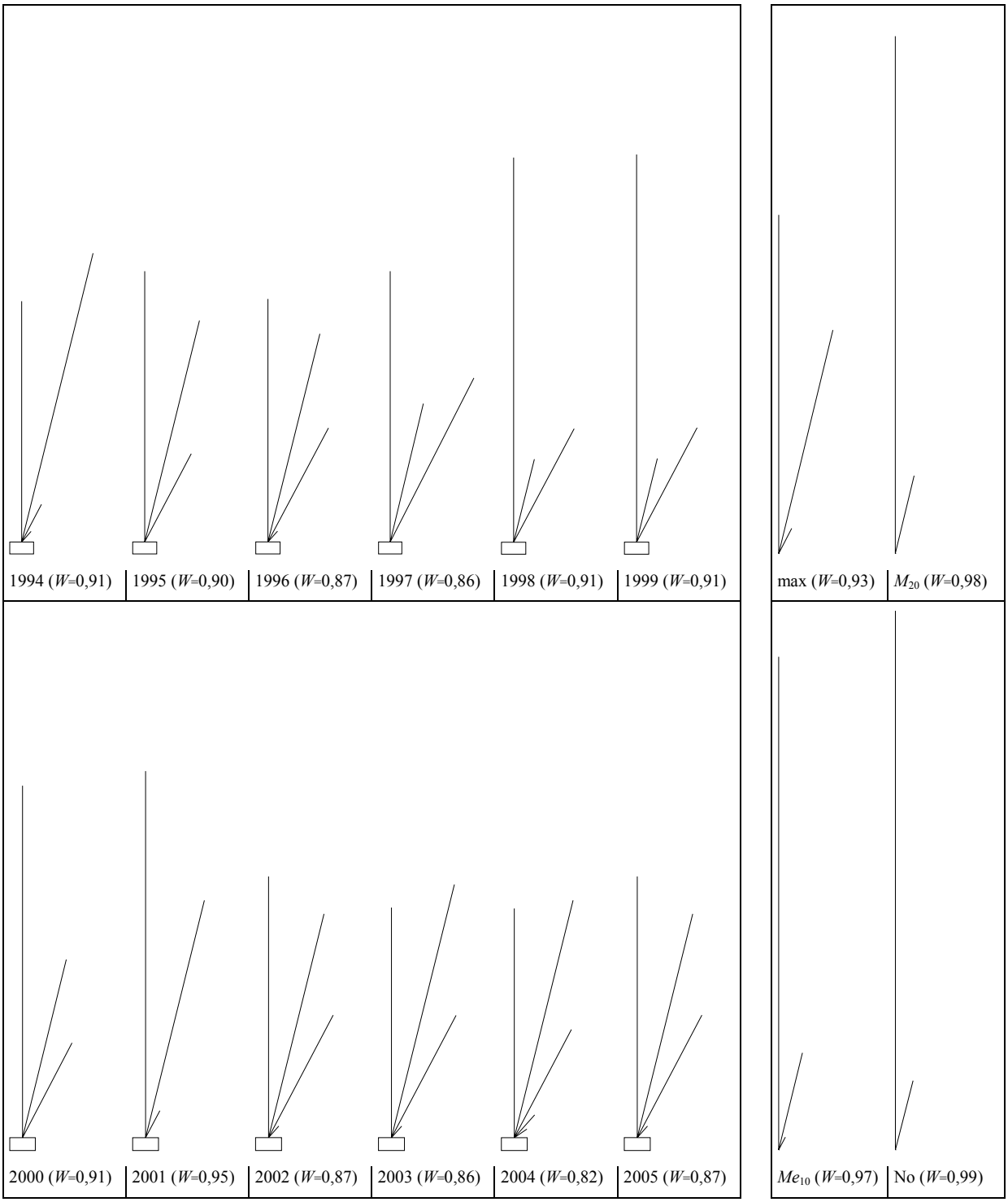
[7] Gre za nadaljevanje dela, ki je pri nas pionirsko z vidika mednarodne objave (Hudomalj in Vidmar, 2003).

***Tabela 1**: Iskalne zahteve za zbiranje podatkov o vseh časopisih z danega podorčja v zbirki JCR. (1994-2005).*

področje	**iskalna zahteva v polju Kategorija**	**kategorije (CA=)**
agronomija	ag*	AGRICULTURAL ECONOMICS & POLICY AGRICULTURAL ENGINEERING AGRICULTURE AGRICULTURE, DAIRY & ANIMAL SCIENCE AGRICULTURE, MULTIDISCIPLINARY AGRICULTURE, SOIL SCIENCE AGRONOMY
fizika	physic*	PHYSICS PHYSICS, APPLIED PHYSICS, ATOMIC, MOLECULAR & CHEMICAL PHYSICS, CONDENSED MATTER PHYSICS, FLUIDS & PLASMAS PHYSICS, MATHEMATICAL PHYSICS, MULTIDISCIPLINARY PHYSICS, NUCLEAR PHYSICS, PARTICLES & FIELDS
kemija	ch*	CHEMISTRY CHEMISTRY, ANALYTICAL CHEMISTRY, APPLIED CHEMISTRY, CLINICAL & MEDICINAL CHEMISTRY, INORGANIC & NUCLEAR CHEMISTRY, MEDICINAL CHEMISTRY, MULTIDISCIPLINARY CHEMISTRY, ORGANIC CHEMISTRY, PHYSICAL
medicina	medic*	MEDICAL ETHICS MEDICAL INFORMATICS MEDICAL LABORATORY TECHNOLOGY MEDICINE, GENERAL & INTERNAL MEDICINE, LEGAL MEDICINE, MISCELLANEOUS MEDICINE, RESEARCH & EXPERIMENTAL
psihologija	psycho*	PSYCHOLOGY PSYCHOLOGY, APPLIED PSYCHOLOGY, BIOLOGICAL PSYCHOLOGY, CLINICAL PSYCHOLOGY, DEVELOPMENTAL PSYCHOLOGY, EDUCATIONAL PSYCHOLOGY, EXPERIMENTAL PSYCHOLOGY, MATHEMATICAL PSYCHOLOGY, MULTIDISCIPLINARY PSYCHOLOGY, PSYCHOANALYSIS PSYCHOLOGY, SOCIAL
računalništvo	comp*	COMPUTER APPLICATIONS & CYBERNETICS COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE COMPUTER SCIENCE, CYBERNETICS COMPUTER SCIENCE, HARDWARE & ARCHITECTURE COMPUTER SCIENCE, INFORMATION SYSTEMS COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS COMPUTER SCIENCE, SOFTWARE ENGINEERING COMPUTER SCIENCE, SOFTWARE, GRAPHICS, PROGRAMMING COMPUTER SCIENCE, THEORY & METHODS
statistika	st*	STATISTICS & PROBABILITY

***Tabela 2**: Rang raziskovalnih področij glede na različne mere, dobljene na podlagi faktorjev vpliva iz zbirke JCR, za leta 1994 do 2005.*

leto	mera	agronomija	fizika	kemija	medicina	psihologija	računalništvo	statistika
1994	max	7	2	3	1	4	5	6
	M_{20}	7	3	2	1	4	5	6
	Me_{10}	7	3	2	1	4	5	6
	No	6	5	1	2	3	4	7
1995	max	7	3	2	1	4	6	5
	M_{20}	6	3	2	1	4	5	7
	Me_10	6	3	2	1	4	5	7
	No	6	5	1	2	3	4	7
1996	max	6	2	3	1	4	5	7
	M_{20}	6	3	2	1	4	5	7
	Me_{10}	7	3	2	1	4	5	6
	No	6	5	1	3	2	4	7
1997	max	7	3	2	1	4	6	5
	M_{20}	6	3	2	1	4	5	7
	Me_{10}	6	3	2	1	4	5	7
	No	6	5	1	3	2	4	7
1998	max	6	3	2	1	4	5	7
	M_{20}	6	3	2	1	4	5	7
	Me_{10}	6	3	2	1	4	5	7
	No	6	5	1	3	2	4	7
1999	max	6	3	2	1	4	5	7
	M_{20}	6	3	2	1	4	5	7
	Me_{10}	6	3	2	1	4	5	7
	No	6	5	1	3	2	4	7
2000	max	6	4	2	1	3	5	7
	M_{20}	6	3	2	1	4	5	7
	Me_{10}	6	3	2	1	4	5	7
	No	6	5	1	3	2	4	7
2001	max	6	4	2	1	3	5	7
	M_{20}	6	3	2	1	4	5	7
	Me_{10}	6	3	2	1	4	5	7
	No	6	5	1	2	3	4	7
2002	max	7	2	3	1	4	5	6
	M_{20}	6	3	2	1	4	5	7
	Me_{10}	6	3	2	1	4	5	7
	No	6	5	1	3	2	4	7
2003	max	7	2	3	1	4	5	6
	M_{20}	7	3	2	1	4	5	6
	Me_{10}	6	3	2	1	4	5	7
	No	6	5	1	3	2	4	7
2004	max	7	2	3	1	4	5	6
	M_{20}	6	2	3	1	4	5	7
	Me_{10}	7	1	3	2	4	5	6
	No	6	5	1	3	2	4	7
2005	max	7	2	3	1	4	5	6
	M_{20}	6	3	2	1	4	5	7
	Me_{10}	6	3	2	1	4	5	7
	No	6	5	1	3	2	4	7



***Slika 18**: Skladnost rangiranja sedmih znanstvenih področij (k=7) glede na podatke o faktorjih vpliva (JCR 1994-2005): skladnost med merami v različnih letih (m=4; leva skupina diagramov) in skladnost med leti na podlagi izbrane mere (m=12; desna skupina diagramov). Za pojasnila glej besedilo in tabeli 1 in 2.*

4.4. OPIS PREDLAGANIH DIAGRAMOV NA PODLAGI WILKINSONOVE GRAFIČNE SLOVNICE

4.4.1. OPIS RAZŠIRJENIH KONVEKSNOLUPINSKIH DIAGRAMOV

Opisali smo diagrame na slikah 14, 15 in 16. Vsakokrat je najprej opisan razsevni diagram in nato (pri slikah 14 in 15 razširjeni) konveksnolupinski diagram. Pri prvem opisu pomeni spremenljivka *doh* dohodnino in *ssm* starost ob smrti, pomen ostalih spremenljivk pa je samoumeven.

Prvi opis razširjenega konveksnolupinskega diagrama predstavlja izhodišče za razumevanje in "gradnjo" vseh ostalih opisov. Pri prvem in tretjem opisu je za oba diagrama uporabljena logaritemska lestvica. Pri prvem in drugem opisu razširjenega konveksnolupinskega diagrama je bistvena prvina prikaz robnih porazdelitev z uporabo grafične algebre, oba prikaza pa uporabljata tudi statistike (ocenjevanje univariatne oziroma bivariatne gostote verjetnosti), torej kompleksne grafične objekte. Grafična algebra je uporabljena tudi za določitev položaja legende pri prvem in tretjem opisu (razširjenega) konveksnolupinskega diagrama.

Opisi so narejeni v skladu s splošno obliko Wilkinsonove grafične slovnice, saj obstoječa implementacija jezika GPL še nima vseh potrebnih zmožnosti in tudi morebitni prenos v drugo okolje (npr. v R s paketom *ggplot*) je lažji iz splošnejše oblike. Opis prvega konveksnolupinskega diagrama ne obsega izračuna in izpisa bivariatne mere razpršenosti, za kar bi bilo potrebno grafični slovnici dodati ustrezno bivariatno statistično transformacijo, nato pa bi lahko spremenljivko, ki bi ji jo priredili, uporabili v funkciji *label*.

Opis diagramov na sliki 14:

levo (razsevni diagram)
FRAME: doh*ssm
SCALE: log(dim2,10)
GRAPH: point(shape(spol))
GUIDE: axis1(label("dohodnina (SIT)"))
GUIDE: axis2(label("starost ob smrti (leta)"))
desno (razširjeni konveksnolupinski diagram)
FRAME: doh*ssm
SCALE: log(dim2,10)
GRAPH: point(size(0))
GRAPH: link.edge.hull()
GRAPH: line.bin.rect(position.stack(doh*1),granularity(spol))
GRAPH: line.bin.rect(position.stack(1*ssm),granularity(spol))
GUIDE: axis1(label("dohodnina (SIT)"))
GUIDE: axis2(label("starost ob smrti (leta)"))
GUIDE: legend(granularity(spol),position.stack(doh*ssm*1))

Opis diagramov na sliki 15:

levo (razsevni diagram)
FRAME: krvavenje*izcedek
GRAPH: point(shape(tehnika))
GUIDE: axis1(label("krvavenje (standardizirana vrednost)"))
GUIDE: axis2(label("izcedek (standardizirana vrednost)"))
desno (razširjeni konveksnolupinski diagram)
FRAME: krvavenje*izcedek
GRAPH: point(size(0))
GRAPH: cotour.density.kernel.normal.joint(0.95)
GRAPH: line.density.kernel.normal(position.stack(krvavenje*1),granularity(tehnika))
GRAPH: line.density.kernel.normal(position.izcedek(1*ssm))
GUIDE: axis1(label("krvavenje (standardizirana vrednost)"))
GUIDE: axis2(label("izcedek (standardizirana vrednost)"))
GUIDE: legend.interior(granularity(tehnika),position(–1.5,2.5))

Opis diagramov na sliki 16:

levo (razsevni diagram)
TRANS: velikost=(petallength*petalwidth)^0.5
TRANS: podolgovatost=petallength/petalwidth
FRAME: velikost*podolgovatost
SCALE: log(dim1,2)
SCALE: log(dim2,2)
GRAPH: point(position.jitter(),color(species))
GUIDE: axis1(label("velikost (cm)"))
GUIDE: axis2(label("raztresena podolgovatost (razmerje)"))
GUIDE: form.line(position((0.4,2),(0.4,16)))
GUIDE: form.line(position((1.46,2),(1.46,16)))
GUIDE: legend.horizontal(color(species),position.stack(velikost*1))
desno (konveksnolupinski diagram)
TRANS: velikost=(petallength*petalwidth)^0.5
TRANS: podolgovatost=petallength/petalwidth
FRAME: velikost*podolgovatost
SCALE: log(dim1,2)
SCALE: log(dim2,2)
GRAPH: point(size(0))
GRAPH: link.edge.hull(color(species))
GUIDE: axis1(label("velikost (cm)"))
GUIDE: axis2(label("raztresena podolgovatost (razmerje)"))
GUIDE: form.line(position((0.4,2),(0.4,16)))
GUIDE: form.line(position((1.46,2),(1.46,16)))
GUIDE: legend.interior(color(species).position(1.5,10))

4.4.2. OPIS KONKORDANČNIH DIAGRAMOV

Opisali smo vse štiri predlagane diagrame za prikaz konkordance. Pri opisu blazinice z bucikami je za "blazinico" izbrana bolj "groba", a preprostejša možnost z vnaprej pripravljenim likom, ki je v polarnih koordinatah ustrezne oblike in velikosti; zahtevnejša možnost bi bila uporaba podob (stavka *facet*) in združitev glavnega grafa v polarnih z dodatnim grafom v pravokotnih koordinatah.

V skladu z odlikami Wilkinsonove grafične slovnice je opis enak za vse diagrame istega tipa (pri prvem diagramu je oznaka vrednosti *W* zato navedena v splošni obliki "n,nn"). Opisi so nekoliko manj podrobni kot v prejšnjem razdelku, a vsebujejo vse ključne prvine.

Opis konkordančnega mehurčnega diagrama predpostavlja, da so podatki pripravljeni v obliki dvojic {rang, srednji rang}, za ostale opise pa morajo biti podatki v obliki vseh dvojic rangov (vrstni red dvojic je lahko poljuben, urejenost rangov v dvojici po velikosti ni potrebna). Vse opise bi se sicer dalo izdelati tako, da bi delovali na vhodnih podatkih v obliki matrike z *m* vrsticami (ocenjevalci) in *k* stolpci (objekti), a za to bi bilo potrebnih veliko stavkov za pretvorbo spremenljivk in več uporabe algebre, zaradi česar bi bila slovnična struktura grafik manj jasno razvidna.

Opis konkordančnega mehurčnega diagrama (sliki 6 in 17):

```
DATA: f=count()
FRAME: rang*srednjirang
GRAPH: point(size(f))
GUIDE: axis1(label("rang"))
GUIDE: axis2(label("srednji rang"))
GUIDE: axis2(rule(srednjirang),label(objekt))
GUIDE: form.line(position((1,1),(4,4)),granularity(2))
GUIDE: text("W=n,nn",position(1,4))
```

Opis konkordančnega diagrama z vzporednima osema (slika 7):

```
DATA: rm=string("R<")
DATA: rv=string("R>")
TRANS: manjsirang=min(rang1,rang2)
TRANS: vecjirang=max(rang1,rang2)
FRAME: rm*manjsirang+rv*vecjirang
GRAPH: line(size(count()))
```

Opis konkordančnega stolpčnega diagrama (slika 8):

```
TRANS: absrazlika=abs(diff(rang1,rang2))
FRAME: absrazlika
GRAPH: histobar()
```

Opis konkordančnega diagrama blazinice z bucikami (sliki 9 in 18):

```
DATA: blazinica=link("pravokotnik2x1")
DATA: nic=constant(0)
TRANS: absrazlika=abs(diff(rang1,rang2))
TRANS: f=count()
FRAME: absrazlika*(f+nic)
COORD: polar()
COORD: transpose()
GRAPH: link.edge.join()
GUIDE: image(position(0,0),shape(blazinica))
```

5. RAZPRAVLJANJE

V statistični praksi se s podatki, ki jih je primerno prikazati z razširjenimi konveksnolupinskimi diagrami, srečujemo razmeroma pogosto. Algoritmi za iskanje najmanjše konveksne lupine dane množice točk so že dolgo znani in natančno preučeni (npr. Efron, 1965) in tudi izboljševanje oziroma dopolnjevanje razsevnih diagramov (predvsem s postopki glajenja) je že nekaj časa predmet statističnih raziskav (Cleveland, 1993; Cleveland in McGill, 1984b; Lewandowsky in Spence, 1989), toda diagrami, kakršne predlagamo, so novost. V prvih letih prikaza in analize podatkov z osebnimi računalniki so se konveksne lupine sicer uveljavile za prikaz rezultatov korespondenčne analize (Greenacre, 1984), a so kasneje ostale na obrobju zanimanja statistikov. Podobno velja za sorodne postopke računske geometrije, kot so Voronojevi in Delaunayevi diagrami, ki bi si po razmeroma zgodnji vključitvi v standardni nabor metod računalniške analize in prikaza podatkov (zlasti v okviru prostorske statistike – Ripley, 1981) zaslužili več pozornosti raziskovalcev, h čemur bi lahko pripomoglo tudi pričujoče delo.

