UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

DIPLOMSKO DELO

UPORABA PODATKOVNEGA RUDARJENJA V GEOGRAFSKIH INFORMACIJSKIH SISTEMIH

Ljubljana, avgust 2008 **KATJA HOZJAN**

IZJAVA

Študentka KATJA HOZJAN izjavljam, da sem avtorica tega diplomskega dela, ki sem ga napisala pod mentorstvom dr. JURIJA JAKLIČA, in da dovolim njegovo objavo na fakultetnih spletnih straneh.

V Ljubljani, dne 13. 8. 2008 Podpis:_____

KAZALO

KAZALO SLIK

KAZALO TABEL

KAZALO PRILOG

UVOD

Danes je na voljo veliko podatkov, ki nam lahko prinesejo poslovne in druge koristi. Ti podatki pa postanejo uporabni šele, ko jih učinkovito analiziramo in pretvorimo v informacije. Uspešna podjetja se od preostalih ločijo po tem, da znajo pravočasno predvideti spremembe v okolju in se nanje hitro odzvati. Pravilne in hitre odločitve pa temeljijo na informacijah, ki so nam v danem trenutku na voljo.

Napredna tehnologija omogoča zbiranje podatkov na različnih področjih. Po eni strani ti podatki pomenijo potencial za odkrivanje uporabnih informacij in znanja; po drugi strani smo omejeni v naši zmožnosti procesiranja ogromnih količin podatkov za pridobitev informacij. Zato podatkovno rudarjenje zahteva preproste tehnike in orodja. Podatkovno rudarjenje je pomemben del poslovnega odločanja in pomeni proces analize velikih količin podatkov z namenom odkriti zanimive vzorce in pravila. Koristno je na vseh področjih, kjer obstajajo velike količine podatkov, iz katerih lahko pridobimo kakovostne informacije. Znanje, ki ga dobimo s podatkovnim rudarjenjem, je v poslovnem svetu koristno, če poveča dobiček. Uporabljamo ga lahko na več področjih: kot raziskovalno orodje v farmacevtski industriji za analizo določenih kemikalij, za izboljšanje poslovnih procesov, v marketingu za iskanje povezav med prodajo izdelkov in segmentacijo, za odkrivanje prevar in drugo (Berry & Linoff, 1998, str. 7–11).

Večina poslovnih odločitev je povezanih s krajem, z regijo ali državo. Ker ima večina vseh poslovnih podatkov vgrajeno prostorsko komponento kraja, regije ali države, je poslovno odločanje v sodobnih podjetjih odvisno od posebnih analiz na podlagi prostorske razsežnosti. Geografsko usmerjene podatke pa ureja, analizira in prikazuje geografski informacijski sistem (GIS). Gre za računalniško podprt sistem, ki omogoča uporabo podatkov, ki opisujejo realen prostor. GIS-tehnologija postaja vedno bolj dostopna in prijazna za uporabo, ne le na tradicionalnih področjih, kot sta geografija in demografija, ampak tudi v antropologiji, zdravstvu in sociologiji (Steinberg & Steinberg, 2006, str. 6).

Diplomska naloga prikazuje, kako je podatkovno rudarjenje uporabno pri raziskovanju geografskih podatkov. Zahteve podatkovnega rudarjenja geografskih podatkov se razlikujejo od klasičnega podatkovnega rudarjenja. Geografski podatki so opisani z geografskim prostorom in so torej prostorsko odvisni. Te značilnosti pomenijo, da je klasično podatkovno rudarjenje neprimerno za prostorske podatke. Pridobivanje znanja iz teh podatkov zahteva posebne pristope.

NAMEN IN CILJI

Namen diplomske naloge je povezati podatkovno rudarjenje z vizualizacijskimi raziskovalnimi metodami, da bi olajšali raziskavo geografskih podatkov. Namen je prikazati, da lahko s to kombinacijo tehnik hitreje in uspešneje pridobimo znanje iz geografskih podatkov ter odkrijemo vzorce, ki jih s klasičnimi tehnikami ni mogoče najti.

Cilj diplomske naloge je predstaviti glavne značilnosti in tehnike podatkovnega rudarjenja, odkrivanja zakonitosti v podatkih ter osnove geografskih informacijskih sistemov. Poleg tega želim prikazati, kako lahko z geografskim podatkovnim rudarjenjem odkrivamo novo znanje, s katerim lahko pridobimo potencialno uporabne informacije za podjetje. Zato bom v diplomskem delu predstavila teoretičen del in primer uporabe podatkovnega rudarjenja v geografskih informacijskih sistemih.

STRUKTURA

Diplomska naloga je razdeljena v tri sklope. Na začetku so predstavljena problematika, namen in struktura diplome.

V prvem delu je opredeljen pojem podatkovnega rudarjenja. Predstavljene so najpogosteje uporabljene tehnike in orodja podatkovnega rudarjenja. To so odločitvena drevesa, asociativna pravila, razvrščanje v skupine, nevronske mreže in genetski algoritmi. Podanih je tudi nekaj načinov urejanja podatkov in pomen vizualizacije v podatkovnem rudarjenju.

Drugi del je namenjen teoretičnim spoznanjem s področja geografskih informacijskih sistemov. Predstavljene so značilnosti, oblika geografskih podatkov ter načini pridobivanja in prikazovanja podatkov.

