UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

ROK PLATINOVŠEK

METODE ZLIVANJA PODATKOV

DIPLOMSKO DELO

Ljubljana, 2008

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

ROK PLATINOVŠEK

MENTOR: red. prof. dr. Vasja Vehovar

METODE ZLIVANJA PODATKOV

DIPLOMSKO DELO

Ljubljana, 2008

Avtor bi se rad zahvalil mentorju za pomoč in nasvete,
zahvala pa gre tudi podjetju TNS Radar na Finskem,
ki je avtorja v času od oktobra 2007 do julija 2008
gostilo na študentski praksi.

Metode zlivanja podatkov

V pričujočem besedilu predstavljamo metode zlivanja podatkov, pri čemer poudarjamo razlike med parametričnimi in neparametričnimi metodami. Temeljne koncepte uvedemo z obravnavo metode najbližjega soseda, nato pa razpravljamo o nedoločenosti problema zlivanja podatkov ter predpostavki pogojne neodvisnosti, ki jo ob uporabi metode najbližjega soseda naredimo implicitno.

Preden predstavimo sodobne metode osvetlimo širše teoretično ozadje s poudarkom na Bayesovem pristopu k statističnemu sklepanju. Zlivanje podatkov nato redefiniramo kot problem vstavljanja manjkajočih vrednosti. Omenimo EM algoritem, ki v primeru zlivanja podatkov zaradi nedoločenosti ne pripelje do ene same rešitve. Le-to dosežemo šele z uporabo algoritma plemenitenja podatkov, katerega delovanje je mogoče stabilizirati z uporabo informativne apriorne porazdelitve.

Parametrične in neparametrične metode apliciramo na podatke, pridobljene z anketo RIS-IKT 2005, ter poudarimo njihove prednosti in pomankljivosti: v večini primerov so bolj primerne parametrične metode, ki jih odlikuje učinkovitost pri uporabi dodatnih informacij ter dejstvo, da zaradi večkratnega vstavljanja ponazarjajo negotovost pri sklepanju o korelacijah med spremenljivkami, ki niso nikoli opazovane skupaj.

Ključne besede: zlivanje podatkov, metoda najbližjega soseda, predpostavka pogojne neodvisnosti, vstavljanje manjkajočih vrednosti, bayesiansko sklepanje

Methods of data fusion

The present text introduces methods of data fusion and stresses the differences between parametric and nonparametric methods. Fundamental concepts of data fusion are introduced through the nearest neighbour method. The discussion is then turned to the identification problem and the assumption of conditional independence that is made implicitly when applying the nearest neighbour method.

A wider theoretical background with an emphasis on Bayesian inference is presented before introducing the modern methods of data fusion. Data fusion is then redefined as a problem of missing data imputation. A mention is made of the EM algorithm that fails to result in a unique solution. A unique solution, however, is reached when running the data augmentation algorithm which can be stabilized by applying an informative prior.

Parametric and nonparametric methods are applied to the RIS-IKT 2005 survey dataset. Their advantages and disadvantages are then emphasized: parametric methods are found to be more advantageous in most cases as they are well adept to handling additional information and are able to represent the uncertainty concerning inference on the variables never jointly observed through multiple imputations.

Keywords: data fusion, nearest neighbour method, conditional independence assumption, missing data imputation, Bayesian inference

Kazalo

Kazalo slik

Kazalo tabel

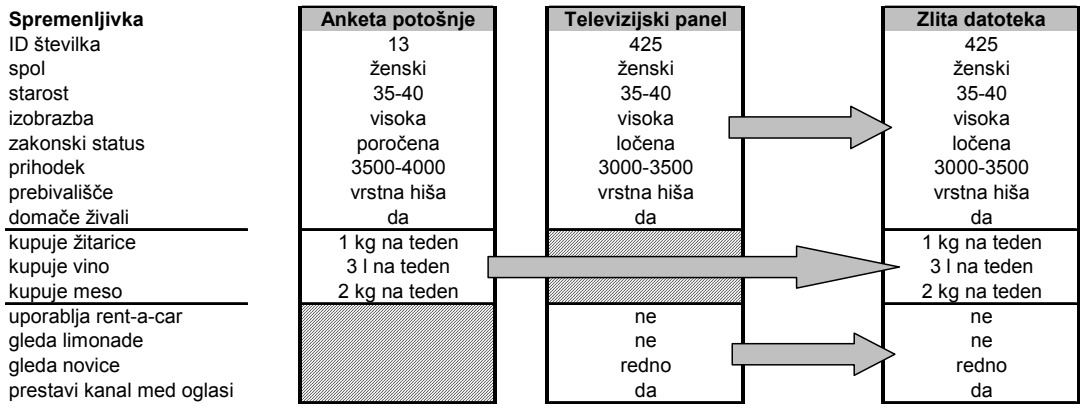
1 Uvod

V pričujočem besedilu bomo z izrazom *zlivanje podatkov* imenovali postopke, ki imajo cilj združiti informacijo, pridobljeno z različnimi neodvisnimi vzorčnimi anketami. V angleškem jeziku se za zlivanje podatkov uporabljata izraza *data fusion* in *statistical matching*, prvi pogosteje v Evropi, drugi pa v ZDA. Zlivanje podatkov je relativno novo področje raziskovanja, ki zaradi naraščajoče količine podatkov, ki je dandanes na voljo, pridobiva vse več pozornosti (D'Orazio *in drugi* 2006: ix).

Interes, da bi skozi medije učinkoviteje dosegli potrošnike na podlagi poznavanja njihovega potrošniškega vedenja, je imel pomembno vlogo za razvoj področja zlivanja podatkov. Idealno sredstvo za tovrstno proučevanje je anketa, ki obsega podrobna vprašanja tako glede posameznikovega potrošniškega vedenja, kot tudi vprašanja, ki zadevajo njegovo spremljanje medijskih vsebin. Ker pa je izvedba tako obsežne ankete povezana z visokimi stroški, predstavlja vabljivo alternativo uporaba postopkov zlivanja podatkov, s katerimi lahko podatke TV panela združimo z že obstoječo anketo potrošnje (Rässler 2002: 2).

Da bi bili kar se da konkretni, razvijmo omenjeni primer naprej. Denimo, da imamo anketo potrošnje ter televizijski panel s spremenljivkami, ki so prikazane na spodnji sliki. Pomembno je opažanje, da obe datoteki vsebujeta po nekaj *specifičnih* spremenljivk, osem demografskih spremenljivk pa je *skupnih* obema datotekama.

Slika 1.1: Poenostavljena ilustracija zlivanja podatkov.



(vir: Rässler 2002: 3)

Pri tipičnem postopku zlivanja podatkov dodelimo eni od datotek vlogo *darovalca* (*angl*. donor), drugi pa *prejemnika* (*angl*. recipient) nato pa vsaki enoti iz prejemne datoteke poiščemo primerno darovalno enoto iz darovalne datoteke. Pri iskanju

najprimernejše darovalne enote navadno uporabimo določeno mero podobnosti: poiščemo tisto enoto iz darovalne datoteke (denimo, da je v tem primeru to anketa potrošnje), ki je po skupnih spremenljivkah najbolj podobna prejemni enoti (iz televizijskega panela). Podatke o potrošnji nato pripojimo prejemni enoti. Ko vsaki enoti iz prejemne datoteke najdemo primerno darovalno enoto, dobimo novo, popolnjeno *zlito* datoteko (Rässler 2002: 2-3).

V nadaljevanju bomo podrobneje predstavili omenjeno ter njej sorodne metode, ki jih bomo imenovali *neparametrične*, nato pa pozornost namenili še sodobnejšim alternativam zlivanja podatkov.

1.1 Motivacija

Cilj pričujočega diplomskega dela je poljudno in razumljivo predstaviti osnovne koncepte zlivanja podatkov. V končnem delu besedila želimo izpostaviti kontrast med neparametričnimi in parametričnimi metodami zlivanja podatkov ter poudariti prednosti in slabosti obeh.

Avtorju je bila tema zlivanja podatkov dodeljena s strani podjetja, kjer je opravljal študentsko prakso. Začetno seznanjanje s tematiko je bilo dokaj težavno, saj tipično besedilo s tega področja predvideva, da bralec tematiko zlivanja podatkov že obvlada ter da je dobro seznanjen z določenimi statističnimi koncepti ter metodami, predvsem z bayesianskim pristopom k statističnemu sklepanju ter ocenjevanju parametrov s simulacijo. Omenjena besedila so navadno abstraktne in tehnične narave, kar nepoučenega bralca oddalji od tematike ter mu pusti občutek, da je zlivanje podatkov neosvojljivo področje na katerem strokovnjaki pišejo za druge strokovnjake.

Pričujoče besedilo se želi oddaljiti od omenjenega načina podajanja znanja. Področje zlivanja podatkov bomo vsekakor skušali predstaviti korektno. Ker pa je tematika zelo abstraktna, bomo uporabljali mnogo konkretnih primerov, matematične izraze pa bomo skušali čimbolj razumljivo obrazložiti, pa tudi izpustiti, kjer bomo mnenja, da niso potrebni. Privoščili si bomo tudi daljši odmik od ožjega področja zlivanja podatkov ter predstavili širšo teoretično podlago, ki je potrebna za razumevanje parametričnih metod zlivanja podatkov.

Od tipičnega besedila s področja zlivanja podatkov se bomo skušali oddaljiti tudi v osrednjem empiričnem delu. V literaturi se za ilustracijo predstavljene metode pogosto

uporabi nerealen simuliran primer, npr. tisoč enot iz večrazsežnostne normalne porazdelitve. Poleg neživljenjskosti ter abstraktnosti – govora je npr. o spremenljivkah x_1, x_2 itd. – smo mnenja, da z uporabo takšnega primera avtor ne more dobro predstaviti *pomanjkljivosti* predstavljenih metod zlivanja podatkov, saj jih uporabimo za zlivanje podatkov, na kakršnih se obnesejo najbolje. Namesto tega bomo v empiričnem delu uporabili podatke, ki izhajajo iz dejanske ankete ter reševali s tem povezane težave.

1.2 Struktura naloge

V drugem poglavju uvedemo oznake in terminologijo zlivanja podatkov. Predstavimo naključno vstavljanje obstoječih vrednosti ter metodo najbližjega soseda, nato pa razpravljamo o problemu nedoločenosti ter predpostavki pogojne neodvisnosti. Poglavje zaključimo s kratkim pregledom zgodovine zlivanja podatkov.

V četrtem poglavju obravnavamo parametrične metode zlivanja podatkov, ki so konceptualno dokaj kompleksne, zato tretje poglavje namenimo predstavitvi teorije, potrebne za njihovo razumevanje: Bayesovemu pristopu k statističnemu sklepanju. Preden obravnavamo Bayesov pristop, pa v prvem delu poglavja najprej ponovimo osnovne statistične koncepte ter obravnavamo porazdelitve več spremenljivk nato pa opišemo ocenjevanje parametrov po metodi največjega verjetja, ki predstavlja protipol Bayesovemu pristopu.

V četrtem poglavju zlivanje podatkov prevedemo na problem vstavljanja manjkajočih vrednosti ter obravnavamo zanemarljivost mehanizma manjkajočih vrednosti. Opišemo idejo in delovanje EM algoritma, za katerega pa se izkaže, da za zlivanje podatkov ni uporaben. Bolj uspešna je njegova bayesianska prilagoditev, algoritem plemenitenja podatkov, ki skozi večkratno vstavljanje ponazarja negotovost pri sklepanju o manjkajočih vrednostih. V zaključnem delu poglavja opišemo implementacijo plemenitenja podatkov v konkretnem programskem paketu ter omenimo stopnje veljavnosti, ki jih pri zlivanju podatkov lahko dosežemo.

V petem poglavju vse opisane metode preizkusimo na konkretnih podatkih. Začnemo z opisom testnih podatkov ter statistik, nato pa predstavimo rezultate vsake od metod. V zaključnem poglavju rezultate ovrednotimo, izpostavimo prednosti in slabosti metod, ter razmišljamo o praktičnih rabah opisanih postopkov zlivanja podatkov.

2 Neparametrične metode zlivanja podatkov

V tem poglavju bomo predstavili osnovne koncepte zlivanja podatkov ter vpeljali potrebno terminologijo in oznake. Mnogo besed bomo namenili problemu nedoločenosti in predpostavki pogojne neodvisnosti, ki dajeta področju zlivanja podatkov poseben pečat. V zadnjem razdelku bomo kratko orisali zgodovino razvoja postopkov zlivanja podatkov.

2.1 Formalizacija problema ter terminologija

Za problem zlivanja podatkov sta značilna dva anketna vzorca, ki ju želimo zliti v enotno datoteko. V nadaljevanju besedila bomo vseskozi predpostavljali, da imamo opraviti z dvema *neodvisnima* vzorcema, ki sta bila naključno izbrana iz iste populacije (D'Orazio in drugi 2006: 3-4). Predpostavili bomo tudi, da je število posameznikov, ki so sodelovali v obeh anketah, zanemarljivo, ali pa takšni posamezniki sploh ne obstajajo (Rässler 2002: 3).

Datoteki, imenujmo ju A in B, vsebujeta nekaj *skupnih spremenljivk* (*angl*. common variables), nekaj spremenljivk pa je specifičnih. Predpostavili bomo, da je pretežen del spremenljivk številskega tipa[1]. Z oznako *X* bomo označevali spremenljivke, ki jih vsebujeta obe datoteki; z *Y* specifične spremenljivke, ki jih vsebuje zgolj datoteka A; z *Z* pa spremenljivke, specifične datoteki B. Imamo torej razmere, kot jih prikazuje Tabela 2.1:

Tabela 2.1: Vzorčni podatki pri problemu zlivanja podatkov. Osvetljene celice označujejo neopazovane spremenljivke.

	X_1	X_2	...	X_P	Y_1	Y_2	...	Y_Q	Z_1	Z_2	...	Z_R
datoteka A	x_{11}	x_{12}	...	x_{1P}	y_{11}	y_{12}	...	y_{1Q}				
	x_{21}	x_{22}	...	x_{2P}	y_{21}	y_{22}	...	y_{2Q}				
	...	...		...	...	...		...				
	x_{n1}	x_{n2}	...	x_{nP}	y_{n1}	y_{n2}	...	y_{nQ}				
datoteka B	x_{11}	x_{12}	...	x_{1P}					y_{11}	y_{12}	...	y_{1R}
	x_{21}	x_{22}	...	x_{2P}					y_{21}	y_{22}	...	y_{2R}
	...	...		...					...	...		...
	x_{m1}	x_{m2}	...	x_{mP}					y_{n1}	y_{n2}	...	y_{nR}

(vir: D'Orazio in drugi 2006: 5)

[1] Za metode zlivanja podatkov s kategoričnimi spremenljivkami glej npr. D'Orazio in drugi 2004; Kamakura in Wedel 1997 ali Singh in drugi 1988.

Metode, ki jih predstavljamo v tem poglavju, pristopajo k problemu zlivanja podatkov tako, da praznine v zgornji tabeli zapolnijo z razpoložljivimi vrednostmi drugih enot. Eni od datotek dodelimo vlogo *darovalca*, drugi pa *prejemnika*, nato pa vsaki enoti iz prejemne datoteke najdemo darovalca ter neopazovanim spremenljivkam pripišemo vrednosti ustreznih spremenljivk darovalne enote.

Navadno dodelimo vlogo darovalca datoteki z več enotami. Če bi namreč imela darovalna datoteka občutno manj enot kot prejemna, bi v zliti datoteki iste darovalne enote nastopale večkrat, kar bi umetno zmanjšalo variabilnost zlitih spremenljivk (za daljšo razpravo izbiri darovalca in prejemnika glej D'Orazio in drugi 2006: 35-36).

Za omenjene postopke se je ustalila raba angleškega imena »*hot deck*« (HD), za kar bomo uporabljali slovenski izraz *vstavljanje obstoječih vrednosti*. V nadaljevanju bomo predstavili dve različici tega postopka: naključno vstavljanje obstoječih vrednosti ter metodo najbližjega soseda.

2.1.1 Naključno vstavljanje obstoječih vrednosti

Kot pove ime, gre pri postopku naključnega vstavljanja obstoječih vrednosti za *naključno* izbiro darovalca vsaki prejemni enoti. Včasih se naključna izbira opravi znotraj primerne podskupine: enote obeh datotek uvrstimo v homogene skupine glede na vrednost določene skupne spremenljivke, npr. spola. V tem primeru dobimo dva darovalna razreda (*angl*. donation class): za prejemnika moškega spola so tako primerni darovalci le enote moškega spola in obratno.

Metodo naključnega HD ilustrirajmo na naslednjemu izmišljenemu primeru. Denimo, da imamo datoteko A s specifično spremenljivko, ki opisuje pogostost gledanja televizije, ter datoteko B s specifično spremenljivko o pogostosti nakupovanja.

Tabela 2.2: Seznama enot v datotekah A in B.

datoteka A			
IDA	spol	starost	TV
1	M	18	5
2	M	36	3
3	M	51	5
4	Ž	20	3
5	Ž	36	2
6	Ž	49	4

datoteka B			
IDB	spol	starost	nakupi
1	M	19	4
2	M	28	1
3	M	41	3
4	M	51	2
5	Ž	18	2
6	Ž	27	4
7	Ž	40	4
8	Ž	54	5

Ker je večja, bomo datoteki B dodelili vlogo darovalca. Če zlivanje podatkov opravimo v darovalnih razredih, določenih s *spolom*, je ena od možnih realizacij metode naključnega vstavljanja obstoječih vrednosti naslednja:

Tabela 2.3: Datoteka, zlita po metodi naključnega HD z darovalnima razredoma.

IDA	IDB	spol	starost	TV	nakupi*
1	3	M	18	5	3
2	2	M	36	3	1
3	2	M	51	5	1
4	4	Ž	20	3	2
5	6	Ž	36	2	4
6	7	Ž	49	4	4

Vsaki enoti datoteke A smo torej poiskali naključnega darovalca istega spola, ter na ta način v zliti datoteki ustvarili zlito spremenljivko *nakupi**. Opozorimo, da lahko določena enota tudi večkrat nastopi kot darovalec.

2.1.2 Metoda najbližjega soseda

Navadno imamo na voljo mnogo spremenljivk, ki so skupne obema datotekama. V tem primeru pogosto uporabimo metodo najbližjega soseda, po kateri je najprimernejši darovalec za enoto *a* enota, ki ji je po skupnih spremenljivkah *najbliže*. Opredeliti moramo torej *mero razdalje*, s katero lahko izračunamo razdaljo med poljubnima enotama. Najpogosteje uporabljene mere razdalje so Evklidska, Mahalanobisova (glej De Maesschalk in drugi 2000) in razdalja Manhattan (za alternativne mere razdalje glej Abdi 2007). Če se zgodi, da je več potencialnih darovalcev enako oddaljenih od prejemne enote, eno izmed njih izberemo naključno (D'Orazio in drugi 2006: 41).

Tudi pri uporabi metode najbližjega soseda lahko opredelimo darovalne razrede, če želimo, da se darovalec s prejemnikom ujema v neki kritični spremenljivki. Tako lahko npr. določimo, da mora biti darovalec vedno istega spola kot prejemnik, znotraj s spolom opredeljenih darovalnih razredov pa poiščemo najbližjega soseda. Če ta postopek uporabimo na primeru iz prejšnjega razdelka, dobimo naslednjo zlito datoteko:

Tabela 2.4: Datoteka, zlita po metodi najbližjega soseda z darovalnima razredoma.

IDA	IDB	spol	starost	TV	nakupi*
1	1	M	18	5	4
2	3	M	36	3	3
3	4	M	51	5	2
4	5	Ž	20	3	2
5	6	Ž	36	2	4
6	8	Ž	49	4	5

Ideja, ki botruje uporabi metode najbližjega soseda je, da če sta si prejemna in darovalna enota podobni v skupnih spremenljivkah, je verjetno, da sta si podobni tudi v specifičnih spremenljivkah. Seveda se izkaže, da temu ni vedno tako: nakazani problem obravnavamo v naslednjem razdelku.

2.2 Problem nedoločenosti in predpostavka pogojne neodvisnosti

Problematika, ki jo bomo obravnavali v tem razdelku, zaznamuje celotno področje zlivanja podatkov, zato bomo razpravi namenili nekoliko več besed. Predpostavko pogojne neodvisnosti, o kateri bo govora, je težko razložiti brez spuščanja v tehnične podrobnosti, zato bomo že takoj na začetku povedali glavno ugotovitev. Za ilustracijo bomo nato navedli konkreten primer, na koncu pa bomo predpostavko pogojne neodvisnosti skušali še intuitivno razložiti ter razmišljali o njenih posledicah.

S področjem zlivanja podatkov je nerazdružljivo povezan problem nedoločenosti asociacije[2] spremenljivk, ki niso nikoli opazovane skupaj (*angl*. variables never jointly observed). Pogojne asociacije spremenljivk, ki nista nikoli opazovani skupaj, ni moč oceniti iz podatkov (Rässler 2004: 3; za dokaz glej Rubin 1974). Direktna posledica tega dejstva je ugotovitev, da preproste metode zlivanja podatkov v zliti datoteki vedno *vzpostavijo pogojno neodvisnost* spremenljivk, ki niso nikoli opazovane skupaj, pri danih skupnih spremenljivkah (za empirično študijo glej Rässler in Fleischer 1998).

Če se zaradi enostavnosti omejimo na linearno povezanost spremenljivk in se vrnemo na primer, to pomeni, da bo v zliti datoteki *parcialni*[3] korelacijski koeficient med

[2] Izraz *asociacija* je nadpomenka izraza *korelacija*. Korelacija se nanša na *linearno* povezanost spremenljivk, asociacija pa na kakršnokoli (linearno ali nelinearno) povezanost.

[3] Parcialni korelacijski koeficient opisuje korelacijo med spremenljivkama, »ki ostane« po tem, ko njuno korelacijo kontroliramo po skupnih spremenljivkah.

spremenljivkama *TV* in *nakupi* pri danih skupnih spremenljivkah vedno enak nič ne glede na to, katere skupne spremenljivke smo uporabili za zlivanje podatkov. Ker med navadnim (nekontroliranim) korelacijskim koeficientom ter parcialnim korelacijskim koeficientom velja naslednja zveza:

$$r_{yz\cdot x} = \frac{r_{yz} - r_{yx} \cdot r_{zx}}{\sqrt{(1-r_{yx})^2} \cdot \sqrt{(1-r_{zx})^2}}, \tag{1.1}$$

pa to pomeni, da med spremenljivkama *TV* in *nakupi* ustvarimo različne vrednosti (nekontrolirane) korelacije, kadar uporabimo različne skupne spremenljivke. Korelacija *TV-nakupi* lahko namreč zavzame različne vrednosti, ker je podatki ne določajo, oz. ne izražajo nobene informacije o njeni vrednosti.

Omenjena korelacija pa vseeno ne more zavzeti popolnoma poljubne vrednosti z intervala [-1, 1]. Omejujejo jo namreč korelacije s skupnimi spremenljivkami, saj mora biti korelacijska matrika vseh spremenljivk (t.j. vseh skupnih in vseh specifičnih) *pozitivno semi-definitna*, kar pomeni, da mora biti njena determinanta večja ali enaka nič (Rässler 2004: 3).

Denimo, da znaša korelacija med spremenljivkama *starost* in *TV* 0,9, korelacija *starost-nakupi* pa 0,8[4]. Determinanta korelacijske matrike

$$\operatorname{cov}(x, y, z) = \begin{bmatrix} 1 & 0,9 & 0,8 \\ 0,9 & 1 & r_{yz} \\ 0,8 & r_{yz} & 1 \end{bmatrix}$$

mora biti nenegativna, kar nas pripelje do naslednjega izraza:

$$\det(\operatorname{cov}(x, y, z)) = -r_{yz}^2 + 2 \cdot 0,72 \cdot r_{yz} - 0,45 .$$

Korena kvadratne enačbe izražata meji intervala dopustnih vrednosti korelacije *TV-nakupi:* ta korelacija lahko torej zavzame vrednosti z intervala [0,46; 0,98]. Splošna formula za interval dopustnih vrednosti pri eni skupni spremenljivki *x* je:

$$r_{yz} \in r_{xy} \cdot r_{xz} \pm \sqrt{1 + r_{xy}^2 \cdot r_{xz}^2 - r_{xy}^2 - r_{xz}^2}$$

[4] Opomnimo, da je tako visoke korelacije med spremenljivkami v družboslovju skuraj nemogoče opaziti. Primer, ki smo ga povzeli po Rässler (2004: 4), navajamo zgolj ilustrativno.

Korelacija $r_{xy} \cdot r_{xz}$, ki leži na sredini dopustnega intervala, ima naslednji pomen: to je korelacija, ki jo dobimo z zlivanjem podatkov po spremenljivki *x*, saj je pri tej vrednosti parcialni korelacijski koeficient enak nič[5]. To vidimo tako, da levo stran enačbe (1.1) izenačimo z nič ter izpeljemo r_{yz}. Z zlivanjem podatkov po spremenljivki *starost* bi dobili v zliti datoteki vrednost korelacije *TV-nakupi* 0,72.

Denimo, da imamo na voljo še eno skupno spremenljivko, *dohodek*, katere korelacija s spremenljivkama *TV* in *nakupi* je 0,5 oz. 0,8. Ti korelaciji do manjše mere omejujeta korelacijo *TV-nakupi*, ki tokrat lahko zavzame vrednosti od 0,13 do 0,67. Če bi torej kot skupno spremenljivko pri zlivanju podatkov uporabili zgolj *dohodek,* bi dobili korelacijo *TV-nakupi* 0,4.

Seveda bi v praksi uporabili *obe* skupni spremenljivki ter tako dodatno zmanjšali interval dopustnih vrednosti korelacije *TV-nakupi*. V splošnem primeru imamo večje število skupnih in specifičnih spremenljivk in takrat velja, da je prostor dopustnih vrednosti korelacij med spremenljivkami, ki niso nikoli opazovane skupaj, večdimenzionalni elipsoid, katerega sredinska točka ustreza pogojni neodvisnosti pri danih skupnih spremenljivkah (Kiesl in Rässler 2006).

V razpravi se nismo mogli ogniti tehničnim podrobnostim, zdaj pa v bolj poljudnem tonu ponovimo glavno ugotovitev. Ker v podatkih ni enot z vrednostmi na vseh spremenljivkah, so korelacije med spremenljivkami, ki niso nikoli opazovane skupaj, nedoločene. S preprostimi metodami zlivanja podatkov, ki smo jih opisali doslej, bomo nedoločljive korelacije pravilno poustvarili le v primeru, da so spremenljivke, ki niso nikoli opazovane skupaj, pogojno neodvisne pri danih skupnih spremenljivkah (Rässler in Fleischer 1998: 332).

Pri uporabi preprostih postopkov zlivanja podatkov torej implicitno *predpostavljamo* pogojno neodvisnost, te predpostavke pa pri danih podatkih ne moremo preveriti. V tem leži temeljna »tragedija« zlivanja podatkov: veljavnost rezultatov je odvisna od nepreverljive predpostavke.

Pri uporabi metode najbližjega soseda se moramo zavedati možnosti, da so njeni rezultati lahko zavajajoči. Čeprav dobimo kot rezultat popolnjeno zlito datoteko v kateri

[5] S to trditvijo smo si privoščili poenostavitev, saj smo asociacije zreducirali na korelacije, oz. predpostavili, da so vse spremenljivke povezane zgolj linearno.

lahko izračunamo neocenljive korelacije, ne smemo pozabiti, da noben postopek zlivanja podatkov manjkajoče informacije o neocenljivih korelacijah ne more *ustvariti*. V rezultatih se zrcali zgolj predpostavka, ki smo jo naredili v postopku (Rodgers in DeVol 1982: 129).