Razširjeni konveksnolupinski diagrami tako predstavljajo primer učinkovitega prenosa zahtevnih matematičnih algoritmov, kot sta iskanje konveksnih lupin in ocenjevanje bivariatne gostote, v vsakdanjo prakso dela s podatki. Predlagamo jih predvsem kot nadomestilo za večskupinske razsevne diagrame (multi-group scatter plots), ki pri večjem številu podatkov in znatnejšem prekrivanju skupin postanejo prenatrpani oziroma povsem nepregledni. Če točke na razsevnem diagramu zamenjamo s konveksnimi lupinami, pridobimo znotraj njih prostor za jasen prikaz opisnih statistik z daljicami (error bars) ali elipsami zaupanja (confidence ellipses). Ne glede na obliko skupne porazdelitve in robnih porazdelitev tako nazorno prikažemo srednje vrednosti in razpršenost v dveh razsežnostih, če uporabimo elipse zaupanja, pa tudi medsebojno povezanost razsežnosti. Drugi učinkovit dodatek je bivariatna mera razpršenosti, ki jo izračunamo tako, da ploščino znotraj konveksne lupine delimo s številom enot (velikostjo skupine). Razširjeni konveksnolupinski diagram dopolnimo s prikazom robnih porazdelitev ob robu diagrama. Če so v podatkih prisotne odstopajoče vrednosti (osamelci – outliers), lahko namesto konveksnih lupin uporabimo bivariatne obrobe gostote verjetnosti. Razširjeni konveksnolupinski diagrami torej na pregleden način prikazujejo veliko količino informacij – za več velikih množic enot jasno

prikazujejo srednjo vrednost, razpršenost in obliko skupne porazdelitve dveh spremenljivk ter robno porazdelitev vsake od spremenljivk.

Kot vsi prikazi podatkov, ki temeljijo na razsevnem diagramu, tudi konveksne lupine oziroma bivariatne obrobe gostote verjetnosti niso omejene na surove podatke. Koordinate točk lahko dobimo iz katerekoli multivariatne statistične metode zmanjševanja števila razsežnosti (analiza glavnih komponent, faktorska analiza, večrazsežno lestvičenje, združevanje v skupine idr.), ali pa izvorne razsežnosti pretvorimo (uporabimo transformacijske funkcije). Na ta način lahko postanejo razširjeni konveksnolupinski diagrami, ki jih enostavno izdelamo z ustreznim programjem, standardno ali celo avtomatizirano dopolnilo množično uporabljanim multivariatnim eksploratornim statističnim analizam.

Predlagana bivariatna mera razpršenosti, t.j. ploščina konveksne lupine na statistično enoto, je preprosto izračunljiva in jasno predstavljiva in se jo da včasih tudi neposredno interpretirati. Programski paket *chplot* je skupaj z njo že dokazal svojo praktično uporabnost za preučevanje dnevnih in tedenskih delovnih migracij (Buliung in Roorda, 2005, 2006), na podoben način pa bi se ga dalo uporabiti tudi za preučevanje prostorske in/ali časovne razširjenosti izbranih pojavov ali vedenj na področjih javnega zdravja, epidemiologije in javne higiene.

Seveda imajo tudi razširjeni konveksnolupinski diagrami svoje slabosti oziroma omejitve. Glavno omejitev predstavlja število skupin, saj z naraščanjem števila skupin postanejo nepregledni (po naših izkušnja najkasneje pri desetih skupinah). Imajo tudi "konkurenco": bivariatni zaboj z ročaji (Rousseeuw, Ruts in Tukey, 1999) in sorodni prikazi (Becketti in Gould, 1987; Goldber in Iglewicz, 1992; Hyndman, 1996; Tongkumchum, 2005; Zani, Riani in Corbellini, 1998) so nekoliko težje razumljivi, a prav tako prikazujejo srednje vrednosti, razpršenost in obliko porazdelitev in jih je moč dopolniti s poljubnim prikazom robnih porazdelitev (npr. z običajnimi zaboji z ročaji). Hkrati v primerjavi z običajnimi prikazi bivariatnih porazdelitev ohranjajo nekatere prednosti, ki jih ima zaboj z ročaji pred drugimi prikazi univariatnih porazdelitev z vidika robustnosti (prim. Hoaglin, Mosteller in Tukey, 1983). A namenjeni so predvsem za iskanje osamelcev in mišljeni predvsem kot dopolnilo, ne pa kot nadomestilo razsevnega diagrama. V taki obliki torej niso primerni za prikaz velike količine podatkov in tudi brez prikaza točk imajo v primerjavi z razširjenimi

konveksnolupinskimi diagrami več grafičnih elementov, zato postanejo natrpani oziroma nejasni že pri manjšem številu skupin.

Pri prikazu konkordance nam je pomemben motiv predstavljal naraščajoči pomen analize urejenostnih (ordinalnih) podatkov v biomedicini (npr. Nelson in Pepe, 2000). Hkrati se v statistiki uveljavlja prepričanje, da so številne raziskovalne hipoteze, predvsem na področju psihologije, sociologije in drugih družbenih ved, pa tudi z njimi tesno povezanih medicinskih ved (epidemiologija, javno zdravje), v resnici urejenostne narave (Cliff, 1996a). To pomeni, da se nanašajo na velikostno zaporedje (udeležencev eksperimentov, bolnikov, meritev ipd.) in ne na same opažene oziroma izmerjene vrednosti številskih spremenljivk. Po nizu člankov je ta problematika dobila tudi sistematičen monografski pregled (Cliff, 1996b), vendar so na področju ustreznega prikaza podatkov dosežki manj celoviti.

Urejenostni podatki sodijo v širši okvir opisnih podatkov in prikaz le-teh je v zadnjih letih sicer doživel razcvet (Friendly, 2000), toda problem prikaza konkordance doslej ni bil obravnavan. Zanimiv je predvsem v povezavi s primerjavo konkordance med dvema skupinama in večskupinsko analizo konkordance (razdelek 1.2.2), ki smo je po skoraj treh desetletjih od njenega nastanka nedavno prenesli v raziskovalno prakso in skušali metodološko dopolniti (Vidmar in Černigoj, 2004). Hkrati so koeficienti konkordance v zadnjem času pritegnili pozornost enega od vodilnih raziskovalcev na področju uporabne statistike (Legendre, 2005; Legendre in Lapointe, 2004).

Različni konkordančni diagrami, ki jih predlagamo, imajo različne dobre in slabe lastnosti. Konkordančni mehurčni diagram je med konkordančnima diagramoma na podlagi dodeljenih rangov tisti, ki ima višje razmerje med podatki in črnilom. Odlikuje ga tudi prikaz povprečnih rangov objektov. Istovrsten diagram bi lahko izdelali z naključnim raztresenjem, a pri velikem številu ocenjevalcev bi postal nejasen zaradi prekrivanja točk med sosednjimi rangi. Konkordančni diagram z vzporednimi osmi je med vsemi predlaganimi konkordančnimi diagrami najmanj jasen in razmerje med podatki in črnilom ima nizko. Poleg tega ga lahko razumemo le, če smo dovolj pozorni na oznaki osi, vseeno pa si zasluži predstavitev iz estetskih nagibov in zaradi zanimivih lastnosti vzporednih koordinatnih osi.

Konkordančna diagrama na podlagi razlik rangov se izogneta vplivu vrstnega reda ocenjevalcev v matriki podatkov tako, da prikazujeta absolutne razlike med dvojicami rangov, ki so dodeljene istemu objektu. Njun izgled je v primerjavi z diagramoma na podlagi rangov preprostejši, toda tako kot konkordančni diagram z vzporednima osema ne prikazujeta medsebojnega odnosa objektov. Konkordančni diagram blazinice z bucikami ima izmed vseh prikazov konkordance največje razmerje med podatki in črnilom, zato je uporaben predvsem za primerjavo konkordance med večjim številom skupin ocenjevalcev.

Prvi dodatni primer uporabe konkordančnih diagramov (razdelek 4.3.2) predstavlja izrazito uspešno evalvacijo. Dane podatke bi se seveda dalo prikazati tudi drugače, npr. z diagramom Nelsona in Pepeja (2000), a konkordančni diagrami so se izkazali kot orodje, ki bi zdravnikom in drugim raziskovalcem, vključenim v študijo, lahko neposredno koristilo oziroma olajšalo razumevanje podatkov ter dopolnilo objave. Pri drugem primeru uporabe konkordančnih diagramov je potrebno poudariti, da študija še ni dokončana, metodološka in vsebinska scientometrična razprava o njej pa tudi ne bi sodila v okvir pričujočega dela. Z vidika prikaza konkordance je bistveno, da s konkordančnimi diagrami blazinice z bucikami natančno razločujemo med stopnjami konkordance, ki so vse zelo visoke, pri čemer nam pomaga medsebojna primerjava diagramov po višini in po širini.

Predlaganim konkordančnim diagramom bi lahko očitali, da so "osnovne zaznavne naloge", na katerih temeljijo, zaznavanje dolžine, ploščine, smeri in kota, te pa niso pri vrhu "hierarhije natančnosti opazovalčevega presojanja" (Cleveland, 1994; Cleveland in McGill, 1984a). Toda sodobna kognitivna psihologija je dokazala, da primerjava zaznavnih lastnosti izven konteksta ni možna (Lockhead, 1992, 1995), hkrati pa se je uveljavilo prepričanje, da razumevanje kakovostnih znanstvenih grafik zahteva veliko časa in pozornosti (Tufte, 1983). Zato smo konkordančne diagrame zasnovali tako, da prikazujejo jasno razpoznaven vzorec za dano stopnjo konkordance, pri čemer vsebujejo dovolj informacije za nadaljnji podroben pregled. Zlasti pri konkordančnem mehurčnem diagramu elementi konteksta (osi in oznake srednjih rangov objektov) dejansko pretvorijo "manj natančne zaznavne naloge" v "najbolj natančno zazanavno nalogo", namreč presojanje položaja na isti lestvici.

Konkordančni diagrami so zasnovani črno-belo, saj v skladu z načeli dobre grafične prakse (npr. Wainer in Thissen, 1981) uporabo barv zahtevajo le informacijske razsežnosti, ki bi sicer ostale neprikazane ali nerazločne. Omejitev predlaganih konkordančnih diagramov (izvzemši stolpčnega) je, da postanejo nejasni z velikim številom objektov (mehurčni diagram in diagram blazinice z bucikami) oziroma ocenjevalcev (diagram z vzporednima osema). Toda to ni huda pomanjkljivost, saj v študijah konkordance redko nastopa več kot deset objektov ali ocenjevalcev, vsaj znotraj posamezne skupine. Predstavljeni primeri konkordančnih diagramov ne prikazujejo vezanih rangov, a ti ne predstavljajo težave za nobeno od predlaganih metod. V celoti vzeto so predstavljeni konkordančni diagrami zanimivi in sledijo načelom smiselnega in učinkovitega prikaza podatkov. Številni primeri vizualizacijskih metod iz preteklosti, ki so uveljavljene v znanstveni literaturi, a se v vsakdanji statistični praksi sploh ne uporabljajo (npr. obrazi za prikaz večrazsežnih podatkov – Chernoff, 1973), nas resda učijo previdnosti glede sodb o praktični uporabnosti. Vseeno menimo, da bi vsaj konkordančni mehurčni diagram in konkordančni diagram blazinice z bucikami lahko našla pot do uporabnikov na različnih področjih od ekologije do marketinga, kjer preučujejo konkordanco.

V razpravljanju o konkordančnih diagramih ne moremo mimo zamisli o prikazu večrazsežnih porazdelitev urejenostnih spremenljivk s permutacijskimi politopi (Baggerly, 1995; Thompson, 1994). Na tak način se sicer da prikazati konkordančne podatke, tudi za ogromno število ocenjevalcev, a kljub svoji teoretični in računski kompleksnosti je tak prikaz omejen zgolj na tri ali štiri objekte, pri petih objektih pa je že povsem nepraktičen. Čeprav področje permutacijskih politopov presega namen in matematično raven pričujočega dela, ga moramo vsaj omeniti zato, ker ima daljnosežen pomen, ker ustvarja nenavadne in vznemirljive grafične prikaze in ker združuje dve osrednji temi pričujočega dela – konveksne lupine (ki omejujejo politopske "histograme" večrazsežnih urejenostnih podatkov) in prikaz podatkov v obliki rangov.

Anketa pri potencialnih uporabnikih priča v prid uporabnosti razširjenih konveksnolupinskih diagramov. Polovični odziv glede na zahtevnost in trajanje izpolnjevanja ankete ni nizek. Uspešnost respondentov pri uporabi programskega paketa *chplot* potrjuje njegovo prijaznost do uporabnika. Objektivno gledano je ta torej morda celo izrazitejša, kot kažejo odgovori, še

zlasti, če upoštevamo skoraj popolno nepoznavanje zahtevnejših oziroma manj vsakdanjih multivariatnih grafičnih metod med respondenti. Če bi nerespondentom pripisali prevladujoč odklonilen odnos do predlaganih diagramov in programja zanje, bi prevladujoči vtis še vedno ostal pozitiven. Ob tem se je potrebno zavedati, da je bilo že za anketo težko določiti ustrezno ciljno populacijo in obliko izvedbe, ki je zagotovila sprejemljivo posplošljivost izsledkov, za zaznavno-kognitivni eksperiment pa bi bilo to še težje. Zanj nismo našli zasnove, ki bi zagotavljala kriterijsko in vsebinsko veljavnost ugotovitev. Pri konkordančnih diagramih so zaradi ozke specifičnosti problematike pomisleki glede ankete ali eksperimenta še toliko močnejši, zato se tovrstnega vrednotenja – upoštevaje tudi izkušnje nekaterih vodilnih raziskovalcev, da je znanstvena vrednost študij uporabnikov na področju prikaza podatkov često vprašljiva (Kosara, Healey, Interrante, Laidlaw in Ware, 2003) – raje nismo lotili.

Pri uporabi Wilkinsonove grafične slovnice gre zgolj za poskus prikaza njene uporabnosti, ne za pravi raziskovalni prispevek, a tudi tovrstni poskusi so koristni v začetnem obdobju uveljavljanja velikih znanstvenih dosežkov. O izjemnem pomenu in potencialu Wilkinsonovega dela namreč mimo vseh etičnih, pravnih in gospodarskih dilem, ki spremljajo problematiko patentov na področju računalništva, priča tudi patent, ki mu je bil nedavno dodeljen za "računalniški postopek in napravo za izdelavo vidne grafike z uporabo algebre grafov" (U.S. Patent No. 7023453, 2006).

6. ZAKLJUČEK

Znano je, da med raznovrstnostjo, kompleksnostjo in kakovostjo razvitih in v statističnih znanstvenih publikacijah predstavljenih statističnih metod ter dejansko prevladujočo uporabo statističnih metod vlada ogromen razkorak, še posebej v biomedicinskih vedah in družboslovju. Razkorak se z vse hitrejšim razvojem statistike in hkrati vse bolj ozko izobrazbeno in interesno specializiranostjo raziskovalcev še povečuje. Prepričani smo, da ga je potrebno in moč zmanjšati z razvojem najširše dostopnega programja, ki je sprejemljivo tudi za uporabnike brez poglobljene matematične oziroma statistične izobrazbe. Zato so metode, ki smo jih razvili, z vidika matematične statistike preproste in pomenijo zgolj poseben primer oziroma nov način uporabe obstoječih, bolj temeljnih metod. Dostopnost predlaganih metod pa smo zagotovili tako, da smo jih implementirali z odprtokodno in prosto dostopno programsko opremo, ki deluje v različnih operacijsko-sistemskih okoljih.

Uspeli smo izpolniti konkretne cilje, ki smo si jih zadali, in pritrdilno odgovorili na vsa zastavljena raziskovalna vprašanja. K uresničitvi dolgoročnega cilja, torej uveljavitve prikaza podatkov kot samostojne znanstvene discipline in širjenja dobrih praks prikaza podatkov na področju biomedicine, smo skušali prispevati s povezavo statističnih in psiholoških spoznanj, računalniško-informacijske tehnologije in biostatistično-svetovalnih izkušenj, pa tudi z vključitvijo bodočih strokovnjakov s področja biostatistike v ovrednotenje izdelanega programja. Zaradi zelo skromnega poznavanja zahtevnejših oziroma sodobnejših metod prikaza večrazsežnih podatkov celo med študenti statistike je ta cilj še toliko pomembnejši.