V tretjem delu so predstavljene osnove podatkovnega rudarjenja v geografskih informacijskih sistemih in različni pristopi k podatkovnemu rudarjenju v GIS. Prestavljen je primer uporabe tehnike razvrščanja v skupine z uporabo metode SOM (angl. s*elf-organizing maps*) v geografskih informacijskih sistemih. V primeru so uporabljeni podatki o gospodinjstvih v Lizboni in s pomočjo tehnik podatkovnega rudarjenja ter geografskih informacijskih sistemov predstavljeni rezultati. Opisane so uporabljene metode in orodja ter rezultati, ki sem jih dobila s programom ArcView, ki sem ga uporabila za grafični prikaz podatkov na zemljevidu.

Na koncu diplomske naloge sem poskušala vsa spoznanja zaokrožiti v smiselno celoto.

1 OSNOVE PODATKOVNEGA RUDARJENJA

Izraz podatkovno rudarjenje je izpeljan iz podobnosti med iskanjem koristnih informacij v velikih bazah podatkov – na primer iskanje povezanih produktov v trgovini – in iskanjem rudne žile v gori. Oba procesa zahtevata veliko materiala in inteligentno raziskovanje, z namenom najti prav tisto, kar ima vrednost. Če imamo podatkovno bazo, ki je dovolj velika in kvalitetna, lahko podatkovno rudarjenje odpre nove poslovne priložnosti (Rhodes, 2005, str. 280).

1.1 OPREDELITEV POJMA PODATKOVNO RUDARJENJE

V teoriji ni enotne definicije pojma **podatkovno rudarjenje** (angl. *data mining*) , saj posamezni avtorji podajajo različne definicije. V literaturi se pojavlja tudi izraz izkopavanje podatkov, oba pojma pa pomenita enako.

Razlag je vsaj toliko, kolikor je uporabnih vidikov podatkovnega rudarjenja. Najbolj pogosto in napačno mnenje je, da podatkovno rudarjenje vključuje pretok velike količine podatkov prek inteligentnih sistemov, ki samostojno najdejo vzorce in dajo magično rešitev problema. Kot najpogostejša pravilna definicija se v literaturi pojavlja opredelitev podatkovnega rudarjenja kot odkrivanje vzorcev in pravil v velikih količinah podatkov. Berry in Linoff (1998, str. 7) opredeljujeta pojem kot proces avtomatskega ali polavtomatskega analiziranja velikih količin podatkov, pri čemer je namen odkriti nove, zanimive, uporabne vzorce in pravila. Grobelnik (1999, str. 1) meni, da bi najbolj splošno opredelili, da proces podatkovnega rudarjenja vključuje različne za industrijo značilne koncepte in ideje iz poslovnega sveta in trženja, upravljanja baz podatkov, računalništva in statistike. Podatkovno rudarjenje je interaktiven in ponavljajoč se proces, kjer je treba uporabiti tako znanje problemskega področja kot tudi naprednih tehnologij, da bi ugotovili prikrite povezave in značilnosti podatkov. Ni le množica tehnik, kot se pogosto navaja, čeprav so analize podatkov, izdelava modelov in strojno učenje eni izmed pomembnejših delov procesa podatkovnega rudarjenja. Kovačič, Jaklič, Štemberger in Groznik (2004, str. 244) definirajo podatkovno rudarjenje kot proces avtomatskega iskanja veljavnih, prej nepoznanih vzorcev, pravil in povezav v veliki količini podatkov oziroma napovedovanje na podlagi obstoječih podatkov z namenom pridobitve boljših rezultatov pri odločanju.

Nekateri uporabljajo tudi širši pojem KDD *(*angl. *knowledge discovery from databases)*, ki zajema tudi “pre-”in “postprocesiranje”. Koraki v modelu KDD so (Leskošek, 2006, str. 6):

- Definicija (poslovnega) problema
- Izgradnja baze

- Pojasnjevanje podatkov
- Priprava podatkov za modeliranje
- Izgradnja modela
- Ocena modela

1.1.1 Zgodovina

V evoluciji od poslovnih podatkov do poslovnih informacij je vsak nov korak zgrajen na prejšnjem. Z uporabnikovega vidika so v Tabeli 1 predstavljeni štirje revolucionarni koraki v razvoju podatkovnega rudarjenja, saj dajejo na nova poslovna vprašanja hiter in natančen odgovor (Thearling, 2008).

Tabela 1: Koraki v razvoju podatkovnega rudarjenja

Korak	*Poslovno vprašanje*	*Tehnologije*	*Značilnosti*
Zbiranje podatkov (1960)	Kakšen je bil prihodek v prejšnjem letu?	Računalniki	Statično pridobivanje podatkov
Dostop do podatkov (1980)	Kakšna je bila prodaja na določenem oddelku lanskega marca?	Relacijske baze, SQL	Dinamično pridobivanje podatkov
Podpora odločanju (1990)	Koliko izdelkov smo prodali v določen kraj iz določenega oddelka lanskega marca?	OLAP, podatkovna skladišča	Dinamično pridobivanje podatkov na več nivojih
Podatkovno rudarjenje (2000–danes)	Kaj pričakujemo, da se bo zgodilo na določenem oddelku naslednji mesec? Zakaj?	Velike baze podatkov, napredni algoritmi	Proaktivno pridobivanje informacij

Vir: K. Thearling, An Introduction to Data Mining, 2008.