2.3 Uporaba dodatnih informacij

Predpostavke, da so specifične spremenljivke pogojno neodvisne pri danih skupnih spremenljivkah, ob podatkih, ki jih imamo na voljo, ne moremo preveriti. Izkaže se tudi, da v praksi predpostavka pogojne neodvisnosti pogosto ni zadovoljena (D'Orazio in drugi 2006: 65). Posledice zlivanja podatkov ob nepravilno predpostavljeni pogojni neodvisnosti opisuje npr. že Kadane (1978).

Da bi premagali problem nedoločenosti, moramo v postopek zlivanja podatkov vnesti dodatno informacijo. Kadar je le mogoče, to naredimo tako, da si pomagamo z »zunanjo« enovito datoteko[6]: lahko gre za zastaran popis prebivalstva, administrativni register ali pa izvedemo manjšo *ad hoc* anketo prav z namenom pridobiti manjkajočo informacijo (D'Orazio in drugi 2006: 67).

Za zlivanje podatkov z dodatnim virom informacije so zelo primerne parametrične metode, ki jih obravnavamo v nadaljevanju, v tem razdelku pa bomo opisali neparametrično metodo, ki jo predlagajo Singh in drugi (Singh in drugi v D'Orazio in drugi 2006: 84). Postopek lahko opišemo kot metodo najbližjega soseda v dveh korakih. Če dodelimo datoteki A vlogo prejemnika, datoteki B vlogo darovalca ter dodatno datoteko imenujemo datoteka C, imamo naslednji postopek (D'Orazio in drugi 2006: 84-85):

1. Vsaki enoti datoteke A poiščemo darovalca iz datoteke C pri čemer razdaljo izračunamo na spremenljivkah X in Y (vseh spremenljivkah, ki so skupne datotekama A in C).
2. V datoteko A smo tako vstavili vrednosti spremenljivk Z_C. Končne vrednosti spremenljivk Z_B vstavimo iz datoteke B pri čemer razdaljo računamo na spremenljivkah X in Z.

[6] Dodatno informacijo lahko sicer predstavlja tudi zgolj poznavanje pravilnih *vrednosti* korelacij med spremenljivkami, ki niso nikoli opazovane skupaj.

V prvem koraku torej vstavljamo vrednosti iz datoteke C, ki je navadno nižje kvalitete kot datoteka B. Ker pa pri izračunu razdalje v prvem koraku nastopajo tudi spremenljivke Y, imajo lahko vstavljene spremenljivke Z_C s spremenljivkami Y korelacije, ki odstopajo od pogojne neodvisnosti. Nadejamo se torej, da smo s prvim korakom premagali problem nedoločenosti, v drugem koraku pa nato vrednosti spremenljivk Z_C nadomestimo z »živimi« vrednostmi iz datoteke B.

2.4 Zgodovina rabe zlivanja podatkov

V naslednjem poglavju se bomo nekoliko odmaknili od področja zlivanja podatkov, ter postavili teoretične temelje za sodobne metode, ki jih obravnavamo v četrtem poglavju. Trenutno poglavje končujemo s kratkim opisom zgodovine zlivanja podatkov.

Zlivanje podatkov ima v Evropi drugačno tradicijo ter ime kot v ZDA in Kanadi: v Evropi se te postopke navadno označuje z angleškim izrazom *data fusion*, v Ameriki pa z izrazom *statistical matching*. Razlika je tudi v glavni gonilni sili, ki je bila v Evropi tržno-raziskovalna industrija, v Ameriki pa zvezni uradi (Rässler 2002: 1, 49). Avtorji D'Orazio in drugi razločujejo med tremi velikimi področji uporabe zlivanja podatkov: mikrosimulacijske študije (glej npr. Kovacevic in Liu 1994; Sutherland in drugi 2002), uradne statistike (Coli in Tartamella 2000; Yoshizoe in Araki 1999) ter tržno raziskovanje (Baker in drugi 1989; Soong in de Montigny 2001). Prvi dve področji uporabe sta značilni bolj za ameriško, zadnje pa za evropsko tradicijo.

Prvi poskusi zlivanja podatkov so se pojavili neodvisno v Nemčiji, Franciji in Veliki Britaniji med letoma 1963 in 1965. Zaradi tehnoloških omejitev so bili ti prvi poskusi pretežno neuspešni. V Veliki Britaniji tržno-raziskovalna industrija postopkov zlivanja podatkov sprva ni želela uporabljati, nasprotno pa se je njihova raba hitro razširila v Nemčiji in Franciji, kjer so združili znanje, pridobljeno s prvimi poskusi (Rässler 2002: 45). Metode, ki so bile takrat v rabi, so bile različice metode najbližjega soseda, avtorji pa so kot pogoj za uspešno zlivanje podatkov poudarjali predvsem »dobre« skupne spremenljivke (Rässler 2002: 49).

V ZDA in Kanadi so se s postopki zlivanja podatkov začeli ukvarjati veliki zvezni uradi. Glavna motivacija za uporabo zlivanja podatkov je bila v 70ih letih sprejeta zakonodaja, ki je prepovedala povezovanje baz podatkov preko osebnih identifikacijskih številk (*angl*. record linkage), dovoljevala pa povezovanje s pomočjo

skupnih spremenljivk (*angl*. statistical matching). Za razliko od Evrope so v Ameriki postopke zlivanja podatkov uporabljali za združevanje informacij iz baz podatkov z velikim številom enot (tudi več milijonov enot), zaradi česar so bili prisiljeni uporabljati manj skupnih spremenljivk kot njihovi evropski kolegi. S takratno tehnologijo za tako veliko število enot namreč ni bilo moč v uglednem času izračunati razdalj na velikem številu skupnih spremenljivk (Rässler 2002: 49-51).

3 Teoretično ozadje parametričnih metod

Vsebina, predstavljena v tem poglavju, ni tesno povezana s področjem zlivanja podatkov. Ker pa bomo v nadaljevanju predstavili sodobne metode zlivanja podatkov, ki temeljijo na bayesianskih načelih, se nam zdi pomembno kratko predstaviti tudi sam Bayesov pristop k statističnemu sklepanju. Ker bayesiansko sklepanje predstavlja protipol klasičnemu, bomo najprej opisali načela najbolj pogosto uporabljane klasične metode statističnega sklepanja: metode največjega verjetja.

3.1 Osnovni koncepti verjetnosti

Da bi temeljito pojasnili nekatere pojme, ki jih uporabljamo v nadaljevanju, se bomo pomudili še pri temeljnih konceptih verjetnosti. Pomemben je predvsem zadnji razdelek poglavja, v katerem je govora o porazdelitvah več spremenljivk, skupnih ter robnih porazdelitvah.

3.1.1 Verjetnost izida

Razlika med klasičnim ter bayesianskim pristopom se prične že pri opredelitvi verjetnosti. Klasični pristop verjetnost definira s pomočjo dolge serije poskusov: trditev, ”verjetnost, da v enem metu kovanca pade glava, znaša 1/2" se po klasičnem pristopu utemeljuje z ugotovitvijo, da bi v neskončni seriji poskusov padla glava približno v 50 %. Bayesianski pristop pa verjetnost definira bolj subjektivno ter poudarja, da pri omenjeni trditvi o verjetnosti delamo vrsto predpostavk med drugim, da je kovanec ”pošten” ter da smo se iz poprejšnjih izkušenj naučili, da pade glava v približno polovici poskusov (Lynch 2007: 9).

Čeprav je sámo verjetnost težko opredeliti, pa so splošni aksiomi o verjetnosti znani ter splošno sprejeti (Lynch 2007: 10). Verjetnost, da se določen izid E zgodi, bomo označili s $p(E)$. Vsi možni izidi, ki se lahko zgodijo v poskusu, sestavljajo *vzorčni prostor* (*angl*. sample space) tega poskusa, vsota verjetnosti vseh izidov v vzorčnem prostoru pa je enaka ena:

$$\sum_{\forall E \in S} p(E) = 1$$

Idejo verjetnosti lahko razširimo na več hkratnih ali zaporednih poskusov. Verjetnost, da se skupaj zgodita izida dveh poskusov, denimo A in B, imenujemo *skupna* verjetnost in označimo s $p(A, B)$. Verjetnost, da se zgodi izid *A ali* izid *B* (ali oba skupaj) pa bomo označili s $p(A \text{ ali } B)$. Sledi preprosta lastnost:

$$p(A \text{ ali } B) = p(A) + p(A) - p(A, B) .$$

Na desni odštejemo skupno verjetnost izidov A in B $p(A, B)$, ker smo jo s seštevanjem $p(A)$ in $p(B)$ prišteli dvakrat (Lynch 2007: 10).

Pomembna sta naslednja izraza. Skupna verjetnost izidov A in B je enaka produktu posameznih izidov, torej:

$$p(A, B) = p(A)p(A), \tag{2.1}$$

če, in samo če, sta poskusa A in B *neodvisna* (Lynch 2007: 10). Neodvisnost poskusov pomeni, da izid enega poskusa prav v ničemer ni odvisen od izida drugega. Dva zaporedna meta kovanca sta npr. neodvisna poskusa, saj izid prvega v ničemer ne vpliva na izid drugega. V primeru, da poskusa nista neodvisna, pa velja izraz:

$$p(A, B) = p(A \mid B)p(B) .$$

Oznaka "|" tu izraža pogojenost, preberemo pa jo kot "pri pogoju", torej: skupna verjetnost izidov A in B je enaka verjetnosti, da se zgodi B, pomnoženi z verjetnostjo, da se zgodi A pri pogoju, da se je B že zgodil (Lynch 2007: 11). Verjetnost $p(A|B)$ imenujemo *pogojna* verjetnost, ter izrazimo z:

$$p(A \mid B) = \frac{p(A, B)}{p(B)} . \tag{2.2}$$

Zgornji lastnosti lahko razširimo na poljubno število poskusov: verjetnost, da opazimo izide treh neodvisnih poskusov A, B in C je $p(A,B,C)=p(A)p(B)p(C)$.

3.1.2 Verjetnostna gostota

Primer meta kovanca je mogoče opisati z zgornjimi elementarnimi izrazi, ker vzorčni prostor sestoji iz zgolj dveh izidov. Če bi s predstavljenimi izrazi skušali opisati večje vzorčne prostore, npr. 100 metov kovanca, pa bi kmalu zašli v težave. Zaradi tega v takem primeru uporabimo *funkcijo*, ki vsakemu izidu v vzorčnem prostoru priredi

verjetnost oz. relativno frekvenco. Taksna funkcija poleg slučajne spremenljivke vsebuje *parametre*, ki določajo obliko krivulje funkcije. To funkcijo imenujemo *verjetnostna gostota* v primeru, da opisuje zvezno porazdeljene izide, ter *verjetnostna shema*, če so izidi porazdeljeni diskretno (Lynch 2007: 12). Ime *gostota* se nanaša na lastnost, da taksna funkcija opisuje, kje v vzorčnem prostoru so zgoščeni najbolj verjetni izidi.

V nadaljevanju bomo namesto obeh izrazov uporabljali kar izraz *porazdelitev*, lastnost, da se slučajna spremenljivka x porazdeljuje po porazdelitvi $g(.)$, pa bomo označili z $x \sim g(.)$. Pika v oklepaju pri tem pomeni parametre porazdelitve: normalna porazdelitev ima npr. dva parametra – povprečje μ in varianco σ^2. Če ima spremenljivka relativne frekvence, ki sledijo normalni porazdelitvi, to lastnost označimo z $x \sim N(\mu, \sigma^2)$, sámo porazdelitev pa opišemo z izrazom:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}.$$

Pri tem je bistveno to, da je porazdelitev algebraična funkcija, ki ob danih vrednostih parametrov (zgoraj μ in σ^2) priredi relativno frekvenco katerikoli vrednosti iz vzorčnega prostora. V splošnem lahko katerokoli funkcijo obravnavamo kot porazdelitev, v kolikor izpolnjuje pogoja, da 1) so vse njene vrednosti ne-negativne ter 2) je njen integral po celotnem vzorčnem prostoru enak ena (glej Lynch 2007: 13).

Zgoraj smo namesto izraza *verjetnost* raje uporabili izraz *relativna frekvenca*, saj ima, striktno rečeno, v zvezni porazdelitvi vsaka konkretna vrednost x verjetnost nič. Kadar imamo opravka z zvezno porazdelitvijo, torej ni smiselno govoriti o verjetnosti, da opazimo konkretno vrednost, pač pa govorimo pa verjetnosti, da opazimo vrednost z določenega *intervala*. V ta namen definiramo *kumulativno* porazdelitveno funkcijo kot integral porazdelitve od najmanjše vrednosti, ki jo lahko zavzame x, pa do neke konkretne vrednosti x_i:

$$F(X) = p(x < X) = \int_{-\infty}^{x_i} f(x)dx .$$

Če nas pri dani porazdelitvi in parametrih zanima verjetnost, da opazimo vrednost x z intervala med x_1 in x_2, to verjetnost izračunamo z odštevanjem ustreznih integralov oz. integriranjem $f(x)$ od x_1 do x_2 (pri tem smo predpostavili, da je $x_1 < x_2$).

$$p(x_1 < x < x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx \,. \tag{2.3}$$

Geometrijsko pomeni zgornji integral površino pod krivuljo porazdelitve med vrednostma x_1 in x_2.

3.1.3 Opisne statistike porazdelitev

Pogosto nas ne zanima oblika porazdelitve z vsemi podrobnostmi, pač pa zgolj sumarni opis s statistikami kot so povprečje, varianca, mediana. Te količine za zvezno porazdelitev izračunamo z integriranjem, za diskretno porazdelitev pa s seštevanjem (Lynch 2007: 18). V nadaljevanju bo naveden zgolj integralski zapis, ki bo v primeru zvezne porazdelitve pomenil seštevanje.

Prvo od opisnih količin, povprečje, definiramo kot:

$$\mu_x = \int_{x \in S} x \cdot f(x)dx \,.$$

Povprečje pogosto imenujemo tudi pričakovana vrednost, ter označujemo z $E(x)$. Varianco definiramo kot:

$$\sigma_x^2 = \int_{x \in S} (x - \mu_x)^2 \cdot f(x)dx \,.$$

Varianco lahko razumemo tudi kot pričakovano vrednost kvadratov odmikov od povprečja $E((x - \mu_x)^2)$. Tudi kvantile, vključno z mediano, izračunamo z integriranjem. Mediana je tako definirana kot tista vrednost Q, ki zadovolji pogoj:

$$0{,}5 = \int_{-\infty}^{Q} f(x)dx \,.$$

Kot glavno ugotovitev razdelka poudarimo, da moramo biti sposobni vsako porazdelitev integrirati, da jo lahko opišemo z opisnimi statistikami.

3.1.4 Porazdelitve več spremenljivk

Namen tega razdelka je predstaviti koncepte *skupne porazdelitve*, *robne porazdelitve* ter *pogojne porazdelitve* ter prikazati, kako so med seboj povezani. Da bi bili kar se da nazorni, bomo omenjene koncepte ponazorili na konkretnem primeru (povzeto po Lynch 2007: 19-25).

Gostoto, ki zadeva več kot eno slučajno spremenljivko, imenujemo *skupna* gostota ali *večrazsežnostna* porazdelitev. Kot konkreten primer vzemimo porazdelitev:

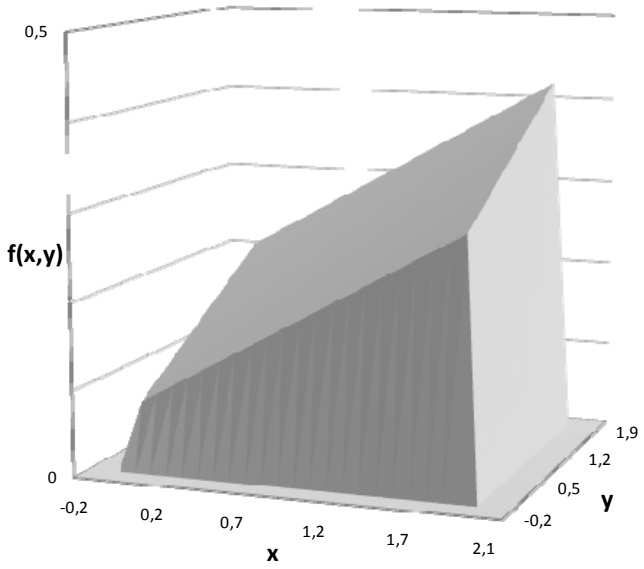
$$f(x) = \begin{cases} c(2x+3y+2) & ko\ je \quad 0 < x < 2; \quad 0 < y < 3 \\ 0 & sicer. \end{cases}$$

Pri tem sta x in y dimenziji dvorazsežnostne slučajne spremenljivke, $f(x, y)$ pa je "višina" gostote, ki izraža relativno frekvenco/verjetnost pri določenih vrednostih x in y. Konstanta c je normalizacijska konstanta, ki zagotavlja, da je integral gostote po celotnem vzorčnem prostoru enak ena (v konkretnem primeru je $c = 1/28$).

Kot prikazuje Slika 3.1, je porazdelitev gostote $f(x, y)$ izsek ravnine. Ravnina narašča hitreje v smeri spremenljivke y, ki ima višjo vrednost koeficienta. Porazdelitev lahko prikažemo tudi s tabelo, v katere celicah so verjetnosti, da opazimo vrednost z danega dvorazsežnostnega intervala. V Tabeli 3.1 je so prikazane verjetnosti za 25 intervalov porazdelitve $f(x, y)$, ki so bile izračunane s posplošitvijo izraza (2.3) na dvorazsežnostni primer:

$$p(x < X, y < Y) = F(x, y) \int_{-\infty}^{X} \int_{-\infty}^{Y} f(x, y) dxdy \, .$$

Slika 3.1: Primer dvorazsežnostne verjetnostne gostote: izsek ravnine.



(vir: Lynch 2007: 20)

Čeprav je včasih pomembno poznati verjetnost konkretnega predela večrazsežnostne spremenljivke, kot prikazuje Tabela 3.1, pa nas pogosteje zanima določen *izsek* dimenzij večrazsežnostne spremenljivke. Navadno potrebujemo dve vrsti izsekov: *robne* porazdelitve ter *pogojne* porazdelitve.

Robna porazdelitev spremenljivke x je v diskretizirani obliki prikazana v najnižji vrstici Tabele 3.1. Razumemo jo lahko kot seštevanje preko kategorij druge spremenljivke. Robno porazdelitev si lahko predstavljamo tudi kot porazdelitev, ki bi jo dobili, če bi gostoto na Sliki 3.1 sploščili v smeri spremenljivke y. V splošnem velja, da dobimo robno porazdelitev z *integriranjem preko preostalih spremenljivk* v porazdelitvi. Za dvorazsežnostni primer torej:

$$f(y) = \int_{x \in S} f(x, y) dy \,. \tag{2.4}$$

Pogojna porazdelitev pa je izrez iz dvorazsežnostne porazdelitve pri določeni vrednosti ene komponente. Osvetljeni stolpec iz Tabele 3.1 lahko razumemo kot diskretizirano robno porazdelitev spremenljivke x pri pogoju $y \in [0{,}8;\ 1{,}2]$. Da bi omenjena vrstica predstavljala pravo porazdelitev, jo moramo normalizirati, kar dosežemo z deljenjem z vrednostjo 20 %. To pa je tudi *verjetnost pogoja*, ki smo ga postavili za spremenljivko y.

Tabela 3.1: Verjetnostna shema linearne dvorazsežnostne porazdelitve.

x\y	**0,0 - 0,4**	**0,4 - 0,8**	**0,8 - 1,2**	**1,2 - 1,6**	**1,6 - 2,0**	*skupaj*
0,0 - 0,4	1,7%	2,4%	3,1%	3,8%	4,5%	15,4%
0,4 - 0,8	2,2%	2,9%	3,5%	4,2%	4,9%	17,7%
0,8 - 1,2	2,6%	3,3%	4,0%	4,7%	5,4%	20,0%
1,2 - 1,6	3,1%	3,8%	4,5%	5,1%	5,8%	22,3%
1,6 - 2,0	3,5%	4,2%	4,9%	5,6%	6,3%	24,6%
skupaj	13,1%	16,6%	20,0%	23,4%	26,9%	100,0%

Omenjena ugotovitev velja tudi v splošnem. Izraz (2.2), ki smo ga zapisali za diskretne verjetnosti, velja namreč tudi za verjetnostne gostote:

$$f(x \mid y) = \frac{f(x, y)}{f(y)} \,.$$

Zapisana enačba pravi, da dobimo pogojno porazdelitev tako, da skupno porazdelitev delimo z robno porazdelitvijo, torej integralom (2.4). Za obravnavani primer zapišimo obe robni ter obe pogojni porazdelitvi:

$$f(x) = \int_{y=0}^{2} (1/28)(2x + 3y + 2)dy = (1/28)(4x + 10)$$

$$f(y) = \int_{x=0}^{2} (1/28)(2x + 3y + 2)dx = (1/28)(6y + 8)$$

$$f(x \mid y) = \frac{(1/28)(2x + 3y + 2)}{(1/28)(6y + 8)}$$

$$f(y \mid x) = \frac{(1/28)(2x + 3y + 2)}{(1/28)(4x + 10)}.$$

Pomembno je naslednje opažanje: izraz za *robno* porazdelitev vsake spremenljivke ni odvisen od druge spremenljivke. Izraz za *pogojno* porazdelitev pa vseeno vsebuje tudi drugo spremenljivko: izraz za pogojno verjetnost x je npr. odvisen tudi od y. Ko pa izberemo *konkretno vrednost* za y, bo preostali izraz za pogojno porazdelitev x odvisen le še od same slučajne spremenljivke x.

3.2 Ocenjevanje parametrov po metodi največjega verjetja

V tem poglavju se obračamo k nalogi, ki jo navadno razumemo pod imenom »statistika«: statističnemu sklepanju o parametrih. Za statično sklepanje lahko rečemo, da poteka v nasprotni smeri kot sklepanje o verjetnosti (Lynch 2007: 35). Pri verjetnosti gre za to, da iz dane porazdelitve ter njenih parametrov *deduciramo* verjetnosti posameznih izidov: dan imamo model, ki generira izide; znana nam je tako oblika (matematična funkcija) kot parametri modela, sprašujemo pa se o verjetnosti posameznih izidov. Pri statističnem sklepanju je naloga obratna: dan imamo nabor izidov (podatke), zanimajo pa nas vrednosti parametrov porazdelitve, ki jih je generirala.

Klasični pristop k statističnem sklepanju obsega dve osnovni stopnji 1) ocenjevanje modela in 2) sklepanje. Ocenjevanje parametrov po metodi največjega verjetja (v nadaljevanju ML, *angl*. Maximum Likelihood) je v družboslovju najpogosteje uporabljena metoda s katero ocenimo tako parametre (prva stopnja), kot tudi napako ocene, ki jo potrebujemo za drugo stopnjo (Lynch 2007: 35). Klasično statistično sklepanje, pri katerem si pomagamo z metodo največjega verjetja, lahko razdelimo v naslednje korake:

1. Najprej je potrebno določiti obliko – matematično funkcijo – modela, ki naj bi ustvaril dane podatke, pri čemer se opremo na poznavanje proučevanega pojava. Zapišemo torej funkcijo, ki izraža verjetnost izidov v odvisnosti od vrednosti parametrov v predpostavljenem modelu.
2. Nato pa uporabimo temeljno idejo ocenjevanja po ML: pri dani obliki modela so najboljša ocena njegovih parametrov tiste konkretne vrednosti, pri katerih je najverjetneje, da »so se zgodili« prav podatki, ki smo jih opazili v vzorcu. Za zapisano funkcijo, ki izraža verjetnost opaženih podatkov, moramo najti tisto vrednost parametra, ki maksimizira to verjetnost – ta vrednost parametra je ocena po metodi največjega verjetja.
3. Določiti je potrebno napako ML ocene, ki jo potrebujemo za zadnji korak.
4. Sklepanje o zadanem problemu.

Razdelki tega poglavja bodo sledili zgoraj predstavljenim korakom. V prvem razdelku si bomo zadali problem, ki ga želimo rešiti, ter namenili nekaj besed binomski porazdelitvi, s katero bomo opisali model podatkov.

3.2.1 Določitev primernega modela ter funkcije verjetja

V neki državi bodo izvedene volitve na katerih nastopata dva kandidata. V predvolilni anketi se je 556 anketirancev opredelilo za kandidata A, 511 pa za kandidata B. Zanima nas, ali lahko na podlagi rezultatov ankete napovemo, da bo na volitvah zmagal kandidat A.

Za opis danih podatkov je primerna *binomska* porazdelitev. Binomska porazdelitev opisuje verjetnost, da se v seriji n poskusov zgodi x uspehov ob dani verjetnosti uspeha K (Lynch 2007: 25). Gre torej za to, da pri danem številu poskusov razlikujemo zgolj med dvema izidoma, za katera se je ustalila raba izrazov *uspeh* in *neuspeh*. V konkretnem primeru bomo kot uspeh označili oddajo glasu za kandidata A. Če se število uspehov x porazdeljuje po binomski porazdelitvi, lahko zapišemo:

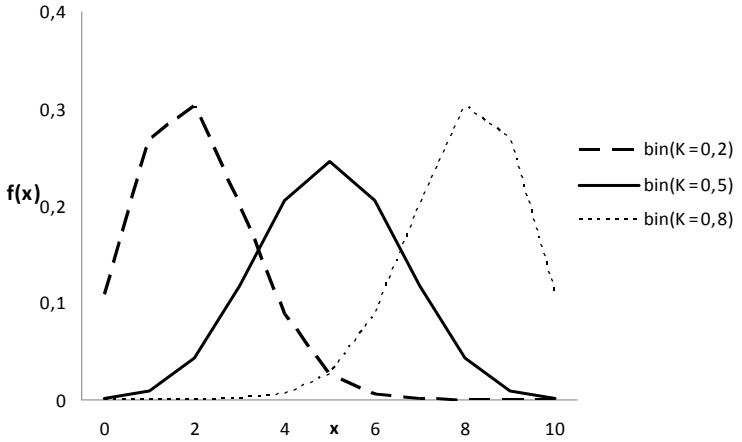
$$f(x \mid n, K) = \binom{n}{x} K^x (1-K)^{n-x}. \tag{2.5}$$

Faktoriela v zgornjem izrazu predstavlja zgolj normalizacijsko konstanto, ki je potrebna zaradi dejstva, da se lahko x uspehov zgodi v različnem vrstnem redu. Kadar sta

parametra K in n znana, lahko z izrazom (2.5) izračunamo verjetnost za katerokoli vrednost x.

Ker se bomo z njo srečali tudi v naslednjem poglavju, si oglejmo nekatere značilnosti binomske porazdelitve. Njeno povprečje je nK, varianca pa $nK(K - 1)$. Če je $K > 0{,}5$, je porazdelitev asimetrična v levo, pri $K < 0{,}5$ v desno, pri $K = 0{,}5$ pa je porazdelitev simetrična. Na Sliki 3.2 so prikazane porazdelitve za tri vrednosti parametra K ter za $n = 10$.

Slika 3.2: Tri binomske funkcije za *n* = 10 ter različne vrednosti parametra *K*.



(vir: Lynch 2007: 27)

Zaradi nazornosti smo porazdelitve na zgornji sliki ponazorili s črtami. Ker je binomska porazdelitev diskretna, bi jo, strogo rečeno, morali ponazoriti s histogramom, saj so verjetnosti za ne-cela števila nedefinirane.