Delo seveda pušča vrsto odprtih vprašanj in nalog za prihodnost. V okviru okolja R bi se dalo izdelati celovit programski paket za analizo in prikaz konkordance (npr. z nadgradnjo paketa *concord*) ali pa preučiti uporabnost trorazsežnih razširjenih konveksnolupinskih diagramov (z uporabo paketov *Scatterplot3d* in *rgl*). Priložnost za uveljavljavitev kakovostnega prikaza podatkov v slovenskem prostoru bi v prihodnosti lahko nudilo izboljšanje rutinskega prikaza podatkov v bolnišnični informatiki in javnozdravstvenem poročanju. Prikaz podatkov v povezavi s statističnim svetovanjem bi se lahko uveljavil tudi kot samostojen predmet v okviru študija statistike na Univerzi v Ljubljani – obstoječega podiplomskega ali prihodnjega dodiplomskega. Če verjamemo zamisli o družbi znanja, pa je zavest o pomenu ustreznega

prikaza podatkov potrebno širiti tudi preko meja statistike in zdravstva – na primer z razvojem sodobnega, zmogljivega in javno dostopnega dodatka za kakovosten prikaz statističnih podatkov za elektronsko preglednico Excel, ki ga bodo lahko uporabljali tudi nematematiki in nestatistiki.

7. LITERATURA

Artnik, B., Vidmar, G., Javornik, J., Laaser, U. (2005). Premature mortality in Slovenia in relation to selected biological, socioeconomic, and geographical determinants. *Croatian Medical Journal*, *47* (1), 103-113.

Anderson, E.A. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, *59*, 2-5.

Baggerly, K.A. (1995). *Visual estimation of structure in ranked data* (doktorsko delo). Houston: Rice University.

Becker, R.A., Chambers, J.M. (1984). *S: an interactive environment for data analysis and graphics*. Pacific Grove: Wadsworth & Brooks Cole.

Beckett, J., Schucany, W.R. (1975). ANACONDA: Analysis of concordance of g groups of judges. V *Proceedings of Social Statistics Section of the American Statistical Association* (str. 311-313).

Beckett, J., Schucany, W.R. (1979). Concordance among categorized groups of judges. *Journal of Educational Statistics*, *4* (2), 125-137.

Becketti, S., Gould, W. (1987). Rangefinder box plots: a note. *The American Statistician*, *41* (2), 149.

Bertin, J. (1983). *Semiology of graphics* (prev. W.J. Berg in H. Wainer). Madison: University of Wisconsin Press.

Bertin, J. (1981). *Graphics and graphic information-processing* (prev. W.J. Berg in P. Scott). Berlin: Walter de Gruyter.

Buliung, R.N., Roorda, M.J. (2005). Exploring the spatial stability of activity-travel behaviour: initial results from the Toronto Travel-Activity Panel Survey. PROCESSUS Second International Colloquium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications. URL http://www.civ.utoronto.ca/sect/traeng/ilute/processus2005/PaperSession/Paper03_Buliung-Roorda_ExploringSpatialMeasures_CD.pdf

Buliung, R.N., Roorda, M.J. (2006). Spatial variety in weekly, weekday-to-weekend, and day-to-day patterns of activity-travel behaviour: initial results from the Toronto Travel-Activity Panel Survey. Transportation Research Board 2006 Annual Meeting. URL http://www.mdt.mt.gov/research/docs/trb_cd/Files/06-0765.pdf

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, *68*, 361-368.

Cleveland, W.S. (1993). *Visualizing data*. Summit: Hobart Press.

Cleveland, W.S. (1994). *The elements of graphing data*. Summit: Hobart Press.

Cleveland, W.S., McGill, R. (1984a). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*, 531-554.

Cleveland, W.S., McGill, R. (1984b). The many faces of a scatterplot. *Journal of the American Statistical Association*, *79*, 807-812.

Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. Multivariate Behavioral Research, *31* (3), 331-350.

Cliff, N. (1996b). *Ordinal methods for behavioral data analysis*. Mahwah: Lawrence Erlbaum.

Corsten, L.C.A., Gabriel, K.R. (1976). Graphical exploration in comparing variance matrices. Biometrics, *32*, 851-863.

Efron, B. (1965). The Convex Hull of a Random Set of Points. *Biometrika*, *52*, 331-453.

Ehrenberg, A.S.C. (1952). On sampling from a population of rankers. *Biometrika*, *39*, 82-87.

Everitt, B.S. (1978). *Graphical techniques for multivariate data*. London: Heinemann.

Everitt, B., Rabe-Hesketh, S. (2001). *Analyzing medical data using S-PLUS*. New York: Springer.

Fua, Y.-H., Ward, M.O., Rundensteiner, E.A. (1999). Hierarchical parallel coordinates for exploration of large datasets. V *Proceedings of the 10th IEEE Visualization 1999 Conference* (str.43-50)..

Fisher, R.A.(1936). The use of multiple measurements in taxonomic problem. *Annals of Eugenics Part II*, 179-188.

Friendly, M. (2000). *Visualizing Categorical Data*. Cary: SAS Institute.

Friendly, M. (2002). Corrgrams: exploratory displays for correlation matrices. *The American Statistician*, *56* (4), 316-324.

Friendly, M., Kwan, E. (2003). Effect ordering for data displays. *Computational Statistics & Data Analysis*, *43*, 509-539.

Gescheider, G.A. (1985). *Psychophysics: method, theory, and application* (2. izd.). Hillsdale: Lawrence Erlbaum.

Goldberg, K.M., Iglewicz, B. (1992). Bivariate extensions of the boxplot. *Technometrics*, *34* (3), 307-320.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Harris, R.L. (2000). *Information graphics: a comprehensive illustrated reference*. New York: Oxford University Press.

Heiser, D.A. (2006). *Microsoft Excel 2000 and 2003 faults, problems, workarounds and fixes*. URL http://www.daheiser.info/excel/frontpage.html

Hoaglin, D.C., Mosteller, F., Tukey, J.W. (ur.) (1983). *Understanding robust and exploratory data analysis*. New York: John Wiley.

Hollander, M., Sethuraman, J. (1978). Testing for agreement between two groups of judges. *Biometrika*, *65* (2), 403-411.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Hudomalj, E., Vidmar, G. (2003). OLAP and bibliographic databases. *Scientometrics*; *58* (3), 609-622.

Hyndman, R.J. (1996). Computing and graphing highest density regions. *The American Statistician*, *50* (2), 120-126.

Ihaka R., Gentleman R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*, 299-314

Iman, R.L., Conover, W.J. (1987). A measure of top-down correlation. *Technometrics*, *29* (3), 351-357.

Inselberg, A. (1985). Plane with parallel coordinates. *Visual Computer*, *1*, 69-97.

Inselberg, A. (1998). Visual data mining with parallel coordinates. *Computational Statistics*, *13* (1), 47-63.

Inselberg, A., Dimsdale, B. (1990). Parallel coordinates: a tool for visualizing multi-dimensional geometry. V *Proceedings of the First IEEE Conference on Visualization, 1990* (str. 361-378).

Jacoby, W.G. (1997). *Statistical graphics for univariate and bivariate data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 117. Thousand Oaks: Sage.

Jacoby, W.G. (1998). *Statistical graphics for visualizing multivariate data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 120. Thousand Oaks: Sage.

Javornik, J., Korošec, V. (ur.). (2003). *Poročilo o človekovem razvoju Slovenija 2002/2003: človekov razvoj in zdravje*. Ljubljana: UMAR, UNDP.

Kendall, M., Babbington Smith, B. (1939). The problem of m rankings. *Annals of Mathematical Statistics*, *10*, 275-287.

Kendall, M., Dickinson Gibbons, J. (1990). *Rank correlation methods* (5. izd.). London: Edward Arnold.

Kosara R., Healey C.G., Interrante V., Laidlaw D.H., Ware C. (2003). Thoughts on user studies: why, how, and when. *IEEE Computer Graphics & Applications*, *Visualization Viewpoints*, 23 (4), 20-25.

Kosslyn, S.M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, *3* (3), 185-225.

Kosslyn, S.M. (1994). *Elements of graph design*. New York: W.H. Freeman.

Legendre, P., Lapointe, F.-J. (2004). Assessing congruence among distance matrices: Single-malt Scotch whiskies revisited. *Australian and New Zealand Journal of Statistics*, *46* (4), 615-629.

Legendre, P. (2005). Species associations: the Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics*, *10* (2), 226-245.

de Levie, R. (2004). *Advanced Excel for scientific data analysis*. New York: Oxford University Press.

Lewandowsky, S., Spence, I. (1989). Discriminating strata in scatterplots. *Journal of the American Statistical Association*, *84*, 682-688.

Lockhead, G.R. (1992). Psychophysical scaling: judgment of attributes or objects? *Behavioral and Brain Sciences*, *15*, 543-558.

Lockhead, G.R. (1995). Psychophysical scaling methods reveal and measure context effects. *Behavioral and Brain Sciences*, *18*, 607-612.

Lyerly, S.B. (1952), The average Spearman rank correlation coefficient. *Psychometrika*, *17*, 421-428.

Murdoch, D.J., Chow, E.D. (1996). A graphical display of large correlation matrices. *The American Statistician*, *50* (2), 178-180.

Nelson, J.C., Pepe, M.S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research*, *9*, 475-496.

Netto, G.J., Watkins, D.L., Williams, J.W., Colby, T.V., dePetris, G., in sod. (2006). Interobserver agreement in hepatitis C grading and staging and in the Banff grading schema for acute cellular rejection: the "hepatitis C 3" multi-institutional trial experience. *Archives of Pathology and Laboratory Medicine*, *130* (8), 1157-1162.

Neuwirth, E., Arganbright, D. (2004). *The active modeler – mathematical modeling with Microsoft Excel*. Belmont: Brooks Cole.

Page, E.B. (1963). Ordered hypotheses for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, *58*, 216-230.

Palachek, A.D., Kerin R.A. (1982). Alternative approaches to two-group concordance problem in brand preference ranking. *Journal of Marketing Research*, *19*, 386-389.

Palachek, A.D., Schucany, W.R. (1984). On approximate confidence intervals for measures of concordance. *Psychometrika*, *49* (1), 133-141.

Playfair, A.W. (2005). *The commercial and political atlas and Statistical breviary* (ur. H. Wainer in I. Spence, barvna reprodukcija 3. izd., London, J. Walsh, 1801). New York: Cambridge University Press.

Randles, R.H., Wolfe, D.A. (1979). *Introduction to the theory of nonparametric statistics*. New York: Wiley.

R Development Core Team (2004). *R: a language and environment for statistical computing*. Dunaj: R Foundation for Statistical Computing. URL http://www.R-project.org

Ripley, B. (1981). *Spatial statistics*. New York: Wiley.

Rousseeuw, P.J., Ruts, I., Tukey, J.W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, *53*, 382-387.

Schucany, W.R., Frawley, W.H. (1973). A rank test for two group concordance. *Psychometrika*, *38* (2), 249-258.

Seber, G.A.F. (1984). *Multivariate observations*. New York: Wiley.

Siegel, S., Castellan, J. (1988). *Nonparametric statistics for the behavioral sciences* (2. izd.). New York: McGraw-Hill.

Spence, R. (2000). *Information visualization*. Harlow: Addison-Wesley.

Thompson, G.L. (1994). Visualising frequency distributions of ranked data. *Computational Statistics*, *9* (1), 1-10.

Tongkumchum, P. (2005). Two-dimensional box plot. *Songklanakarin Journal of Science and Technology*, *27* (4), 859-866.

Trosset, M.W. (2005). Visualizing correlation. *Journal of Computational and Graphical Statistics*, *14* (1), 1-19.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Tufte, E.R. (1983). *The visual display of quantitative information*. Chesire: Graphics Press.

Tufte, E.R. (1990). *Envisioning information*. Chesire: Graphics Press.

Vidmar, G., Černigoj, M. (2004). Studying norms in small groups by means of multi-group concordance analysis, *Horizons of Psychology*, *13* (4), 55-66.

Vidmar, G., Pohar, M. (2005). Augmented convex hull plots: rationale, implementation in R and biomedical applications. *Computer Methods and Programs in Biomedicine*, *78* (1), 69-74.

Vidmar, G., Rode, N. (2007). Visualising concordance. Sprejeto v objavo v *Computational Statistics*.

Wainer, H., Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology*, *32*, 191-241.

Wald, A., Wolfowitz, J. (1944). Statistical tests based on permutations of the observations. *Annals of Mathematical Statistics*, *15*, 358-372.

Wegman, E.J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, *85*, 664-675.

Wegman, E.J. (2003). Visual data mining. *Statistics in Medicine*, *22* (9), 1383-1397.

West, R.W., Wu, Y., Heydt, D. (2004). An introduction to StatCrunch 3.0. *Journal of Statistical Software*, *9* (6). URL http://www.jstatsoft.org/v09/i05/scjss/

Wilkie, D. (1980). Pictorial representation of Kendall's rank correlation coefficient. *Teaching Statistics*, *2*, 76-78.

Wilkinson, L. (1999). *The grammar of graphics*. New York: Springer.

Wilkinson, L., Wills, D., Rope, D., Norton, A., Dubbs, R. (2005). *The grammar of graphics* (2. izd.). New York: Springer.

Wilkinson, L., (2006). *Computer method and apparatus for creating visible graphics by using a graph algebra*. U.S. Patent No. 7023453. Washington, DC: U.S. Patent and Trademark Office.

Zani, S., Riani, M., Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, *28*, 257-270.

Zupančič-Pridgar, A. (2003). Vpliv vaginalne flore na zaplete po konizaciji (magistrsko delo). Ljubljana: Medicinska fakulteta.

PRILOGA 1: OBJAVLJENI ČLANEK O RAZŠIRJENIH KONVEKSNOLUPINSKIH DIAGRAMIH

Computer Methods and Programs in Biomedicine (2005) **78**, 69–74

ELSEVIER

Computer Methods and Programs in Biomedicine

www.intl.elsevierhealth.com/journals/cmpb

Augmented convex hull plots: Rationale, implementation in R and biomedical applications

Gaj Vidmar*, Maja Pohar

University of Ljubljana, Faculty of Medicine, Institute of Biomedical Informatics, Vrazov trg 2, SI-1000 Ljubljana, Slovenia

Received 24 May 2004; received in revised form 20 December 2004; accepted 20 December 2004

KEYWORDS
Multivariate visualization;
Convex hulls;
Bivariate density;
R

Summary The paper addresses the possibility to replace cluttered multi-group scatter-plots with augmented convex hull plots. By replacing scatter-plot points with convex hulls, space is gained for visualization of descriptive statistics with error bars or confidence ellipses within the convex hulls. An informative addition to the plot is calculation of the area of convex hull divided by corresponding group size as a bivariate dispersion measure. Marginal distributions can be depicted on the sides of the main plot in established ways. Bivariate density plots might be used instead of convex hulls in the presence of outliers. Like any scatter-plot type visualization, the technique is not limited to raw data — points can be derived from any dimension reduction technique, or simple functions can be used as axes instead of original dimensions. The limited possibilities for producing such plots in existing software are surveyed, and our general and flexible implementation in R — the publicly available chplot function — is presented. Examples based on our daily biostatistical consulting practice illustrate the technique with various options.

1. Introduction

Convex hull drawing is a well-known computational geometry problem and there is a multitude of algorithms available for solving it. Even though links between computational geometry and statistics have been studied for quite a while [1–3], various tessellation methods are relatively rarely used in statistical visualization practice. An exception is the use of Voronoi diagrams in relation to discriminant analysis [4,5], while convex hull has been applied primarily for classification purposes in the field of pattern recognition [6–8].

Our paper focuses on the use of convex hulls rather than on the algorithms for finding them. There is a comprehensive resource with online Java implementation of different convex hull algorithms [9]. For solving the convex hull problem in *n* dimensions, the detailed *qhull* software package, which is based on the quickhull algorithm [10], has been developed and made publicly available [11].

In the next section, we present our concept of augmented convex hull plots and discuss its

* Corresponding author. Tel.: +386 1 5437783.
E-mail address: gaj.vidmar@mf.uni-lj.si (G. Vidmar).

0169-2607/$ – see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2004.12.003

elements. It should be stressed at the beginning that the basic idea of replacing, or at least accompanying cluttered scatterplots with some type of ''clouds'' is by no means novel. In the 1980s, convex hulls were presented as an aid to visualization of correspondence analysis results [12] and as such, found some application in the field of ecological ordination. At the same time, bivariate radial smoothers were discussed as one of the many possible "faces" one can superimpose on scatterplots [13], or even use instead of original points, although the authors considered that rather a last resort (judging by the fact that it is used in just one of the 20 figures in the article). Nevertheless, subsequent visualization research focused either on the choice of type and color of symbols for maximizing discrimination in scatter-plots [14], or on scatter-plots with smoothers [15], rather than on the idea which we elaborate.

2. Method

We believe that convex hulls are a suitable replacement for scatter-plots if the groups are large and there is considerable overlap of points between them. Our implementation is in two dimensions, but one can envision three-dimensional implementations, possibly incorporated into dynamic visualization in modern statistical, data-mining, and/or visualization software. The proposed convex hull plots are primarily intended for depicting differences between groups on continuous variables, possibly accompanying logistic regression, discriminant analysis, or some other classification technique.

By replacing scatter-plot points with a convex hull, one gains space for clear graphical presentation of descriptive statistics for both axes. For that purpose, we implemented basic "parametric" and "nonparametric" choices in terms of error-bar plots and confidence ellipses. Our choices for error-bar plot are the mean plus/minus multiple of standard deviation or standard error of the mean (i.e., the normal tolerance interval or the confidence interval for the mean), and the median with the interquartile range, while other variants might be various M-estimators and robust measures of dispersion. Alternatively, the user can choose to plot a confidence ellipse within the convex hull, thus also depicting correlation between variables. More complex diagrams depicting the whole distribution, such as notched boxplots or violin plots, should probably not be used as an augmentation to convex hulls because of visual clutter.