1.1.2 Naloge podatkovnega rudarjenja

Izraz podatkovno rudarjenje je sklop nalog, ki vse vsebujejo izločevanje pomembnih novih informacij iz podatkov (Berry & Linoff, 1998, str. 9). Teh šest nalog je:

- Klasifikacija
- Ocena
- Napoved
- Asociativna pravila
- Razvrščanje v skupine
- Opisovanje in vizualizacija

Prve tri naloge – klasifikacija, ocena, in napoved – so primeri **usmerjenega** podatkovnega rudarjenja. V usmerjenem podatkovnem rudarjenju je cilj uporabiti dane podatke in zgraditi model, ki opiše določeno spremenljivko. Klasifikacija vključuje pregledovanje in določanje novega objekta v prej definiran razred; na primer v praksi določanje kupcev v prej določene segmente. Naloga je zgraditi model, ki ga bomo lahko uporabili na neurejenih podatkih z namenom klasificiranja teh podatkov v enega izmed določenih razredov. Klasifikacija se torej ukvarja z ločenimi rezultati: na primer da ali ne. Oceno pa uporabimo, ko imamo določene vstopne podatke in potrebujemo oceno za neznano spremenljivko (na primer dohodek, višino). V praksi je ocena uporabljena za klasifikacijo: banka se odloča, komu bi ponudili posojilo, in ocenijo, katere stranke bi se odzvale na ponudbo, ocenijo število otrok v družini, skupen družinski dohodek, ocenijo vrednosti nepremičnin. Pogosto sta klasifikacija in ocena uporabljeni skupaj. Katerakoli od tehnik, uporabljena za klasifikacijo in oceno, pa je lahko prilagojena za uporabo pri napovedi. Prilagodimo jo z uporabo primerov, kjer je vrednost spremenljivke, ki jo napovedujemo, že znana, in hkrati z zgodovinskimi podatki za te primere.

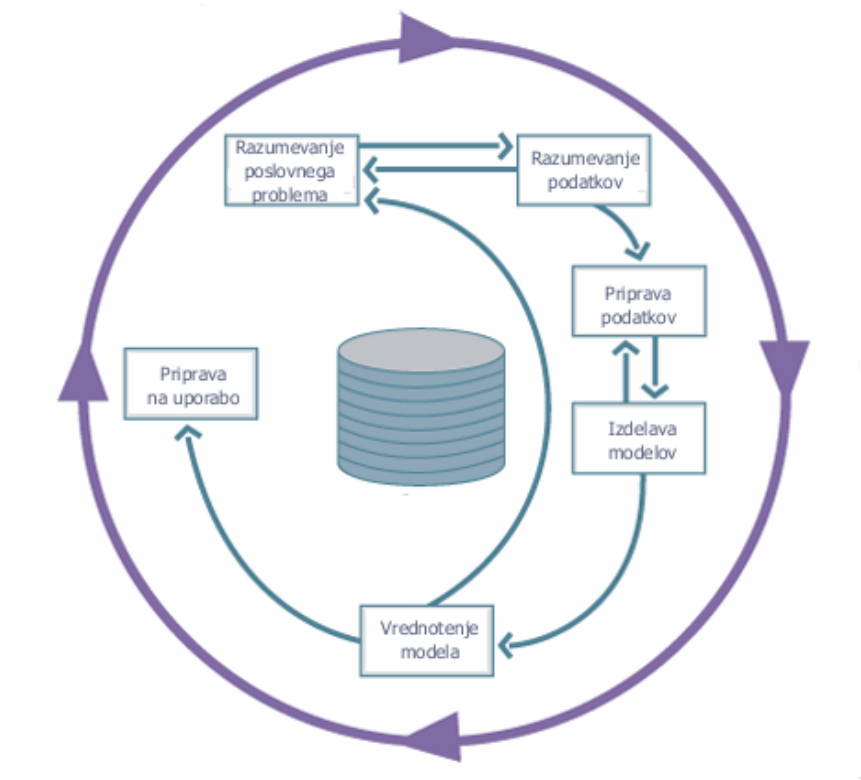
Naslednje tri naloge so primeri **neusmerjenega** podatkovnega rudarjenja. Pri neusmerjenem podatkovnem rudarjenju je cilj dokazati povezave med spremenljivkami. Torej, ciljna spremenljivka ni znana. Asociativna pravila nam povedo, katere spremenljivke sodijo skupaj, na primer pomagajo pri iskanju povezav med prodajo izdelkov. Razvrščanje v skupine pa je naloga segmentacije raznolike skupine v večje število podskupin (angl. *clusters*). Kar razlikuje razvrščanje v skupine od klasifikacije, je to, da pri razvrščanju v skupine razredi niso prej definirani. Tehniki asociativnih pravil in razvrščanja v skupine bosta podrobneje predstavljeni v nadaljevanju tega poglavja. Včasih pa je smoter podatkovnega rudarjenja enostavno opisati, kaj se dogaja v zapleteni bazi podatkov, da bi lažje razumeli procese, produkte in ljudi.

1.1.3 Koraki v modelu podatkovnega rudarjenja

Podatkovno rudarjenje ne pomeni samo uporabe orodja za rudarjenje, temveč ga moramo opazovati skozi celoten proces, kot je prikazan na Sliki 1. Zelo pogosta napaka pri rudarjenju je slabo definiranje poslovnega problema, ki ga želimo z rudarjenjem rešiti. V tem primeru je le malo verjetno, da bo rudarjenje v poslovnem smislu uspešno, čeprav dobimo dobre in veljavne rezultate (Kovačič et al., 2004, str. 256).