Glasove volivcev x_1, $x_2 \ldots x_n$ obravnavamo kot *neodvisne* realizacije slučajne spremenljivke ter predpostavimo, da vsi izidi (glasovi) izhajajo iz *iste* porazdelitve. Kadar lahko za opazovane enote predpostavimo ti dve lastnosti, pravimo, da za njih velja IID porazdelitev (*angl*. Identical Independent Distribution). V danem primeru torej predpostavljamo, da ima vsak volivec enako, vnaprej določeno (vendar neznano) verjetnost, da se odloči za kandidata A, ter da volivci ne vplivajo drug na drugega.

Ob predpostavki IID lahko verjetnost, da smo opazili dane podatke, izrazimo s *produktom posameznih* verjetnosti, pri čemer izraz (2.1) posplošimo na n neodvisnih izidov. Za splošni primer ob predpostavki IID *skupno verjetnost* za opaženi vektor X (dimenzije n) zapišemo kot:

$$p(X \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Pri tem je $f(.)$ predpostavljena porazdelitev, s θ pa označimo njen parameter, ki je lahko tudi večdimenzionalen: v obravnavanem primeru je npr. $\theta = (n, K)$.

Naslednji premislek je značilen za klasični pristop. Oceniti želimo parameter θ pri danem vektorju podatkov X. Ker pa po klasični paradigmi obravnavamo parameter θ kot *vnaprej določen* (vendar nam neznan), ne moremo govoriti o *verjetnosti* parametra. Za namen ocenjevanja parametra pri danih podatkih definiramo koncept *verjetja* (*angl*. likelihood) katerega smisel je obraten konceptu verjetnosti, pri kateri gre za deduciranje verjetnosti izidov pri danih parametrih modela. Verjetje $L(.)$ tako definiramo kot:

$$L(\theta \mid X) \equiv p(X \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

Dani podatki ter ocenjevani parameter v prehodu iz notacije $p(.)$ v notacijo $L(.)$ torej zgolj zamenjajo mesti. V obravnavanem primeru predvolilne ankete funkcijo verjetja izrazimo kot:

$$L(K \mid X) = \binom{1067}{556} K^{556} (1-K)^{511}$$

Tej funkciji bomo naslednjem razdelku poiskali maksimum in tako določili oceno parametra K po metodi največjega verjetja.

3.2.2 Ocena parametra po metodi največjega verjetja

Da bi prišli do ocene populacijskega železa volivcev K, ki podpirajo kandidata A, smo zapisali zgornji izraz za verjetje. Le-ta vključuje podatke iz vzorca, ki ga je generiral model z ocenjevanim parametrom, končna vrednost verjetja pa je odvisna od izbire konkretne vrednosti parametra K. Po logiki metode ML je najboljša ocena za K tista vrednost, ki maksimizira verjetje in pri kateri je torej najbolj verjetno, da smo opazili dane vzorčne podatke.

Da bi izbrali tisto vrednost K, pri kateri bo verjetje največje, si pomagamo z odvodom. Iz osnov analize vemo, da je v najvišji točki funkcije odvod enak nič (glej npr. Vidav 1994: 312-319). Vzeli bomo torej odvod funkcije verjetja na parameter, vrednost tega

izraza postavili na nič, ter odčitali vrednost parametra, pri katerem funkcija doseže maksimum. Iz praktičnih razlogov pred odvajanjem funkcijo verjetja navadno logaritmiramo: z logaritmiranjem prevedemo večkratno množenje v večkratno seštevanje. Logaritem verjetja doseže maksimum v isti točki kot samo verjetje.

$$\ln(L(\theta \mid X)) \equiv l(\theta \mid X) = \sum_{i=1}^{n} \ln(f(x_i \mid \theta)$$

Za obravnavani primer izrazimo logaritem verjetja kot:

$$l(K \mid X) \propto 556 \cdot \ln K + 511 \cdot \ln(1-K) .$$

Da bi našli tisto vrednost *K*, pri kateri zgornji izraz doseže maksimum, vzamemo odvod funkcije na parameter *K*, ga izenačimo z nič, ter odčitamo vrednost *K*. V splošnem je odvod logaritma binomske funkcije (Lynch 2007, 38):

$$\frac{dl(K \mid X,n)}{dK} = \frac{\sum x_i}{K} - \frac{n - \sum x_i}{1-K} .$$

Ko ta izraz izenačimo z nič in razrešimo, dobimo za vrednost *K* naslednji izraz:

$$\hat{K} = \frac{\sum x_i}{n} .$$

Ocena populacijskega deleža *K* po metodi največjega verjetja je za dani primer torej vzorčni delež »uspehov«, oziroma 0,521. Ker je ta vrednost zgolj ocena, jo označimo s $\hat{K}$, oznako *K* pa ohranimo za označevanje dejanskega, a neznanega, populacijskega deleža.

Ta rezultat vsekakor ni presenetljiv, pomembno pa je razumevanje postopka, po katerem smo do njega prišli ter same ideje največjega verjetja. Po metodi ML ocenimo tudi standardno napako ocene parametra, kar si bomo ogledali v naslednjem razdelku.

3.2.3 Ocena standardne napake

Vrednost $\hat{K}$ je zgolj ocena prave vrednosti populacijskega parametra *K*, ki smo jo izračunali na podlagi vzorca. Če bi anketo ponovili, bi dobili več vzorcev z različnimi ocenami parametra *K*, ki niso vse nujno blizu populacijski vrednosti. Potrebujemo torej

način, kako kvantificirati negotovost pri ocenjevanju populacijskega deleža K z vrednostjo $\hat{K}$.

Uporabna lastnost metode ML je, da lahko drugi odvod logaritma verjetja uporabimo pri ocenjevanju variance vzorčne porazdelitve. Izkaže se, da moramo vzeti obratno vrednost negativne pričakovane vrednosti drugega odvoda logaritma verjetja (Lynch 2007: 40):

$$I(\theta)^{-1} = \left(- E\left(\frac{\partial^2 l(\theta \mid X)}{\partial\theta\partial\theta^T} \right) \right)^{-1} .$$

Ko izpeljemo izraz za drugi odvod, vanj vstavimo ML oceno parametra in s tem izrazimo standardno napako ocene. Končni izraz za inverz informacijske matrike za obravnavani primer je (za izpeljavo glej Lynch 2007: 40-41):

$$I(K)^{-1} = \frac{\hat{K}(1-K)}{n} .$$

Kvadratni koren zgornjega izraza nam dá standardno napako. V obravnavanem primeru le-ta torej znaša 0,015.

3.2.4 Statistično sklepanje

Naše vprašanje glede obravnavanega volilnega primera je bilo, ali lahko na podlagi pridobljenega vzorca napovemo rezultat volitev. Z metodo ML smo izračunali oceno populacijskega deleža K volivcev, ki podpirajo kandidata A, ter njeno standardno napako. Po klasičnem pristopu lahko statistično sklepanje izvedemo na dva načina: z intervalom zaupanja ali s testom hipoteze. Prikazali bomo oba pristopa.

Okrog točkovne ocene zgradimo 95 % interval zaupanja. Pri tem se naslonimo na centralni limitni izrek, ki pravi, da je vzorčna porazdelitev normalna ne glede na osnovno porazdelitev, če je le velikost vzorca dovolj velika (glej npr. Fisher 2000). Če je vzorčna porazdelitev normalna, izrazimo 95 % interval zaupanja tako, da točkovni oceni prištejemo in odštejemo dvakratnik standardne napake (Kalton in Vehovar 2001, 20):

$$p\left[K \in \left(\hat{K} \pm 1{,}96 \cdot se(\hat{K})\right)\right] = 0{,}95 .$$

Za obravnavani primer tako dobimo meji 95 % intervala zaupanja [0,492; 0,550]. Spodnja meja intervala je nižja od 0,5, zato ne moremo s 95 % gotovostjo trditi, da bo na volitvah zmagal kandidat A. Podoben rezultat dobimo pri testu hipoteze. Zanikati želimo ničto hipotezo, da je $K < 0{,}5$. Testno statistiko izračunamo po spodnji formuli:

$$t = \frac{(0{,}521 - 0{,}5)}{0{,}015} = 1{,}4\,.$$

Ker vrednost testne statistike ni dovolj visoka (ne presega 1,96), ničte hipoteze ne moremo zanikati.

V pričujočem poglavju smo predstavili ML pristop k statističnemu sklepanju. Z metodo največjega verjetja na podlagi opaženega vzorca najdemo točkovno oceno parametra, ki maksimizira funkcijo verjetja, ter točkovno oceno njene standardne napake. Nato opravimo klasični statistični test: od ML ocene odštejemo določeno predpostavljeno teoretično vrednost parametra ter razliko delimo z oceno standardne napake. Po centralnemu limitnemu izreku je porazdelitev vzorčne statistike asimptotično normalna, zato lahko z uporabo *z*-porazdelitve (ali Studentove *t*-porazdelitve) ocenimo verjetnost, da smo opazili vzorčno statistiko pri predpostavki, da je teoretična vrednost pravilna. Če je verjetnost, da smo pri predpostavljeni teoretični vrednosti opazili vzorčno statistiko, zelo majhna, predpostavljeno teoretično vrednost zavrnemo (Lynch 2007: 78).

3.3 Osnove Bayesove statistike

V naslednjih razdelkih bomo skušali predstaviti Bayesov pristop k statističnemu sklepanju. Začeli bomo z obravnavo Bayesovega izreka za točkovne verjetnosti, nato pa bomo uporabo bayesianskih načel na verjetnostnih porazdelitvah prikazali še z obravnavo volilnega primera iz prejšnjega poglavja.

3.3.1 Bayesov izrek za točkovne verjetnosti

Za ilustracijo osnovnega Bayesovega izreka vzemimo naslednji primer. Neka ženska meni, da obstaja možnost, da je noseča po enem spolnem odnosu, vendar pa glede tega ni gotova. Zato se odloči opraviti test nosečnosti, ki je 90 % natančen, kar pomeni, da pokaže pozitiven rezultat v 90 % primerov, ko gre res za nosečnost. Test dá v njenem primeru pozitiven rezultat. Omenjeno žensko zanima, kakšna je verjetnost, da je noseča,

ob dejstvu, da je test pozitiven. Znana pa ji je neka druga verjetnost: verjetnost pozitivnega izida v primeru, da je res noseča (primer, povzet po Lynch 2007: 47-50).

V tem razdelku bomo pokazali, kako odgovoriti na zadano vprašanje z uporabo Bayesovega izreka, s katerim je moč obrniti pogojne verjetnosti. Prvotni Bayesov izrek se nanaša na *točkovne verjetnosti* (torej verjetnosti izidov, kot predstavljene v razdelku 3.1.1) ter opisuje, kako iz verjetnosti za izid A pri pogoju, da se je zgodil izid B, izračunamo verjetnost izida B pri pogoju da se je zgodil izid A (Lynch 2007: 47). Spodnja enačba sledi neposredno iz izraza (2.2):

$$p(B \mid A) = \frac{p(A \mid B)\,p(B)}{p(A)}\,.$$

Verjetnost $p(B)$ je nepogojna (robna) verjetnost izida, ki nas zanima – v tem primeru nosečnosti. Verjetnost $p(A)$ pa je *robna* verjetnost izida A – pozitivnega rezultata testa. Izračunamo jo kot vsoto pogojnih verjetnosti izida A pri vseh možnih izidih B_i v vzorčnem prostoru. V obravnavanem primeru vzorčni prostor obsega le dva izida: omenjena ženska je ali noseča, ali pa ne. Za splošni diskretni primer robno verjetnost izida A izrazimo kot:

$$p(A) = \sum_{B_i \in S_B} p(A \mid B_i)\,p(B_i)\,.$$

Za izračun te količine potrebujemo dodaten podatek: verjetnost, da dá test nosečnosti *pozitiven* rezultat v primeru, da testirana ženska *ni* noseča. Denimo, da je ta verjetnost 50 %. Da lahko po Bayesovem izreku izračunamo verjetnost, ki nas zanima, moramo to informacijo združiti z določenim *predhodnim poznavanjem* obravnavanega primera.

Omenjeno predhodno znanje, ki ga potrebujemo, je *robna* verjetnost za zanositev ob enem spolnem odnosu. Za to verjetnost pravimo, da je *apriorna*, ker se nanaša na informacijo, ki obstaja, še preden je opravljen test. Denimo, da iz predhodnih raziskav vemo, da omenjena verjetnost znaša približno 15 %. Ob poznavanju vseh potrebnih količin lahko zapišemo spodnjo enačbo:

$$p(N=1 \mid T=1) = \frac{p(T=1 \mid N=1)\,p(N=1)}{p(T=1 \mid N=0)\,p(N=0) + p(T=1 \mid N=1)\,p(N=1)}\,.$$

Ko v izraz vstavimo konkretne vrednosti, dobimo verjetnost nosečnosti ob dejstvu, da je bil rezultat testa pozitiven:

$$p(N=1 \mid T=1) = \frac{0{,}90 \cdot 0{,}15}{0{,}90 \cdot 0{,}15 + 0{,}50 \cdot 0{,}85} = 0{,}241.$$

Omenjena verjetnost znaša zgolj 0,241. Tej verjetnosti ob upoštevanju bayesianskega izrazoslovja pravimo *posteriorna* verjetnost, saj se nanaša na verjetnost nosečnosti po tem, ko smo že opazili podatke (pozitivni rezultat testa)(Lynch 2007: 49).

Če se omenjena ženska zaveda omejitev uporabljenega testa nosečnosti, se lahko odloči, da ga ponovi. V tem primeru lahko uporabi »posodobljeno« verjetnost nosečnosti (p = 0,241) kot novo vrednost za $p(B)$. Apriorno verjetnost nosečnosti torej posodobimo, da odraža rezultat prvega testa (Lynch 2007: 49). Če dá tudi ponovljeni test pozitiven rezultat, je nova posteriorna verjetnost:

$$p(N=1 \mid T=1) = \frac{0{,}90 \cdot 0{,}241}{0{,}90 \cdot 0{,}241 + 0{,}50 \cdot 0{,}759} = 0{,}364.$$

Tudi ta verjetnost še ni prepričljiv dokaz nosečnosti, vendar pa lahko omenjena ženska test ponovi še enkrat, pri čemer se verjetnost nosečnosti dvigne na 0,507, seveda če tudi v tej ponovitvi test pokaže pozitiven rezultat (Lynch 2007: 49).

Prikazani princip ponovnega opravljanja testa ter ponovnega izračuna verjetnosti je temeljen Bayesovemu pristopu k raziskovanju. Po bayesianski paradigmi vedno začnemo z določeno apriorno verjetnostjo, jo kombiniramo z novo informacijo ter izračunamo posteriorno verjetnost. To posteriorno verjetnost lahko uporabimo kot apriorno verjetnost v nadaljnji obravnavi problema. Z bayesianskega gledišča torej, kadar skušamo potrditi ali zavrniti hipotezo, nikoli ne začnemo povsem nevedni, saj imamo zaradi predhodnih raziskav vedno na voljo apriorno znanje (Lynch 2007: 49).

3.3.2 Bayesov izrek za verjetnostne porazdelitve

Vedno vnovična raba Bayesovega izreka na istem problemu ter sam Bayesov izrek z matematičnega vidika nikakor nista sporna. Vendar pa je za Bayesov pristop k statističnem sklepanju značilna uporaba *verjetnostnih porazdelitev* (in ne točkovnih verjetnosti) v Bayesovem izreku (Lynch 2007: 50).

V obravnavanem primeru smo predpostavili, da znaša apriorna verjetnost nosečnosti natančno 0,15. Bolj realistično gledano pa naše poznanje te verjetnosti ni tako natančno, saj gre zgolj za *oceno*, ki tudi sama temelji na predhodnih vzorcih, in nikakor ne za poznano populacijsko vrednost. Z bayesianskega vidika lahko verjetnost 0,15 zamenjamo z verjetnostno *porazdelitvijo*, ki bo izražala našo negotovost glede dejanske vrednosti te količine. Uporaba apriorne verjetnostne *porazdelitve* v Bayesovem izreku pa ima za posledico posteriorno verjetnost, ki tudi sama ni več točkovna, temveč je podana v obliki verjetnostne porazdelitve (Lynch 2007: 50).

Splošno rečeno je cilj bayesianske statistike apriorno negotovost glede modelskih parametrov ponazoriti z verjetnostno porazdelitvijo, kombinirati to apriorno negotovost z opaženimi podatki in tako pridelati posteriorno porazdelitev, ki vsebuje manj negotovosti. Za ta pristop je značilno subjektivno videnje realnosti, saj verjetnost obravnavamo kot mero negotovosti, kar je v nasprotju z klasičnim pojmovanjem verjetnosti. Z bayesianskega vidika je mogoče katerokoli količino, glede katere prave vrednosti smo negotovi, ponazoriti z verjetnostno porazdelitvijo. S klasičnega vidika pa je verjetnostne porazdelitve dopustno uporabljati zgolj za ponazoritev podatkov (izidov), ne pa tudi za parametre. Za le-te se namreč predpostavlja, da so fiksno določene vrednosti (Lynch 2007: 50).

Bayesov izrek, v katerem nastopajo verjetnostne porazdelitve, lahko zapišemo kot:

$$f(\theta \mid Y) = \frac{f(Y \mid \theta) f(\theta)}{f(Y)}. \tag{2.6}$$

V zgornjem izrazu označuje člen $f(\theta \mid Y)$ posteriorno porazdelitev parametra θ, $f(Y \mid \theta)$ pa je *vzorčna gostota* (*angl*. sampling density) podatkov *Y*, ki je sorazmerna funkciji verjetja – od nje se razlikuje zgolj za normalizacijsko konstanto. Oznaka $f(\theta)$ se nanaša na apriorno porazdelitev parametra, s $f(Y)$ pa označimo robno porazdelitev podatkov (Lynch 2007: 50). Če je vzorčni prostor zvezen, to robno porazdelitev izračunamo po:

$$f(Y) = \int f(Y \mid \theta) f(\theta) d\theta,$$

kar je integral vzorčne gostote po celotnem vzorčnem prostoru parametra θ. Ta integral igra v izrazu (2.6) zgolj vlogo normalizacijske konstante, zato bistvo Bayesovega izreka za verjetnostne porazdelitve pogosto poenostavljeno zapišemo kot (Lynch 2007: 51):

$$posteriorna\ porazdelitev \propto verjetje \times apriorna\ porazdelitev\,. \qquad (2.7)$$

Dejstvo, da je posteriorna porazdelitev parametra zgolj sorazmerna (in ne enaka) s produktom verjetja opaženih podatkov ter apriorne porazdelitve parametra, v splošnem ne predstavlja težave (Lynch 2007: 51).

3.3.3 Uporaba Bayesovega izreka za verjetnostne porazdelitve: volilni primer

V tem razdelku se bomo vrnili k primeru predvolilne ankete iz poglavja 3.2 ter ga obravnavali na bayesianski način. Tako kot v poglavju 3.2 se sprašujemo, ali je na podlagi pridobljenega vzorca mogoče napovedati zmagovalca volitev. Pri obravnavi problema bomo postopali po naslednjih korakih:

1. Najprej je treba določiti matematično funkcijo modela, ki naj bi ustvaril opažene podatke, ter zapisati funkcijo verjetja.
2. Naše predhodno poznavanje obravnavanega problema je potrebno izraziti s primerno apriorno porazdelitvijo.
3. Izračun posteriorne porazdelitve.
4. Sklepanje o zadanem problemu, kar v bayesianskem kontekstu pomeni *opisati posteriorno porazdelitev* z opisnimi statistikami.

Verjetje opaženih podatkov (prvi korak) smo zapisali že v razdelku 3.2.1. Predpostavili smo, da glasovi za kandidata A izhajajo iz binomske porazdelitve, ter zapisali funkcijo verjetja, ki izraža *skupno verjetnost* opaženih podatkov pri dani vrednosti modelskega parametra – deleža K.

Proces maksimiranja funkcije verjetja nam pove, kako verjetno je, da smo opazili prav dane podatke pri določeni vrednosti parametra K. Vprašanje, na katero želimo v končni fazi odgovoriti, pa je, ali bo na volitvah zmagal kandidat A pri dejstvu, da smo opazili podatke iz vzorca. Spet gre torej za primer prehajanja iz ene pogojne verjetnosti na drugo za kar uporabimo Bayesov izrek. Ker smo Bayesov izrek poenostavljeno zapisali s sorazmernostjo (glej izraz (2.7)), bomo tudi pri funkciji verjetja izpustili normalizacijsko konstanto ter zapisali:

$$f(Y \mid K) \propto K^{556}(1-K)^{511}.$$

Edina preostala količina, ki jo potrebujemo za določitev bayesianskega modela, je *apriorna porazdelitev*. V predhodnih razdelkih smo kot verjetnost zanositve najprej uporabili točkovno oceno 0,15 nato pa po premisleku predlagali, da bi bila bolj na mestu *porazdelitev*, ki bi obsegala razumne vrednosti verjetnosti zanositve in jim priredila relativne frekvence. Gotovost, da se prava verjetnost nahaja blizu vrednosti 0,15 bi tako izrazili z zelo koničasto krivuljo, negotovost pa s položnejšo krivuljo.

Na podoben način bomo s porazdelitvijo opisali tudi svoje predhodno znanje o podpori kandidatu A. Primerna matematična funkcija, s katero izrazimo delež podpore, je *funkcija beta*[7] (Lynch 2007: 54):

$$f(K \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} K^{\alpha-1}(1-K)^{\beta-1},$$

kjer predstavlja $0 < K < 1$ ocenjevani delež volivcev kandidata A, oznaka $\Gamma(.)$ pa pomeni funkcijo gama[8]. Povprečje in varianca porazdelitve beta sta določena s parametroma α in β:

$$E(K \mid \alpha, \beta) = \frac{\alpha}{\alpha+\beta} \tag{2.8}$$

$$Var(K \mid \alpha, \beta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{2.9}$$

Odgovor na vprašanje, kako izbrati primerni vrednosti parametrov α in β za želeno apriorno porazdelitev, temelji na vsaj dveh dejavnikih namreč 1) koliko znanja o parametru K smo imeli na voljo preden je bila opravljena anketa ter 2) do kakšne mere želimo zaupati tej informaciji (Lynch 2007: 55).

[7] Predstavljamo si jo lahko kot posplošitev binomske funkcije na zvezna števila: parametra α in β, ki ustrezata številu uspehov ter neuspehov, lahko zavzameta tudi ne-cele vrednosti.

[8] Z omenjeno funkcijo se na tem mestu ne bomo ukvarjali, saj bo vedno predstavljala zgolj del normalizacijske konstante. Omenimo le, da jo lahko razumemo kot posplošitev faktoriele na ne-cela števila.

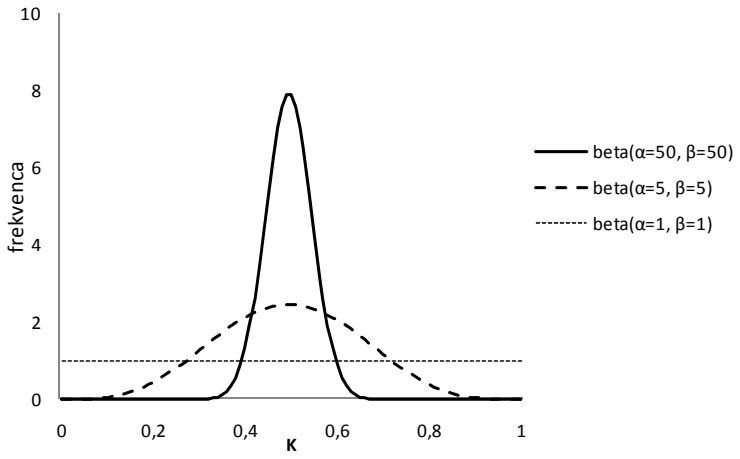
Če pred anketo nismo vedeli o parametru K ničesar, želimo da uporabljena apriorna porazdelitev vpliva na posteriorno v kar najmanjši možni meri. Če npr. postavimo $\alpha = 1$ in $\beta = 1$, je dobljena porazdelitev:

$$f(K \mid \alpha, \beta) \propto K^{1-1}(1-K)^{1-1} = 1$$

na dopustnem intervalu $K \in [0,1]$ sorazmerna z enakomerno porazdelitvijo. Takšna porazdelitev je popolnoma položna, kar pomeni, da ne daje teže nobeni vrednosti K pred drugimi in ima zato izredno majhen vpliv na posteriorno porazdelitev. Takšno porazdelitev zato imenujemo *neinformativna* (Lynch 2007: 55).

V drugi skrajnosti pa je primer, ko o obravnavanem problemu že vnaprej vemo veliko, temu znanju pa želimo z apriorno porazdelitvijo podeliti tudi precej teže. Kot je razvidno iz izraza (2.9), je varianca porazdelitve beta manjša pri višjih vrednostih parametrov α in β (Lynch 2007: 55). Slika 3.3 prikazuje tri primere porazdelitve beta. Vse prikazane krivulje imajo povprečje 0,5, njihove variance – razpršenosti okrog povprečja – pa se razlikujejo zaradi različnih vrednosti parametrov α in β. Funkcija beta ima podobne lastnosti kot binomska funkcija: če imata α in β različni vrednosti, je asimetrična; pri večjem številu »poskusov« (vsota parametrov α in β) je bolj koničasta, pri manjšem pa bolj položna.

Slika 3.3: Tri porazdelitve beta s povprečjem $\alpha/(\alpha + \beta) = 0,5$.



(vir: Lynch 2007: 56)

Vrnimo se na konkretni primer. Predhodno znanje o podpori kandidatu A lahko pridobimo iz rezultatov predvolilnih anket, ki so bile izvedene pred obravnavano anketo. Denimo, da je v treh predhodnih anketah podporo kandidatu A izrazilo skupno

942 anketirancev, kandidatu B pa 1008. To apriorno informacijo upoštevamo v porazdelitvi beta s parametroma $\alpha = 942$ in $\beta = 1008$:

$$f(K \mid \alpha = 942, \beta = 1008) \propto K^{942-1}(1-K)^{1008-1}.$$

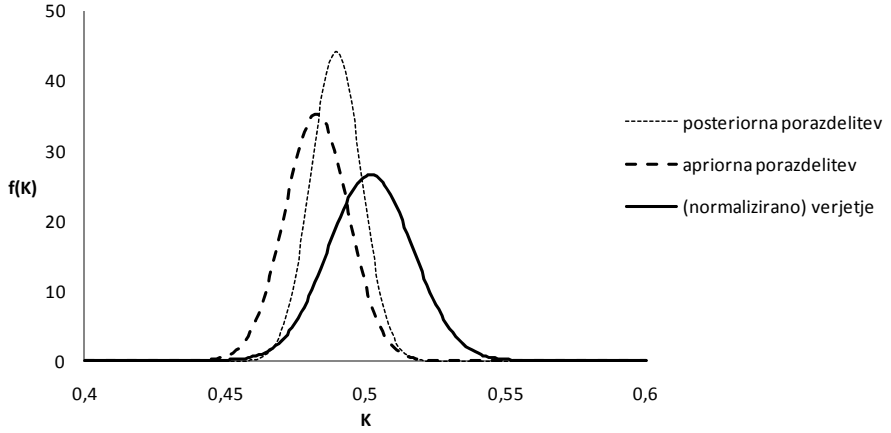
Ko predhodnemu znanju pridružimo funkcijo verjetja, ki vsebuje informacijo iz vzorca zadnje predvolilne ankete, dobimo posteriorno verjetnostno porazdelitev za parameter K:

$$\begin{aligned} f(K \mid \alpha = 942, \beta = 1008, \textit{podatki}) &\propto K^{941}(1-K)^{1007} K^{556}(1-K)^{551} \\ &\propto K^{1497}(1-K)^{1518}. \end{aligned}$$

Tudi posteriorna porazdelitev ima obliko funkcije beta; njena parametra sta $\alpha = 1498$ in $\beta = 1519$. S tem smo ponazorili pomemben koncept bayesianske statistike: *konjugiranost*. Kadar ima dobljena posteriorna porazdelitev isto funkcijsko obliko kot apriorna porazdelitev, pravimo, da sta verjetje ter apriorna porazdelitev *konjugirani* (Lynch 2007: 57). Konjugiranost je bila pred nastopom numeričnih metod, ki jih omenjamo v poglavju 3.4, izredno pomemben koncept. Z uporabo verjetju konjugirane apriorne porazdelitve se je namreč zagotovilo, da je imela posteriorna porazdelitev poznano funkcijsko obliko, ki jo je bilo moč enostavno povzeti in tako odgovoriti na zastavljeno raziskovalno vprašanje (Lynch 2007: 57).