But since marginal distributions are of interest as such, we suggest the established way of plotting them separately, with corresponding line color or line style on the top and right-hand side of a two-dimensional plot, whereby either frequency polygons or kernel density plots are suitable choices if there are more than two groups. Another informative addition to the convex hull plot is the area of the convex hull divided by the group size, which is a dispersion measure related to mean absolute deviation in the sense that it weighs all deviations from the center equally. Instead of convex hulls, bivariate (trivariate in three dimensions) density plots can be used (with corresponding area/volume per point as dispersion measure), which should be preferred in the presence of outliers, whereby a relatively sparse grid (i.e., large bandwidth) produces results resembling convex hulls.

3. Implementation

3.1. Convex hull plotting in existing software

Among major commercial statistical software packages, SYSTAT® [16] is the only one that provides automated convex hull drawing. Convex hull is one of the optional additions to scatter-plots, together with related computational geometry procedures — Voronoi tessellation, Delaunay triangulation, minimum spanning tree and travelling salesman path. Instead of the convex hull, one can also use a nonparametric kernel density estimator with user-specified probability. In the Version 10, which was available to us for testing purposes, multi-group convex hull plots can only be produced in two dimensions. There is a huge choice of univariate density plots that can be displayed on chart's borders, but it is not possible to superimpose error bars or boxplots on the main plot area. As a substitute, "Gaussian bivariate ellipses" can be placed on the plot. An example of such a plot for the famous Iris dataset [17] is provided in Fig. 1. However, possibilities for subsequently enhancing the plot are very limited due to the modest capabilities of the Graph window.

Automated convex hull drawing has also been implementated in Microsoft® Excel: it is included in the *cluster* add-in by Cinquegranni, who has developed an astonishing collection of publicly available statistical and/or visualization tools in Excel [18]. The convex hull drawing procedure is based on advice from Peltier [19], the leading Excel-charting

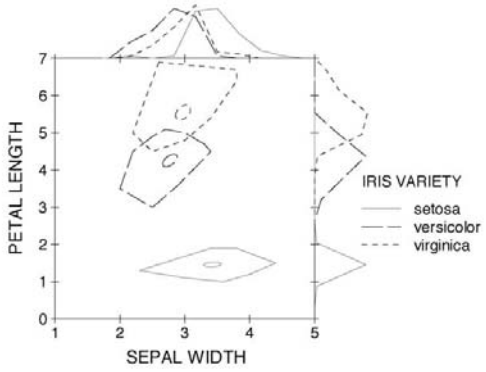


Fig. 1 A convex hull plot of the Iris dataset with confidence ellipses, produced with SYSTAT®.

expert. Although the procedure provides no data-selection or design-related options and just draws the convex hulls around the clusters determined by the clustering algorithm, the VBA code could be used for a more general implementation in line with of our ideas, which might be a worthy challenge for Excel-charting enthusiasts.

On the basis of an extensive search of bibliographic databases and the web, we conclude the list of existing convex hull implementation in statistical and related software with ADE-4 (acronym for Analyses des Données Écologiques), a system for "exploratory and euclidean methods in environmental sciences" [20]. Considering the French tradition of correspondence analysis [21] and ordination methods in ecology [22], it is no surprise that convex hull drawing has been implemented within it.

Because of platform independence, graphical power and flexibility, and because of compatibility of S code [23,24] with the widespread commercial S-PLUS® system [25], we opted for implementation of our ideas in the freely available R system [26]. Our starting point was the *chull* function for convex polygons in two dimensions, which was ported from S-PLUS® to R in 1999 [27]. Recently, convex hulls have been applied within the *multiv* package [28] in the *hierclust* function as the default option for representing clusters in "movie mode", and the ADE-4 system has been ported to R [29], but convenient production of publication-quality convex hull or bivariate density plots with freely chosen data and various options for additional information and marginal distributions call for a self-standing function. Hence, we developed the *chplot* function, which was later expanded into the *chplot* package.

```
chplot (x, y, groups, chull = TRUE,
  clevel = 0.95, band.power = 0.2,
  mar.den = FALSE,
  descriptives = ''mean.sd'',
  dlevel = 0.68, bw = FALSE,
  ratio = 0.75, plot.points = FALSE,
  log = '''', xlab, ylab,
  legend.control (include = TRUE,
  area.in, pos, cex, bty, title),
  ...)

chadd (param, pos, add.fun, ...)
```

3.2. Our implementation in R: the *chplot* package

The package consists of four components: the *chplot* function, the *legend.control* object for detailed legend control, the *chadd* function that enables the user to freely add further elements to the augmented convex hull plot, and the *hdr* dataset presented in the next section. Detailed documentation is included in the package, so here we just list the syntax, which is self-explanatory to a large extent, and give an overview of the functions.

The *chplot* function requires just three vectors of data as input (*x*, *y* and group code), but offers numerous optional parameters (with the most generally useful values as defaults) and retains all the flexibility of R's plotting routines, so that the user can produce a wide variety of plots with precisely controlled features. The first choice for the user to make is whether to draw convex hulls, which is the default, or bivariate density contours. In the latter case, confidence level can be set (the default is 95%) as well as bandwidth (based on [30], the default value is group-size$^{-0.2}$). Next, marginal density plots can be chosen for plotting marginal distributions, whereby the same density scale is automatically used for both dimensions. The default option produces relative frequency polygons with points in the middle of the intervals and the minimum and maximum interval with zero frequency depicted for all the groups on a common relative frequency scale.

The default option for depicting descriptive statistics within convex hulls, or density contours produces a cross with the lines intersecting at the mean of *x* and *y* for each group and depicting the 68% tolerance interval (i.e., stretching one standard deviation in both horizontal and vertical directions). Standard error of the mean

can be chosen instead of the standard deviation; specifying zero-level plots empty convex hulls or density contours, regardless of the variability measure. A non-symmetric alternative is to make the lines depict the first and the third quartile for both axes and cross at the median of *x* and *y*. Instead of the crosses, which are always parallel to the axes of the main plot, confidence ellipses can be selected, which are inclined in accordance with the correlation between *x* and *y* for the given group centered at the respective means and depict the bivariate confidence interval for the mean.

The whole plot can be produced either in color, which is the default, or in black-and-white, in which case a different line pattern is used for each group. Next, the user can specify the ratio of the main plot to the total plot area, which also determines the default legend placement: if the ratio is less or equal to the default value of 0.75, the legend is placed in the top right corner, otherwise it is placed within the main plot and positioned by the user with mouse. If the ratio is 1, marginal distributions are not plotted (and the legend is, of course, placed inside the plot). On the other hand, the user can make the plot more crowded by choosing to plot all the raw data points in addition to the convex hulls or density contours.

The legend can be omitted; if present, as it is by default, its entries are group names (i.e., factor level names in R). If area per point has been calculated, it is reported after group name in parentheses. The area is calculated by default for convex hulls, while for bivariate density plots it can be requested. When the density contours are non-contiguous or even just line fragments, the area per point does not make any sense, and the confidence level and/or bandwidth formula power should be adjusted. One or both variables can be plotted on logarithmic scale; the setting applies to the main plot as well as to the marginal plot(s). If desired, axis titles different from *x* and *y* variable names can be set.

Default legend font size can be overridden, and the same goes for legend frame (which is omitted by default if the legend is placed outside the main plot). When the legend is outside the main plot, the legend title can be set. The ellipsis symbol at the end of the *chplot* function syntax indicates that further arguments related to the legend can be specified in accordance with the R's *plot* function.

Three more characteristics of the *chplot* function are noteworthy: first, cases with missing value on any of the three compulsory variables are excluded from the plot (the function displays a warning to that effect); second, instead of specifying three vectors, only one data-frame (or matrix) with two or three columns can be specified as *x* (the second column being treated as *y* and the third one containing group membership), which makes variable names (if they exist) automatically appear as axis labels and in the legend; third, after producing the plot, the function restores all the graphics parameters to their previous values.

In addition to drawing the plot, the *chplot* function also returns a list object, which is used by the *chadd* function. This function adds crucial functionality, since it allows the user to freely add further elements (from boxes and lines to complex plot) to the augmented convex hull plot.

3.3. Availability

The package has been included in CRAN [31], from where the package source, binaries and reference manual can be downloaded [32]. It was introduced in R version 1.9. The package depends on the *KernSmooth* package [33], since it uses the *bkde2D* function for bivariate kernel density estimation, and on the *ellipse* package [34] for drawing ellipses.

4. Examples

We demonstrate the technique and the capabilities of our software with two datasets from our daily consulting practice. The first one comes from a large collaborative research project on socio-economic determinants of mortality in Slovenia, the results of which have partly been published in the UNDP and government sponsored report on human development and health [35]. The data were obtained by combining the population register (which contains all the census data), the electronic records of compulsory death registration forms and the personal income tax database. The *hdr* dataset contains data on the deceased in 1998 for whom the amount of personal income tax paid could be identified ($N = 9051$). Since missing data issues have not been resolved with the providers yet, the dataset cannot be considered representative of the entire population, but it is very useful for demonstrative purposes. Fig. 2 depicts gender differences in age at death and income tax with convex hulls, whereby the general trend of women living somewhat longer and earning somewhat less than men is clear. Logarithmic axis is chosen for income because the distribution is heavily right-skewed; descriptive statistics (mean $\pm$ standard deviation) are depicted with error bars, marginal distributions are plotted with relative frequency polygons, and information on area of convex hull per point is

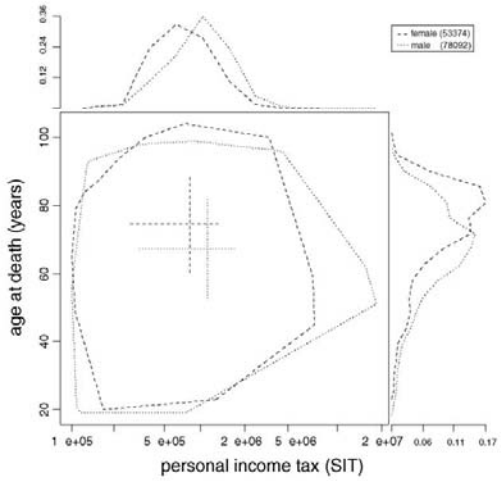


Fig. 2 Augmented convex hull plot of gender differences in age at death and paid personal income tax for deceased in Slovenia in 1998 [35]. R syntax: `chplot(hdr, log=''x'', bw=TRUE, xlab=''personal income tax (SIT)'', ylab='age at death (years)'', title=FALSE, bty=''0'')`.

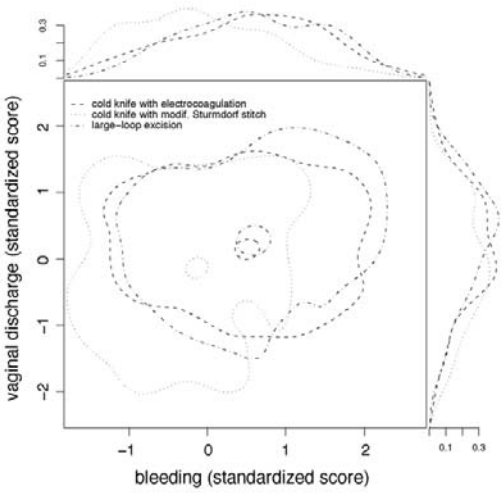


Fig. 3 Augmented bivariate density plot of standardized score of the number of days with patient-reported bleeding and/or vaginal discharge in relation to conization technique [36]. R syntax: `chplot(conization, chull=FALSE, mar.den=TRUE, descriptives='ellipse'', bw=TRUE, pos='in'', bty=''n'', cex=0.7)`.

added (indicating greater variability among men); due to the default main plot-to-total ratio of 0.75, the legend is placed outside the main plot.

The second example is from the field of gynecology, from a study of risk factors for complications after conization [36] in which post-operative outcome was assessed with several "objective" (e.g. increased body temperature) and "subjective" measures (e.g. self-assessed pain) in over 800 women. In Fig. 3, two derived outcome measures (normalized and rescaled number of post-operative days with the patient reporting bleeding and/or vaginal discharge) are plotted for different operation techniques. Descriptive statistics are depicted with confidence ellipses, illustrating the fact that the two chosen dimensions are relatively independent ($r = -0.15$); marginal distributions are plotted with density plots, and since the ratio of the main plot to the total plot area is 0.9, the legend was placed inside the main plot by default and then positioned interactively. The plot indicates what has been confirmed by various multivariate models, namely that one technique (cold-knife conization with modified Sturmdorf stitch) is superior to the other two in terms of fewer post-operative complications.

Of course, the convex hull technique – like any scatter-plot type visualization – is not limited to "raw" data values. As already stressed in the introduction, sets of points can be derived from any multivariate dimension reduction technique, e.g. factor analysis, principal component analysis, multidimensional scaling [37], or correspondence analysis [38], or any functions can be used as axes instead of original dimensions.

5. Conclusion

The proposed augmented convex hull plot with all its variants represents just one of the many possibilities of using computational geometry algorithms for statistical visualization. In that sense, it is at least remotely related to such important and elegant concepts and methods as Voronoi and Delaunay tessellations.

Our basic idea was to eliminate visual clutter from multi-group scatter-plots of large datasets by replacing points with convex hulls or bivariate density contours. That brings about many possibilities for enhancing the exposition of the data: descriptive statistics can be plotted with error bars or confidence ellipses within the convex hulls, a bivariate dispersion measure (area of convex hull or density polygon per point) can be added to the legend, and marginal distributions can be plotted to the side and above the main plot with frequency polygons or density curves.

All these options are easily invoked and controlled with an R function named *chplot*, relieving the user from tedious manual tuning and assuring a coherent look of the chart. Together with the *legend.control* object for detailed legend control, the *chadd* function that enables the user to freely add further elements to the augmented convex hull plot and the *hdr* dataset; it has been included in the *chplot* package in order to make it properly documented and easilly installable through R's packaging mechanism. The package is publicly available at CRAN [31].

Augmented convex hull plots were designed for reducing visual clutter, but as the number of groups increases, even they eventually become unclear. Nevertheless, it is our experience that up to 10 groups can be nicely plotted with the majority of real-life datasets. Even though the expert opinion is that implementation of a 3D convex hull algorithm in R is a difficult task [39], a possibility for the future might be the extension of the procedure to three dimensions, possibly using the *Scatterplot3d* package [40], or even making the visualizations interactive with the *rgl* package [41]. However, as with any newly proposed visualization type, the immediate challenge is to see whether augmented convex hull plots gain wider acceptance.

References

[1] B. Efron, The convex hull of a random set of points, Biometrika 52 (1965) 331–453.
[2] M. Shamos, Geometry and statistics: Problems at the interface, in: J. Traub (Ed.), Recent Results and New Directions in Algorithms and Complexity, Academic Press, New York, 1976, pp. 251–280.
[3] B. Ripley, Spatial Statistics, first ed., Wiley, New York, 1981.
[4] H. Kamiya, A. Takemura, On rankings generated by pairwise linear discriminant analysis of *m* populations, J. Multivariate Anal. 61 (1997) 1–28.
[5] G. Ragozini, A data-driven discriminant rule by Voronoi tessellation, NTTS '98–Seminar on New Techniques and Technologies for Statistics: 4–6 November 1998, Sorrento, Italy, 1998, http://europa.eu.int/en/comm/eurostat/research/conferences/ntts-98/papers/cp/071c.pdf.
[6] Y.U. Degytar, M.Y. Finkelshtein, Classification algorithms based on construction of convex hulls of sets, Eng. Cybermetics 12 (1974) 150–154.
[7] V. Di Gesu, M.C. Maccarone, Description of fuzzy images by convex hull technique, Proceedings of the 8th International Conference on Pattern Recognition (ICPR), International Association for Pattern Recognition, Paris, 1986, pp. 1276–1278.
[8] L.Y. Shan, M. Thonnat, Description of object shapes by apparent boundary and convex-hull, Pattern Recognition 26 (1993) 95–107.
[9] http://www.cse.unsw.edu.au/~lambert/java/3d/hull.html.
[10] C.B. Barber, D.P. Dobkin, H. Huhdanpaa, The quickhull algorithm for convex hulls, ACM Trans. Mathematical Software (TOMS) 22 (1996) 469–483.
[11] http://www.thesa.com/software/qhull/.
[12] M.J. Greenacre, Theory and Applications of Correspondence Analysis, first ed., Academic Press, London, 1984.
[13] W.S. Cleveland, R. McGill, The many faces of a scatterplot, J. Am. Stat. Assoc. 79 (1984) 807–812.
[14] S. Lewandowsky, I. Spence, Discriminating strata in scatterplots, J. Am. Stat. Assoc. 84 (1989) 682–688.
[15] W.S. Cleveland, Visualizing Data, first ed., Hobart Press, Summit, 1993.
[16] http://www.systat.com.
[17] R.A. Fisher, The use of multiple measurements in taxonomic problem, Ann. Eugenics Part II (1936) 179–188.
[18] http://www.prodomosua.it.
[19] http://peltiertech.com.
[20] J. Thioulouse, D. Chessel, S. Dolédec, J.M. Olivier, ADE-4: a multivariate analysis and graphical display software, Statistics Computing 7 (1997) 75–83.
[21] J.-P. Benzecri, L'analyse des données tome 2: l'analyse des correspondances, Bordas, Paris, 1980.
[22] P. Legendre, L. Legendre, Numerical Ecology, second ed., Elsevier, Amsterdam, 1998 (in English).
[23] R.A. Becker, J.M. Chambers, A.R. Wilks, The New S Language, Chapman & Hall, London, 1988.
[24] J.M. Chambers, Programming with Data: A Guide to the S Language, Springer, New York, NY, 1998.
[25] http://www.insightful.com/products/splus.
[26] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2004, ISBN 3-900051-00-3, http://www.R-project.org.
[27] http://maths.newcastle.edu.au/~rking/R/devel/99a/0299.htm.
[28] http://cran.r-project.org/doc/packages/multiv.pdf.
[29] http://cran.r-project.org/doc/packages/ade4.pdf.
[30] B. Everitt, S. Rabe-Hesketh, Analyzing Medical Data using S-PLUS, Springer, New York, NY, 2001.
[31] http://cran.r-project.org/.
[32] http://cran.r-project.org/src/contrib/Descriptions/chplot.html.
[33] http://cran.r-project.org/doc/packages/KernSmooth.pdf.
[34] http://cran.r-project.org/doc/packages/ellipse.pdf.
[35] J. Javornik, V. Korosec (Eds.), Human development report Slovenia 2002/2003: Human development and health, Institute of Macroecomic Analysis and Development, Ljubljana, 2003.
[36] A. Zupancic-Pridgar, Influence of vaginal flora on morbidity after conization, M.Sc. thesis, University of Ljubljana, Faculty of Medicine, Ljubljana, 2003.
[37] E.D. Gallagher, D. Shull, Statistical analyses of Boston Harbor benthos: 1991–1998, University of Massachusetts, Department of Environmental, Coastal and Ocean Sciences, Boston, 1999, http://www.es.umb.edu/faculty/edg/files/pub/bh98rept2.pdf.
[38] T. Pipan, Ecology of copepods (Crustacea: Copepoda) in percolation water of the selected karst caves, Ph.D thesis, University of Ljubljana, Faculty of Biotechnology, Department of Biology, Ljubljana, 2003.
[39] http://maths.newcastle.edu.au/~rking/R/help/01c/1404.html.
[40] U. Ligges, M. Mächler, Scatterplot3d–an R package for visualizing multivariate data, J. Stat. Software 8 (11) (2003), http://www.jstatsoft.org/v08/i11/JSSs3d.pdf.
[41] D. Adler, RGL: 3D visualization device system (OpenGL), http://wsopuppenkiste.wiso.uni-goettingen.de/~dadler/rgl/.