Metodologija CRISP-DM (angl. *cross industry standard process for data mining*) opisuje korake v modelu podatkovnega rudarjenja – kako se spopasti s poslovnim problemom in ga nato preoblikovati v problem podatkovnega rudarjenja. Procesni model, ki ga predpisuje metodologija CRISP-DM, je sestavljen iz šestih korakov (Grobelnik, 1999, str. 2):

Slika 1: Cikel podatkovnega rudarjenja po CRISP-DM



Vir: M. Grobelnik, Data Mining s Clementine (SPSS) – teorija in praksa, 1999, str. 2.

- Razumevanje (poslovnega) problema: Prvi korak se osredotoča na razumevanje ciljev in zahtev projekta z vidika (poslovnega) problema, osredotoča pa se tudi na oblikovanje začetnega programa za doseganje ciljev projektov.
- Razumevanje podatkov: Drugi korak se začne s prvim zbiranjem podatkov ter spoznavanjem z njimi, kar pripomore k lažjemu ugotavljanju morebitnih težav glede kakovosti podatkov, hkrati pa raziskovalec dobi prvi vpogled vanje in zazna zanimive podskupine za oblikovanje hipotez o prekritih informacijah.
- Priprava podatkov: Faza priprave podatkov pokriva vse aktivnosti za oblikovanje končne strukture podatkovne baze, ki sloni na surovih podatkih. Ta podatkovna baza bo pozneje zapolnjena z rezultati, dobljenimi z različnimi modeli. Naloge v tej fazi so: izbira tabele, zapisov in atributov kot tudi transformacije in čiščenje podatkov. Takšni podatki se pozneje uporabijo pri

modeliranju. Te naloge se navadno večkrat ponovijo, vendar v različnem vrstnem redu.

- Izdelava modelov: V tej fazi se izbira in uporablja različne tehnike modeliranja ter popravlja različne parametre za optimizacijo vrednosti. Navadno obstaja več tehnik za reševanje istega problema izkopavanja podatkov. Nekatere tehnike imajo specifične zahteve glede oblike podatkov, zato se je pogosto treba vrniti k fazi priprave podatkov.
- Vrednotenje modela: Pri tem koraku projekta je treba zgraditi model (ali več modelov). Ker se z njimi opravlja različne analize podatkov, je pomembno, da so ti modeli kakovostni. Pred končno uporabo modela je potrebna podrobnejša evalvacija modela ter korakov, ki smo jih uporabili pri oblikovanju končnega modela. S tem zagotovimo dosego cilja. Poiskati moramo še morebitna dejstva, ki pri oblikovanju trenutnih modelov niso bila upoštevana. Ob koncu te faze je cilj izkopavanja podatkov dosežen.
- Priprava za uporabo: Projekt se navadno ne konča z oblikovanjem končnega modela. Čeprav je namen modela povečati znanje o podatkih, bo treba pridobljeno znanje organizirati in predstaviti tako, da ga bodo naročniki/kasnejši uporabniki lahko uporabili. Glede na njihove zahteve lahko ta faza vsebuje le oblikovanje poročila ali pa bolj zapleteno uporabo. Modelov ne bo vedno uporabljal le raziskovalec, temveč bo uporabnik pogosto stranka. Zato je pomembno, da tudi stranka razume, kaj bo treba narediti za učinkovito uporabo dobljenih modelov.

Čeprav je bil model razvit za potrebe večjih, kompleksnejših projektov, je dovolj robusten, da ga je mogoče uporabiti tudi za katerikoli manjši projekt podatkovnega rudarjenja.

1.2 PRIPRAVA PODATKOV

V realnosti podatki niso že vnaprej pripravljeni za podatkovno rudarjenje, zato moramo podatke transformirati v obliko, ki jo potrebujemo. Ker podatke lahko dobimo iz različnih delov organizacij, iz različnih sistemov (transakcijski sistemi, podatkovna skladišča, ankete, zunanji viri podatkov), so ti večinoma neurejeni. Ponavadi so shranjeni v relacijski bazi podatkov. K sreči pa lahko podatke prenesemo, saj podatkovno rudarjenje to tipično zahteva. Za podatkovno rudarjenje praviloma potrebujemo podatke v tabelarni obliki. Vrstice so enote podatkovnega rudarjenja, stolpci pa opisujejo atribute (Berry & Linoff, 1998, str. 132–143).

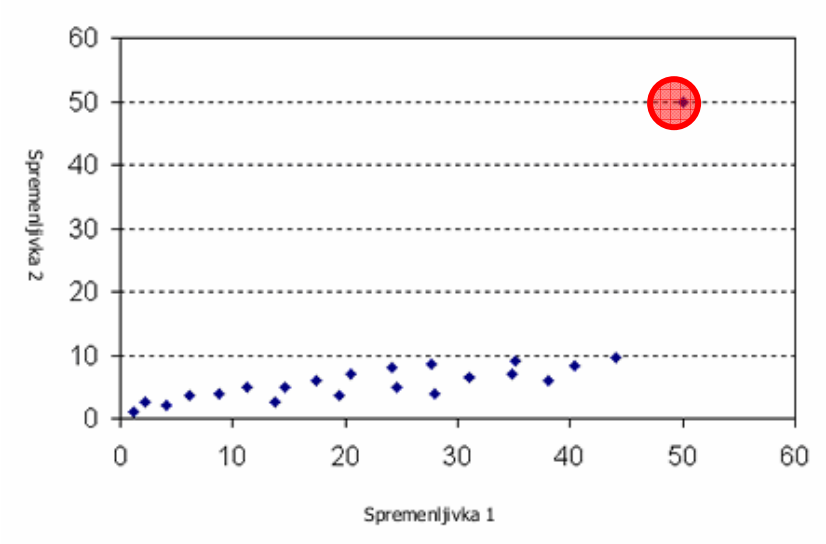
Urejanje podatkov do točke, kjer lahko podatke učinkovito uporabimo za podatkovno rudarjenje, se imenuje priprava podatkov (angl. *data preparation*).