Slika 3.4 prikazuje apriorno porazdelitev, funkcijo verjetja ter posteriorno porazdelitev; prikazane funkcije so normalizirane po parametru K. Iz slike je razvidno, da je posteriorna porazdelitev kompromis med apriorno porazdelitvijo ter funkcijo verjetja. Posteriorna porazdelitev se nahaja med obema krivuljama, vendar je bliže apriorni porazdelitvi, saj smo s apriorno porazdelitvijo ponazorili večje število »poskusov«, kot s funkcijo verjetja, ki se nanaša na zadnjo predvolilno anketo. Pomembno opažanje je tudi, da je posteriorna porazdelitev višja ter ožja kot njena faktorja, saj združuje informacijo iz obeh virov ter tako zmanjšuje negotovost glede ocenjevanega parametra.

Slika 3.4: Apriorna porazdelitev, normalizirana funkcija verjetja ter posteriorna porazdelitev.



(povzeto po Lynch 2007: 58)

Tretji korak (izračun posteriorne porazdelitve) je bil v obravnavanem primeru zaradi konjugiranosti zelo enostaven. Preostane nam še zadnji korak: statistično sklepanje. Pri odgovoru na zastavljeno raziskovalno vprašanje moramo izračunati določene opisne statistike posteriorne porazdelitve. Ker ima posteriorna porazdelitev znano obliko (funkcija beta), z uporabo izrazov (2.8) in (2.9) izračunamo, da je njeno povprečje 0,497, varianca pa 0,00008283 (kar da standardni odklon 0,0091).

Če smo pripravljeni predpostaviti, da lahko porazdelitev beta dobro aproksimiramo z normalno porazdelitvijo, lahko na podlagi normalnosti vzorčne porazdelitve zapišemo 95 % interval ter zaključimo, da ocenjevani delež volivcev kandidata A pade med 0,479 ter 0,515 ($0{,}497 \pm 1{,}96 \cdot 0{,}0091$).

Vendar pa predpostavka normalnosti ni potrebna: vzorčna porazdelitev ima znano obliko (funkcija beta), zato lahko 95 % verjetnostni interval izračunamo tako, da najdemo vrednosti, ki omejujeta želeni interval: najti moramo vrednost, pod katero pade 2,5 % porazdelitve, ter vrednost, nad katero je 2,5 % porazdelitve. To sta vrednosti 0,479 ter 0,514, ki se od mej intervala pod predpostavko normalnosti le malo razlikujeta (Lynch 2007: 58). Tako kot v prejšnjem poglavju, tudi tokrat torej ne moremo trditi, da bo zmagal kandidat A, saj interval vsebuje vrednost 0,5.

Bayesov pristop k statističnemu sklepanju pa nam omogoča več kot klasični: s klasičnim sklepanjem lahko hipotezo, da bo zmagal kandidat A, zgolj potrdimo, ali pa je glede na rezultat ne moremo potrditi. Ne moremo pa odgovoriti na vprašanje, *kakšna* je verjetnost, da bo kandidat A zmagal. Ko imamo v bayesianski paradigmi ocenjene

parametre posteriorne porazdelitve, je odgovor na zastavljeno vprašanje enostaven: ta verjetnost ustreza integralu posteriorne porazdelitve od 0,5 do 1. V našem primeru je verjetnost, da bo zmagal kandidat A, 0,351 (Lynch 2007: 58-59).

3.3.4 Posteriorna porazdelitev predvidenih podatkov

Ko smo ocenili parametre modela, ki naj bi bil ustvaril vzorčne podatke, želimo včasih napovedati prihodnje vrednosti, ki izhajajo iz tega modela – pri danih podatkih želimo torej generirati morebitne prihodnje odločitve volivcev.

Porazdelitev prihodnjih vrednosti imenujemo posteriorna porazdelitev predvidenih podatkov (*angl*. posterior predictive distribution) in ima naslednjo obliko:

$$f(\tilde{y} \mid y) = \int f(\tilde{y} \mid \theta) f(\theta \mid y) d\theta \, .$$

V zgornjem izrazu $\tilde{y}$ zaznamuje predvidene vrednosti. Levi člen v integralu se nanaša na porazdelitev predvidenih vrednosti pri danih modelskih parametrih, v desnemu pa prepoznamo posteriorno porazdelitev pri danih podatkih prvotnega vzorca.

V tem poglavju smo torej prikazali osnove Bayesovega pristopa k statističnemu sklepanju. Sámo sklepanje v zadnjem koraku vedno izvedemo tako, da posteriorno porazdelitev opišemo z opisnimi statistikami, za kar pa moramo izračunati določene integrale posteriorne porazdelitve (glej razdelek 3.1.3). V obravnavanem primeru nam integriranje ni predstavljalo velike težave: uporabili smo verjetju konjugirano apriorno funkcijo in tako dobili posteriorno porazdelitev z znano funkcijsko obliko o kateri »vemo tako rekoč vse« – izpeljave za njeno povprečje, varianco ter kvantile so na voljo npr. v matematičnem priročniku.

Integriranje posteriorne porazdelitve po *analitični* poti pa je v splošnem lahko težavno. Za namen opisovanja posteriorne porazdelitve z opisnimi statistikami so zato zagovorniki Bayesovega pristopa razvili primerne *numerične* metode, ki jih bomo opisali v naslednjem poglavju.

3.4 Simulacijske metode

V prejšnjem poglavju smo v primeru predvolilne ankete uporabili splošno znane integrale porazdelitve beta. Za mnogo porazdelitev, zlasti večrazsežnostnih, pa je integrale težko izračunati. V nekaterih primerih za apriorno porazdelitev ne moremo

predpostaviti verjetju konjugirane funkcijske oblike, včasih pa lahko v modelu nastopajo matematične funkcije, katerih integralov na enostaven način ne moremo izračunati (Lynch 2007: 79).

Opisne statistike zato izračunamo s *simulacijskimi metodami*: iz dane posteriorne porazdelitve generiramo (simuliramo) vzorec velikosti *n*, nato pa iz tako pridobljenih podatkov z diskretnimi formulami izračunamo približke integralom, ki jih potrebujemo.

Za trenutek se vrnimo k primeru iz prejšnjega poglavja. Posteriorna porazdelitev parametra, ki smo jo izpeljali, je imela obliko funkcije beta s parametroma $\alpha = 1498$ in $\beta = 1519$. Za to funkcijo imamo na voljo splošno znani formuli za povprečje ter varianco (izraza (2.8) in (2.9)). Denimo pa, da omenjenih količin za dani primer ne bi znali izračunati: iz porazdelitve bi lahko simulirali dovolj velik vzorec (npr. vzorec velikosti 5000 enot z uporabo ukaza `rbeta(5000,1498,1519)` v programskem jeziku R) ter izračunali njegovo povprečje in varianco (Lynch 2007: 80). Omenjeno simulacijo smo izvedli ter dobili za povprečje vrednost 0,496 za varianco pa 0,00008199. Obe vrednosti sta torej dobra približka analitično izračunanima vrednostma 0,497 ter 0,00008283.

Z uporabo simulacije v postopek statističnega sklepanja vnašamo *simulacijsko* napako. Izračunani približki integralov namreč temeljijo na vzorcu in imajo s tem povezano *vzorčno napako*. Le-ta pa v principu ne predstavlja pomembne slabosti: standardno napako simuliranega vzorca lahko z večanjem velikosti vzorca poljubno zmanjšamo zato so v limiti omenjeni približki ekvivalentni analitično izračunanim vrednostim (Lynch 2007: 80).

Ker je integriranje porazdelitev v splošnem lahko težavno, se torej poslužujemo simulacijskih metod vzorčenja iz posteriornih porazdelitev. Vendar pa tudi za simuliranje vzorcev iz porazdelitev ne obstajajo vedno vzpostavljeni postopki (kot npr. zgoraj omenjena funkcija `rbeta`).

Simuliranje vzorcev je v splošnem enostavno za enorazsežnostne porazdelitve (glej npr. Lynch 2007: 80-88). Stvari pa se zapletejo, ko želimo simulirati vzorec iz *večrazsežnostne* porazdelitve. Te razmere so v praksi zelo pogoste, saj ima npr. že enorazsežnostna normalna porazdelitev dva parametra, posteriorna porazdelitev njenih parametrov pa je torej dvorazsežnostna.

Širšo rabo bayesianskih metod so omogočili šele nedavni napredki na področju simuliranja iz večrazsežnostnih porazdelitev z rabo metod MCMC (Monte Carlo Markovskih verig, *angl*. Markov Chain Monte Carlo). Le-te ciljno porazdelitev razbijejo na bolj obvladljive, navadno enorazsežnostne, komponente ter vzorčijo iz njih. Najbolj široko uporabljena algoritma MCMC sta Gibbsovo vzorčenje (*angl*. Gibbs sampling; glej Lynch 2007: 88-105) ter metoda Metropolis-Hastings (Lynch 2007: 107-129).

Med metode MCMC spada tudi metoda *plemenitenja podatkov* (*angl*. Data Augmentation), ki jo bomo predstavili v nadaljevanju ter uporabili v empiričnem delu naloge. Najprej pa se bomo vrnili na tematiko zlivanja podatkov ter jo obravnavali s pomočjo konceptov, ki smo jih vpeljali v tem poglavju.

4 Parametrične metode zlivanja podatkov

Metode, ki jih bomo predstavili v tem poglavju, so metode za vstavljanje manjkajočih vrednosti. S problemom manjkajočih vrednosti se srečamo npr. v anketni metodologiji, kadar anketiranci zavrnejo odgovor na katerega od anketnih vprašanj. Ker navadno ne želimo zavreči vseh enot, ki jim manjka vrednost na kateri od spremenljivk, je ena od rešitev *vstavljanje* (*angl*. imputation) primernih vrednosti.

Zlivanje podatkov lahko obravnavamo kot poseben primer vstavljanja manjkajočih vrednosti. Za razliko od neparametričnih metod datotekama A in B ne bomo dodelili vloge prejemnika oz. darovalca, ampak bomo podatke obeh vzorcev združili v eno datoteko (datoteko B položili pod datoteko A kot to prikazuje Tabela 2.1) nato pa naenkrat poustvarili vse »manjkajoče« vrednosti (osvetljene celice v Tabeli 2.1).

Najprej bomo predstavili glavne ideje modelskega pristopa k reševanju problema manjkajočih vrednosti ter omenili EM algoritem, nato pa bomo obravnavali še Bayesov pristop k tej problematiki.

4.1 Problem manjkajočih vrednosti

Posledica zavrnitve odgovora na anketno vprašanje je manjkajoča vrednost v podatkovni matriki. V celotni podatkovni matriki, ki jo bomo označili z *Y,* lahko zato razločujemo med vrednosti, za katere razpolagamo z odgovori (označimo jih z Y_{obs}), ter manjkajočimi vrednostmi (Y_{mis}). Vrednosti Y bomo imenovali tudi *popolni* podatki, vrednosti Y_{obs} pa *opaženi* podatki.

Kot v razdelku 3.2 bomo tudi pri reševanju problema manjkajočih vrednosti predpostavili, da je podatke generiral model z določeno funkcijsko obliko (npr. večrazsežnostna normalna porazdelitev). Pod predpostavko IID lahko verjetnost, da imamo prav podatke, ki so se pojavili v našem vzorcu, zapišemo kot:

$$p(Y \mid \theta) = \prod_{i=1}^{m+n} f(y_i \mid \theta), \tag{3.1}$$

pri čemer θ označuje parametre modela, izraz $f(y_i \mid \theta)$ pa verjetnostno funkcijo ene enote (vrstice v podatkovni matriki)(Schafer 1997: 10). Naš cilj je določiti vrednosti parametrov θ, težavo pa povzroča dejstvo, da zaradi manjkajočih vrednosti nimamo na

voljo popolnih podatkov Y, ki nastopajo v zgornjem izrazu, pač pa opažene podatke Y_{obs}.

4.1.1 Mehanizem manjkajočih vrednosti

Tako kot smo to storili za podatke, bomo tudi za mehanizem, ki povzroča manjkajoče vrednosti, postavili ekspliciten model. Najprej moramo definirati *indikator odzivnosti R*: to je matrika, ki za vsako spremenljivko vsake enote označuje, ali je njena vrednost opažena ali manjkajoča. Če je vrednost manjkajoča, ima indikator odzivnosti R vrednost nič, sicer pa vrednost ena. Ko gre za zlivanje podatkov, je torej R matrika enic iste dimenzije kot Tabela 2.1, z ničlami na mestih, ki so v Tabeli 2.1 osvetljena.

Mehanizem manjkajočih vrednosti lahko opišemo z naslednjo verjetnostjo:

$$p(R \mid Y, \xi).$$

Gre torej za verjetnost, da smo pri danih podatkih Y ter parametrih modela manjkajočih vrednosti ξ opazili prav takšno pojavitev manjkajočih vrednosti, kot jo imamo v podatkih. Funkcijska oblika samega modela manjkajočih vrednosti ter parametri ξ, ki ga določajo, pravzaprav niso pomembni: v nadaljevanju nas bo zanimalo le, kdaj lahko mehanizem manjkajočih vrednosti zanemarimo.

Gre namreč za to, da bomo porazdelitev opaženih podatkov obravnavali kot *robno porazdelitev* porazdelitve popolnih podatkov (glej razdelek 3.1.4). Porazdelitev opaženih podatkov Y_{obs} dobimo tako, da porazdelitev Y *integriramo preko manjkajočih vrednosti*:

$$p(Y_{obs} \mid \theta, \xi) = \int p(Y \mid \theta, \xi) dY_{mis} \,. \tag{3.2}$$

Če v zgornjem izrazu ne bi bilo motilnega parametra ξ, bi veljalo, da je porazdelitev opaženih podatkov sorazmerna s porazdelitvijo popolnih podatkov. V naslednjem razdelku se zato sprašujemo, kdaj je parameter ξ dovoljeno zanemariti ter porazdelitev Y_{obs} izraziti z integralom preko manjkajočih vrednosti.

4.1.2 Zanemarljivost mehanizma manjkajočih vrednosti

Izkaže se, da morata biti za zanemarljivost mehanizma manjkajočih vrednosti izpolnjena dva pogoja: ločenost parametrov ter lastnost MAR (Schafer 1997: 10-12).

Koncept ločenosti (*angl*. distinctness) se nanaša na parametra θ in ξ. Ločenost teh parametrov pomeni, da mehanizem manjkajočih vrednosti ni relevanten za sklepanje o parametrih, ki opisujejo popolne podatke (Rässler 2002: 76), oz. da poznavanje enega parametra pove malo o drugem. Z bayesianskega gledišča to pomeni, da mora biti mogoče skupno apriorno porazdelitev parametrov θ in ξ razstaviti na neodvisni robni apriorni porazdelitvi za θ ter ξ (Rubin 1976: 585).

Pojem MAR (*angl*. Missing At Random) je v analizo nepopolnih podatkov vpeljal Rubin (glej Rubin 1987). Za nadaljnjo razpravo je pomembno zgolj, da je lastnost MAR v primeru zlivanja podatkov vedno izpolnjena, saj manjkajoče vrednosti ustrezajo anketnim vprašanjem, ki niso bila vprašana. Gre torej za *načrtovane* manjkajoče vrednosti (*angl*. missing by design) za katere lastnost MAR vedno drži (Rässler 2002: 76).

Lastnost MAR sicer pomeni, da verjetnost, da določena vrednost manjka, ni odvisna od Y_{mis}, ampak le od Y_{obs} (Schafer 1997: 10). Matematično to lastnost izrazimo z naslednjo enačbo:

$$p(R \mid Y, \xi) = p(R \mid Y_{obs}, \xi).$$

Če sta izpolnjeni tako ločenost, kot lastnost MAR, pravimo, da je mehanizem manjkajočih vrednosti *zanemarljiv* (*angl*. ignorable), saj lahko v tem primeru izraz (3.2) zapišemo kot:

$$\begin{aligned} p(Y_{obs} \mid \theta, \xi) &= \int p(Y \mid \theta, \xi) dY_{mis} \\ &= \int p(Y \mid \theta, \xi) \cdot p(Y \mid \theta) dY_{mis} \\ &= p(R \mid Y_{obs}, \xi) \int p(Y \mid \theta) dY_{mis} \\ &= p(R \mid Y_{obs}, \xi) p(Y_{obs} \mid \theta). \end{aligned} \tag{3.3}$$

V zgornjem izrazu je člen, ki vsebuje parameter θ, popolnoma ločen od člena, ki vsebuje parameter ξ. Če je mehanizem manjkajočih vrednosti zanemarljiv, lahko torej zapišemo:

$$L(\theta \mid Y_{obs}) \propto p(Y_{obs} \mid \theta).$$

Povzemimo bistvo dosedanje razprave: začeli smo s postavitvijo modela za popolne podatke (izraz (3.1)). Le-ta izraža razmerje med parametri θ ter podatki Y, zato bi lahko parametre θ ocenili npr. z metodo največjega verjetja (glej razdelek 3.2). Težavo predstavlja dejstvo, da imamo namesto popolnih podatkov Y, ki nastopajo v modelu, na razpolago le opažene podatke Y_{obs}. Izkaže se, da lahko ob določenih tehničnih predpostavkah, ki pa so v primeru zlivanja podatkov vedno izpolnjene, namesto popolnih podatkov Y uporabimo kar opažene podatke Y_{obs}.

4.1.3 EM algoritem

Postopek bi torej lahko nadaljevali tako, da bi ocenili parametre θ, nato pa predpostavljeni model ob ocenah parametrov uporabili za vstavljanje manjkajočih vrednosti. V praksi pa navadno postopamo drugače: če lahko predpostavimo, da je mehanizem manjkajočih vrednosti zanemarljiv, lahko uporabimo iterativni postopek, ki izračuna parametre θ ter *istočasno* vstavi manjkajoče vrednosti. Gre za t.i. EM algoritem (*angl*. Expectation-Maximization, glej Dempster in drugi 1977), ki izkorišča povezanost med manjkajočimi vrednostmi Y_{mis} ter modelskimi parametri θ: Y_{mis} vsebuje informacijo, relevantno za ocenjevanje θ, po drugi strani pa nam θ pomaga najti primerne vrednosti Y_{mis} (Schafer 1997: 38).

Zaradi te medsebojne odvisnosti lahko uporabimo naslednji postopek: na podlagi začetne ocene parametrov θ vstavimo konkretne vrednosti na mesto manjkajočih; parametre θ ponovno ocenimo na podlagi popolnjenih podatkov ter postopek nadaljujemo dokler ocene ne konvergirajo (Schafer 1997: 37).

EM algoritem je pri večini preprostih problemov izredno stabilen in zanesljiv, vendar pa tako kot pri vseh optimizacijskih postopkih tudi pri EM algoritmu ni zagotovila, da bo vedno konvergiral k enemu samemu (*angl*. unique) globalnemu maksimumu. Schafer navaja različne anomalije, ki se lahko pojavijo v postopku (1997: 51-54), mi pa bomo omenili le primer, ko ima funkcija verjetja *greben* (*angl*. likelihood ridge).

Opišimo najprej konkreten primer modela, ki naj bi generiral podatke. Tak model je npr. večrazsežnostni normalni model, ki ga bomo uporabili v empiričnem delu naloge: postulirali bomo, da je naše podatke generirala večrazsežnostna normalna porazdelitev. Le-ta je v splošnem določena z velikim številom parametrov θ: iz podatkov je treba

oceniti *povprečje* vsake spremenljivke ter njen *standardni odklon*, pa tudi vse *korelacije* med spremenljivkami[9].

Kadar imamo med podatki spremenljivke, ki niso nikoli opazovane skupaj, so njihove medsebojne korelacije nedoločljive (*angl*. inestimable), funkcija verjetja pa maksimuma ne doseže v eni točki, temveč na naboru povezanih vrednosti, ki mu pravimo *greben* funkcije verjetja (Schafer 1997: 52-53).

Za problem zlivanja podatkov so značilne prav opisane razmere z nedoločljivimi korelacijami (razen, če uporabljamo dodaten zunanji vir podatkov). Posledica je, da EM algoritem konvergira k različnim točkam na grebenu, kadar ga poženemo iz različnih začetnih vrednosti. Ocenjena vrednost nedoločljivih korelacij torej ni odvisna od podatkov, ki jih imamo, temveč od začetnih vrednosti, s katerimi poženemo EM algoritem (za konkreten primer glej Rässler 2002: 79-84).

Da bi ocenjevanje parametrov stabilizirali in prišli do ene same rešitve, moramo v podatke vključiti določeno število enot z vrednostmi na vseh spremenljivkah, ali pa v postopek ocenjevanja vpeljati določeno apriorno informacijo (Rässler 2002: 84). Slednjo možnost, ki je značilna za Bayesov pristop, si bomo ogledali v naslednjem poglavju.

4.2 Bayesov pristop k problemu manjkajočih vrednosti

Kot rečeno, se pri Bayesovem pristopu vse statistično sklepanje vrti okrog posteriorne porazdelitve modelskih parametrov pri danih podatkih, ki jo lahko ob upoštevanju oznak iz razdelka 4.1 zapišemo kot:

$$p(\theta, \xi \mid Y_{obs}, R) \propto p(R, Y_{obs} \mid \theta, \xi) \cdot \pi(\theta, \xi),$$

kjer izraz $\pi(\theta, \xi)$ označuje skupno apriorno porazdelitev parametrov θ in ξ. Pod predpostavko, da velja lastnost MAR, lahko v zgornji izraz vstavimo izraz (3.3)[10]:

$$p(\theta, \xi \mid Y_{obs}, R) \propto p(R \mid Y_{obs}, R) p(Y_{obs} \mid \theta) \cdot \pi(\theta, \xi).$$

[9] Če je v modelu n spremenljivk, je torej število parametrov, ki jih je treba oceniti, $2n + n(n\text{-}1)/2$.

[10] V tem razdelku sledimo izpeljavi, podani v Schafer (1997: 17).

Za sklepanje o parametrih θ (ločeno od parametrov modela manjkajočih vrednosti ξ) pa potrebujemo *robno* porazdelitev, ki jo dobimo z integriranjem zgornje porazdelitve preko motečega parametra ξ (*angl*. nuisance parameter). Tako kot v razdelku 4.1 bomo tudi v naslednjih odstavkih pokazali, kdaj smemo *integrirati čez manjkajoče vrednosti* torej kdaj lahko varno zanemarimo mehanizem manjkajočih vrednosti.

Če sta parametra θ in ξ ločena, velja, da lahko njuno apriorno porazdelitev razstavimo na:

$$\pi(\theta,\xi) = \pi_\theta(\theta)\cdot\pi_\xi(\xi)\,.$$

Želeno robno porazdelitev parametra θ lahko zato pod predpostavkama MAR in ločenosti parametrov zapišemo kot:

$$\begin{aligned} p(\theta \mid Y_{obs}, R) &= \int p(\theta,\xi \mid Y_{obs}, R)d\xi \\ &\propto p(Y_{obs} \mid \theta)\pi_\theta(\theta)\cdot\int p(R \mid Y_{obs},\xi)\pi_\xi(\xi)d\xi \\ &\propto L(\theta \mid Y_{obs})\pi_\theta(\theta). \end{aligned} \tag{3.4}$$

Ker v zgornjem izrazu ne nastopa parameter ξ, smo s tem pokazali, da lastnosti MAR in ločenost tudi v bayesianski paradigmi pomenita zanemarljivost mehanizma manjkajočih vrednosti.

Izraz (3.4) se nanaša na posteriorno porazdelitev modelskih *parametrov*: iz te porazdelitve lahko simuliramo vzorec naključnih vrednosti parametrov. Cilj zlivanja podatkov pa je najti primerne vrednosti za »manjkajoče« podatke. Ko ocenimo posteriorno porazdelitev parametrov, lahko manjkajoče vrednosti simuliramo iz posteriorne porazdelitve predvidenih podatkov (glej razdelek 3.3.4):

$$p(\tilde{Y} \mid Y_{obs}) = \int p(\tilde{Y} \mid \theta)p(\theta \mid Y_{obs})d\theta\,.$$

Vendar pa tudi v bayesianski paradigmi navadno za vstavljanje manjkajočih vrednosti uporabimo numerično metodo: v razdelku 4.2.2 bomo predstavili postopek, analogen EM algoritmu, pred tem pa bomo nekaj besed namenili konceptu večkratnega vstavljanja.

4.2.1 Večkratno vstavljanje

Enostavni postopki vstavljanja manjkajočih vrednosti (kot npr. metoda najbližjega soseda) ustvarijo datoteko, v kateri zapolnijo manjkajoče vrednosti. Kot smo že omenili, je ta rezultat zavajajoč, saj s postopkom vstavljanja manjkajočih vrednosti (oz. zlivanja podatkov) nismo *ustvarili* manjkajoče informacije, o pravilnosti vstavljenih vrednosti pa nismo tako prepričani, kot to izraža popolnjena datoteka.

Postopek večkratnega vstavljanja (*angl*. multiple imputation) je predlagal Rubin (1987), ideja tega postopka pa je naslednja: manjkajoče vrednosti ustvarimo s simuliranjem iz posteriorne porazdelitve predvidenih podatkov[11], datoteke pa ne popolnimo le enkrat, temveč *večkrat* (Rässler 2002: 88).

Razlike med ustvarjenimi popolnimi datotekami izražajo negotovost glede vstavljenih vrednosti: če so si datoteke med seboj zelo različne, to kaže na veliko negotovost glede vstavljenih vrednosti. Rezultat večkratnega vstavljanja je zato manj zavajajoč, saj ne daje vtisa, da je postopek ustvaril manjkajočo informacijo, temveč ponazarja negotovost (Rubin 1988).

4.2.2 Plemenitenje podatkov

Tako kot ocenjevanje manjkajočih vrednosti po metodi največjega verjetja v praksi raje izvedemo s pomočjo EM algoritma, si tudi v bayesianski paradigmi pomagamo z numeričnim postopkom, algoritmom plemenitenja podatkov[12] (*angl*. Data Augmentation).

Postopek poteka v dveh korakih (Rässler 2002: 107):

- najprej manjkajoče vrednosti simuliramo iz posteriorne porazdelitve predvidenih podatkov $p(Y_{mis} \mid Y_{obs}, \theta^{(t)})$ ob konkretnih vrednostih parametrov $\theta^{(t)}$;
- nato neznane parametre simuliramo iz posteriorne porazdelitve *popolnih* podatkov ob trenutnih vstavljenih vrednostih $p(\theta \mid Y_{obs}, Y_{mis}^{(t)})$.

[11] Rubinov postopek pravzaprav dovoljuje simuliranje manjkajočih vrednosti iz *katerekoli* verjetnostne porazdelitve, ki opisuje verjetnost simuliranih podatkov pri danih opaženih podatkih (Rässler 2002: 88).

[12] Ime izhaja iz rabe algoritma DA za Bayesiansko sklepanje ob nepopolnih podatkih. V mnogo primerih je delo s posteriorno porazdelitvijo $p(\theta|Y_{obs})$ zaradi manjkajočih vrednosti težavno. Če pa opažene podatke Y_{obs} »oplemenitimo« (*angl*. to augment) s predvidenimi vrednostmi Y_{mis}, je dobljena posteriorna porazdelitev popolnih podatkov mnogo bolj obvladljiva (Schafer 1997: 72).