PRILOGA 2: PRIROČNIK ZA PAKET CHPLOT ZA OKOLJE R

The chplot Package

July 12, 2005

Title Augmented Convex Hull Plots

Version 1.2

Author Maja Pohar, Gaj Vidmar

Maintainer <maja.pohar@mf.uni-lj.si>

Depends ellipse, KernSmooth, lattice

Description Informative and nice plots for grouped bivariate data.

License GPL version 2 or newer

URL http://www.mf.uni-lj.si/ibmi-english

R topics documented:

`chadd` *Add elements to a chplot*

Description

`chadd` is a function that adds any further elements to a plot produced with the function `chplot`.

Usage

```
chadd(param,pos,add.fun,...)
```

Arguments

`param`	the parameters specifying the plotting regions. As obtained from `chplot`.
`pos`	the plotting region to which an element is to be added: `1` for the main plot, `2` for top left (marginal distribution plot of the x-variable), `3` for top right (the legend), `4` for bottom right (marginal distribution plot of the y-variable).
`add.fun`	the function to be applied.
`...`	optional parameters to `add.fun`, separated by commas.

2 chplot

See Also

chplot.

Examples

```
data(hdr)
# hdr dataset
param<-chplot(age~income|gender,data=hdr,log="x")
# box around the legend region
chadd(param,3,box,"figure")
# tickmark for overall mean in marginal distribution plots
chadd(param,2,lines,c(mean(hdr$income),mean(hdr$income)),c(0,.05))
chadd(param,4,lines,c(0,.025),c(mean(hdr$age),mean(hdr$age)))
```

`chplot` *Augmented Convex Hull Plot*

Description

Plots 2D convex hulls or bivariate density contours, one for each group of data. Descriptive statistics are plotted as error bars or confidence ellipses within convex hulls. Marginal distributions as well as a special legend are added by default. Additionally, area of convex hull per point can be displayed.

Usage

```
chplot(formula,data,chull=TRUE,clevel=0.95,band.power=.2,
mar.den=FALSE,descriptives="mean.sd",dlevel=0.68,bw=FALSE,ratio=.75,
plot.points=FALSE,log="",xlab,ylab,col,lty,legend,...)
```

Arguments

`formula`	a formula describing the form of conditioning plot. The formula is generally of the form `y ~ x | g1`, indicating that plots of `y` (on the y axis) versus `x` (on the x axis) should be produced conditional on the variable `g1`. However, the conditioning variable `g1` may be omitted. The names of the `y`, `x` and `g1` variables are used for the axes and legend title.
`data`	a data frame containing values for any variables in the formula.
`chull`	logical; if `TRUE` (default), convex hulls are plotted, otherwise density contours are drawn.
`clevel`	the confidence level for the density plot if `chull=FALSE` (default is 0.95).
`band.power`	applies if `chull=FALSE`; bandwidth for bivariate density estimation is calculated as `(group size)^(-band.power)` for each group (default is 0.2).
`mar.den`	logical; defines the type of marginal distribution plots. If `FALSE` (default), relative frequency polygons are plotted, otherwise density plots are drawn.
`descriptives`	the option to be used for depicting descriptive statistics. The default value is `mean.sd`, which produces a cross with the lines intersecting at the mean of `x` and `y` for each group and depicting the 68-percent tolerance interval (i.e., stretching one standard deviation in each direction) with default `dlevel` setting. Option `mean.se` does the same with standard errors of the means, while `median` makes the lines one quartile long in each direction and cross at the median. Option `ellipse` plots confidence ellipses.

`dlevel`	the tolerance/confidence level applied if `descriptives=mean.sd` or `mean.se`.
`bw`	logical; if `TRUE`, the plot is produced in black-and-white. The default is `FALSE`, which plots in colour.
`ratio`	the ratio of the main plot to the whole figure region. The default value is 0.75. If equal to 1, the marginal distributions are not plotted. If the ratio is less or equal to 0.75, the default legend position is outside the main plot (i.e., in the top right corner), otherwise it is within the main plot (i.e., to be selected by the user with mouse).
`plot.points`	are the points added to the original plot? Default is `FALSE`.
`log`	the character strings `"x"`, `"y"` or `"xy"` makes a specific (or both) axes logarithmic; the default, which does nothing, is `""`.
`xlab`	a title for the x axis in the main plot; the default is the name of the x variable.
`ylab`	a title for the y axis in the main plot; the default is the name of the y variable.
`col`	the plotting colors; vector of length equalling the number of groups.
`lty`	the type of line; vector of length equalling the number of groups.
`legend`	either logical, in which case the default is `TRUE` and a legend is drawn with the default settings, or a list of legend parameters. For legend parameters, see the `legend` function; two additional parameters are allowed: `area.in` (logical; specifies whether the area of convex hull per point is displayed; default is `TRUE` if `chull=TRUE` and `FALSE` otherwise), and `pos` (`in` or `out`; default depends on `ratio` value; if `in` is chosen, the legend is positioned by the user with mouse, while `out` places the legend in the top right corner of the plotting area). Note that the default legend parameters are not the same as in the `legend` function: default `title` is the name of the grouping variable, default `bty` depends on legend position (`"o"` if inside and `"n"` if outside), while default `cex` is calculated on the basis of `ratio`, and the `cex` option is the multiplier of that default.
`...`	other arguments will be passed to the main plotting region and will affect points if `plot.points=TRUE`.

Details

The relative frequency polygons chosen with `mar.den=FALSE` connect the points in the middle of the intervals and the starting and ending interval with zero frequency in each group. All the frequencies are rescaled in order to make the plots immediately comparable (the same is true for the density curves).

The default density contour might not fit into the plot - this can be avoided by decreasing the `clevel`. In case of broken contours, the area per point is not a sensible measure.

Value

A list with components:

`area`	the area of convex hull per point for each group.
`usrc`	the limits of the central plotting region.
`usru`	the limits of the top left plotting region.
`usrr`	the limits of the bottom right plotting region.
`ratio`	the ratio used.
`is.xlog`	logical, denoting whether either `log="x"` or `log="xy"` was used.
`is.ylog`	logical, denoting whether either `log="y"` or `log="xy"` was used.

References

Vidmar, G., and Pohar, M. Augmented convex hull plots: rationale, implementation in R and biomedical applications. *Computer Methods and Programs in Biomedicine*, 2005, 78, 69-74.

See Also

`chadd`, `chull`, `bkde2D`.

Examples

```
# the hdr dataset
data(hdr)
chplot(age~income|gender,data=hdr,log="x")
# the iris dataset
data(iris)
chplot(Sepal.Length~Sepal.Width|Species,data=iris,bw=TRUE,
       legend=list(cex=.6))
chplot(Petal.Length ~ Petal.Width | Species, data = iris,
      legend = list(cex = 0.6),plot.points=TRUE,pch = 18, cex = 0.5)
```

`hdr` *Deceased in Slovenia in 1998*

Description

Personal income tax payed (in SIT), age (in years) and gender for deceased in Slovenia in 1998.

Usage

```
data(hdr)
```

Format

`hdr` is a list with 9051 cases (rows) and 3 variables (columns) named `income`, `age` and `gender`.

Source

Statistical Office of the Republic of Slovenia.

References

Javornik, J., and Korosec, V. Eds. (2003) *Human development report Slovenia 2002/2003: Human development and health*. Ljubljana: Institute of Macroecomic Analysis and Development.

Index

5

PRILOGA 3: ANKETA O RAZŠIRJENIH KONVEKSNOLUPINSKIH DIAGRAMIH

Spostovana kolegica oziroma kolega,

lepo te prosim za pomoc pri izdelavi mojega doktorskega dela!

- Ne, ne bo treba narediti nicesar namesto mene, pac pa te prosim, da si vzames dobre pol ure casa za sodelovanje v anketi. Da, prav si prebral[a], dobre pol ure, in to zbranega dela, in se R bo treba uporabljati! In za to ne bos nic placan[a]!

Zakaj naj se torej trudis, namesto da tole sporocilo takoj zbrises? - Zato, ker verjamem, da ti bo nekaj pomenila moja obljuba, da ti uslugo vrnem, ce bos v podobnem polozaju pri izdelavi lastnega magisterija oziroma doktorata, in ker verjamem, da bos ob tem zvedel[a] nekaj koristnega in se naucil[a] uporabljati paket za R, za katerega verjamem, da je zelo preprost in koristen za vsakogar, ki se ukvarja s statistiko :)

Malo bolj formalno te torej lepo prosim, da sledis spodnjim navodilom, nato pa mi svoje odgovore posljes kot odgovor na tole sporocilo (ker te ne podcenjujem, ti popolnoma zaupam, da ti jih bo uspelo tako ali drugace vnesti), za kar se ti ze vnaprej od srca zahvaljujem!

Statisticni pozdrav,

Gaj Vidmar
mlajsi veteran z IBMI

Gre za prikaz podatkov z razsirjenimi konveksnolupinskimi diagrami oziroma paketom chplot za R. Ampak se zdalec ni tako zapleteno, kot se je zdajle slisalo!

Konveksno lupino si je najlazje predstavljati kot crto, ki v mnozici tock povezuje tiste, ki so "najbolj zunanje". V statistiki jo lahko uporabimo za prikaz skupne porazdelitve dveh spremenljivk. Ker pa nam pokaze le lego skrajnjih vrednosti, ne pa tudi srednje vrednosti in obicajnih mer razprsenosti, ki nas najveckrat najbolj zanimajo, jo je smiselno dopolniti s "krizem", ki ima sredisce v srednji vrednosti (aritmeticni sredini ali mediani spremenljivk, ki ju nanasamo na koordinatnih oseh), z dolzino "krakov" pa prikazuje razprsenost (npr. standardni odklon, interval zaupanja za aritmeticno sredino ali pa interkvartilni razmik).

Se bolj celovito sliko dobimo, ce dodamo prikaz robnih porazdelitev: porazdelitev spremenljivke, ki doloca vodoravno os, narisemo zgoraj, spremenljivke, ki doloca navpicno os, pa na desni strani. Da bi videli celotno obliko porazdelitve, uporabimo "frekvencni mnogokotnik" (s crto povezemo tocke, ki bi bile na sredini vrhnjih stranic stolpcev histograma). Sliko, ki jo tako dobimo, sem poimenoval razsirjeni konveksnolupinski diagram. Namenjen je prikazu porazdelitve dveh stevilskih spremenljivk pri vecih skupinah (t.j. pogojno glede na eno opisno spremenljivko), ce imamo opravka z zelo veliko podatki in bi bil obicajni razsevni diagram (scatterplot z razlicnimi barvami ali simboli) nepregleden.

```
Podobno, kot lahko v vsakdanjem zivljenju s kanckom domisljije marsikaj
razdelimo v kategoriji "oglato" in "okroglo", lahko tudi konveksne lupine s
krizi in zagastimi crtami nadomestimo z obrobami ocenjene gostote
(bivariate density contours), elipsami zaupanja (confidence ellipses – s
svojima osema kazejo bivariatni interval zaupanja za povprecje, nagnjene pa
so glede na korelacijo med obravnavanima spremenljivkama) in zglajenimi
diagrami gostote verjetnosti posamezne spremenljivke.

- Uuuh, se vedno se slisi zapleteno, ne? In za narisati je najbrz se
tezje!? - Ne, ni! Za hitro in res preprosto izdelavo takih diagramov v R-u
sva namrec s kolegico Majo Pohar naredila paket chplot. Ko ga bos zacel[a]
uporabljati, ti bo ob lepih slikicah hitro vse jasno :) Verjemi mi, da res
ne bo nic zapletenega, saj sem si vse skupaj izmislil jaz, ki nisem
matematik, pa se pri izdelavi paketa sem sodeloval in ga tudi sam
uporabljam, ceprav se sicer Ra se vedno bojim in raje klikam po bolj
neumnih programih.

------------------------------------------------------------

Skratka, dovolj teorije, na delo! Lepo prosim, pozeni R in si namesti paket
chplot. Ker beres e-posto, imas povezavo do spleta, zato ne bo tezav, saj
je paket chplot vkljucen v CRAN:

      Packages -> Install package(s)...

Potem ga "nalozi" (chplot pa bo sam nalozil se pakete ellipse, KernSmooth
in lattice, ki jih potrebuje za svoje delo):

      library (chplot)

Sedaj si vzemi 10 minut casa in preberi, kako se uporablja funkcijo chplot:

      ?chplot

Seveda si ni treba zapomniti vseh opcij, le osnovno sintakso:

chplot(formula kot pri glm-ju, data=podatki, opcije)

Ker bo tvoj prvi razsirjeni konveksnolupinski diagram prikazoval podatkovje
hdr (ki je prislo s paketom), se moras z njim vsaj na hitro seznaniti:

      ?hdr

Med desettisoc pikami bi se bilo tezko znajti, ne? Zato naredi, kot pravita
pomoc in prirocnik za paket chplot:

      data(hdr)
      chplot(age~income|gender,data=hdr,log="x")

------------------------------------------------------------
In tu je ze PRVO VPRASANJE (vsega skupaj jih je samo sedem):

1. Kaj kazejo podatki hdr? (izberi en odgovor; predpostavimo, da placati
vec dohodnine pomeni vec zasluziti)

      a) zenske v splosnem zivijo dlje in vec zasluzijo
      b) moski v splosnem zivijo dlje, a zasluzijo manj
      c) moski v splosnem prej umrejo, a zasluzijo vec
      d) zenske v splosnem prej umrejo in manj zasluzijo
------------------------------------------------------------
```

Tudi v nadaljevanju, prosim, skrbno prenasaj ukaze iz tega sporocila v R in opazuj dogajanje na zaslonu :) Najprej pripeto datoteko sim.r shrani na prirocno mesto (v Oknih npr. na namizje ali v mapo C:\Temp, kot predpostavljajo tale navodila), nato pa jo odvleci v R ali uporabi ukaz

```
source("C:/Temp/sim.r")
```

Gre za 40000 trorazseznih podatkov – po 10000 iz vsake od stirih skupin. Razsevni diagram narisi takole:

```
attach(sim)
plot(x,y,col=z+2)
```

Zapomnisi si, kaj vidis (eee, pravzaprav ...), nato pa zadevo narisi z razsirjenim saj_ves_cim:

```
chplot(y~x|z)
```

Sedaj ti lahko izdam, da so podatki nazrebani iz stirih asimetricnih porazdelitev enake oblike (torej enake razprsenosti), ki pa so "premaknjene", torej imajo razlicna povprecja.

DRUGO VPRASANJE bi tako moralo biti se lazje od prvega:

2. Katera skupina je bila (najverjetneje) nazrebana iz porazdelitve z najvecjim povprecjem x in y?

Zdaj, ko ze malo obvladas zadevo, pa se lahko spoprimes z velikani statistike! Prikazati moras namrec znamenite Fisherjeve perunike (iris dataset, ki se namesti z Rom), pri cemer mora biti tvoj diagram se jasnejsi od tistega, ki ga je izdelal William Cleveland v svoji knjigi Visualizing Data (1993). V vzorcu je po 50 perunik vsake od treh vrst, znano pa je, da jih najbolje locimo, ce v logaritemskem merilu prikazemo odvisnost razmerja dolzine in sirine vencnih listov (podolgovatost=petal length/petal width) od njunega zmnozka (velikost=petal length*petal width). Cleveland je to pokazal takole (navpicni crti ponazarjata najboljse pravilo za razvrscanje):

```
data(iris)
library(lattice)
set.seed(19)
petal.length <- iris[,3,]
petal.width <- iris[,4,]
variety <- factor(iris[,5,])
n <- length(levels(variety))
mea <- (log(petal.length,2)+log(petal.width,2))/2
dif <- jitter(log(petal.length,2)-log(petal.width,2), 2)
xyplot(dif ~ mea,
 panel = function(...){
 panel.superpose(...) panel.abline(v = c(0.4, 1.46))},
 groups = variety, aspect = 1,
 xlab = "velikost (log 2 cm)",
 ylab = "raztresena podolgovatost (log 2 razmerja)",
 key = list(points = Rows(trellis.par.get("superpose.symbol"), 1:n),
 text = list(paste("Iris", levels(variety))), columns = n))
```

S chplotom pa zadosca tole (R bo pred risanjem navpicnih crt pocakal, da s klikom na levi gumb misi postavis legendo):

```
x<-log(sqrt(iris[,3]*iris[,4]))/log(2)
y<-log(iris[,3]/iris[,4])/log(2)
variety <- factor(paste("Iris",iris[,5,]))
param<-chplot(y ~ x | variety, legend=list(area.in=F),
 xlab="velikost (log2 cm)",ylab="podolgovatost (log2 razmerja)",
 dlevel=0,ratio=1)
chadd(param,1,abline,v=1.46)
chadd(param,1,abline,v=.4)
```

---

3. Ti je uspelo?

Ce ti je, si zdrzal[a] ze skoraj do konca! Sledi se ena naloga, nato pa le se tri res kratka vprasanja :) Je pa naloga nekoliko tezja. Ce se ti je ne ljubi narediti ali ti je v desetih minutah ne uspe narediti, jo preskoci, le napisi mi, prosim, kaj je bil razlog.