Bação in Lobo (2005, str. 3) in Berry & Linoff (1998, str. 161) menijo, da je postopkov priprave podatkov veliko in razvitih je bilo več različnih tehnik za obravnavanje teh problemov, zato naj naštejem le najbolj pogosto uporabljene:

Z **obravnavanjem manjkajočih vrednosti** nadomestimo manjkajoče vrednosti s pametnim ugibanjem o pravi vrednosti. Razlogov za manjkajočo vrednost je lahko več, zato poskušamo razumeti, zakaj vrednosti ni. Če odstranimo celoten zapis, lahko izgubimo pomembne informacije. Ker nekateri algoritmi dovoljujejo manjkajoče vrednosti, lahko te ignoriramo. Možno je tudi napovedati vrednosti, redko pa se uporablja tudi nadomestitev s povprečno vrednostjo spremenljivke.

Odstranjevanje napak in odstopajočih vrednosti (angl. *outliers*) predstavlja odkrivanje vzorcev v podatkih, ki zaradi različnih razlogov ne predstavljajo zahtevanih vrednosti. Če uporabimo katerokoli tehniko podatkovnega rudarjenja na takšnih podatkih, bomo dobili rezultate, ki bodo vsebovali napake. Najti te nepravilne podatke (na primer človek, velik tri metre) pa ponavadi ni enostavno. Vrednost, ki odstopa, je lahko napaka ali pa prava (rekordna) vrednost. Kot kaže Slika 2, je razpršitveni diagram uporabno orodje za razumevanje porazdelitve spremenljivke. Včasih je lahko nezanesljiv pri določanju odstopajočih vrednosti, še posebej takrat, ko je zadnji razred določen kot odprt interval. V takšnih primerih uporabimo dodatne grafe. Nekatere odstopajoče vrednosti se pojavijo le v kombinaciji z drugo spremenljivko, zato lahko uporabimo tudi histograme.

Slika 2: Primer odstopajoče vrednosti



Vir: F. Bação, Predmet Data Mining, prosojnice predavanj, 2007.

Ko najdemo odstopajočo vrednost, najprej poskušamo ugotoviti, kaj jo je povzročilo, nato pa uporabimo enega od načinov pri delu z odstopajočimi vrednostmi:

- Ne naredimo ničesar, saj pri nekaterih tehnikah podatkovnega rudarjenja te vrednosti nimajo velikega učinka (odločitvena drevesa). Po drugi strani pa lahko nekaj odstopajočih vrednosti prinese popolnoma napačne rezultate (nevronske mreže).
- Filtriramo vrstico, ki vsebuje odstopajočo vrednost. To je v nekaterih primerih slaba odločitev, saj s tem izgubimo pomembne podatke.
- Izbrišemo celotno spremenljivko, torej izbrišemo celoten stolpec in ga nadomestimo z opisom spremenljivke.
- Nadomestimo vrednost s primerno vrednostjo. Ta način se najpogosteje uporablja, vrednost pa lahko nadomestimo z 0 ali s povprečno vrednostjo spremenljivke. V nekaterih primerih je smiselno vrednost napovedati, če imamo na voljo primerne podatke, iz katerih lahko napovemo približno pravilno vrednost.

Pri **normalizaciji podatkov** gre za preoblikovanje slučajne spremenljivke, ki ni normalno porazdeljena, v slučajno spremenljivko, ki je vsaj približno normalno porazdeljena. Ta postopek bo preprečil prevladovanje ene spremenljivke nad drugimi in je najbolj pomemben pri uporabi tehnik podatkovnega rudarjenja, kot so najbližji sosed, razvrščanje v skupine in večina nevronskih mrež.

1.3 TEHNIKE

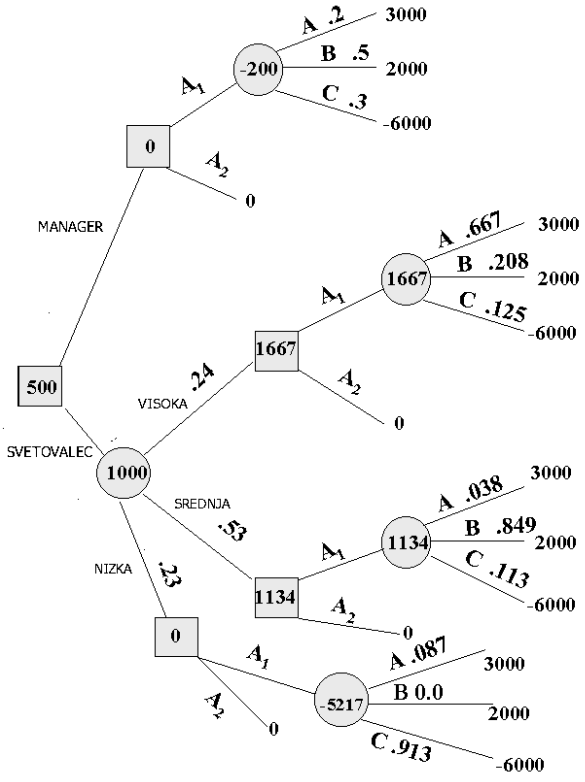
V nadaljevanju bom predstavila najpogosteje uporabljene tehnike podatkovnega rudarjenja: odločitvena drevesa, asociativna pravila, razvrščanje v skupine, nevronske mreže in genetske algoritme.