Ob primernih začetnih vrednostih ta postopek opredeljuje Markovsko verigo in pod dokaj splošnimi pogoji konvergira k *stacionarni porazdelitvi* (Rässler 2002: 107-108). Za razliko od EM algoritma, ki konvergira k *točki* v vzorčnemu prostoru parametrov, za algoritme MCMC namreč pravimo, da so konvergirali, ko se simulirane porazdelitve ustalijo in algoritem vzorči iz stacionarne *porazdelitve* (Schafer 1997: 80).

Konvergenco algoritmov MCMC pa lahko definiramo tudi na drugačen način: kot odsotnost avtokorelacij simuliranih vrednosti (Schafer in Olsen 1998: 16). Gre namreč za to, da je v Markovski verigi vsaka simulirana vrednost nekega parametra odvisna od predhodne vrednosti (zato tudi ime *veriga*). Če bi za večkratno vstavljanje uporabili kar vrednosti parametrov iz *zaporednih* iteracij, bi s tem dobili pet popolnjenih datotek, ki bi se zaradi avtokorelacije med seboj razlikovale v manjši meri kot bi se sicer. Pravilno večkratno vstavljanje pa zahteva *neodvisne* vstavljene vrednosti, zato vstavljanje manjkajočih vrednosti Y_{mis} v praksi izvedemo tako, da ne vzamemo zaporednih vrednosti parametrov v Markovski verigi, ampak vrednosti, ki so razmaknjene za določeno dovolj visoko število iteracij (za podrobnosti glej Lynch 2007: 131-153).

4.2.3 Plemenitenje podatkov v programu NORM

V empiričnem delu naloge bomo plemenitenje podatkov izvedli v programu[13] NORM (Schafer 1999), ki predpostavlja, da podatki, ki mu jih posredujemo, izhajajo iz večrazsežnostne normalne porazdelitve. Tehnične podrobnosti podrobno navaja Schafer (1997: 147-163), na tem mestu pa zgolj ponovimo, da večrazsežnostno normalno porazdelitev v splošnem določa veliko število parametrov: n-razsežnostna normalna porazdelitev je določena z n povprečji, n standardnimi odkloni ter $n(n - 1)/2$ korelacijami spremenljivk. Ti parametri (povprečja, standardni odkloni ter korelacije) skupaj predstavljajo parametre modela, ki smo jih doslej označevali s θ.

Pri uporabi programa NORM bomo upoštevali napotke, navedene v Schafer in Olsen (1998), ter postopali po naslednjih korakih:

1. Na podatkih bomo pognali EM algoritem ter shranili ocene parametrov.

[13] Za temeljit pregled alternativne programske opreme za vstavljanje manjkajočih vrednosti glej npr. Harel in Zhou (2006) ali Horton in Kleinman (2007).

2. Te vrednosti bomo uporabili kot začetne vrednosti algoritma DA, ki ga bomo pognali za nekaj tisoč iteracij ter shranili vrednosti vseh parametrov v vseh iteracijah.
3. Shranjene vrednosti bomo proučili in določili tisto dovolj visoko število iteracij *k*, ki jih mora algoritem opraviti, da so simulirane vrednosti parametrov neodvisne.
4. Algoritem DA bomo nato pognali še enkrat ter manjkajoče podatke vstavili vsakih *k* iteracij.

Če imamo na voljo dodatno datoteko z vrednostmi na vseh spremenljivkah, lahko v stopnji plemenitenja podatkov uporabimo kar neinformativno apriorno porazdelitev: zaradi dodatnih popolnoma opazovanih enot je moč namreč iz podatkov oceniti tudi korelacije, ki bi bile brez njih nedoločljive.

Drugače pa moramo postopati, kadar imamo v podatkih spremenljivke, ki niso nikoli opazovane skupaj. Da bi algoritem plemenitenja podatkov deloval stabilno, moramo uporabiti *informativno* apriorno porazdelitev: v programu NORM imamo na voljo t.i. *grebensko* apriorno porazdelitev (*angl*. ridge prior; za podrobnosti glej Schafer 1997: 151-152). Preko le-te v model vpeljemo apriorno »informacijo«, da so modelske spremenljivke brezpogojno neodvisne (Rässler 2002: 173, 197). Ta »informacija« je seveda prav lahko napačna, vendar pa to ni bistvenega pomena: apriorno porazdelitev uporabljamo le kot izhod v sili, da stabiliziramo algoritem plemenitenja podatkov. V postopek lahko vpeljemo *kakršno koli* apriorno informacijo, ter ji damo v primerjavi s podatki tako majhno težo, da apriorna porazdelitev na končno rešitev ne bo imela tako rekoč nobenega vpliva. Sama funkcijska oblika informativne apriorne porazdelitve za zlivanje podatkov torej ni pomembna, zato je ustrezna tudi omenjena rešitev v programu NORM (Rässler 2002: 110).

Z opisom algoritma plemenitenja podatkov ter programa NORM zaključujemo obravnavo Bayesovega pristopa k problemu manjkajočih podatkov. Preden metode zlivanja podatkov uporabimo na konkretnih podatkih, pa namenjamo nekaj besed še stopnjam veljavnosti, ki jih z metodami zlivanja podatkov lahko dosežemo.

4.3 Stopnje veljavnosti zlivanja podatkov

Kot smo omenili, je cilj zlivanja podatkov ustvariti popolnjeno datoteko, ki vsebuje informacijo o vseh spremenljivkah. Uporabljeni postopek zlivanja je lahko pri tem bolj ali manj uspešen, kar je v veliki meri odvisno od kakovosti skupnih spremenljivk, ki so na voljo. Avtorica Susanne Rässler (2002: 29-33) predlaga ločevanje med štirimi stopnjami veljavnosti, ki jih postopek zlivanja podatkov lahko doseže.

4.3.1 Prva stopnja: ohranitev posameznih vrednosti

Prvo stopnjo veljavnosti dosežemo, če s postopkom zlivanja podatkov uspešno poustvarimo pravilne (a neznane) vrednosti zlitih spremenljivk na enotah. Pravilno poustvarjeno vrednost imenujemo *zadetek*, v prejemni datoteki pa lahko izračunamo *stopnjo zadetkov* (*angl*. hit-rate). Avtorica posvari, da nas stopnja zadetkov lahko zavede, če jo uporabljamo kot kriterij uspešnosti zlivanja podatkov. Njen argument ponazorimo s primerom:

Imamo kovanec, ki je v resnici nepošten, saj v 60 % metov pade glava. Ta »pravilni« model imenujmo model A, primerjamo pa ga z modelom B, ki predpostavlja, da ima kovanec na obeh straneh glavo. Če za zlivanje podatkov (napovedovanje rezultata prihodnjih metov) uporabimo model A, dobimo stopnjo zadetkov $0{,}6 \cdot 0{,}6 + 0{,}4 \cdot 0{,}4 = 0{,}52$; z modelom B pa stopnjo zadetkov 0,60. Kljub višji stopnji zadetkov seveda ne moremo trditi, da je model B boljši, saj je naš cilj veljavno sklepanje o deležu glav (Rubin 1996: 475).

Pri zlivanju podatkov si torej ne prizadevamo doseči prve stopnje veljavnosti. Poustvarjene zlite vrednosti namreč ne skušajo odsevati pravih vrednosti, njihova interpretacija na nivoju mikro podatkov pa je brez pomena (Rässler 2004: 6).

4.3.2 Druga stopnja: ohranitev skupne porazdelitve

Ponovimo, da se oznaki Y in Z nanašata na specifične, X pa na skupne spremenljivke. Drugo stopnjo veljavnosti dosežemo, če nam uspe s postopkom zlivanja podatkov v zliti datoteki ustvariti pravilno skupno porazdelitev vseh spremenljivk in torej velja $\tilde{f}_{X,Y,Z} = f_{X,Y,Z}$ (s $\tilde{f}$ smo označili zlito porazdelitev, s f pa pravilno porazdelitev) (Rässler 2004: 7).

Najpomembnejši cilj zlivanja podatkov je ustvariti zlito datoteko, ki jo lahko uporabimo kot enovit (*angl*. single source) vir podatkov o porazdelitvi $f_{X,Y,Z}$. Pri tem ne želimo pravilno poustvariti posameznih vrednosti, ampak omogočiti veljavno statistično sklepanje o porazdelitvi $f_{X,Y,Z}$ (Rässler 2004: 7). Susanne Rässler pokaže, da je brez uporabe dodanih zunanjih informacij to mogoče le, kadar so specifične spremenljivke Y in Z pogojno neodvisne pri danih skupnih spremenljivkah X (2002: 21-22).

4.3.3 Tretja stopnja: ohranitev korelacijske strukture

Včasih osebo, ki zlito datoteko uporablja za analize, zanimajo le korelacije med spremenljivkami. Tretjo stopnjo veljavnosti postopek zlivanja podatkov doseže, če pravilno poustvari korelacijsko strukturo ter višje momente porazdelitev spremenljivk (Rässler 2004: 7). Pravilno morajo biti poustvarjene tudi robne frekvence, tako da velja $\tilde{f}_{Y,X} = f_{Y,X}$ ter $\tilde{f}_{Y,Z} = f_{Y,Z}$.

Brez uporabe zunanjih informacij lahko tretjo stopnjo veljavnosti zlivanja podatkov dosežemo le, kadar so spremenljivke v povprečju pogojno nekorelirane (Rässler in Fleischer 1998: 323).

4.3.4 Četrta stopnja: ohranitev robnih porazdelitev

Da imamo zlivanje podatkov lahko za uspešno, mora biti izpolnjen najmanj pogoj, da so v zliti datoteki robne porazdelitve poustvarjene pravilno. Kadar poustvarjamo vrednosti spremenljivk Y, mora torej veljati $\tilde{f}_Y = f_Y$ ter $\tilde{f}_{Y,X} = f_{Y,X}$. Razlike med porazdelitvama izvornega in zlitega vzorca ne smejo biti večje kot pričakovane razlike med dvema neodvisnima vzorcema iz iste populacije (Rässler 2004: 7-8).

V realnem primeru zlivanja podatkov lahko preverimo zgolj četrto stopnjo veljavnosti, saj nimamo podatka o pravilnih vrednostih, skupnih porazdelitvah ali korelacijski strukturi spremenljivk. V naslednjem poglavju bomo zato metode zlivanja podatkov preizkusili na podatkih, za katere bomo omenjene »pravilne rešitve« poznali vnaprej, saj nas poleg robnih porazdelitev zanima predvsem, kako uspešni bodo algoritmi pri poustvarjanju nedoločljivih korelacij.

5 Uporaba metod zlivanja podatkov

V tem poglavju bomo predstavili rezultate različnih metod zlivanja podatkov, s katerimi smo skušali rešiti isti problem. V prvem delu bomo opisali zadani problem ter statistike, s katerimi smo ocenjevali uspešnost metod.

5.1 Testni podatki ter statistike

Kot testne podatke bomo uporabili podatke, zbrane z anketo RIS-IKT, ki jo je spomladi leta 2005 v okviru projekta Raba interneta v Sloveniji (RIS) izvedel Center za metodologijo in informatiko Fakultete za družbene vede. Kot vzorčni okvir je bil uporabljen Centralni register prebivalstva Republike Slovenije iz katerega je bilo naključno izbranih 2000 oseb, starih od 10 do 74 let. Posamezniki so bili anketirani na domu, dosežena je bila 78 odstotna stopnja odgovorov. Anketna vprašanja so zadevala predvsem posameznikovo rabo informacijskih tehnologij s poudarkom na rabi mobilnega telefona (več o anketi na http://surs.ris.org; glej tudi Vehovar 2007).

Ideja preizkusa, na podlagi katerega bomo ocenili metode zlivanja podatkov, je naslednja: datoteko s podatki o anketirancih najprej umetno razdelimo na dva dela, imenujmo ju datoteki A in B. V datoteki A nato nekaj spremenljivk izbrišemo: te spremenljivke lahko torej obravnavamo, kot da so *specifične* datoteki B, saj jih datoteka A ne vsebuje. Analogno tudi v datoteki A izbrišemo določene spremenljivke ter tako ustvarimo spremenljivke, specifične datoteki B. Umetno ustvarjeni praznini v datotekah nato zapolnimo z zlivanjem podatkov ter dobljene spremenljivke primerjamo s prvotnimi. Opisani preizkus je seveda povsem pedagoške narave, saj že vnaprej poznamo ”pravilni rezultat”, prav to pa nam tudi omogoča, da ocenimo, do kakšne mere so bile preizkušene metode uspešne.

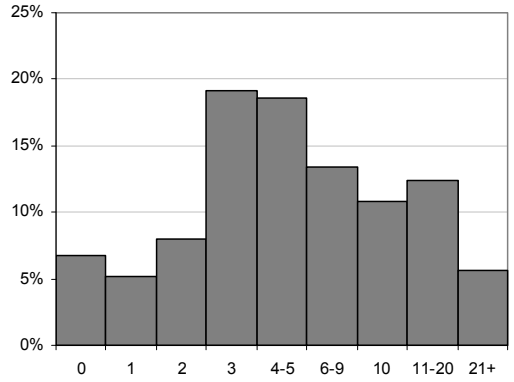
5.1.1 Opis spremenljivk

Najprej moramo torej izbrati spremenljivke, ki jih bomo obravnavali kot specifične, vsem ostalim spremenljivkam pa prepustimo vlogo potencialnih skupnih spremenljivk. Da bi bilo kakršnokoli zlivanje podatkov uspešno, morajo skupne spremenljivke do visoke mere pojasnjevati specifične. K problemu izbire primernih specifičnih spremenljivk smo zato pristopili tako, da smo za vsako spremenljivko opravili linearno regresijo na vse ostale spremenljivke ter zabeležili determinacijski koeficient. Kot

specifične spremenljivke smo izbrali spremenljivke z najvišjimi determinacijskimi koeficienti:

Število klicev prek mobilnega telefona v tipičnem delovnem dnevu (to spremenljivko v nadaljevanju imenujemo *klici*). Da bi dosegli vsaj približno normalno porazdelitev, smo spremenljivko rekodirali po dokaj arbitrarnem ključu (glej Sliko 5.1), kar je bilo potrebno zaradi zgostitev okrog okroglih vrednosti[14] (deset klicev ipd.).

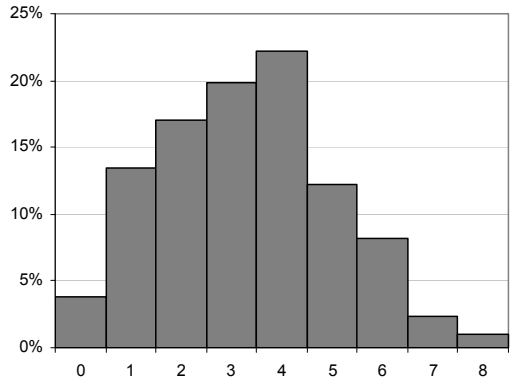
Slika 5.1: Porazdelitev spremenljivke *klici* (datoteka A, *n* = 500).



Število različnih uporabljanih storitev mobilnega telefona (odslej *storitve*). Anketiranci so odgovarjali, kako pogosto uporabljajo storitve kot npr. preusmeritev klicev, opomnik, budilka ipd. (na seznamu je bilo osem storitev, za podrobnosti glej Trček in Platinovšek 2007: 152). Anketiranec, ki je določeno storitev uporabljal *vsaj občasno*, je pri spremenljivki *storitve* dobil "točko". Anketiranci, ki uporabljajo velik nabor različnih storitev svojega mobilnega telefona, imajo torej na tej spremenljivki visoko vrednost in obratno.

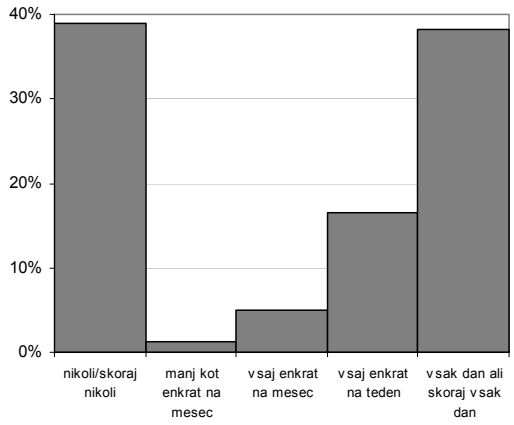
[14] Kadar mora anketiranec v spomin priklicati preteklo porabo dobrin ali storitev, je zaokroževanje pogost pojav (glej Battistin in drugi 2000).

Slika 5.2: Porazdelitev spremenljivke *storitve* (datoteka B, *n* = 500).



Pogostost rabe interneta (spremenljivka *internet*) ima pet vrednosti: od "nikoli ali skoraj nikoli" do "vsak dan ali skoraj vsak dan". Kot prikazuje Slika 5.3, se spremenljivka *internet* porazdeljuje po U-porazdelitvi, ki ni niti približno podobna normalni porazdelitvi. Ker bomo pri uporabi parametričnih metod zlivanja podatkov predpostavili normalnost, pričakujemo, da bo ta spremenljivka dober pokazatelj, kako robustne so parametrične metode na kršitev omenjene predpostavke.

Slika 5.3: Porazdelitev spremenljivke *internet* (datoteka B, *n* = 500).



V končno datoteko smo vključili le tiste skupne spremenljivke, ki dobro pojasnjujejo zgoraj opisane specifične spremenljivke: za vsako od treh specifičnih spremenljivk smo opravili linearno regresijo na vse razpoložljive skupne spremenljivke, nato pa obdržali le tiste skupne spremenljivke, ki so imele vsaj v enem od treh regresijskih modelov statistično značilen koeficient[15].

[15] Takšen pristop predlagajo D'Orazio in drugi (2007: 170). Za možne alternativne pristope pri izbiri primernih skupnih spremenljivk za zlivanje podatkov glej D'Orazio in drugi (2007: 167-171).

Opis petnajstih skupnih spremenljivk, ki smo jih s tem dobili, je naveden v Prilogi A, rezultati linearne regresije pa v Prilogi B. Determinacijski koeficient pri regresiji na omenjenih petnajst spremenljivk je za vsako od treh specifičnih spremenljivk višji od 45%. Izbrane skupne spremenljivke torej v smislu linearne regresije do visoke mere pojasnjujejo specifične spremenljivke, kar pomeni, da se lahko pri uporabi zlivanja podatkov nadejamo dobrih rezultatov.

Tabela 5.1 prikazuje korelacije med specifičnimi spremenljivkami (za celotno korelacijsko matriko glej Prilogo C). Kot smo omenili v razdelku 2.2, pa je za zlivanje podatkov pomembno, ali so specifične spremenljivke korelirane po tem, ko *korelacije kontroliramo po skupnih spremenljivkah*. V obravnavanem primeru je pomembno opažanje, da je parcialna korelacija med spremenljivkama *klici* ter *internet* majhna ter statistično neznačilna, korelacija *klici-storitve* pa se po parcializaciji zmanjša, vendar ostane pozitivna in statistično značilna.

Tabela 5.1: Matriki korelacij ter parcialnih korelacij specifičnih spremenljivk.

korelacije

	klici	storitve	internet
klici	1	0,443	0,384
storitve	0,443***	1	0,516
internet	0,384***	0,516***	1

parcialne korelacije (kontrolirano po skupnih spremenljivkah)

	klici	storitve	internet
klici	1	0,141	0,026
storitve	**0,141*** **	1	0,157
internet	**0,026**	0,157***	1

*** p < 0,01

Korelacijo *klic-internet* skupne spremenljivke v celoti pojasnijo: poenostavljeno lahko trdimo, da sta spremenljivki *klici* in *internet* pri danih skupnih spremenljivkah pogojno neodvisni[16]. Po drugi strani pa pogojna neodvisnost ne velja za spremenljivki *klici* in *storitve*, zato lahko pričakujemo, da bodo metode, ki zlivanje podatkov opravijo ob predpostavljeni pogojni neodvisnosti, korelacijo *klici-storitve* v zliti datoteki poustvarile nepravilno.

5.1.2 Delovne datoteke

Iz prvotne datoteke RIS-IKT smo obdržali 1092 enot, ki ustrezajo anketirancem, ki so navedli odgovor na vsako izmed vprašanj, ki ustrezajo specifičnim spremenljivkam[17].

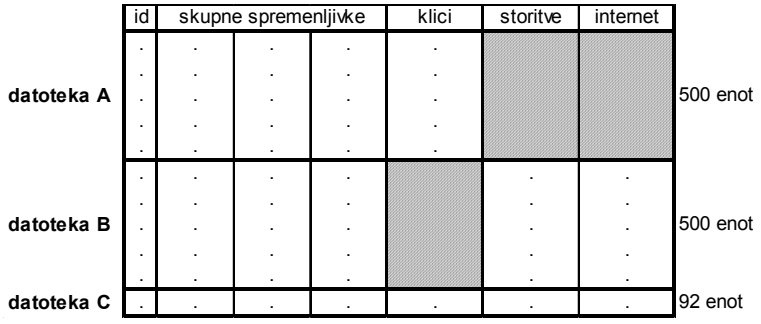
[16] Pri tem smo asociacijo spremenljivk zopet poenostavili na korelacijo, torej *linearno* povezanost.

[17] Izločili smo 908 enot, ki večinoma ustrezajo neuporabnikom mobilnega telefona, ki na vprašanja o *klicih* in *storitvah* sploh niso odgovarjali.

Pri specifičnih spremenljivkah torej ne dopuščamo manjkajočih vrednosti, manj strogi pa smo pri skupnih spremenljivkah, pri katerih manjkajoče vrednosti dopuščamo.

Datoteko smo nato razdelili na tri dele[18]: datoteko A in datoteko B s po 500 enotami, ter datoteko C s preostalimi 92 enotami. V datoteki A smo nato izbrisali spremenljivki *storitve* in *internet*, v datoteki B pa spremenljivko *klici*. V datoteki C nismo izbrisali ničesar: v nadaljevanju jo bomo obravnavali kot popoln zunanji vir informacij oz. *enovito anketo* (*angl*. single source survey) manjšega obsega. Shemo testne datoteke prikazuje Slika 5.4.

Slika 5.4: Shema datotek A, B in C.



Vsako od metod zlivanja podatkov bomo najprej uporabili zgolj na datotekah A in B – torej brez dodatnega »zunanjega« vira informacij – nato pa bomo v postopek vključili še informacijo preostalih 92 enot z vrednostmi na vseh spremenljivkah.

5.1.3 Testne statistike

Praznine v datoteki bomo torej zapolnili z različnimi metodami, s čimer bomo dobili zlite datoteke. Na njih bomo opravili naslednje analize:

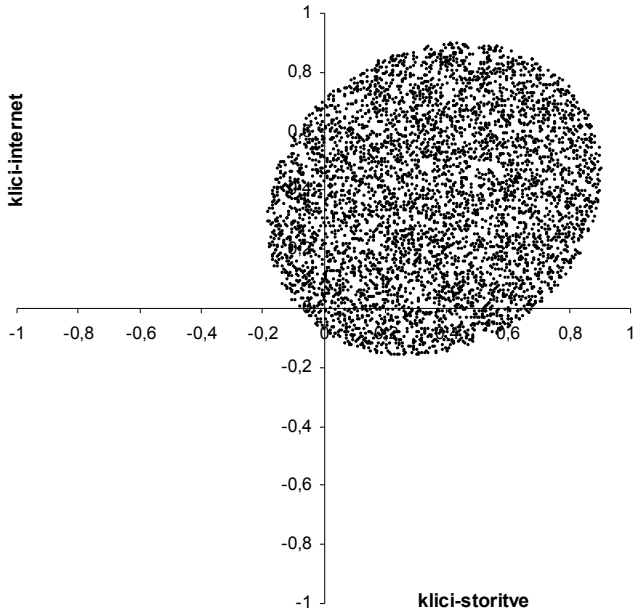
- Ugotavljali bomo, ali je metoda zlivanja podatkov ohranila *robne porazdelitve* specifičnih spremenljivk. S histogramom bomo primerjali obliko porazdelitve, za dobljeno porazdelitev bomo izračunali tudi povprečje ter standardni odklon ter ju primerjali z ustreznima parametroma ”pravilne” porazdelitve. Z omenjenimi testi preverjamo četrti nivo veljavnosti, kot ga definira Susanne Rässler (glej razdelek 4.3).

[18] Pred tem smo v datoteki naključno preuredili vrstni red enot. S tem smo izključili nevarnost, da bi se datoteke A, B in C bistveno razlikovale npr. zaradi vpliva anketarja ali koderja. V prvotni datoteki so bile enote namreč urejene glede na zaporedno številko anketarja.

- Ugotavljali bomo, ali je metoda ohranila pravilno *korelacijsko strukturo* (tretji nivo veljavnosti). Pozorni bomo predvsem na korelaciji *klici-storitve* ter *klici-internet*, pa tudi na ustrezni *parcialni* korelaciji.

Podatek o korelacijah *klici-storitve* in *klici-internet* zaradi zasnove preizkusa manjka, zato bosta prav vrednosti teh korelacij pomemben preizkusni kamen vsake metode. Omenjeni korelaciji ne moreta zavzeti povsem poljubne vrednosti z intervala [-1, 1], saj ju do neke mere omejujejo korelacije z ostalimi spremenljivkami: celotna korelacijska matrika mora biti pozitivno semi-definitna (glej razdelek 2.2 ter Prilogo C). Ob upoštevanju omenjenega pogoja lahko s pomočjo simulacije izrišemo območje vrednosti, ki jih lahko zavzameta korelaciji *klici-storitve* ter *klici-internet*[19].

Slika 5.5: Elipsa, ki omejuje dovoljene vrednosti korelacij *klici-storitve* ter *klici-internet*.



Slika 5.5 prikazuje elipso dovoljenih vrednosti korelacij *klici-storitve* ter *klici-internet*: kot smo omenili v razdelku 2.2, ima območje dovoljenih vrednosti vedno obliko elipsoida. Dovoljeno območje korelacij *klici-storitve* ter *klici-internet* je dokaj obsežno ter leži pretežno v prvem kvadrantu. Natančno na sredini elipse ležita vrednosti 0,36 (*klici-storitve*) ter 0,37 (*klici-internet*), ki ustrezata pogojni neodvisnosti (glej razdelek

[19] Sliko 5.5 smo dobili tako, da smo v programu R simulirali vrednosti obeh korelacij. Par naključno generiranih vrednosti z intervala [-1,1] smo vstavili na primerni mesti v celotni korelacijski matriki, nato pa preverili, ali je je dobljena matrika pozitivno semidefinitna. Pare korelacij, za katere ta pogoj ni bil izpolnjen, smo zavrgli. Slika prikazuje 2000 točk z dovoljenega območja.

2.2). Izračunali smo ju kot povprečje najvišje ter najnižje simulirane vrednosti vsake od spremenljivk, pri katerih je bila korelacijska matrika še pozitivno semi-definitna.

5.2 Neparametrične metode

V tem poglavju bomo na opisanih podatkih uporabili metodo najbližjega soseda (*angl*. Nearest Neighbour oz. NN), ki smo jo opisali v razdelku 2.1.2. Ker omenjena metoda v zliti datoteki vzpostavi pogojno neodvisnost spremenljivk, ki niso nikoli opazovane skupaj, bomo preizkusili še njeno različico, ki si pri preseganju pogojne neodvisnosti pomaga z datoteko zunanjih informacij (glej razdelek 2.3).

Zlivanje podatkov po metodi najbližjega soseda smo izvedli v programu R (http://www.r-project.org), uporabili pa smo algoritme avtorjev D'Orazio in drugi (2006: 223-244; v elektronski obliki dostopno na http://www.wiley.com/go/matching).