Torej, z razirjenimi konveksnolupinskimi diagrami moras prikazati pripete podatke o porodih (datoteka porodi.r). Najprej jih shrani na disk in nalozi v R (priporocam tudi attach). Kot vidis, obsegajo sest spremenljivk:
- pto = porodna teza otroka (v kilogramih)
- spol = spol otroka (M ali Z)
- gestac = gestacijska starost (v tednih; t.j. trajanje nosecnosti)
- starost = starost matere ob porodu (v letih)
- kajenje = kajenje matere med nosecnostjo (ne, obcasno, pogosto)
- vm = vzdraznost maternice med nosecnostjo (da ali ne)

NALOGA:

4. Za te podatke smiselno uporabi funkcijo chplot, pri cemer poskusi
- uporabiti obrobe gostote verjetnosti (density contours namesto konveksnih lupin): ce rises pto v odvisnosti od starosti, priporocam visoko stopnjo zaupanja, npr. 0.99, in majhen paramater glajenja, lahko celo 0.03; ce pa v odvisnosti od gestac, je bistveno dolociti dovolj majhen parameter glajenja in ne uporabiti robnih gostot verjetnosti;
- uporabiti elipse zaupanja (confidence ellipses namesto prikaza opisnih statistik s krizi) s stopnjo zaupanja 0.99;
- nastaviti barvo ali tip crte.

Ce ti je uspelo, prosim, v enem stavku cisto na kratko opisi oziroma interpretiraj dobljeni diagram! (Ce ti ni uspelo, prosim, napisi, zakaj.)

Zdaj pa povzemi svojo izkusnjo s chplotom in OZNACI, KAJ SI MISLIS:

5. V primerjavi z ostalimi paketi za okolje R je chplot
   a) izrazito preprost
   b) razmeroma preprost
   c) podobno zahtevnen kot vecina drugih paketov v Ru
   d) razmeroma zahteven
   e) izrazito zahteven

6. Celotna zamisel o razsirjenih konveksnolupinskih diagramih se mi zdi
   a) izjemno koristna
   b) zmerno koristna
   c) nekaj srednjega (ni cisto brezzvezna, kaj posebnega pa tudi ni)
   d) v glavnem nekoristna
   e) popolnoma nekoristna

In se ZADNJE VPRASANJE:

7. Ce poznas se kaksno drugo graficno metodo, ki jo lahko namesto razsirjenih konveksnolupinskih diagramov uporabis za prikaz vrednosti dveh stevilskih spremenljivk glede na vrednost ene opisne spremenljivke (skupine, kategorije), jo (ju, jih) prosim, navedi:

Tako - upam, da je bilo cisto zares zanimivo in poucno.

Lahko te potolazim, da bi te lahko ze na zacetku obremenil se z literaturo, pa te nisem :) No, ce te razsirjeni konveksnolupinski diagrami se zanimajo, si seveda lahko preberes clanek o njih (G.Vidmar, M.Pohar: Augmented convex hull plots: rationale, implementation in R and biomedical applications, Computer Methods and Programs in Biomedicine, 2005, letn. 78, stev. 1, str. 69-74), ki je dostopen preko spletnih strani IBMI (Biostatisticni center -> Raziskovalno delo -> Izbrane objave). Ce te zanima prikaz podatkov nasploh, pa sem ti seveda vedno z veseljem pripravljen svetovati oziroma posredovati tiskano in elektronsko literaturo s tega podrocja (ki je je dobesedno ogromno).

Za konec te lepo prosim samo se za dva podatka o tebi:

STAROST _____
DIPLOMIRAL[A] IZ _____

- in H V A L A ! ! ! ! ! !

gaj.vidmar@mf.uni-lj.si

PRILOGA 4: V OBJAVO SPREJETI ČLANEK O KONKORDANČNIH DIAGRAMIH

Visualising Concordance

Gaj Vidmar[1] and Nino Rode[2]

[1] Institute of Biomedical Informatics, Faculty of Medicine, University of Ljubljana, Vrazov trg 2, SI-1000 Ljubljana, Slovenia
[2] Faculty of Social Work, University of Ljubljana, Topniska 33, SI-1000 Ljubljana, Slovenia

Summary

Concordance describes the agreement between m rankings of k objects. Despite the long history of measures of concordance and the recently revived interest in comparison of concordance (c.f. Legendre 2005), the task of visualising concordance remained virtually unaddressed. We first show how to depict concordance by simply plotting raw data in parallel coordinates. Then we review further possibilities for depicting concordance using the recently developed plots of inter-rater variability in ordinal ratings (Nelson & Pepe 2000) and plots of correlation matrices (Trosset 2005). Next, we propose two novel concordance plots. The concordance bubble-plot is based on raw rank data, while the pin-cushion plot depicts rank differences in polar coordinates. We present visualisations of artificial and real-life datasets with different degree of concordance and identify strong and weak points of the proposed plots. In conclusion, we review some other work related to visualisation of concordance and discuss some other options for constructing novel concordance plots.

Keywords: Concordance coefficients, Graphical methods, Ranks, Parallel coordinates, Polar coordinates, Bubble plot, Pin-cushion plot

1 Introduction

The problem of concordance pertains to m rankings of k objects. Such data are frequently gathered in human resources management, education, marketing and elsewhere, when job applicants, promotion candidates, products, political parties or other subjects or objects are ranked by executives, experts, focus groups, other human observers or even automated algorithms. There are different definitions and measures of concordance, so before considering the issue of visualising concordance, we briefly review the various concordance coefficients.

A number of sources can serve as the basis for such a review (Palachek & Schucany 1984, Siegel & Castellan 1988, Legendre & Lapointe 2004). Several concordance measures have been proposed, but with the exception of top-down correlation (Iman & Conover 1987), they are all based on bivariate ordinal correlation – either Spearman's ρ or Kendall's τ. The best-known and most widely used solution is the one proposed by Kendall & Babington-Smith (1939). Their coefficient of concordance W is basically the average ρ of all possible pairs of rankings (1), and it is related to the Friedman's test (repeated-measures analysis of variance using ranks) in the sense that Friedman's χ^2 statistic provides the means for testing significance of W against the null-hypothesis of no concordance:

$$W = [(m-1)\,\mathrm{E}(\rho) + 1] / m;\ \chi^2 = W\,m\,(k-1),\quad df = k-1. \tag{1}$$

Kendall's coefficient of agreement is obtained the same way as W except that τ is used instead of ρ, which can be written as $u_K = W_\tau$. Ehrenberg (1952) proposed the simple uncorrected average of τ's as a coefficient of agreement, so the resulting coefficient (2) is linearly related to u_K:

$$u_E = \mathrm{E}(\tau_{\text{ranking-ranking}}) = [u_K\,(m/(m-1)) - 1] / (m-1). \tag{2}$$

As first proposed by Lyerly (1952), by changing τ back to ρ (and hence u_K back to W) in the above formula, one obtains an analogous concordance measure as the expected value of ρ's ($u_L = \mathrm{E}(\rho_{\text{ranking-ranking}})$). Finally, if a selected ranking is considered a target or benchmark with respect to the other rankings, the appropriate measure is the correlation between a group of judges and a criterion, which is defined as the average of individual judge-criterion correlations ($T_C = \mathrm{E}(\tau_{\text{ranking-criterion}})$). This coefficient is also computationally linked to some of the methods for comparing concordance between groups, reviewed in Section 3.

3

2 Existing Possibilities to Depict Concordance

Having absorbed the statistical notion of concordance, one can start looking for its graphical representation. However, we did not find any indication that the task of visualising concordance has been directly addressed so far. Nevertheless, without inventing any new statistical graph type, the rank data for concordance analysis can be directly depicted using parallel coordinates (Inselberg 1985). As shown by Wilkie (1980), Kendall's τ can be calculated from and depicted by the number of intersections of lines between corresponding points in the two rankings plotted in parallel coordinates. Similarly, the more intersections we see if we plot concordance data (with k lines) in parallel coordinates with each axis (of the m axes) representing one judge, the less there is concordance, and vice versa.

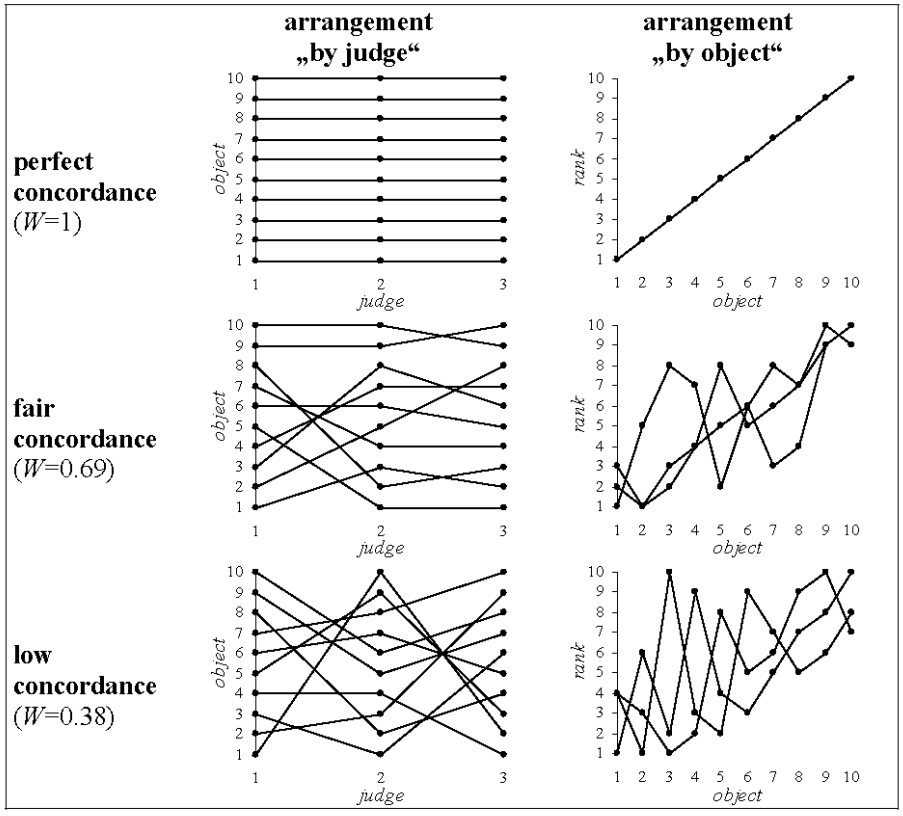


Figure 1: Representation of concordance data (ranking of k=10 objects by m=3 judges) using parallel coordinates for different levels of concordance (perfect, fair and low): parallel axes can be judges (left) or objects (right, sorted by increasing mean rank).

Such arrangement „by judge“ is one of the two possibilities for depicting raw concordance data in parallel coordinates; the other is „by object“, i.e., using one line (of the m lines) for each judge with each axis (of the k axes) representing one object, whereby perfect concordance corresponds to completely overlapping lines. In the arrangement „by judge“, judges can be ordered on the basis of mutual similarity determined by *a posteriori* tests and clustering proposed by Legendre (2005), whereas in the arrangement „by object“ it is useful to order objects by mean rank. Different degrees of concordance are depicted using both arrangements in Figure 1.

The closest match to concordance plots among special-purpose visualisation is graphical description of inter-rater variability in ordinal categorical ratings in medical setting (Nelson & Pepe, 2000), but that method has slightly different scope. Its main idea, i.e., that the mean-variance relationship of ordinal data requires rating distributions to be examined separately for different mean rating levels in order to get a full description of inter-rater variability, does also apply to mean ranking levels and inter-judge variability, and we do apply it in the concordance bubble-plots introduced in Section 4.1. However, the principle to simply add a third dimension for the mean rating level to the simple histogram display of total variability for all objects is more effective for sample sizes much larger than those usually encountered in concordance studies. Furthermore, the method extends naturally to more possible values (i.e., continuous data, with the distributions possibly smoothed), rather than to fewer possible values (i.e., rank data, especially in case of ties).

The third existing possibility to depict concordance data is offered by visualisation of correlation matrices proposed by Trosset (2005). He based his method on angular representation of product-moment correlation using h-plots (Corsten & Gabriel 1976, Seber 1984), and extended it to arbitrary correlation matrices using multidimensional scaling. If applied to matrices of Spearman rank-correlation coefficients, the method seems to be the natural graphical companion to Legendre's method of testing concordance (i.e., clustering of judges), but it turns out to have notable drawbacks. The method depicts perfect concordance correctly with overlapping lines at three o'clock position (and practically unambiguously, since the lines always overlap and the attained minimum of the objective function is always below 10^{-8}), but the algorithm fails to converge to a correct and unique representation of zero concordance. We found that with the default parameters, the code (http://www.math.wm.edu/~trosset/Research/MDS/vc.s), which we ported to R (R Development Core Team, 2004), also failed to converge even with statistically significant concordance (W=0.675, p=0.033) for correlation matrix based on k=m=4. It kept converging, though, for lower and non-significant concordance with a somewhat larger dataset (W=0.367, p=0.088; objective function 2.692; k=4, m=6; data on French wine experts from Schucany & Frawley 1973, presented in more detail in Section 4.1). In any case, moderate concordance

with small k and n is to be expected in real-life applications, and the method can fail with such data. Moreover, the angular correlation-matrix plots provide no information on objects and their mean ranks. Nevertheless, our approach shares a basic idea with them, namely polar co-ordinates, which we use in pin-cushion plots (Section 4.2).

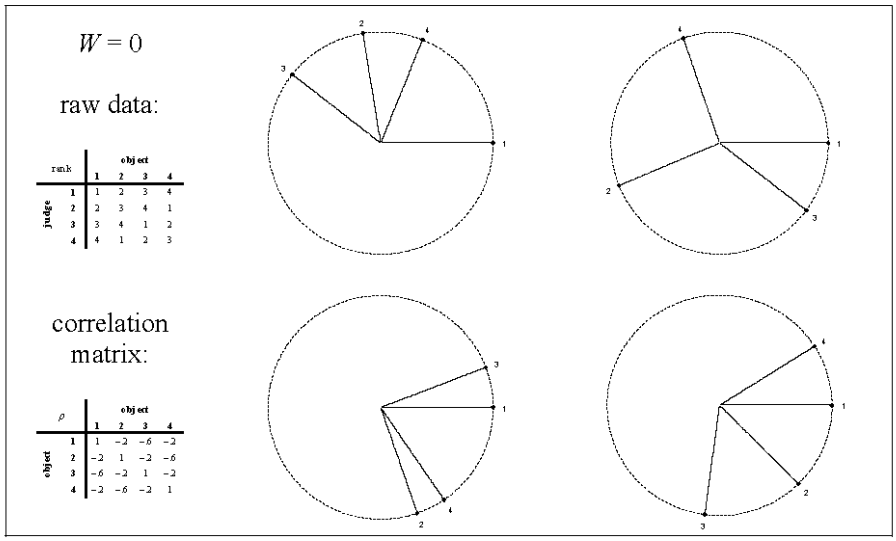


Figure 2: Representation of concordance data (k=m=4) using visualisation of Spearman rank-correlation matrices according to Trosset (2005): for zero concordance, the method fails to converge (raw data, correlation matrix and four examples of resulting plots given).

The deficiencies of the exisitng possibilities motivated us to develop new ways of visualising concordance. A further motive was that in addition to depicting a dataset corresponding to a single concordance coefficient, visualisation of concordance should also serve as a means for quickly and efficiently comparing concordance between groups or samples, or at least to clarify the associated statistics, and none of the exisitng possibilities is particularly suitable for that purpose. Like with concordance coefficients, we briefly review the statistics for concordance comparison first, and then proceed with visualisation.