1.3.1 Odločitvena drevesa

Že samo ime odločitvena drevesa (angl. *decision trees*) pove, da gre za napovedovalne modele, ki so zgrajeni v obliki dreves. Tehnika je bila prvotno razvita za potrebe statistike, danes pa je uporabljena za avtomatizacijo procesov odločanja. Ime po drevesu je tehnika dobila zaradi prikaza v obliki diagrama, ki je videti kot drevo. Diagram je sestavljen iz treh bistvenih komponent, in sicer vozlišč odločanja, vej in listov (Zidar, 2005, str. 7). Sicer pa odločitvena drevesa spadajo v širši razred metod induktivnega učenja, ki je podmnožica avtomatskega učenja iz rešenih primerov. Učenje z indukcijo je oblika učenja z množico primerov, ki so opisani z vhodi in izhodi. Cilj je najti hipotezo, ki temelji na učnih primerih in je sposobna uspešno predvideti izhode še nerešenih primerov. Njihov pristop je hkrati enostaven in učinkovit, predvsem pa je uporaben, ker skupaj z odločitvijo poda tudi popolno obrazložitev poti (sprehod po posameznih vejah odločitvenega drevesa), po kateri pridemo do ponujene odločitve (Mertik & Jevčič, 2001, str. 3).

Reševanja se s pomočjo odločitvenih dreves lotimo tako, da pripravimo množico rešenih primerov. Množico razdelimo na učno, s katero bomo gradili odločitveno drevo, in testno, s katero bomo preverjali zanesljivost zgrajenega drevesa. Najprej določimo množico atributov, ki opisujejo posamezen primer (vhodni podatki), in izberemo en odločitveni atribut, ki pomeni rešitev zastavljene naloge (izhodni podatek). Vsem atributom nato določimo razrede možnih vrednosti. Pri atributih z diskretnimi vrednostmi predstavlja vsaka diskretna vrednost svoj razred, pri atributih z numeričnimi vrednostmi pa razrede določimo po intervalih. Pri gradnji drevesa lahko vsak atribut predstavlja eno notranje vozlišče drevesa, ki ga imenujemo tudi atributno vozlišče. Takšno atributno vozlišče ima natančno toliko vej, kot ima atribut različnih vrednostnih razredov. Listi drevesa pa predstavljajo odločitve in so eden izmed vrednostnih razredov odločitvenega atributa. Ko iščemo rešitev za nek nov, še nerešen primer, začnemo v korenu drevesa in se v vsakem atributnem vozlišču odločimo za vejo glede na vrednost pripadajočega atributa v nerešenem primeru, vse dokler ne pridemo do lista in s tem do predlagane rešitve. Primer drevesa vidimo na Sliki 3 (Mertik & Jevčič, 2001, str. 5).

Slika 3: Primer odločitvenega drevesa



Vir: H. Arshan, Tools for decision analysis, 2008.

Na Sliki 3 je prikazan primer odločitvenega drevesa, ki bi podjetju pomagal pri odločitvi, ali prodajati določen izdelek ali ne. Prva možnost je, da se menedžer odloči sam. A_1

pomeni odločitev za prodajo, A_2 pa, da izdelka ne prodaja. Druga možnost je najem svetovalca in nato čakanje na njegovo analizo. Če analiza napove visoko ali srednjo povečanje prodaje, nato sledi prodaja določenega izdelka. Če analiza napove nizko prodajo, določenega izdelka ne prodajamo. Če se odloči za prodajo, potem sledijo verjetnosti za določeno višino prodaje. A pomeni visoko dejansko prodajo, B srednjo in C nizko. Z uporabo odločitvenega drevesa lahko izračunamo, katera možnost se najbolj splača.

1.3.2 Asociativna pravila

Asociativna pravila imajo podoben namen kot drevesa odločanja, le da so v obliki pravil »če-potem« (angl. *if-then*). Asociativna pravila so povezave oziroma vzorci, ki jih v podatkih odkrijemo s podatkovnim rudarjenjem. Odkrivanje takšnih pravil temelji na konceptu transakcij, in sicer: če se zgodi dogodek A, potem lahko z določeno verjetnostjo predvidimo dogodek B. Pravila so oblike (Kovačič et al., 2004, str. 250):

"Če se zgodi A1...An, potem se v okviru iste transakcije pogosto zgodi tudi B1...Bm."

Primer takšnega pravila je: če kupec kupi rolerje, bo verjetno kupil tudi ščitnike za kolena, zato se asociativna pravila uporabljajo pri analizi nakupne košarice.