Omenjeni algoritmi kot mero podobnosti med enotami uporabljajo evklidsko razdaljo. Dejstvo, da so skupne spremenljivke v splošnem merjene na različnih lestvicah, je upoštevano tako, da se razlika na posamezni spremenljivki deli z njenim variacijskim razponom. Razdalja med prejemno enoto a ter potencialnimi darovalci se torej izračuna po naslednji formuli:

$$\mathrm{dist}(a,b) = \sqrt{\sum_{i=1}^{P} \frac{1}{\mathrm{rng}(x_i)} (x_{a,i} - x_{b,i})} \,, \tag{4.1}$$

kjer rng(x_i) označuje variacijski razpon spremenljivke x_i (razliko med max(x_i) in min(x_i)).

Avtor je kodo algoritma nekoliko spremenil, tako da je algoritem sposoben rokovati z manjkajočimi vrednostmi skupnih spremenljivk. Primer, ko ima ena od enot, med katerima se računa razdalja (ali obe), na neki skupni spremenljivki manjkajočo vrednost, algoritem obravnava, kot da se obravnavani enoti na tej spremenljivki razlikujeta v največji možni meri, torej za celoten variacijski razpon spremenljivke. Z drugimi besedami to pomeni, da v vsoti (4.1) za to spremenljivko prištejemo vrednost 1.

5.2.1 Metoda najbližjega soseda

Z metodo najbližjega soseda smo zapolnili vrzeli v testni datoteki: najprej smo enotam v datoteki A poiskali primerne darovalce iz datoteke B ter tako v datoteki A poustvarili

spremenljivki *storitve* in *internet*. Nato smo prejemnikom v datoteki B poiskali darovalce iz datoteke A in s tem zapolnili vrzel na spremenljivki *klici*.

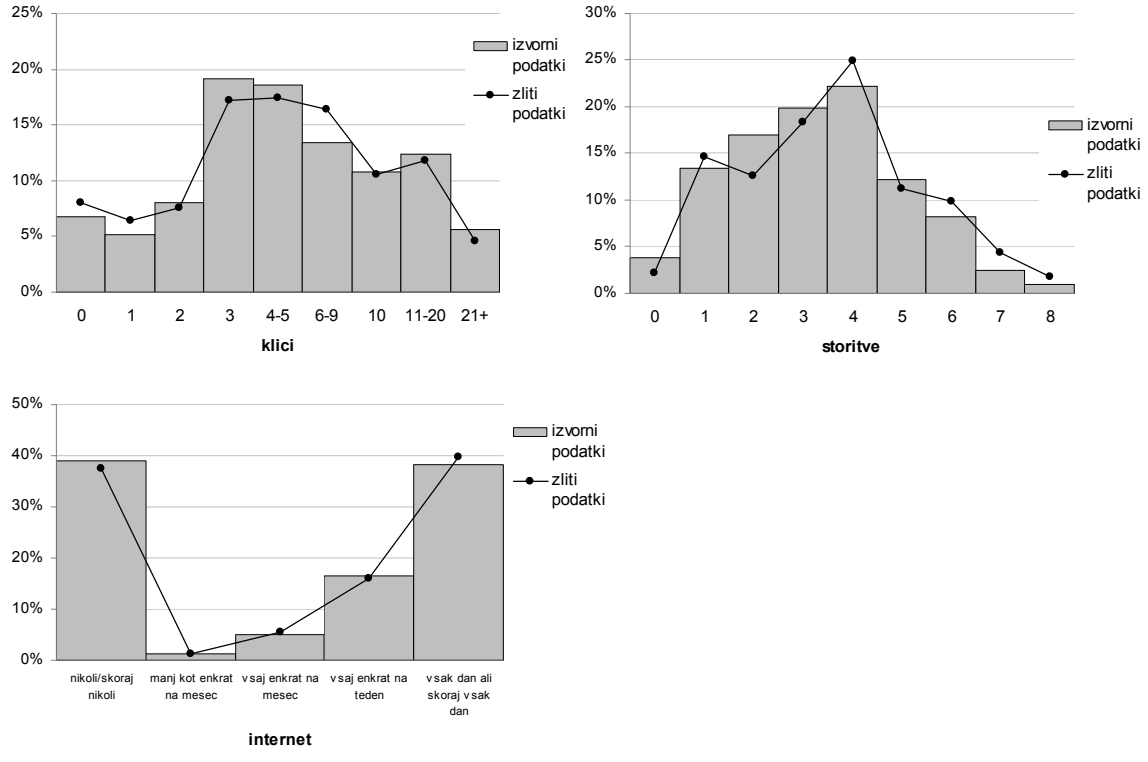
Določene enote so se izkazale kot »priljubljeni« darovalci in so bile v vlogi darovalca večkrat (največ sedemkrat). Po drugi strani zato visok delež enot nikoli ni bil v vlogi darovalca, kar je nezaželena lastnost metode NN, saj smo na ta način *zavrgli* informacijo, ki jo vsebujejo te enote. Enot, ki nikoli niso bile darovalci, je bilo približno dve petini. Tabela 5.2 prikazuje večkratno rabo enot darovalcev.

Tabela 5.2: Večkratna raba darovalcev; metoda najbližjega soseda.

	0x	1x	2x	3x	4x	5x	6x	7x
datoteka A	208	163	70	45	8	6		
datoteka B	210	161	70	44	11	2	1	1

Na Sliki 5.6 so s histogramom prikazane izvorne porazdelitve specifičnih spremenljivk, lomljene črte na njih pa označujejo podatke, ki smo jih poustvarili z metodo NN. Razen manjših odstopanj zliti podatki tesno sledijo porazdelitvi izvornih podatkov, le pri spremenljivki *storitve* opazimo večje odstopanje pri vrednosti 2. Slika 5.6 torej nakazuje, da je metoda najbližjega soseda dobro poustvarila robne porazdelitve.

Slika 5.6: Prikaz robnih porazdelitev specifičnih spremenljivk s histogrami; metoda najbližjega soseda.



Navedeno ugotovitev potrdijo tudi statistični testi[20], ki jih prikazuje Tabela 5.3. Za vsako od specifičnih spremenljivk smo primerjali povprečje in standardni odklon zlite porazdelitve z izvorno (izvorne količine so navedene v oklepaju). Izračunali smo testne statistike razlik povprečij in standardnih odklonov, pri čemer smo oba vzorca obravnavali kot slučajna. Statistično značilna je zgolj razlika povprečij za spremenljivko *storitve*, katere zlita vrednost je precenjena.

Tabela 5.3: Povprečja ter standardni odkloni specifičnih spremenljivk; metoda najbližjega soseda.

	(povp.)	povp.	t	sig	(st.dev)	st.dev	F	sig
klici	4,17	4,08	0,64	0,52	2,15	2,17	1,01	0,46
storitve	3,31	3,55	2,16	**0,03**	1,74	1,83	1,05	0,30
internet	3,14	3,19	0,47	0,64	1,80	1,80	1,00	0,49

Spodnji tabeli prikazujeta Pearsonove korelacijske koeficiente ter parcialne korelacijske koeficiente med spremenljivkami, ki niso nikoli opazovane skupaj. Ker poznamo »pravilni« vrednosti korelacij *klici-storitve* ter *klici-internet* (0,44 ter 0,38), ju lahko direktno primerjamo s poustvarjenimi. Tabela 5.4 prikazuje statistični test razlike med poustvarjenimi ter izvornimi korelacijskimi koeficienti[21].

Tabela 5.4: Korelaciji *klici-storitve* ter *klici-internet*; metoda najbližjega soseda.

izvorne (A in B)

	klici
storitve	0,44
internet	0,38

datoteka A

	klici	sig
storitve*	**0,34**	**0,06**
internet*	0,40	0,37

datoteka B

	klici*	sig
storitve	0,37	0,14
internet	0,32	0,22

Statistično značilno razliko opazimo v datoteki A pri korelaciji *klici-storitve**, ki je podcenjena (z zvezdico označujemo zlito spremenljivko). Prenos informacije iz datoteke B v datoteko A je bil torej manj uspešen kot prenos v nasprotni smeri, vendar pa je podcenjena tudi korelacija *klici**-*storitve* v datoteki B.

[20] V nadaljevanju smo pri izračunu povprečij, standardnih odklonov ter korelacij predpostavili, da so spremenljivke merjene na številski lestvici. Spremenljivki *klici* in *storitve* tako zavzemata vrednosti od 0 do 8, spremenljivka *internet* pa od 1 do 5.

[21] Korelacijske koeficiente smo transformirali s Fisherjevo transformacijo, nato pa opravili t-test. Za podrobnosti glej Rässler (2003: 70).

Navedene ugotovitve so v skladu s pričakovanji, saj vemo, da metoda najbližjega soseda vzpostavi pogojno neodvisnost. Za spremenljivki *klici* in *storitve* smo ugotovili, da sta pogojno korelirani (glej Tabelo 5.1): korelirani sta torej tudi, ko njuno korelacijo kontroliramo po skupnih spremenljivkah. Spodnja tabela prikazuje parcialne korelacije.

Tabela 5.5: Parcialni korelaciji *klici-storitve* ter *klici-internet*; metoda najbližjega soseda.

izvorne (A in B)		
	klici	sig
storitve	0,14	0,00
internet	0,03	0,40

datoteka A		
	klici	sig
storitve*	0,00	0,97
internet*	0,03	0,53

datoteka B		
	klici*	sig
storitve	-0,03	0,52
internet	-0,03	0,50

Za razliko od Tabele 5.4 se statistične značilnosti tu nanašajo na test razlike parcialnega korelacijskega koeficienta od nič. Skladno s pričakovanji ugotovimo, da je metoda najbližjega soseda v zliti datoteki vzpostavila pogojno neodvisnost med spremenljivkami, ki nikoli niso opazovane skupaj, glede na skupne spremenljivke. To je tudi razlog za podcenjeno korelacijo *klici-storitve** v Tabeli 5.4.

5.2.2 Uporaba metode NN z zunanjim virom informacij

V razdelku 2.3 smo omenili nadgraditev metode najbližjega soseda, ki z uporabo datoteke zunanjih informacij skuša preseči pogojno neodvisnost:

1. Najprej enotam datoteke A poiščemo darovalce iz datoteke C. Ker datoteka C vsebuje podatke o *vseh* spremenljivkah, lahko pri računanju razdalje poleg skupnih spremenljivk upoštevamo tudi spremenljivko *klici*.
2. V vmesni datoteki imamo zdaj *vse* spremenljivke, le da sta spremenljivki *storitve* in *internet* »informacijsko revni«, saj izhajata iz občutno manjše datoteke C, ki vsebuje le 92 enot. V drugem koraku v izračun razdalje vključimo omenjeni spremenljivki *storitve* in *internet* ter s tem v vmesno datoteko prenesemo »informacijsko bogati« spremenljivki *storitve* in *internet* iz datoteke B.

Analogen postopek nato opravimo še za prenos spremenljivke *klici* iz datoteke A v datoteko B pri datoteki zunanjih informacij C. Izkaže se, da v primerjavi z osnovno metodo NN še večji delež potencialnih darovalnih enot nikoli ni v vlogi darovalca. Delež zavrženih enot se s tem, kot kaže Tabela 5.6, približa polovici. V primerjavi z osnovno metodo najbližjega soseda metoda NN z dodatnim korakom tudi slabše

poustvari robne porazdelitve specifičnih spremenljivk. Na Sliki 5.7 opazimo večja ter bolj pogosta odstopanja lomljene črte od ustrezajočega histograma kot na Sliki 5.6.

Tabela 5.6: Večkratna uporaba darovalcev v drugem koraku; metoda NN z zunanjim virom informacij.

	0x	1x	2x	3x	4x	5x	6x	7x	8x
datoteka A	222	139	83	37	15	2	1		1
datoteka B	234	126	80	35	18	5	2		

Kljub temu pa rezultati v Tabeli 5.7 kažejo, da glede povprečij in standardnih odklonov ni statistično značilnih razlik. V obravnavanem primeru so se odstopanja torej očitno do neke mere izničila med sabo.

Tabela 5.7: Povprečja ter standardni odkloni specifičnih spremenljivk; metoda NN z zunanjim virom informacij.

	(povp.)	povp.	t	sig	(st.dev)	st.dev	F	sig
klici	4,17	4,22	0,38	0,70	2,15	2,14	1,00	0,48
storitve	3,31	3,41	0,95	0,34	1,74	1,71	1,02	0,42
internet	3,14	3,24	0,92	0,36	1,80	1,83	1,02	0,42

Poglavitna prednost, ki smo so ji obetali od uporabe datoteke zunanjih informacij, je, da nadgrajena metoda NN v zliti datoteki ne bo vzpostavila pogojne neodvisnosti spremenljivk, ki niso nikoli opazovane skupaj, glede na skupne spremenljivke, in bo torej bolj pravilno poustvarila neocenljivi korelaciji. Rezultati v Tabeli 5.8 kažejo, da je bila obravnavana metoda pri poustvaritvi korelacije *klici-storitve* res bolj uspešna od navadne metode NN.

Tabela 5.8: Korelaciji *klici-storitve* ter *klici-internet*; metoda NN z zunanjim virom informacij.

izvorne (A in B)

	klici
storitve	0,44
internet	0,38

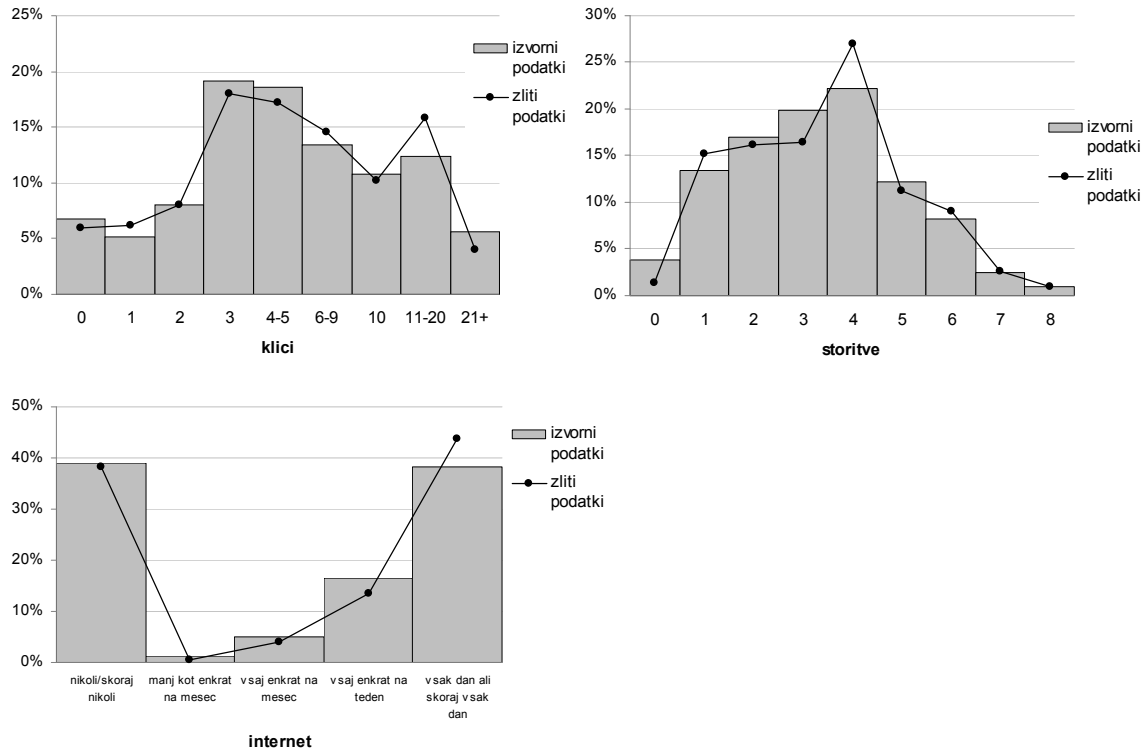
datoteka A

	klici	sig
storitve*	0,39	0,23
internet*	0,45	0,16

datoteka B

	klici*	sig
storitve	0,45	0,39
internet	**0,51**	**0,01**

Slika 5.7: Robne porazdelitve specifičnih spremenljivk; metoda NN z zunanjim virom informacij.



Po drugi strani pa je korelacija *klici*-internet* občutno previsoka glede na poznano »pravilno« korelacijo. Kot je razvidno iz Tabele 5.9, je parcialna korelacija med omenjenima spremenljivkama v datoteki B statistično značilno različna od nič, čeprav sta v izvornih podatkih ti spremenljivki pogojno neodvisni.

Tabela 5.9: Parcialni korelaciji *klici-storitve* ter *klici-internet*; metoda NN z zunanjim virom informacij.

izvorne (A in B)

	klici	sig
storitve	0,14	0,00
internet	0,03	0,40

datoteka A

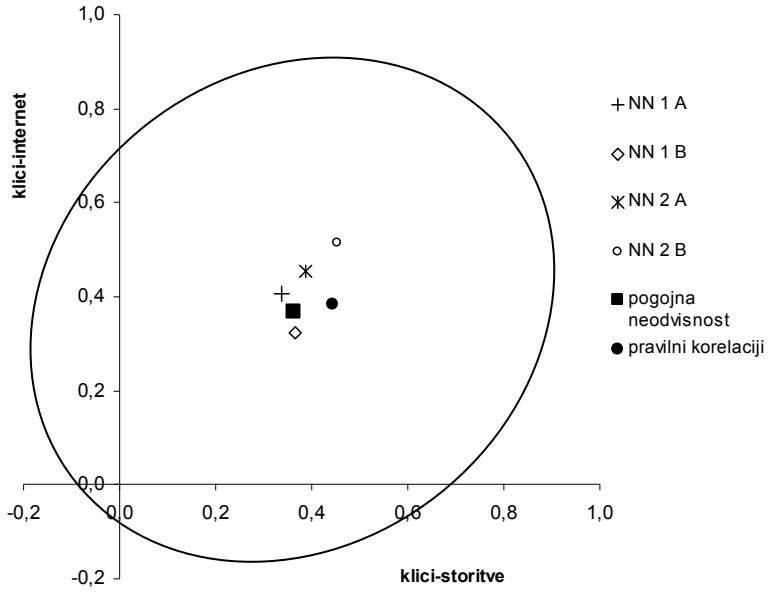
	klici	sig
storitve*	0,02	0,66
internet*	0,05	0,25

datoteka B

	klici*	sig
storitve	0,06	0,22
internet	**0,18**	**0,00**

Na sliki 5.8 prikazujemo poustvarjene korelacije še grafično: pari korelacij so prikazani v koordinatnem sistemu, elipsa pa označuje mejo prostora dovoljenih vrednosti. Kvadrat v sredini elipse označuje par korelacij, pri katerih velja pogojna neodvisnost, s krožcem pa je označena dvojica *pravilnih* korelacij, ki jo poznamo iz izvornih podatkov.

Slika 5.8: Prikaz parov korelacijskih koeficientov na dopustnem območju za neparametrični metodi zlivanja podatkov.



Slika 5.8 nazorno prikazuje, da se je korelacija *klici-storitve* z uporabo dodatnih informacij premaknila v pravo smer (njena vrednost se je zvišala), obenem pa se je zvišala tudi korelacija *klici-internet*, ki je zaradi uporabe dodatnega koraka previsoka. Z uporabo datoteke dodatnih informacij se je torej metoda najbližjega soseda res oddaljila od pogojne neodvisnosti, vendar ne v povsem pravo smer.

5.3 Parametrične metode

V tem poglavju bomo ilustrirali uporabo metode plemenitenja podatkov (*angl*. Data Augmentation oz. DA). Ta metoda je izrazito modelska, saj moramo na samem začetku postulirati model porazdelitve, ki naj bi bila ustvarila opažene podatke. Ker imamo opraviti z zveznimi spremenljivkami, bomo v našem primeru predpostavili, da opaženi podatki izhajajo iz *večrazsežnostne normalne porazdelitve*. Ker se spremenljivka *internet* porazdeljuje po U-porazdelitvi, vemo, da je predpostavka normalnosti grobo kršena. Z omenjeno spremenljivko želimo preizkusiti do kakšne mere je metoda DA res robustna na odmike od predpostavljenega modela, kot to zatrjujejo avtorji (npr. Schafer in Olsen 1998: 9; Rässler 2003: 73).

Pri plemenitenju podatkov gre za uporabo metodologije večkratnega vstavljanja (glej Rubin 1996), zato bo rezultat zlivanja podatkov po tej metodi več popolnjenih datotek in ne zgolj ena, kot v primeru metode najbližjega soseda. V našem primeru bomo vrzeli

v podatkih zapolnili *petkrat*. Za večkratno vstavljanje bomo uporabili program NORM (Schafer 1999).

5.3.1 Metoda plemenitenja podatkov

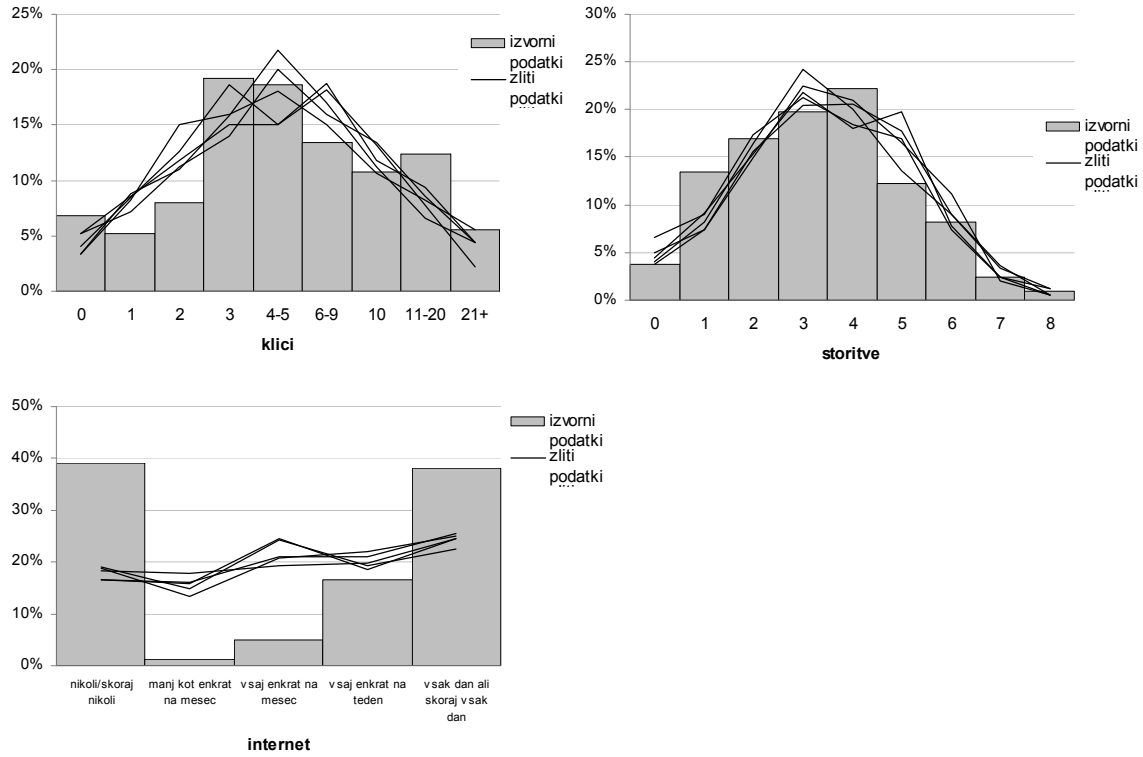
Najprej smo z metodo DA poustvarili izvorne podatke brez uporabe zunanjih informacij. Izziv pri tem je naslednji: da bi določil predpostavljeni model, moramo oceniti 189 parametrov: 18 povprečij, 18 standardnih odklonov ter 153 korelacij, ki skupaj predstavljajo parametre večrazsežnostne normalne porazdelitve osemnajstih spremenljivk. Večino omenjenih parametrov je moč brez težav oceniti iz podatkov, dva izmed njih pa sta z uporabljenimi podatki povsem nedoločena: ker spremenljivki *storitve* in *internet* nista nikoli opazovani ob spremenljivki *klici*, sta korelaciji *klici-storitve* ter *klici-internet* nedoločeni.

Kot smo omenili v razdelku 4.2.2, je prednost metode plemenitenja podatkov v tem, da omogoča, da problem zlivanja podatkov stabiliziramo z uporabo informativne apriorne porazdelitve. Za ilustracijo smo algoritem DA najprej pognali z neinformativno apriorno porazdelitvijo, skrajšane rezultate pa predstavljamo v Prilogi Č. Kot začetne vrednosti parametrov smo vzeli njihove ocene po metodi največjega verjetja, ocenjene z algoritmom EM. Kontrolne statistike so kazale na konvergenco že po prvih nekaj sto iteracijah, vseeno pa smo manjkajoče podatke vstavili prvič šele po 500 iteracijah, naslednje pri 1000 itn. Iz slik je razvidno, da je algoritem »zašel« prav na rob dovoljenega območja korelacij *klici-storitve* ter *klici-internet*. Na nepravilnosti opozorita tudi dejstvo, da imata omenjeni korelaciji iste vrednosti pri vseh petih vstavljanjih, ter povsem zgrešena oblika porazdelitve spremenljivke *klici*.

Kot prikazujeta spodnji sliki, se rezultati povsem spremenijo že, če v postopek vnesemo zelo majhno količino apriorne informacije: uporabili smo grebensko apriorno porazdelitev z vrednostjo hiperparametra ena. Preden razložimo pomen te vrednosti, spomnimo, da po Bayesianskem pristopu kombiniramo informacijo iz vzorca s predhodno informacijo. V obravnavanem primeru imamo tako vzorec velikosti 1000 enot (enote *obeh* datotek), za apriorno porazdelitev pa lahko poenostavljeno rečemo, da se nanaša na namišljeni vzorec z eno enoto (glej Schafer 1997: 155-156). Ker ima apriorna porazdelitev relativno tako majhno težo, njena oblika ni pomembna.

Ker je bilo zaznati zelo visoko avtokoreliranost parametrov, smo algoritem neodvisno pognali petkrat ter manjkajoče vrednosti vstavili šele po 20 000 iteracijah. Slika 5.9 prikazuje rezultate vseh petih vstavljanj naenkrat: pet lomljenih črt ponazarja podatke petih zlitih datotek.

Slika 5.9: Robne porazdelitve specifičnih spremenljivk; metoda plemenitenja podatkov.



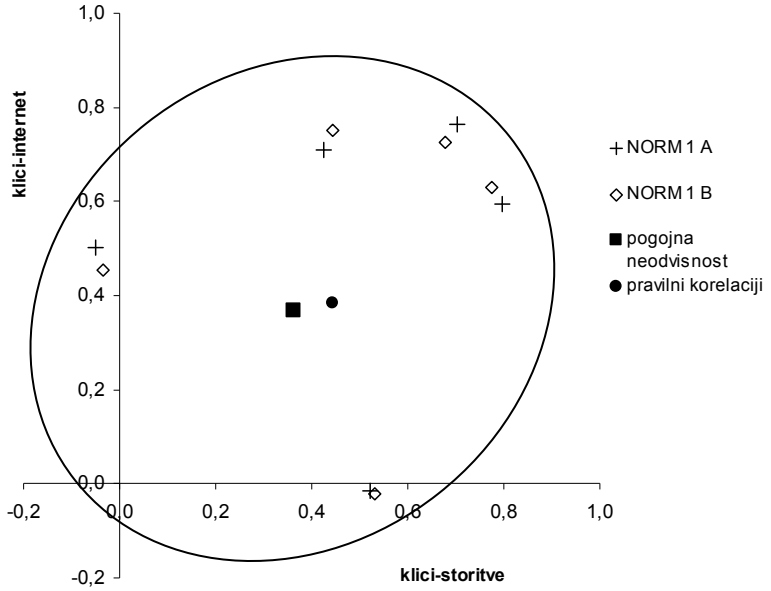
Poustvarjene porazdelitve za spremenljivki *klici* in *storitve* se do visoke mere prilegajo izvornim podatkom. Lomljene črte stolpcem histograma vseeno ne sledijo povsem, pač pa je opaziti značilno zvonasto obliko, saj zliti podatki izhajajo iz normalne porazdelitve.