3 Comparison of Concordance between Groups

The basis of the majority of the multi-group concordance analysis methods is the two-group $\mathscr{L}$ statistic (3), introduced by Schucany & Frawley (1973). It is a generalization of the logic of the test for ordered alternatives and its L statistic

(Page 1963). The goal is to test concordance both within and between two groups of judges (of size m_1 and m_2, respectively) ranking k objects. If the judges (rankings) are represented by m_1 and m_2 rows and the objects are represented by k columns of the two respective tables of ranks, and $\boldsymbol{s}$ and $\boldsymbol{t}$ denote vectors of column sums in those two tables (with elements S_i and T_i, respectively), the statistic is $\mathscr{L} = \Sigma_{i=1..k} S_i T_i = \boldsymbol{s}^{\mathrm{T}} \boldsymbol{t}$. The appropriate statistic for testing the null hypothesis of no concordance between groups is $\mathscr{L}^*$ (3), which is based upon the expected value and variance of $\mathscr{L}$ (4), which, in turn, are derived from $\mathrm{E}(S_i)$, $\mathrm{Var}(S_i)$ and $\mathrm{Cov}(S_i, S_j)$:

$$\mathscr{L}^* = [\mathscr{L} - \mathrm{E}(\mathscr{L})] / \mathrm{Var}(\mathscr{L}); \tag{3}$$

$$\mathrm{E}(\mathscr{L}) = m_1 m_2 k (k+1)^2 / 4;\ \mathrm{Var}(\mathscr{L}) = m_1 m_2 (k-1) k^2 (k+1)^2 / 144. \tag{4}$$

After appraising the normal approximation of $\mathscr{L}^*$, Schucany and Frawley proposed a standardization of $\mathscr{L}$ due to its finite range (6). The resulting statistic (5) is confined to the usual interval of a correlation measure: $\mathscr{W} = 0$ indicates no concordance within and between groups; $\mathscr{W} = 1$ when there is complete agreement within each group on the same rank ordering; and $\mathscr{W} = -1$ when there is perfect agreement within each group but the two groups manifest opposite ordering:

$$\mathscr{W} = [\mathscr{L} - \mathrm{E}(\mathscr{L})] / [\max(\mathscr{L}) - \mathrm{E}(\mathscr{L})] \in [-1, 1]; \tag{5}$$

$$\min(\mathscr{L}) = m_1 m_2 k (k+1) (k+2) / 6;\ \max(\mathscr{L}) = m_1 m_2 k (k+1) (2k+1) / 6. \tag{6}$$

The two-group test was soon extended to multiple groups: Beckett & Schucany (1975, 1979) introduced ANACONDA, multi-group analysis of concordance, named on the basis of the similarity of the concept to ANOVA. However, the method found very little application. Recently, an attempt was made to revive and improve it (Vidmar & Cernigoj 2004), and we refer the interested reader to that paper (as well as the two original papers) for explanation and discussion of the ANACONDA procedure.

Legendre (2005) took the most modern approach towards concordance analysis, motivated by the problem of species association in ecology. He devised a permutation test of W that allows for *a posteriori* tests to determine which of the individual judges are concordant with one or several of the other judges. To preserve a (approximately) correct experiment-wise error rate, p-value adjustment using the Holm (1979) procedure is applied in these tests. The paper provides a clear and detailed explanation of the permutation testing procedures with an excellent illustrative example, and it is accompanied by publicly available software (from http://www.bio.umontreal.ca/casgrain/en/labo/kendall_w.html).

4 New Plots

We introduce two new plots. The concordance bubble-plot is based on raw rank data, while the pin-cushion plot depicts rank differences. Each plot is described in a separate section below.

4.1 Concordance Bubble-Plot

Analogously to the inter-rater variability plots of Nelson & Peppe (2000), concordance bubble-plots display the frequency of assigned ranks as a function of mean rank (i.e., object). They are therefore a particular type of scatter-plots with circle size used to code the number of identical points and object names used as labels for the mean ranks. Representations of different degrees of concordance with concordance bubble-plots are given in Figure 3.

Perfect concordance is represented with all the circles on the main diagonal (upper left panel), and the main diagonal is therefore drawn in each plot (dashed line) to facilitate judgment of discordance among judges. Without tied ranks, for many combinations of m and k the minimum possible concordance is greater than zero, so the mean ranks are not equal, but they are still very close to each other (upper right panel) compared to the situation of perfect concordance in which they are as far apart as possible. The plots in the bottom two panels depict the now classic concordance data of Schucany & Frawley (1973) on two groups of wine experts ranking four wines, whereby the nine American experts tend to agree more among themselves (W=0.60, bottom left panel) than the six French experts (W=0.37, bottom right panel). In general, concordance bubble-plots have a high data-ink ratio and provide crucial additional information for concordance data.

The issue of circle size deserves some additional consideration. Namely, had the area of the circles been made proportional to the number of rank-pairs, the radius for frequency f should have been square root of f times the radius for frequency 1, while with the radius proportional to f, one might dismiss the resulting graph as a typical example of the „lie factor“ (Tufte 1998). However, it is well known from the huge body of research in perception and psychophysics (c.f. Gescheider 1985) that for area as a visual stimulus, a concave increasing function is a good first approximation for the general relationship between actual and perceived magnitude (call it Fechner's Law, or Stevens' Law with power less than 1), so a circle with radius 2 is actually perceived as having not much more than twice the area of a circle with unit radius. Hence, without resorting to a complex psychophysical scaling study of questionable benefit, we produced the sample concordance bubble-plots with radius proportional to f, while it is up to the user to decide what radius scaling to apply in visualisation practice.

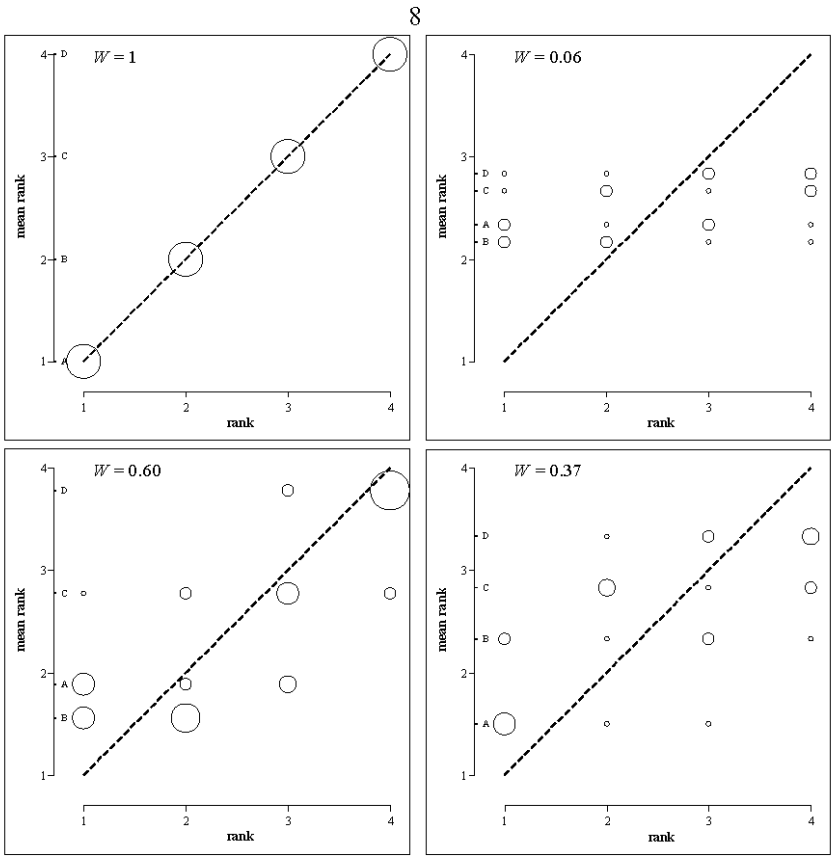


Figure 3: Different degrees of concordance (perfect, minimum possible, fair, low) depicted with concordance bubble-plots (k=*4* for all plots; *m*=9 with *W*=0.60, otherwise *m*=6); artificial data for top row, data from Schucany & Frawley (1973) for bottom row.

4.2 Pin-Cushion plot

Since each of the k objects is assigned $m\,(m–1)\,/\,2$ pairs of ranks, there are $k\,m\,(m–1)\,/\,2$ pairs of same-object ranks in a concordance dataset. Concordance plot can be constructed from the differences between rankings within each of those pairs. We propose representing them by angle, using vertical line as the basis representing no difference (i.e., a pair of identical ranks assigned to an object), and plotting rank difference clockwise (the bigger the difference, the larger the angle). Dependence on the order in which the sets of ranks are placed in the data matrix is avoided by plotting the absolute values of the differences. The maximum possible absolute difference, which is $k – 1$, is represented by the

horizontal axis. The angle from the horizontal axis (ϕ) representing a given absolute rank difference is computed as $\phi = 90° (1 - |d| / (k - 1))$.

If a difference is not present in the dataset, the corresponding line is not drawn; otherwise, the number of occurrences of a difference is represented by line length. Hence, larger concordance is represented by longer upright pins and shorter inclined pins with respect to smaller concordance. A base shape can provide space for the value of concordance coefficient, or a rectangle is drawn to anchor the perception of line length by means of its height matching the line of length one (i.e., $|d|=1$) and its width matching $|d|=2$ (as in Figure 4).

The resulting graph resembles a pin-cushion, as can be seen in Figure 4, which depicts experimental data from social psychology (Vidmar & Cernigoj 2004). Putting aside the original research design and hypothesis testing for brevity, one can assess the efficacy of concordance visualisation through visual impression by considering that the value of W for comparing conditions X and Y within group A is 0.75, while W for comparing groups X and Y within B is 0.56. In general, pin-cushion plot is compact and has a high data-ink ratio, so we consider it particularly suitable for visually comparing concordance between a relatively large number of groups of judges.

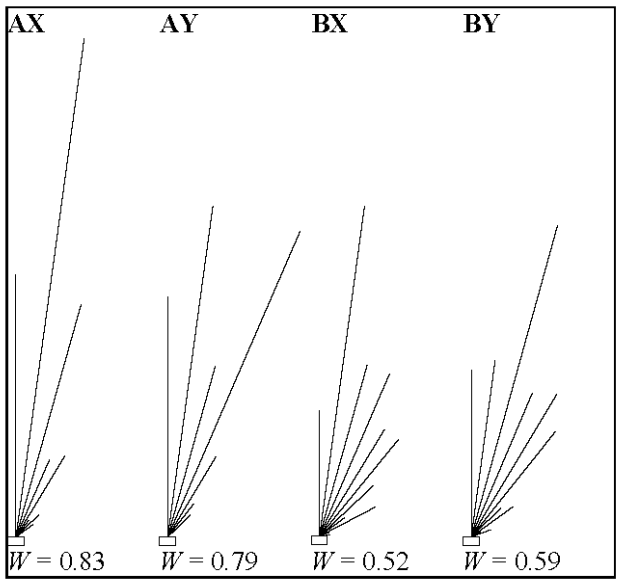


Figure 4: Pin-cushion plots of concordance data (k=12, m=5) from two-factor (groups A&B under conditions X&Y) experiment (Vidmar & Cernigoj 2004).

5 Discussion

Firstly, there are at least two further types of concordance plots that we can think of. They are not as successful as the presented two, but they deserve to be mentioned because of novelty and because they shed additional light on the qualities and downsides of concordance bubble-plot and pin-cushion plot.

- A straight-forward way to construct a concordance graph from the same-object rank pairs is to plot them, preferably in parallel coordinates. The key principle of such a plot would be that for each pair of ranks for a given object, the smaller of the two ranks (labelled R<) would be projected onto the first axis and the larger (R>) onto the second axis, with both axes oriented vertically. The number of identical {R<, R>} pairs could be represented by line thickness. We produced a number of such plots and found that perfect and minimum concordance form easily recognisable patterns, but the in-between levels of concordance are difficult to distinguish. Furthermore, the data-ink ratio is low, and special attention must be paid to axes in order to understand what is being plotted. On the positive side, such „concordance parallel-coordinates plots" have an aesthetic appeal, and they share the special properties of parallel coordinates.

- The shortest summary of the same-object rank pairs is the distribution of absolute differences ($|d|$) of rank pairs for any given object, which is naturally visualised by a bar-plot. To allow comparison between datasets with different k and/or m, it is preferable to plot relative frequencies. In general, the less concordance there is among the judges, the larger the centre of the distribution and the smaller its skewness. The major deficiency of such „concordance bar-plots" is that they – like angular correlation-matrix plots, concordance parallel-coordinates plots and pin-cushion plots – do not provide information on the mean ranks.

For completeness, we must also mention visualisation of frequency distributions of ranked data using permutation polytopes (Thompson 1994, Baggerly 1995). Though that field far exceeds the scope and level of our paper, we would just like to note that despite their theoretical and representational complexity, such visualisations completely represent concordance data only with three or four objects. Nevertheless, we refer the intrigued reader to the references, since it is a fascinating and far-reaching topic.

A possible critique of the presented plots is that area, direction, length and angle are the „elementary perceptual tasks" upon which they are based, and these do not rank high in the hierarchy of accuracy of observer judgments (Cleveland 1994, Cleveland & McGill 1984). However, it is an established fact that perception and therefore also comparison of attributes devoid of context is impossible (Lockhead

1992, 1995), and that understanding good scientific graphics requires time and attention (Tufte 1998). Hence, we designed concordance plots with the aim of presenting a clear and distinguishable pattern for a given degree of concordance, while also providing the necessary elements for subsequent detailed examination. Furthermore, at least in the concordance bubble-plot, the context elements (axes, labels for mean object ranks and diagonal line) actually transform the generally less accurate perceptual tasks into the generally more accurate one, namely judging position along a common scale.

A general concern with both proposed plots is that they become less clear with a large number of objects, but that is not a major limitation because more than a dozen objects or judges are seldom used in actual concordance studies. Considering concordance bubble-plots, had we used jitter instead of circle size to represent frequency, clutter would appear much sooner.

We designed all plots in black and white, since according to good graphical practice guidelines (e.g. Wainer & Thissen 1981), only information dimensions intended to be presented but left unpictured or indiscernible would have required the use of colour. As a final point, even though we have not depicted tied ranks, it is evident that they do not present a problem to the proposed plots.

6 Conclusion

Even though examples of highly contested and seldom used inventions from the history of data visualisation call for caution, we believe that concordance bubble-plot and pin-cushion plot can find application in the various fields where concordance is studied, ranging from ecology to marketing.

Their software implementation is simple – we produced Figure 3 with Microsoft® Excel, and Figure 4 with the freely available graphical package jsplot (http://ourworld.compuserve.com/homepages/jsieberer). The files are available for public download (from http://www.mf.uni-lj.si/ibmi-english/biostat-center, Software section). We are planning to implement concordance visualisation in R.

References

Baggerly, K.A. (1995), *Visual Estimation of Structure in Ranked Data*, PhD thesis, Rice University, Houston.

Beckett, J. & Schucany, W. R. (1975), ‚ANACONDA: Analysis of concordance of g groups of judges', *Proceedings of Social Statistics Section of the American Statistical Association*, 311-313.

Beckett, J. & Schucany, W. R. (1979), ‚Concordance among categorized groups of judges', *Journal of Educational Statistics* **4**(2), 125-137.

Cleveland, W. S. (1994), *The Elements of Graphing Data*, Hobart Press, Summit.

Cleveland, W. S. & McGill, R. (1984), ‚Graphical perception: Theory, experimentation, and application to the development of graphical methods', Journal of the American Statistical Association **79**(387), 531-554.

Corsten, L. C. A., & Gabriel, K. R. (1976), ‚Graphical exploration in comparing variance matrices', *Biometrics* **32**, 851-863.

Ehrenberg, A. S. C. (1952), ‚On sampling from a population of rankers', *Biometrika* **39**, 82-87.

Gescheider, G. A. (1985), *Psychophysics: Method, Theory, and Application*, 2nd ed., Lawrence Erlbaum, Hillsdale.

Holm, S. (1979), ‚A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**, 65-70.

Inselberg, A. (1985), ‚Plane with parallel coordinates', *Visual Computer* **1**, 69-97.

Iman, R. L. & Conover, W. J. (1987), ‚A measure of top-down correlation', *Technometrics* **29**(3), 351-357.

Kendall, M., & Babington Smith, B. (1939), ‚The problem of *m* rankings', *Annals of Mathematical Statistics* **10**, 275-287.

Legendre, P. & Lapointe, F-J. (2004), ‚Assessing congruence among distance matrices: Single-malt Scotch whiskies revised', *Australian & New Zealand Journal of Statistics* **46**(4), 615-629.

Legendre, P. (2005), ‚Species associations: The Kendall coefficient of concordance revisited', *Journal of Agricultural, Biological, and Environmental Statistics* **10**(2), 226-245.

Lockhead, G. R. (1992), ‚Psychophysical scaling: Judgment of attributes or objects?', *Behavioral and Brain Sciences* **15**, 543-558.

Lockhead, G. R. (1995), ‚Psychophysical scaling methods reveal and measure context effects', *Behavioral and Brain Sciences* **18**, 607-612.

Lyerly, S. B. (1952), ‚The average Spearman rank correlation coefficient', *Pychometrika* **17**, 421-428.

Nelson, J. C. & Pepe, M. S. (2000), ‚Statistical description of interrater variability in ordinal ratings', *Statistical Methods in Medical Research* **9**, 475-496.

Page, E. B. (1963), ‚Ordered hypotheses for multiple treatments: A significance test for linear ranks', Journal of the American Statistical Association **58**, 216-230.

Palachek, A. D. & Schucany, W. R. (1984), ‚On approximate confidence intervals for measures of concordance', *Psychometrika* **49**(1), 133-141.

R Development Core Team (2004), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3, http://www.r-project.org.

Schucany, W. R. & Frawley, W. H. (1973), ‚A rank test for two group concordance', *Psychometrika* **38**(2), 249-258.

Seber, G. A. F. (1984), *Multivariate Observations*, Wiley, New York.

Siegel, S. & Castellan, J. (1988), *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed., McGraw-Hill, New York.

Thompson, G. L. (1994), ‚Visualising frequency distributions of ranked data'. *Computational Statistics* **9**(1), 1-10.

Trosset, M. W. (2005), ‚Visualizing correlation', *Journal of Computational and Graphical Statistics* **14**(1), 1-19.

Tufte, E. R. (1998), *The Visual Display of Quantitative Information*, 16th printing, Graphics Press, Chesire.

Vidmar, G. & Cernigoj, M. (2004), ‚Studying norms in small groups by means of multi-group concordance analysis', *Horizons of Psychology* **13**(4), 55-66.

Wainer, H. & Thissen, D. (1981), ‚Graphical data analysis', *Annual Review of Psychology* **32**, 191-241.