Asociativna pravila se med seboj vsebinsko lahko razlikujejo in lahko jih tudi številčno ovrednotimo, pri čemer lahko merimo kazalnike, ki so prikazani v Tabeli 2:

Tabela 2: Razlaga kazalnikov asociativnih pravil

Podpora (angl. *support*)	Napove verjetnost, da transakcija vsebuje A in B	$P(A \& B)$
Zaupanje (angl. *confidence*)	Napove verjetnost, da transakcija, ki vsebuje A, vsebuje tudi B	$\frac{P(A \& B)}{P(A)}$
Izboljšanje (angl. *improvement*)	Napove, koliko boljše je pravilo pri napovedovanju, kakor je bila naključna izbira	$\frac{P(A \& B)}{P(A) \cdot P(B)}$

Vir: J. Jaklič, Predmet Tehnologija za podporo odločanju, prosojnice predavanj, 2006.

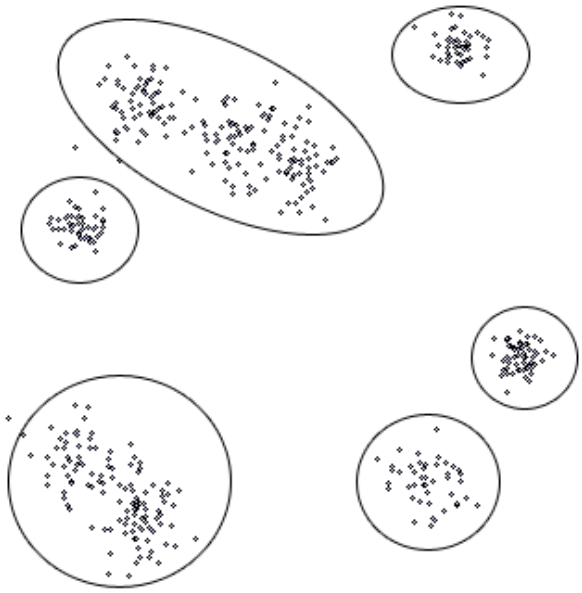
1.3.3 Razvrščanje v skupine

Razvrščanje v skupine je neusmerjeno klasificiranje podatkov v skupine (angl. *clusters*) glede na njihovo podobnost. Namen razvrščanja enot v skupine je uvrstiti enote v skupine po principu podobnosti tako, da so znotraj posamezne skupine enote, ki so si glede na vnaprej določen kriterij podobne, znotraj različnih skupin pa enote, ki so si glede na ta kriterij različne. Vsaka enota je uvrščena v eno samo skupino; torej se skupine ne prekrivajo (Košmelj & Breskvar, 2006, str. 299). Slika 4 nam kaže primer razvrščanja v skupine.

Postopek razvrščanja v skupine ima naslednje korake (Košmelj & Breskvar, 2006, str. 299):

- Izbira enot, izbira njihovih lastnosti
- Standardizacija spremenljivk, če je ta potrebna
- Izbira ustrezne razdalje (različnosti) med enotami, ki je odvisna od vrste
- podatkov in od kriterija podobnosti
- Uporaba različnih metod razvrščanja
- Analiza rezultatov

Slika 4: Primer razvrščanja v skupine



Vir: F. Bação, Predmet Data Mining, prosojnice predavanj 2007.

Metode razvrščanja v skupine predstavljajo pomemben del podatkovnega rudarjenja. V veliko primerih želimo pridobiti novo znanje o problemu, tudi če imamo malo predvidevanj o podatkih ali celo problemu. V teh primerih nam takšne metode olajšajo

delo ter obogatijo naše znanje. Metode delimo na hierarhične in nehierarhične. Kratko bom predstavila dve nehierarhični metodi za segmentacijo podatkov: metodo voditeljev (angl. *K-means*) in metodo SOM (angl. *self-organizing maps*).

Metoda voditeljev (angl. *K-means clustering*)

Program izbere *k* točk slučajno, te točke so začetni voditelji. *K* definiramo sami in pomeni, koliko skupin želimo imeti. Vsaka enota se nato doda najbližjemu voditelju. Ko so vse enote dodane in s tem narejene nove skupine, se izračunajo težišča novonastalih skupin, ki postanejo novi voditelji. Ta postopek se ponavlja, dokler se voditelji ne premikajo več. Tedaj se izračuna vrednost kriterijske funkcije. Dobljena razvrstitev je odvisna od začetne pozicije voditeljev, zato se ta postopek velikokrat ponavlja z različnimi začetnimi konfiguracijami voditeljev. Kot najboljšo razvrstitev se vzame tisto, ki ima najmanjšo vrednost kriterijske funkcije (Košmelj & Breskvar, 2006, str. 299).