Kot rečeno, smo pričakovali, da bo spremenljivka *internet* zaradi svoje porazdelitve predstavljala največji izziv metodi DA, ki predpostavlja normalnost. Res se poustvarjena porazdelitev ne ujema z izvorno, vendar pa se vseeno nakazuje porazdelitev v obliki črke U. Najmanjša ter največja vrednost sta v zlitih podatkih zastopani redkeje kot v izhodiščnih, zato je standardni odklon spremenljivke *internet* močno podcenjen (glej tabelo v Prilogi D).

Na Sliki 5.10 so prikazani pari korelacijskih koeficientov *klici-storitve* ter *klici-internet*, ki jih na podlagi podatkov ni bilo mogoče oceniti. Točke so dokaj enakomerno

razporejene po celotnem dopustnem območju, kar kaže na veliko negotovost glede njihove pravilne vrednosti.

Slika 5.10: Pari korelacijskih koeficientov na dopustnem območju; metoda plemenitenja podatkov.



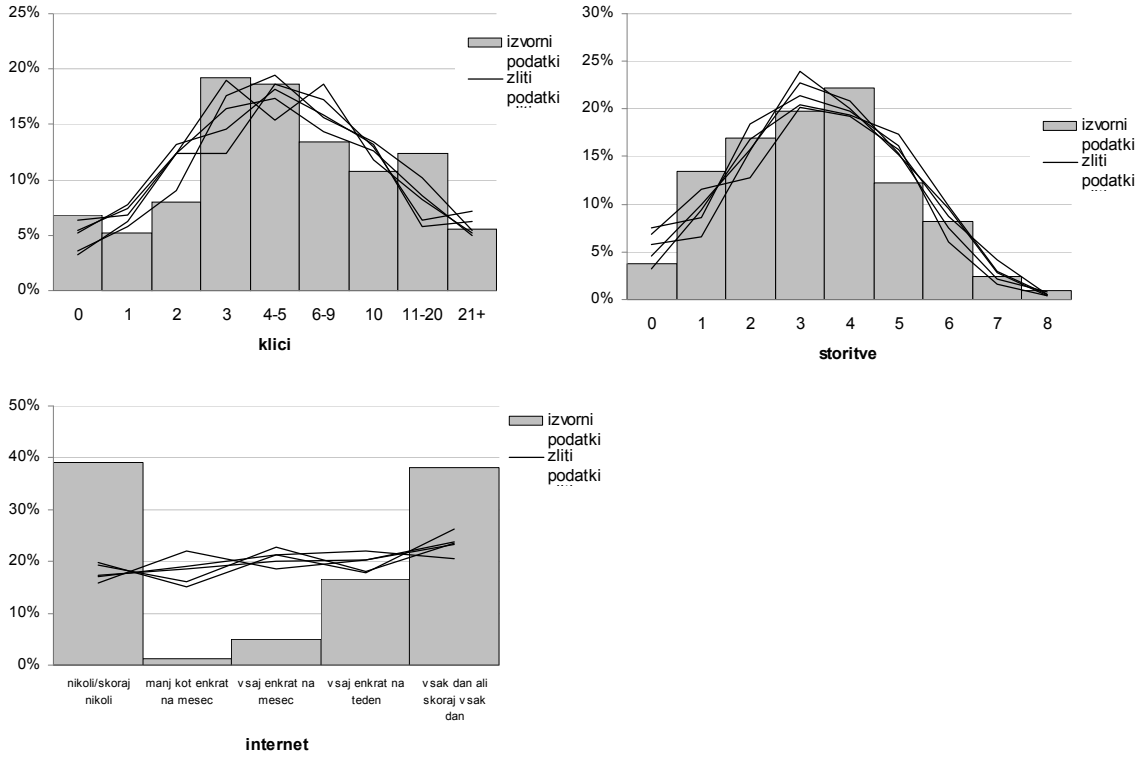
V realnem primeru rabe zlivanja podatkov, ko pravilnih vrednosti neocenljivih korelacij ne poznamo, bi nas razporeditev točk na zgornji sliki opozorila na negotovost, ki je povezana s sklepanjem o teh parametrih, ter obenem podala grobo informacijo o njihovih dovoljenih vrednostih. Ker podatki niso podajali nobene informacije o korelacijah *klici-storitve* ter *klici-internet*, so prikazani rezultati pričakovani.

5.3.2 Metoda DA z uporabo datoteke zunanjih informacij

Datotekama A in B smo nato pridružili popolno datoteko C ter na vzorcu 1092 enot pognali algoritem plemenitenja podatkov. Uporabili smo neinformativno apriorno porazdelitev, saj smo z dodatnimi 92 enotami zagotovili dovolj informacije o korelacijah *klici-spremenljivke* in *klici-internet*, da je algoritem deloval stabilno. Manjkajoče vrednosti smo vstavljali na vsakih 500 iteracij algoritma.

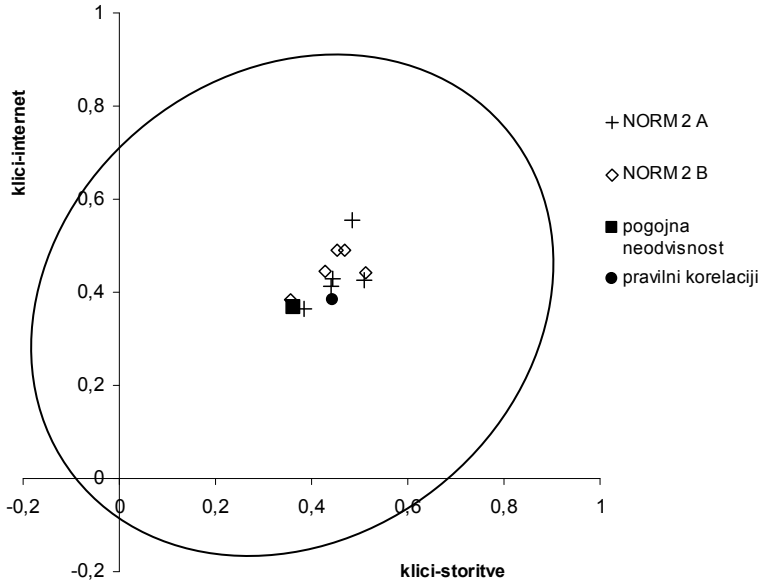
Kot kaže Slika 5.11, vključitev dodatnih enot ni vplivala na obliko robnih porazdelitev. V obliki lomljenih črt še vedno prepoznamo predpostavko normalnosti. Le-ta seveda drži tudi za spremenljivko *internet*, zato je v zlitih podatkih spet preveč vrednosti s sredine njene lestvice in premalo z obeh koncev. Zaradi tega je njen standardni odklon podcenjen, kot je razvidno iz tabele v Prilogi E.

Slika 5.11: Robne porazdelitve specifičnih spremenljivk; metoda DA z uporabo datoteke zunanjih informacij.



Slika 5.12 prikazuje pare korelacijskih koeficientov: točke so tokrat mnogo bolj zgoščene kot na Sliki 5.10, nahajajo pa se v bližini točke pravilnih korelacij. Na Sliki 5.12 se torej še vedno zrcali negotovost glede pravilnih vrednosti korelacij, vendar pa je le-ta občutno manjša kot v primeru zlivanja podatkov brez dodatnih informacij.

Slika 5.12: Pari korelacijskih koeficientov na dopustnem območju; metoda DA z uporabo datoteke zunanjih informacij.



Točke, ki ustrezajo parom poustvarjenih korelacij, na sliki 5.12 večinoma ležijo previsoko, v vodoravni smeri pa so razporejene približno enakomerno okrog pravilne korelacije. V grobem lahko torej rečemo, da je metoda DA z dodatno datoteko informacij korelacijo *klici-spremenljivke* dobro poustvarila, korelacijo *klici-internet* pa nekoliko precenila.

5.3.3 Metoda DA z vstavljanjem parametrov zunanjih informacij

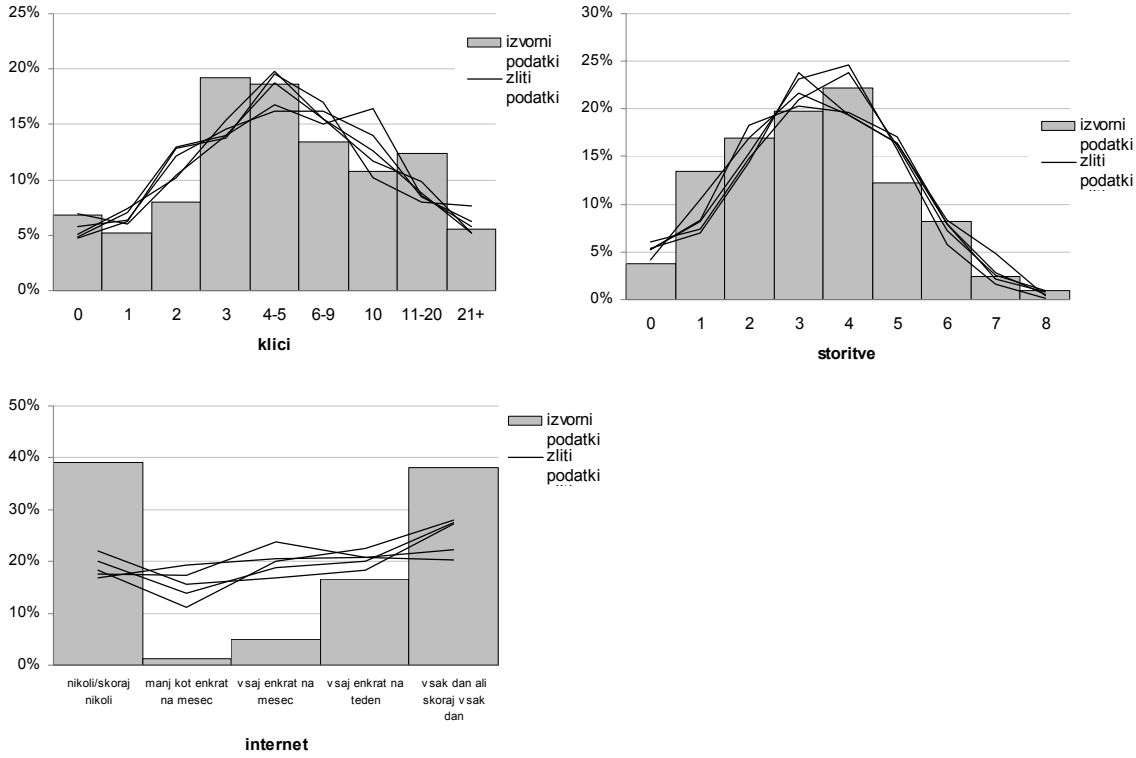
Na koncu smo podatke zlili še z metodo DA, ki smo ji zunanjo informacijo o nedoločljivih korelacijah *klici-storitve* ter *klici-internet* posredovali tako, da smo pravilni vrednosti teh dveh parametrov vključili neposredno v ocenjevani model. Algoritmu DA torej nismo dali na razpolago *enot* iz datoteke C, pač pa le pravilni *vrednosti* omenjenih korelacij.

Podobno kot pri metodi iz prejšnjega razdelka smo najprej pognali algoritem EM, ki je parametre normalnega modela ocenil iz podatkov, vrednosti korelacij *klici-storitve* ter *klici-internet* pa je ocenil na 0,33 in 0,33. Ti vrednosti smo popravili na pravilni vrednosti 0,44 in 0,38 ter s temi izhodiščnimi parametri pognali algoritem DA[22]. Ker naše poznavanje pravilnih vrednosti izhaja iz vzorca 92 enot, smo uporabili grebensko apriorno porazdelitev z vrednostjo hiperparametra 92.

Kot prikazuje Slika 5.13, so rezultati zelo podobni rezultatom iz prejšnjega razdelka: porazdelitvi spremenljivk *klici* in *storitve* se prilegata podatkom, opazno pa je, da so bili zliti podatki ustvarjeni pod predpostavko normalnosti. Porazdelitev spremenljivke *internet* tudi v tem primeru ne sledi U-obliki, zato je njen standardni odklon podcenjen (za tabelo testnih statistik glej Prilogo F).

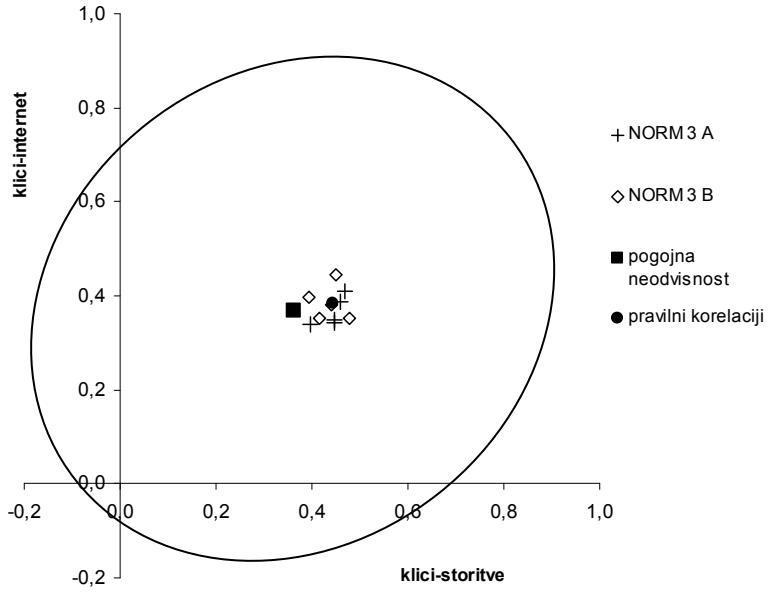
[22] Na tem mestu moramo oporekati Susanne Rässler, ki pravi, da samostojna verzija programa NORM za Windows™ ne omogoča vključitve zunanje informacije v obliki parametrov v ocenjevani model (Rässler 2004: 14). To lahko izvedemo tako, da »ročno« spremenimo izhodno datoteko parametrov algoritma EM.

Slika 5.13: Robne porazdelitve specifičnih spremenljivk; metoda DA z vstavljanjem parametrov zunanjih informacij.



Korelaciji med spremenljivkami, ki niso nikoli opazovane skupaj, je obravnavana metoda poustvarila najbolje. Kot prikazuje Slika 5.14, so točke, ki ponazarjajo pare korelacij, gosto razporejene okrog točke pravilnih korelacij.

Slika 5.14: Pari korelacijskih koeficientov na dopustnem območju; metoda DA z vstavljanjem parametrov zunanjih informacij.



Ta rezultat ni presenetljiv, saj je algoritem DA zgolj prevzel začetni vrednosti korelacij *klici-storitve* in *klici-internet* ter jima dodal določeno slučajno napako. Tekom iteracij se vrednosti omenjenih parametrov nista mogli dramatično spremeniti, saj uporabljeni podatki niso vsebovali nobene informacije o njiju.

S slikami v tem poglavju smo torej ilustrirali, kako uporaba metod večkratnega vstavljanja omogoča ponazoritev negotovosti glede pravilnih vrednosti neocenljivih korelacij. Kot bi pričakovali, je negotovost glede neocenljivih korelacij zelo visoka, kadar ne vnesemo nobene dodatne informacije. Negotovost se občutno zmanjša, če uporabimo datoteko dodatnih informacij, najmanjša pa je, ko informacijo o neocenljivih korelacijah v model vstavimo v obliki parametrov.

6 Zaključek

V zaključku bomo kratko povzeli glavno misel besedila, izpostavili prednosti in slabosti predstavljenih metod ter razmišljali o možnih praktičnih rabah postopkov zlivanja podatkov.

Problem zlivanja podatkov se nanaša na sklepanje na osnovi skupne porazdelitve dveh sklopov slučajnih spremenljivk, pri čemer imamo na voljo le informacijo o robni porazdelitvi vsakega sklopa. Pri tipičnem primeru imamo na eni strani spremenljivke, ki se nanašajo na potrošnjo, na drugi strani spremenljivke o spremljanju TV programov, zanima pa nas sklepanje o potrošnji pri danih vrednostih TV spremenljivk. Le-to je mogoče le na podlagi skupne porazdelitve potrošnje ter TV spremenljivk (Gilula in drugi 2004: 1).

Problem sklepanja o skupni porazdelitvi na podlagi robnih porazdelitev v splošnem ni rešljiv: s konkretnim parom robnih porazdelitev je konsistentnih mnogo skupnih porazdelitev, zato pravimo, da skupna porazdelitev z robnima porazdelitvama ni določena. Da bi problem rešili, moramo imeti na voljo dodatno informacijo. Zlivanje podatkov je mogoče zaradi spremenljivk, ki so skupne obema sklopoma: v omenjenem primeru imamo tako v anketi potrošnje kot v TV anketi na voljo demografske spremenljivke. Metode zlivanja podatkov to skupno informacijo izkoriščajo za sklepanje o skupni porazdelitvi specifičnih spremenljivk (Gilula in drugi 2004: 1).

Kot smo pokazali v razdelku 2.2, pa sama prisotnost skupnih spremenljivk vseeno ni dovolj za enolično določitev problema. Da dosežemo določljivost, moramo v postopkih zlivanja podatkov narediti dodatne predpostavke (Gilula in drugi 2004: 1-2). Kadar lahko asociacije med sklopoma specifičnih spremenljivk popolnoma pojasnimo s skupnimi spremenljivkami, je na problem zlivanja podatkov najbolj naravno gledati kot na problem povezovanja enot (*angl*. matching), ki ga rešujemo z različnimi metodami vstavljanja obstoječih vrednosti.

Na podatkih, ki jih imamo na voljo, pa predpostavke pogojne neodvisnosti ne moremo preveriti, zato ne moremo biti nikoli prepričani, ali smo z uporabo metod vstavljanja obstoječih vrednosti prišli do veljavnega rezultata. Sodobne metode zlivanja podatkov zato k problemu pristopajo povsem drugače ter zlivanje podatkov obravnavajo kot problem manjkajočih vrednosti. Za manjkajoče vrednosti smemo predpostaviti, da

manjkajo naključno (MAR), saj ustrezajo anketnim vprašanjem, ki niso bila vprašana. Mehanizem manjkajočih vrednosti lahko zato obravnavamo kot zanemarljiv, s čimer postanejo najbolj privlačno orodje reševanja problema metode večkratnega vstavljanja (Rässler 2002: 200).

Zlivanje podatkov pa se od klasičnega problema manjkajočih vrednosti razlikuje po problemu nedoločenosti, saj podatki ne določajo asociacij med spremenljivkami, ki niso nikoli opazovane skupaj. Z uporabo EM algoritma, ki je sicer eno najzanesljivejših ter najpogosteje uporabljenih orodij vstavljanja manjkajočih vrednosti, zato ne pridemo do ene same rešitve.

Ob upoštevanju pristopa, ki ga predlaga Rubin (1987) ter razvije Schafer (1997), prestopimo v bayesiansko paradigmo ter uporabimo informativno apriorno porazdelitev. Z njo stabiliziramo delovanje algoritma plemenitenja podatkov, ki predstavlja bayesiansko analogijo EM algoritmu. Večkratno vstavljanje nam omogoča, da ponazorimo negotovost glede vstavljenih vrednosti ter neocenljivih korelacij.

Sodobne modelske metode zlivanja podatkov so v nekaterih pogledih nedvoumno primernejše od metod vstavljanja obstoječih vrednosti, v določenih razmerah pa pridejo do izraza tudi njihove slabosti. Avtorja Soong in De Montigny v besedilu z naslovom »No free lunch in data fusion/integration« (2004) dokazujeta, da ne obstaja ena sama najboljša metoda zlivanja podatkov, saj lahko tudi najbolj napredne metode v določenih razmerah odpovejo. V naslednjih odstavkih si zato oglejmo prednosti in slabosti parametričnih in neparametričnih metod.

Prednosti metod vstavljanja obstoječih vrednosti so predvsem:

- enostavna izvedba ter intuitivno razumevanje delovanja;
- za uporabo ni potrebno predpostaviti modela, ki naj bi ustvaril podatke, zato so metode vstavljanja obstoječih vrednosti zmožne pravilno poustvariti robne frekvence tudi za neobičajne oblike porazdelitev.

Te metode pa je potrebno uporabljati z zadostno mero preudarnosti, saj:

- so rezultati zlivanja podatkov veljavni samo, kadar so specifične spremenljivke pogojno neodvisne pri danih skupnih spremenljivkah, predpostavke pogojne neodvisnosti pa na danih podatkih ni mogoče preveriti;

- so rezultati metod vstavljanja obstoječih vrednosti zavajajoči, ker ne ponazarjajo negotovosti glede zlitih spremenljivk;
- metode vstavljanja obstoječih vrednosti niso sposobne izkoristiti dodatne informacije, sploh kadar je le-ta podana v parametrični obliki;
- kadar določene enote iz darovalne datoteke nikoli ne nastopajo kot darovalci, del informacije, ki jo imamo na voljo, zavržemo.

Po drugi strani parametrične metode namesto predpostavke pogojne neodvisnosti zahtevajo, da vnaprej predpostavimo model podatkov. Te metode glede predpostavke pogojne neodvisnosti naredijo velik korak naprej, saj:

- Ponazarjajo negotovost glede zlitih vrednosti skozi večkratno vstavljene datoteke: če se vstavljene datoteke v določenem pogledu (npr. korelaciji med spremenljivkama) med seboj razlikujejo v veliki meri, nas to opozori na veliko negotovost glede pravilne vrednosti.
- So sposobne učinkovito izkoristiti dodatno informacijo, ki jo uvedemo v postopek zlivanja podatkov, kar lahko naredimo tako v obliki dodatnih popolnih enot, kot v parametrični obliki.

Vseeno pa imajo parametrične metode slabosti, da:

- Za nekatere spremenljivke ne moremo vedno predpostaviti, da se porazdeljujejo po postuliranemu modelu. Za spremenljivke, pri katerih je predpostavka o obliki porazdelitve očitno kršena, bo robna porazdelitev poustvarjena nepravilno[23].
- V primerjavi z metodami vstavljanja obstoječih vrednosti je izvedba zlivanja podatkov s parametričnimi metodami bolj zahtevna: za njihovo rabo je potrebno poznavanje širše teoretične podlage, zaradi večkratnega vstavljanja pa je več dela v fazi analize zlitih datotek.

Kot smo omenili, je končni cilj zlivanja podatkov ustvariti datoteko, ki jo bodo lahko morebitni uporabniki analizirali kot enovito datoteko, brez da bi se zavedali njene »zlite« narave. Kot eno zaključnih misli poudarimo, da ta cilj v splošnem ni dosegljiv: ob visokem številu predpostavk, ki jih moramo nujno narediti v postopku zlivanja (ne

[23] Takšne spremenljivke lahko transformiramo, ter jim s tem spremenimo obliko porazdelitve. S transformacijami spremenljivk se v tem besedilu ne ukvarjamo, za podrobnosti glej Rässler 2002: 134-137.

glede na to, katero metodo uporabimo), bi bilo lahko neodgovorno, če oseba, ki zlito datoteko analizira, ni tudi sama opravila zlivanja podatkov in se ne zaveda vseh pomanjkljivosti uporabljenih metod ali pa je bila z njimi vsaj izčrpno seznanjena.

Odgovorne rabe metod vstavljanja obstoječih vrednosti v praksi si avtor ne more predstavljati brez preverjanja, ali so specifične spremenljivke pogojno neodvisne pri danih skupnih spremenljivkah. Če predpostavka pogojne neodvisnosti ni izpolnjena, bodo asociacije v zliti datoteki namreč poustvarjene nepravilno.

Nekoliko paradoksalna je naslednja ugotovitev, ki govori v prid uporabi parametričnih metod. Za preverjanje predpostavke pogojne neodvisnosti je nujen dodatni vir informacij, na katerem izvedemo potrebne analize. Kadar pa razpolagamo z zunanjim virom informacij, je bolj naravna uporaba parametričnih metod, saj so dano informacijo zmožne izkoristiti mnogo bolje. Rabo metod vstavljanja obstoječih vrednosti lahko torej priporočimo zgolj v primeru, da na zunanjem viru informacij ugotovimo pogojno neodvisnost ter da imajo porazdelitve spremenljivk funkcijske oblike, za katere bi težko postavili primeren obvladljiv model. V vseh ostalih primerih priporočamo uporabo parametričnih metod.

Zlivanju podatkov se je pravzaprav najbolje popolnoma izogniti ter, kadar je le mogoče, zagotoviti enovit vir informacij, saj manjkajoče informacije o spremenljivkah, ki niso nikoli opazovane skupaj, ne moremo *ustvariti* z nobeno metodo. Ker uporaba enovitega vira zaradi previsokih stroškov včasih vseeno ni mogoča, navedimo dva primera, v katerih uporaba v tem besedilu opisanih metod predstavlja edino možnost za sklepanje o količinah, ki nas zanimajo:

- Sklopa spremenljivk, katere skupna porazdelitev nas zanima, sta preobsežna, kot v primeru TV ankete ter ankete potrošnje. Učinkovito zlivanje podatkov dosežemo z vključitvijo t.i. *specifičnih skupnih spremenljivk* (*angl*. specific common variables) v oba vprašalnika. Kot navaja Rässler (2002: 35), se v nemškem tržnem raziskovanju v takšnih razmerah navadno v vprašalnik TV ankete vključi nekaj vprašanj o potrošnji, v vprašalnik ankete potrošnje pa nekaj spremenljivk o spremljanju TV vsebin. Zlivanje podatkov je ob prisotnosti takšnih »specifičnih skupnih« spremenljivk mnogo bolj učinkovito kot zlivanje zgolj z uporabo demografskih ter socioekonomskih spremenljivk.

- Enovito anketo lahko razbijemo na sklope vprašanj ter vsem anketirancem zastavimo le najpomembnejša vprašanja. Preostala vprašanja zastavimo le deležu anketirancev, pri čemer pazimo, da je s tem ustvarjeni vzorec »manjkajočih vrednosti« takšen, da so korelacije med vsemi spremenljivkami ocenljive. V tem primeru zato ne gre za tipičen problem zlivanja podatkov pač pa za t.i. *split questionnaire design*. Manjkajoče vrednosti vstavimo z eno izmed opisanih metod. Z zmanjšanjem števila vprašanj na anketiranca je obremenitev posameznega anketiranca manjša, stroški ankete pa nižji (za podrobnosti glej Rässler in drugi 2002).

Problem zlivanja podatkov – ocenjevanje neocenljivih korelacij – je torej najprimerneje rešiti tako, da poskrbimo za dodatno informacijo o spremenljivkah, ki niso nikoli opazovane skupaj. S tem zlivanje podatkov prevedemo na vstavljanje manjkajočih vrednosti. Če dodatne informacije nikakor ni mogoče pridobiti, nam uporaba parametričnih metod omogoči ponazoritev tveganja pri sklepanju o korelacijah med omenjenimi spremenljivkami. Uporaba metod vstavljanja obstoječih vrednosti brez preverjanja predpostavke pogojne neodvisnosti pa je lahko neodgovorna.