Wilkie, D. (1980), ‚Pictorial representation of Kendall's rank correlation coefficient', *Teaching Statistics* **2**, 76-78.

PRILOGA 5: KODA ZA IZRIS KONKORDANČNIH MEHURČNIH DIAGRAMOV Z ELEKTRONSKO PREGLEDNICO MICROSOFT® EXCEL

```
' ConcordanceBubblePlot Macro
' Gaj Vidmar (IBMI, University of Ljubljana, Slovenia), 2006
' Shortcut: Ctrl+Shift+B
Option Explicit
Sub ConcordanceBubblePlot()
    Dim dataRange As Range
    Dim tt As Worksheet
    Dim meanRanks(1 To 100) As Variant, rankFreq(1 To 100, 1 To 100) As Variant
    Dim i, j, k, n As Integer
    Dim t As Double
    Dim ch As ChartObject
    If Selection.Areas.Count > 1 Then
        MsgBox "Non-contiguous area selected!", vbExclamation
        Exit Sub
    End If
    Set dataRange = Selection
    k = dataRange.Columns.Count()
    n = dataRange.Rows.Count() - 1
    If (n < 2) Or (k < 2) Then
        MsgBox "No or not enough data!", vbExclamation
        Exit Sub
    ElseIf (n > 99) Or (k > 99) Then
        MsgBox "Too much data!", vbExclamation
        Exit Sub
    End If
    On Error Resume Next
    Set tt = Sheets("ConcordanceBubblePlot")
    If tt Is Nothing Then GoTo nn
    Application.DisplayAlerts = False
    Sheets("ConcordanceBubblePlot").Delete
    Application.DisplayAlerts = True
nn: Sheets.Add after:=Sheets(Sheets.Count)
    With ActiveSheet
        .Name = "ConcordanceBubblePlot"
        .Cells.Font.Name = "Arial"
        .Cells.Font.Size = 8
        .Columns(3).ColumnWidth = 4
        .Columns(5).ColumnWidth = 4
        .Columns(6).ColumnWidth = 4
        With .PageSetup
            .TopMargin = Application.CentimetersToPoints(2)
            .BottomMargin = Application.CentimetersToPoints(2)
            .LeftMargin = Application.CentimetersToPoints(1)
            .RightMargin = Application.CentimetersToPoints(1)
            .HeaderMargin = Application.CentimetersToPoints(1)
            .FooterMargin = Application.CentimetersToPoints(1)
        End With
    End With
    Range("A1").Value = "W"
    Range("A1").Font.Bold = True
    Range("A1").Font.Italic = True
    Range("A1").HorizontalAlignment = xlRight
    Range("A3").Value = "rank"
    Range("B3").Value = "mean rank"
    Range("C3").Value = "f"
    Range("A3:C3").Font.Bold = True
    Range("A3:C3").HorizontalAlignment = xlRight
    t = 0
    For i = 1 To k
        meanRanks(i) = Application.WorksheetFunction.Average(Range(dataRange.Cells(2, i), _
          dataRange.Cells(n + 1, i)))
        t = t + ((meanRanks(i) - ((k + 1) / 2)) ^ 2)
        For j = 1 To k
            rankFreq(i, j) = Application.WorksheetFunction.CountIf(Range(dataRange.Cells(2, i), _
              dataRange.Cells(n + 1, i)), j)
        Next j
    Next i
    Range("B1").Value = t / ((k * ((k ^ 2) - 1)) / 12)
    Range("B1").NumberFormat = "0.00"
```

```
'main chart data
    i = 1
    t = 0
    Do
        j = 1
        Do
            If rankFreq(i, j) > 0 Then
                t = t + 1
                Cells(3 + t, 1).Value = j
                Cells(3 + t, 2).Value = meanRanks(i)
                Cells(3 + t, 3).Value = rankFreq(i, j)
            End If
            j = j + 1
        Loop Until j = k + 1
        i = i + 1
    Loop Until i = k + 1
    Range("E3").Value = "main"
    Range("E4").Value = "chart"
    Range("E5").Value = "data"
    Range("E3:E5").HorizontalAlignment = xlRight
    With Range(Cells(3, 5), Cells(3 + (k * k), 5)).Borders(xlEdgeTop)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(3, 5), Cells(3 + (k * k), 5)).Borders(xlEdgeRight)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(3, 5), Cells(3 + (k * k), 5)).Borders(xlEdgeBottom)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
'diag. line data
    Cells(5 + (k * k), 1).Value = 1
    Cells(5 + (k * k), 2).Value = 1
    Cells(6 + (k * k), 1).Value = k
    Cells(6 + (k * k), 2).Value = k
    Cells(5 + (k * k), 5).Value = "diag."
    Cells(6 + (k * k), 5).Value = "line"
    Range(Cells(5 + (k * k), 5), Cells(6 + (k * k), 5)).HorizontalAlignment = xlRight
    With Range(Cells(5 + (k * k), 5), Cells(6 + (k * k), 5)).Borders(xlEdgeTop)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(5 + (k * k), 5), Cells(6 + (k * k), 5)).Borders(xlEdgeRight)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(5 + (k * k), 5), Cells(6 + (k * k), 5)).Borders(xlEdgeBottom)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
'Y axis data
    For i = 1 To k
        Cells(7 + i + (k * k), 1) = 0.6
        Cells(7 + i + (k * k), 2) = i
        Cells(7 + i + (k * k), 3) = 0.1
        Cells(7 + i + (k * k), 4) = i
    Next i
    Cells(8 + (k * k), 5).Value = "Y"
    Cells(9 + (k * k), 5).Value = "axis"
    Range(Cells(8 + (k * k), 5), Cells(9 + (k * k), 5)).HorizontalAlignment = xlRight
    With Range(Cells(8 + (k * k), 5), Cells(7 + ((k + 1) * k), 5)).Borders(xlEdgeTop)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(8 + (k * k), 5), Cells(7 + ((k + 1) * k), 5)).Borders(xlEdgeRight)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(8 + (k * k), 5), Cells(7 + ((k + 1) * k), 5)).Borders(xlEdgeBottom)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
'X axis data
    For i = 1 To k
        Cells(8 + ((k + 1) * k) + i, 1) = i
        Cells(8 + ((k + 1) * k) + i, 2) = 0.75
        Cells(8 + ((k + 1) * k) + i, 3) = 0.1
        Cells(8 + ((k + 1) * k) + i, 4) = i
    Next i
    Cells(9 + ((k + 1) * k), 5).Value = "X"
    Cells(10 + ((k + 1) * k), 5).Value = "axis"
    Range(Cells(9 + ((k + 1) * k), 5), Cells(10 + ((k + 1) * k), 5)).HorizontalAlignment = xlRight
    With Range(Cells(9 + ((k + 1) * k), 5), Cells(8 + ((k + 2) * k), 5)).Borders(xlEdgeTop)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(9 + ((k + 1) * k), 5), Cells(8 + ((k + 2) * k), 5)).Borders(xlEdgeRight)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(9 + ((k + 1) * k), 5), Cells(8 + ((k + 2) * k), 5)).Borders(xlEdgeBottom)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
```

```
'mean ranks data
    For i = 1 To k
        Cells(9 + ((k + 2) * k) + i, 1) = 0.6
        Cells(9 + ((k + 2) * k) + i, 2) = meanRanks(i)
        Cells(9 + ((k + 2) * k) + i, 4) = dataRange.Cells(1, i)
    Next i
    Cells(10 + ((k + 2) * k), 5).Value = "mean"
    Cells(11 + ((k + 2) * k), 5).Value = "ranks"
    Range(Cells(10 + ((k + 2) * k), 5), Cells(11 + ((k + 2) * k), 5)).HorizontalAlignment = xlRight
    With Range(Cells(10 + ((k + 2) * k), 5), Cells(9 + ((k + 3) * k), 5)).Borders(xlEdgeTop)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(10 + ((k + 2) * k), 5), Cells(9 + ((k + 3) * k), 5)).Borders(xlEdgeRight)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(10 + ((k + 2) * k), 5), Cells(9 + ((k + 3) * k), 5)).Borders(xlEdgeBottom)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
'axis titles
    Cells(11 + ((k + 3) * k), 1).Value = (k + 1) / 2
    Cells(11 + ((k + 3) * k), 2).Value = 0.85
    Cells(12 + ((k + 3) * k), 1).Value = 0.8
    Cells(12 + ((k + 3) * k), 2).Value = k + 0.25
    Cells(11 + ((k + 3) * k), 4).Value = "rank"
    Cells(12 + ((k + 3) * k), 4).Value = "mean rank"
    Cells(11 + ((k + 3) * k), 5).Value = "axis"
    Cells(12 + ((k + 3) * k), 5).Value = "titles"
    Range(Cells(11 + ((k + 3) * k), 5), Cells(12 + ((k + 3) * k), 5)).HorizontalAlignment = xlRight
    With Range(Cells(11 + ((k + 3) * k), 5), Cells(12 + ((k + 3) * k), 5)).Borders(xlEdgeTop)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(11 + ((k + 3) * k), 5), Cells(12 + ((k + 3) * k), 5)).Borders(xlEdgeRight)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
    With Range(Cells(11 + ((k + 3) * k), 5), Cells(12 + ((k + 3) * k), 5)).Borders(xlEdgeBottom)
        .LineStyle = xlContinuous
        .Weight = xlThin
    End With
'chart drawing
    Set ch = Worksheets("ConcordanceBubblePlot").ChartObjects.Add(225, 25, 275, 275)
    ch.Chart.ChartType = xlXYScatter
    ch.Chart.SeriesCollection.Add Source:=Range(Cells(4, 2), Cells(3 + t, 2)), Rowcol:=xlColumns
    ch.Chart.SeriesCollection.Add Source:=Range(Cells(5 + (k * k), 1), _
      Cells(6 + (k * k), 2)), Rowcol:=xlColumns
    ch.Chart.SeriesCollection.Add Source:=Range(Cells(8 + (k * k), 1), _
      Cells(7 + ((k + 1) * k), 2)), Rowcol:=xlColumns
    ch.Chart.SeriesCollection.Add Source:=Range(Cells(9 + ((k + 1) * k), 1), _
      Cells(8 + ((k + 2) * k), 2)), Rowcol:=xlColumns
    ch.Chart.SeriesCollection.Add Source:=Range(Cells(10 + ((k + 2) * k), 1), _
      Cells(9 + ((k + 3) * k), 2)), Rowcol:=xlColumns
    ch.Chart.SeriesCollection.Add Source:=Range(Cells(11 + ((k + 3) * k), 1), _
      Cells(12 + ((k + 3) * k), 2)), Rowcol:=xlColumns
    ch.Chart.HasLegend = False
    ch.Chart.PlotArea.Interior.ColorIndex = xlNone
    ch.Chart.PlotArea.Border.LineStyle = xlNone
    With ch.Chart.Axes(xlCategory)
        .MinimumScale = 0.5
        .MaximumScale = k + 0.25
        .MajorUnit = 1
        .CrossesAt = 0.5
        .HasMajorGridlines = False
        .MajorTickMark = xlTickMarkNone
        .TickLabelPosition = xlTickLabelPositionNone
        .Border.LineStyle = xlNone
    End With
    With ch.Chart.Axes(xlValue)
        .MinimumScale = 0.5
        .MaximumScale = k + 0.25
        .MajorUnit = 1
        .CrossesAt = 0.5
        .HasMajorGridlines = False
        .MajorTickMark = xlTickMarkNone
        .TickLabelPosition = xlTickLabelPositionNone
        .Border.LineStyle = xlNone
    End With
    With ch.Chart.SeriesCollection(1)
        .XValues = Range(Cells(4, 1), Cells(3 + t, 1))
        .Border.LineStyle = xlNone
        .MarkerStyle = xlMarkerStyleCircle
        .MarkerSize = 5
        .MarkerForegroundColor = vbBlack
        .MarkerBackgroundColorIndex = xlNone
    End With
    For i = 1 To t
        ch.Chart.SeriesCollection(1).Points(i).MarkerSize = Cells(3 + i, 3).Value * 5
    Next
    With ch.Chart.SeriesCollection(2)
        .Border.LineStyle = xlDot
        .MarkerStyle = xlNone
        .Border.Weight = xlThin
        .Border.Color = 1
    End With
```

```
    With ch.Chart.SeriesCollection(3)
        .Border.LineStyle = xlContinuous
        .MarkerStyle = xlNone
        .Border.Weight = xlThin
        .Border.Color = 1
        .ErrorBar Direction:=xlX, Include:=xlErrorBarIncludeMinusValues, Type:=xlErrorBarTypeCustom, _
            MinusValues:=Range(Cells(8 + (k * k), 3), Cells(7 + ((k + 1) * k), 3))
        .ErrorBars.EndStyle = xlNoCap
        .HasDataLabels = True
        .ApplyDataLabels Type:=xlDataLabelsShowValue
        With .DataLabels
            .Font.Name = "Arial"
            .Font.Size = 9
            .AutoScaleFont = False
            .Position = xlLabelPositionLeft
        End With
    End With
    With ch.Chart.SeriesCollection(4)
        .Border.LineStyle = xlContinuous
        .MarkerStyle = xlNone
        .Border.Weight = xlThin
        .Border.Color = 1
        .ErrorBar Direction:=xlY, Include:=xlErrorBarIncludeMinusValues, Type:=xlErrorBarTypeCustom, _
            MinusValues:=Range(Cells(9 + ((k + 1) * k), 3), Cells(8 + ((k + 2) * k), 3))
        .ErrorBars.EndStyle = xlNoCap
        .HasDataLabels = True
        .ApplyDataLabels Type:=xlDataLabelsShowLabel
        With .DataLabels
            .Font.Name = "Arial"
            .Font.Size = 9
            .AutoScaleFont = False
            .Position = xlLabelPositionBelow
        End With
    End With
    With ch.Chart.SeriesCollection(5)
        .Border.LineStyle = xlNone
        .MarkerStyle = xlMarkerStyleDot
        .MarkerSize = 4
        .MarkerForegroundColor = vbBlack
        .MarkerBackgroundColorIndex = xlNone
        For i = 1 To k
            .Points(i).HasDataLabel = True
            .Points(i).DataLabel.Text = Cells(9 + ((k + 2) * k) + i, 4).Value
        Next
        With .DataLabels
            .Font.Name = "Arial"
            .Font.Size = 7
            .AutoScaleFont = False
            .Position = xlLabelPositionRight
        End With
    End With
    With ch.Chart.SeriesCollection(6)
        .Border.LineStyle = xlNone
        .MarkerStyle = xlNone
        .Points(1).HasDataLabel = True
        .Points(1).DataLabel.Text = Cells(11 + ((k + 3) * k), 4).Value
        .Points(2).HasDataLabel = True
        .Points(2).DataLabel.Text = Cells(12 + ((k + 3) * k), 4).Value
        With .DataLabels
            .Font.Name = "Arial"
            .Font.Size = 8
            .Font.Bold = True
            .AutoScaleFont = False
            .Position = xlLabelPositionCenter
        End With
    End With
    Range("G1").Select
End Sub
```

PRILOGA 6: KODA ZA IZRIS KONKORDANČNIH DIAGRAMOV

BLAZINICE Z BUCIKAMI S PROGRAMSKIM PAKETOM JSPLOT

Interaktivni način:

```
# concordance pin-cushion plot, Gaj Vidmar, 2005 #
Page(200.0,200.0);
pi :=  3.14159265359;

# get data #
  s := false;
  while not(s) do
    begin
      e := [["File", "file", ["","*.dat"]],
           ["Judges (m)", "edit", "", 4],
           ["Objects (k)", "edit", "", 4]];
      s := MultiPromptDialog(e, r, :title "Concordance pin-cushion plot")
           and (length(r[2])>0) and (length(r[3])>0);
    end;
  if not(access(r[1],"r")) then
    begin
      error("Could not open file!");
      exit(1);
    end;
  m := atoi(r[2]);
  k := atoi(r[3]);
  n := m*(m-1)*k/2;
  data := new array(m);
  for i := 1 to m do
    data[i] := new array(k);
  f := fopen(r[1],"r");
  for i := 1 to m do
    for j := 1 to k do
      begin
        s := fgeti(f, data[i][j]);
        if not(s) then
          begin
            error("Error while reading file!");
            fclose(f);
            exit(1);
          end;
      end;
  fclose(f);

# rank differences #
  dif := new array(k);
  for i := 1 to k do
    dif[i] := 0;
  for i := 1 to m-1 do
    for j := i+1 to m do
      for h := 1 to k do
       begin
        x := abs(data[i][h] - data[j][h]);
        dif[x+1] := dif[x+1] + 1;
       end;

# transformation matrix, the constant part #
  trmat := new array(9);
  trmat [3] := 25.0;
  trmat [6] := 25.0;
  trmat [7] := 0.0;
  trmat [8] := 0.0;
  trmat [9] := 1.0;

# drawing #
  for i := 1 to k do
    if (dif[i]>0) then
     begin
      j := i - 1;
      trmat[1] := cos(-j*pi/(2*k-1));
      trmat[4] := sin(-j*pi/(2*k-1));
      trmat[5] := trmat[1];
      trmat[2] := -trmat[4];
      Transform(trmat);
        Line(0.0, 0.0, 0.0, 150.0*dif[i]/n);
      EndTransform();
     end;
  Rectangle(25.0-150/n, 25.0, 25.0+150/n, 25.0-150/n, :fillcolor "White");

EndPage();
```

Samodejni način:

```
...

# get data #
  f := fopen("cbat.dat","r");
  s := fgeti(f, m);
  s := fgeti(f, k);
  n := m*(m-1)*k/2;
  data := new array(m);
  for i := 1 to m do
    data[i] := new array(k);
  for i := 1 to m do
    for j := 1 to k do
      s := fgeti(f, data[i][j]);
  fclose(f);

...
```