Za konec omenimo še etične vidike zlivanja podatkov. Uporaba postopkov zlivanja podatkov je lahko etično sporna, saj anketiranci, ki so odgovarjali npr. na TV anketo, niso nujno seznanjeni z možnostjo, da se bodo njihovi podatki uporabili tudi za sklepanje o potrošnji pri danih vrednostih TV spremenljivk. Nekateri svetovni statistični uradi so zato že oblikovali etične kodekse, ki obravnavajo zlivanje podatkov (glej npr. Office of National Statistics 2004 ali Statistics New Zealand 2008). V njih poudarjajo npr. obvezno nedvoumno dovoljenje anketirancev, da se njihovi podatki lahko uporabijo za zlivanje podatkov, zaupnost zlitih virov ter obvezno objavo uporabljene metodologije zlivanja podatkov ob morebitni objavi zlitega vira.

Literatura

Abdi, H. (2007): Distance. V Salkind, N. (ur.): *Encyclopedia of Measurement and Statistics*. Thousand Oaks: Sage.

Baker, K., P. Harris in J. O'Brien (1989): Data fusion: an appraisal and experimental evaluation. *Journal of the Market Research Society* 31(2), 153-212.

Battistin, E., R. Miniaci in G. Weber (2000): What Do We Learn from Recall Consumption Data? *Journal of Human Resources* 38(2), 354-385.

Coli, A. in F. Tartamella (2000): A pilot social accounting matrix for Italy with a focus on households. *26th General Conference of the International Association for Research in Income and Wealth, Cracow*.

De Maesschalk, R., D. Jouan-Rimbaud in D.L. Massart (2000): The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* 50, 1-18.

Dempster, A.P., N.M. Laird in D.B. Rubin (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society* B39, 1-38.

D'Orazio, M., M.Di Zio in M.Scanu (2004): Statistical matching and the likelihood principle: uncertainty and logical constraints. *ISTAT Technical Report* 1/2004.

D'Orazio, M., M.Di Zio in M.Scanu (2006): *Statistical Matching: Theory and Practice.* West Sussex: John Wiley & Sons Ltd.

Fisher, H. (2000): *The Central Limit Theorem from Laplace to Cauchy: Changes in Stochastic Objectives and In Analytical Methods*. Dostopno na http://mathsrv.ku-eichstaett.de/MGF/homes/didmath/seite/1850.pdf (12. september 2008).

Gilula, Z., R.E. Mcculloch in P.E. Rossi (2004): A Direct Approach to Data Fusion. *Journal of Marketing Research* 43, 73-83.

Harel, O. in X.H. Zhou (2006): Multiple imputation: Review of theory, implementation and software. *UW Biostatistics Working Paper Series* Working Paper 297. Dostopno na http://www.bepress.com/uwbiostat/paper297 (20. avgust 2008).

Horton, N. in K.P Kleinman (2007): Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician* 61(1), 79-90.

Kadane, J.B. (1978): Some statistical problems in merging data files. V *Compendium of Tax Research, Office of Tax Analysis, Department of the Treasury*, 159-171. Washington, DC: U.S. Government Printing Office. Ponatisnjeno v *Journal of Official Statistics* 17 (2001), 423-433.

Kalton, G. in V. Vehovar (2001): *Vzorčenje v anketah*. Ljubljana: FDV.

Kamakura W.A. in M. Wedel (1997): Statistical Data Fusion for Cross-Tabulation. *Journal of Marketing Research* 34(4), 485-498.

Kiesl, H. in S. Rässler (2006): Quality in Data Fusion. *Proceedings of the European Conference on Quality in Survey Statistics*. Dostopno na http://www.statistics.gov.uk/events/q2006/downloads/W03_Rassler.doc (8. avgust 2008).

Kovacevic M.S. in T.P. Liu (1994): Statistical Matching of Survey Datafiles: A Simulation Study. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 479-484.

Lynch, S.M. (2007): *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer.

Office of National Statistics (2004): *Protocol on data matching*. Dostopno na http://www.statistics.gov.uk/about/national_statistics/cop/downloads/NSCoPDatamatching.pdf (14. september 2008).

Rässler, S. (2002): *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. New York: Springer.

Rässler, S. (2003): A non-iterative Bayesian approach to statistical matching. *Statistica Neerlandica* 57(1), 58-74.

Rässler, S. (2004): Data fusion: Identification Problems, Validity, and Multiple Imputation. *Austrian Journal of Statistics* 33(1-2), 153-171.

Rässler, S. in K. Fleischer (1998): Aspects concerning data fusion techniques. *ZUMA Nachrichten Spezial 4*, 317-333.

Rässler, S., F. Koller in C. Mäenpää (2002): A split Questionnaire Design applied to German Media and Consumer Surveys. *Proceedings of the International Conference on Improving Surveys, Copenhagen*.

Rodgers, W.L. in E. DeVol (1982): An evaluation of statistical matching. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-132. Washington, DC: American Statistical Association.

Rubin, D.B. (1976): Inference and missing data. *Biometrika* 63, 581-592.

Rubin, D.B. (1987): *Multiple Imputation for Nonresponse Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1988): An overview of multiple imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 79-84. Washington, DC: American Statistical Association.

Rubin, D.B. (1996): Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91, 473-489.

Schafer, J.L. (1997): *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (1999): *NORM: Multiple imputation of incomplete multivariate data under a normal model,* verzija 2, programska oprema za Windows 95/98/NT. Dostopno na http://www.stat.psu.edu/~jls/misoftwa.html (20. julij 2008).

Schafer, J.L. in M.K. Olsen (1998): Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research* 33, 545-571.

Singh, A.C., J. Armstrong in G.E. Lemaitre (1988): Statistical matching using log-linear imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 672-677.

Soong, R. in M. de Montigny (2001): The anatomy of data fusion. V *Worldwide Readership Research Symposium, Venice*, 87-109. Dostopno na http://www.zonalatina.com/WRRSfusion.pdf (23. avgust 2008).

Soong, R. in M. de Montigny (2004): No free lunch in data fusion/integration. *Proceedings of the ARF/ESOMAR »Week of Audience Measurment«, Genova*. Dostopno na http://www.zonalatina.com/WAM2004.pdf (23. avgust 2008).

Statistics New Zealand (2008): *Data integration policy*. Dostopno na http://www.stats.govt.nz/about-us/policies-and-guidelines/data-integration-policy/default.htm (14. september 2008).

Sutherland, H., R. Taylor in J. Gomulka (2002): Combining household income and expenditure data in policy Simulations. *Review of Income and Wealth* 48, 517-536.

Trček, F. in R. Platinovšek (2007): Uporabniki mobilnih telefonov med pragmatičnostjo in ekspresivnostjo. V Vehovar, V. (ur.): *Mobilne refleksije*, 147-165. Ljubljana: FDV.

Vehovar, V. (2007): *Mobilne refleksije*. Ljubljana: FDV.

Vidav, I. (1994): *Višja matematika 1*. Ljubljana: Društvo matematikov, fizikov in astronomov Slovenije.

Yoshizoe, Y. in M. Araki (1999): Use of statistical matching for household surveys in Japan. *52nd Session of the International Statistical Institute, Helsinki*.

Priloga A: Opis spremenljivk

spol: **0** ženski; **1** moški.

predplačnik, *»Ali imate za vaš trenutni glavni mobilni telefon sklenjeno naročniško razmerje ali uporabljate predplačniški paket?«:* **0** naročniško razmerje; **1** predplačniški paket.

starost: starost v letih, v anketo so bili vključeni anketiranci v starosti od 10 do 75 let.

izobrazba, *»Katero stopnjo izobrazbe ste doslej dosegli?«:* **1** brez šolske izobrazbe oz. nepopolna osnovna izobrazba, 1-3 razredi; **2** nepopolna izobrazba, 4-7 razredov; **3** osnovna izobrazba; **4** nižja ali srednja poklicna izobrazba; **5** srednja strokovna izobrazba; **6** srednja splošna izobrazba; **7** višja strokovna izobrazba, višješolska izobrazba; **8** visoka strokovna izobrazba; **9** visoka univerzitetna izobrazba; specialistična povisokošolska izobrazba, magisterij, doktorat.

angleščina, *»Kako bi ocenili svoje znanje angleščine?«:* **1** sploh ne znam; **2** znam slabo; **3** delno; **4** v glavnem znam; **5** tekoče in v celoti.

vel_nas, *velikost naselja:* **1** do 500 prebivalcev; **2** nad 500 do 2000 prebivalcev; **3** nad 2000 do 4000 prebivalcev; **4** nad 4000 do 10 000 prebivalcev; **5** nad 10 000 do 50 000 prebivalcev; **6** nad 50 000 prebivalcev.

kako_dolgo, *Kako dolgo že uporabljate mobilni telefon:* **1** manj kot dve leti; **2** 2-3 leta; **3** 4-5 let; **4** 6-7 let; **5** 8 let in več.

Naslednje štiri spremenljivke so v vprašalniku nastopale v sklopu z nagovorom: *»Ocenite, kako pogosto se doma, v službi, v šoli ali kje drugje pogovarjate po mobilnem telefonu o...«*

službene, *»službenih, poslovnih ali šolskih zadevah«;*

vsakdanje, »*vsakdanjih praktičnih stvareh (dogovarjanja, informacije)«;*

neobvezne, *»neobveznih stvareh (klepetanje, pogovor, druženje)«;*

zaupne, *»zaupnih osebnih temah«.*

vse štiri spremenljivke so merjene na lestvici: **1** nikoli; **2** občasno; **3** mesečno; **4** tedensko; **5** dnevno.

Tudi naslednje štiri spremenljivke so nastopale v sklopu: *»S pomočjo 5-stopenjske lestvice ocenite, kako pomembne so za vas naslednje stvari, ko kupujete nov telefon«*

zmogljivost, *»tehnična zmogljivost«;*

cena;

enostavnost, *»enostavnost uporabe«;*

znamka, *»znamka mobilnega telefona«.*

vse štiri spremenljivke so merjene na lestvici od 1 (sploh ni pomembno) do 5 (zelo pomembno).

Priloga B: Rezultati linearne regresije

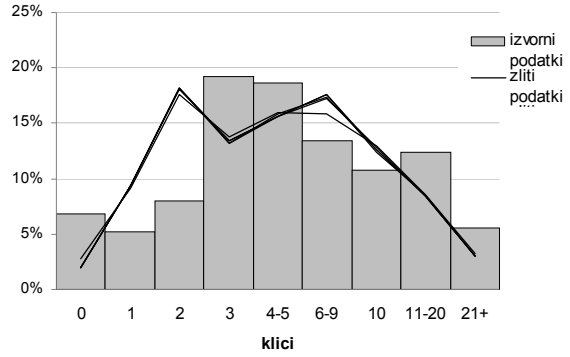
	klici (R^2=0,45)		**storitve** (R^2=0,46)		**internet** (R^2=0,49)	
	beta	t	beta	t	beta	t
konstanta		1,41		5,44		5,36
spol	0,08	3,05	0,01	0,21	0,03	1,42
predplacnik	-0,13	-5,17	-0,10	-3,78	-0,10	-4,06
starost	-0,10	-2,79	-0,32	-9,43	-0,33	-9,76
izobrazba	0,06	2,06	0,04	1,39	0,25	8,79
anglescina	0,04	1,18	0,16	4,90	0,25	7,98
vel_nas	0,04	1,59	0,06	2,40	0,06	2,70
kako_dolgo	0,17	6,78	0,10	4,08	0,01	0,55
poslovne	0,30	11,15	0,07	2,60	0,13	5,08
vsakdanje	0,13	5,06	0,04	1,70	0,02	0,91
neobvezne	0,15	5,18	0,05	1,97	0,04	1,42
zaupne	0,00	-0,10	0,10	3,49	0,01	0,24
zmogljivost	0,05	2,08	0,12	4,59	0,04	1,66
cena	-0,05	-2,13	0,01	0,45	-0,03	-1,20
enostavnost	0,02	0,88	-0,05	-2,16	-0,03	-1,33
znamka	0,02	0,93	0,08	3,27	-0,01	-0,47

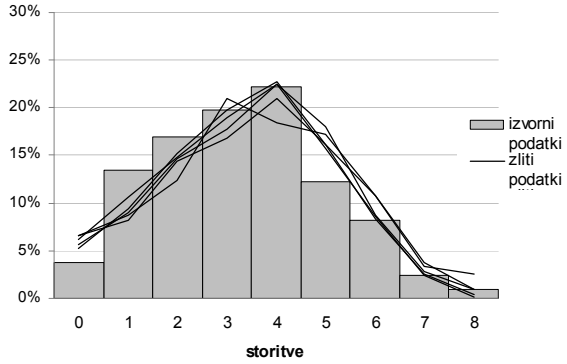
Priloga C: Korelacijska matrika

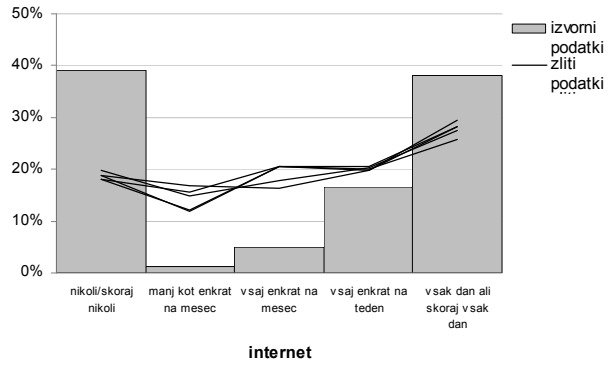
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 spol	1	-0,03	-0,01	-0,10	-0,06	-0,05	0,12	0,14	-0,01	-0,12	-0,11	0,17	-0,07	-0,10	0,14	0,12	0,04	0,03
2 predplacnik	-0,03	1	-0,04	-0,30	-0,15	-0,07	-0,33	-0,24	-0,17	-0,09	-0,11	-0,08	0,00	0,06	-0,06	-0,33	-0,21	-0,26
3 starost	-0,01	-0,04	1	0,29	-0,50	0,11	-0,02	-0,26	-0,20	-0,32	-0,32	-0,31	-0,02	0,28	-0,21	-0,26	-0,52	-0,43
4 izobrazba	-0,10	-0,30	0,29	1	0,28	0,16	0,22	0,22	0,11	-0,01	0,07	-0,04	-0,02	0,07	-0,02	0,21	0,08	0,30
5 anglescina	-0,06	-0,15	-0,50	0,28	1	0,15	0,14	0,32	0,24	0,26	0,28	0,24	0,01	-0,24	0,13	0,33	0,50	0,59
6 vel_nas	-0,05	-0,07	0,11	0,16	0,15	1	0,04	0,02	0,01	0,01	0,00	-0,07	0,02	-0,08	-0,06	0,06	0,06	0,11
7 kako_dolgo	0,12	-0,33	-0,02	0,22	0,14	0,04	1	0,24	0,19	0,12	0,12	0,14	0,03	0,03	0,17	0,37	0,24	0,19
8 poslovne	0,14	-0,24	-0,26	0,22	0,32	0,02	0,24	1	0,32	0,21	0,28	0,22	-0,04	-0,08	0,16	0,53	0,37	0,42
9 vsakdanje	-0,01	-0,17	-0,20	0,11	0,24	0,01	0,19	0,32	1	0,38	0,29	0,12	0,02	-0,02	0,10	0,38	0,28	0,27
10 neobvezne	-0,12	-0,09	-0,32	-0,01	0,26	0,01	0,12	0,21	0,38	1	0,47	0,12	-0,02	-0,14	0,10	0,33	0,32	0,26
11 zaupne	-0,11	-0,11	-0,32	0,07	0,28	0,00	0,12	0,28	0,29	0,47	1	0,21	0,02	-0,06	0,13	0,28	0,36	0,28
12 zmogljivost	0,17	-0,08	-0,31	-0,04	0,24	-0,07	0,14	0,22	0,12	0,12	0,21	1	0,09	-0,02	0,34	0,24	0,35	0,24
13 cena	-0,07	0,00	-0,02	-0,02	0,01	0,02	0,03	-0,04	0,02	-0,02	0,02	0,09	1	0,13	0,09	-0,04	0,03	-0,03
14 enostavnost	-0,10	0,06	0,28	0,07	-0,24	-0,08	0,03	-0,08	-0,02	-0,14	-0,06	-0,02	0,13	1	-0,04	-0,08	-0,21	-0,20
15 znamka	0,14	-0,06	-0,21	-0,02	0,13	-0,06	0,17	0,16	0,10	0,10	0,13	0,34	0,09	-0,04	1	0,18	0,26	0,13
16 klici	0,12	-0,33	-0,26	0,21	0,33	0,06	0,37	0,53	0,38	0,33	0,28	0,24	-0,04	-0,08	0,18	1	0,44	0,38
17 storitve	0,04	-0,21	-0,52	0,08	0,50	0,06	0,24	0,37	0,28	0,32	0,36	0,35	0,03	-0,21	0,26	0,44	1	0,52
18 internet	0,03	-0,26	-0,43	0,30	0,59	0,11	0,19	0,42	0,27	0,26	0,28	0,24	-0,03	-0,20	0,13	0,38	0,52	1

Priloga Č: Metoda DA z neinformativno apriorno porazdelitvijo

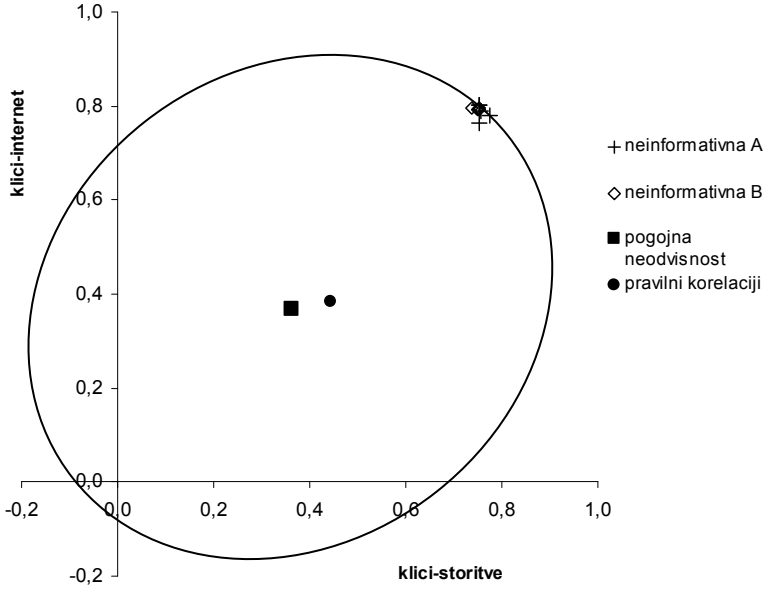
Histogrami porazdelitev







Pari korelacij na dopustnem območju



Priloga D: Metoda plemenitenja podatkov

Povprečja ter standardni odkloni

	(povp.)	povp.	t	sig	(st.dev)	st.dev	F	sig
klici 1	4,17	3,93	1,82	0,07	2,15	1,78	1,02	0,40
klici 2		3,94	1,69	0,09		1,93	1,11	0,13
klici 3		3,94	1,82	0,07		1,75	1,00	0,48
klici 4		3,94	1,73	0,08		1,72	1,01	0,44
klici 5		3,94	1,70	0,09		1,86	1,07	0,23
storitve 1	3,31	3,52	1,88	0,06	1,74	1,78	1,02	0,40
storitve 2		3,62	2,68	0,01		1,93	1,11	0,13
storitve 3		3,36	0,47	0,64		1,75	1,00	0,48
storitve 4		3,43	1,15	0,25		1,72	1,01	0,44
storitve 5		3,56	2,19	0,03		1,86	1,07	0,23
internet 1	3,14	3,22	0,77	0,44	1,80	1,48	1,22	0,01
internet 2		3,21	0,65	0,52		1,48	1,22	0,01
internet 3		3,30	1,60	0,11		1,46	1,24	0,01
internet 4		3,28	1,33	0,18		1,46	1,24	0,01
internet 5		3,20	0,60	0,55		1,44	1,25	0,01

Korelacije specifičnih spremenljivk

izvorne (A in B)

	klici
storitve	0,44
internet	0,38

datoteka A

	klici	sig
storitve* 1	0,76	0,00
storitve* 2	0,75	0,00
storitve* 3	0,75	0,00
storitve* 4	0,75	0,00
storitve* 5	0,78	0,00
internet* 1	0,79	0,00
internet* 2	0,76	0,00
internet* 3	0,80	0,00
internet* 4	0,80	0,00
internet* 5	0,78	0,00

datoteka B

	klici*	sig
storitve 1	0,74	0,00
storitve 2	0,75	0,00
storitve 3	0,75	0,00
storitve 4	0,75	0,00
storitve 5	0,75	0,00
internet 1	0,80	0,00
internet 2	0,79	0,00
internet 3	0,79	0,00
internet 4	0,79	0,00
internet 5	0,79	0,00

Parcialne korelacije specifičnih spremenljivk

izvorne (A in B)

	klici	sig
storitve	0,14	0,00
internet	0,03	0,40

datoteka A

	klici	sig
storitve* 1	0,70	0,00
storitve* 2	0,68	0,00
storitve* 3	0,69	0,00
storitve* 4	0,72	0,00
storitve* 5	0,74	0,00
internet* 1	0,74	0,00
internet* 2	0,73	0,00
internet* 3	0,76	0,00
internet* 4	0,75	0,00
internet* 5	0,75	0,00

datoteka B

	klici*	sig
storitve 1	0,69	0,00
storitve 2	0,70	0,00
storitve 3	0,70	0,00
storitve 4	0,70	0,00
storitve 5	0,70	0,00
internet 1	0,82	0,00
internet 2	0,81	0,00
internet 3	0,81	0,00
internet 4	0,81	0,00
internet 5	0,81	0,00

Priloga E: Metoda DA z uporabo datoteke zunanjih informacij

Povprečja ter standardni odkloni

	(povp.)	povp.	t	sig	(st.dev)	st.dev	F	sig
klici 1	4,17	4,10	0,47	0,64	2,15	1,70	1,02	0,39
klici 2		4,27	0,80	0,42		1,71	1,02	0,41
klici 3		4,12	0,32	0,75		1,87	1,07	0,22
klici 4		3,99	1,29	0,20		1,72	1,01	0,44
klici 5		3,98	1,36	0,17		1,72	1,01	0,43
storitve 1	3,31	3,38	0,64	0,52	1,74	1,70	1,02	0,39
storitve 2		3,20	0,95	0,34		1,71	1,02	0,41
storitve 3		3,41	0,88	0,38		1,87	1,07	0,22
storitve 4		3,53	2,08	0,04		1,72	1,01	0,44
storitve 5		3,46	1,42	0,15		1,72	1,01	0,43
internet 1	3,14	3,16	0,17	0,86	1,80	1,47	1,23	0,01
internet 2		3,13	0,06	0,95		1,40	1,28	0,00
internet 3		3,10	0,35	0,73		1,43	1,26	0,01
internet 4		3,14	0,06	0,95		1,42	1,27	0,00
internet 5		3,10	0,39	0,69		1,38	1,30	0,00

Korelacije specifičnih spremenljivk

prvotne (A in B)

	klici
storitve	0,44
internet	0,38

datoteka A

	klici	sig
storitve* 1	0,51	0,17
storitve* 2	0,38	0,21
storitve* 3	0,44	0,40
storitve* 4	0,49	0,28
storitve* 5	0,44	0,40
internet* 1	0,43	0,29
internet* 2	0,36	0,37
internet* 3	0,41	0,34
internet* 4	0,56	0,00
internet* 5	0,43	0,27

datoteka B

	klici*	sig
storitve 1	0,51	0,15
storitve 2	0,36	0,11
storitve 3	0,43	0,38
storitve 4	0,47	0,35
storitve 5	0,45	0,39
internet 1	0,44	0,22
internet 2	0,38	0,40
internet 3	0,44	0,21
internet 4	0,49	0,04
internet 5	0,49	0,05

Parcialne korelacije specifičnih spremenljivk

prvotne (A in B)

	klici	sig
storitve	0,14	0,00
internet	0,03	0,40

datoteka A

	klici	sig
storitve* 1	0,24	0,00
storitve* 2	0,06	0,22
storitve* 3	0,07	0,13
storitve* 4	0,22	0,00
storitve* 5	0,07	0,11
internet* 1	0,10	0,03
internet* 2	0,02	0,61
internet* 3	0,10	0,03
internet* 4	0,31	0,00
internet* 5	0,14	0,00

datoteka B

	klici*	sig
storitve 1	0,25	0,00
storitve 2	0,00	0,96
storitve 3	0,11	0,02
storitve 4	0,20	0,00
storitve 5	0,15	0,00
internet 1	0,12	0,01
internet 2	0,08	0,07
internet 3	0,12	0,01
internet 4	0,24	0,00
internet 5	0,22	0,00

Priloga F: Metoda DA z vstavljanjem parametrov zunanjih informacij

Povprečja ter standardni odkloni

	(povp.)	povp.	t	sig	(st.dev)	st.dev	F	sig
klici 1	4,17	4,14	0,16	0,87	2,15	1,70	1,03	0,38
klici 2		4,10	0,48	0,63		1,78	1,02	0,41
klici 3		4,11	0,43	0,67		1,71	1,02	0,41
klici 4		3,98	1,34	0,18		1,78	1,02	0,41
klici 5		4,25	0,58	0,56		1,86	1,06	0,25
storitve 1	3,31	3,44	1,23	0,22	1,74	1,70	1,03	0,38
storitve 2		3,33	0,20	0,84		1,78	1,02	0,41
storitve 3		3,44	1,23	0,22		1,71	1,02	0,41
storitve 4		3,42	1,01	0,32		1,78	1,02	0,41
storitve 5		3,45	1,23	0,22		1,86	1,06	0,25
internet 1	3,14	3,17	0,35	0,73	1,80	1,43	1,26	0,01
internet 2		3,27	1,26	0,21		1,44	1,25	0,01
internet 3		3,19	0,46	0,64		1,45	1,24	0,01
internet 4		3,07	0,64	0,52		1,44	1,25	0,01
internet 5		3,08	0,59	0,55		1,48	1,22	0,01

Korelacije specifičnih spremenljivk

izvorne (A in B)

	klici
storitve	0,44
internet	0,38

datoteka A

	klici	sig
storitve* 1	0,40	0,27
storitve* 2	0,47	0,35
storitve* 3	0,45	0,40
storitve* 4	0,45	0,40
storitve* 5	0,46	0,38
internet* 1	0,34	0,28
internet* 2	0,41	0,35
internet* 3	0,34	0,30
internet* 4	0,35	0,33
internet* 5	0,39	0,40

datoteka B

	klici*	sig
storitve 1	0,39	0,25
storitve 2	0,45	0,40
storitve 3	0,48	0,31
storitve 4	0,41	0,34
storitve 5	0,44	0,40
internet 1	0,40	0,39
internet 2	0,44	0,21
internet 3	0,35	0,33
internet 4	0,35	0,34
internet 5	0,38	0,40

Parcialne korelacije specifičnih spremenljivk

izvorne (A in B)

	klici	sig
storitve	0,14	0,00
internet	0,03	0,40

datoteka A

	klici	sig
storitve* 1	0,13	0,00
storitve* 2	0,17	0,00
storitve* 3	0,12	0,01
storitve* 4	0,17	0,00
storitve* 5	0,15	0,00
internet* 1	0,05	0,33
internet* 2	0,09	0,05
internet* 3	-0,02	0,72
internet* 4	-0,01	0,84
internet* 5	0,02	0,64

datoteka B

	klici*	sig
storitve 1	0,06	0,19
storitve 2	0,16	0,00
storitve 3	0,18	0,00
storitve 4	0,08	0,10
storitve 5	0,10	0,03
internet 1	0,02	0,73
internet 2	0,12	0,01
internet 3	0,01	0,80
internet 4	0,05	0,29
internet 5	-0,01	0,77