Metoda SOM

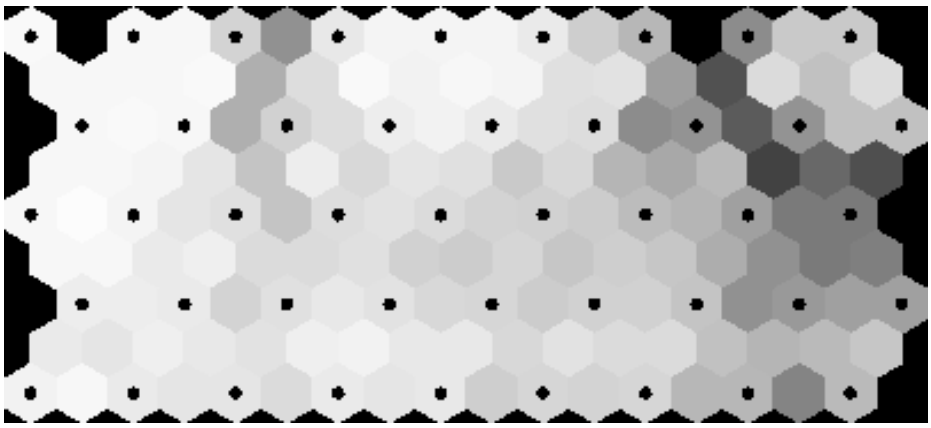
SOM je kratica za *self-organizing maps*, kar pomeni samoorganizirajoče karte. Drugo ime zanjo je Kohonenova mreža (po avtorju Teuvo Kohonen). Uporabna je predvsem za predstavitev strukturiranih statističnih podatkov. Metoda SOM je novejše orodje in je umetna nevronska mreža, ki simulira delovanje funkcij človeških možganov (Bação & Lobo, 2006b). Odkritje SOM temelji na biološkem spoznanju, da podobni dražljaji vzburijo sosednje nevrone. Če pripeljemo na SOM dražljaj, se vzburijo samo področja na mreži, ki ustrezajo podobnim dražljajem. Nevroni se vedejo kot topološka spominska karta, kjer je položaj najbolj vzdraženega nevrona v korelaciji z značilnostmi dražljaja (Kaski & Kohonen, 1995, str. 3).

Vsi vhodi so povezani z vsemi nevroni v Kohonenovi mreži. Vsak vhodni dražljaj posredujemo po povezavah do vseh nevronov tekmovalnega sloja. Uteži povezav med vhodi in nevroni določajo točko v vhodnem prostoru za trenutni vhodni dražljaj.

Bação in Lobo (2006a, str. 17) opozarjata, da namesto zapletenega algoritma lahko uporabimo javno dostopen program SOM_PAK, ki nam podatke predstavi v sivinah. Možnost vizualizacije metode SOM nam omogoča dodaten vpogled v podatke in njihovo sestavo. Eden od načinov vizualizacije rezultatov je uporaba U-matrike (primer je prikazan na Sliki 5 na strani 14). Ta način je tudi rezultat, ki ga dobimo z uporabo programa SOM_PAK. U-matrika predstavi večdimenzionalne podatke v dvodimenzionalni obliki v barvi ali sivinah. Če so razdalje med sosednjimi nevroni majhne, potem ti nevroni predstavljajo skupino vzorcev s podobnimi značilnostmi.[1]

[1] Metoda SOM je podrobneje opisana na primeru v poglavju 3.3.3.

Slika 5: Primer U-matrike

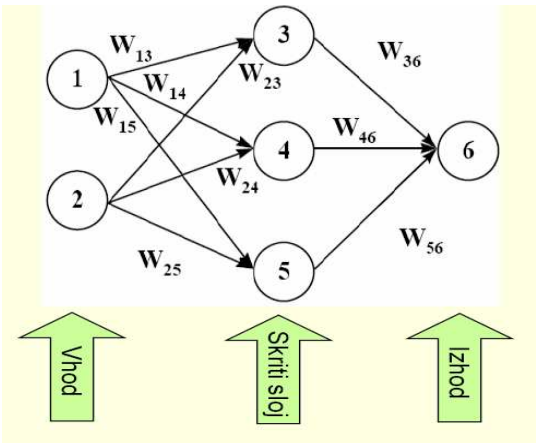


Vir: K. Kaski & T. Kohonen, Exploratory data analysis by the self-organizing map, 1995, str. 501.

1.3.4 Nevronske mreže

Nevronske mreže so ime dobile po nevronih in naj bi posnemale delovanje človeških možganov. Nevroni so organizirani v mreže, od katerih vsaka vsebuje nekaj tisoč celic. Pomembna lastnost takšnih nevronskih mrež je njihova zmožnost učenja in ohranjanja informacij (Berry & Linoff, 2004, str. 286). Nevronske mreže spominjajo na možgane vsaj z dveh pogledov: znanje si pridobivajo s procesom učenja, za shranjevanje znanja pa uporabljajo mednevronske povezave oziroma sinaptične uteži. Nevronsko mrežo lahko naučimo, da opravlja določeno funkcijo s spreminjanjem vrednosti povezav oziroma uteži med elementi (Batagelj, 2002, str. 1). Primer nevronske mreže je prikazan na Sliki 6.

Slika 6: Primer nevronske mreže



Vir: B. Leskošek, Informacijske tehnologije za podporo odločanju: odločitveni in ekspertni sistemi, 2006.

Elementi nevronske mreže so (Si & Nelson, 2003, str. 42):

- Vhodni sloj: vrednost spremenljivk.
- Skriti sloj: funkcije nad utežnimi vsotami vhodov (slojev je lahko več, le pri preprosti nevronski mreži imamo le en sloj).
- Izhodni sloj: vrednost odvisne spremenljivke .
- Uteži: dobimo jih kot rezultat učenja nevronske mreže (na Sliki 6 so označene kot W_i).

Mreža se najprej uči na podatkih, za katere je vrednost napovedi že znana. Potem uporabimo drugo množico, testne podatke, za katere prav tako poznamo vrednost napovedi in mrežo preizkusimo (Kovačič et al., 2004, str. 253). Postopek učenja poteka v več korakih:

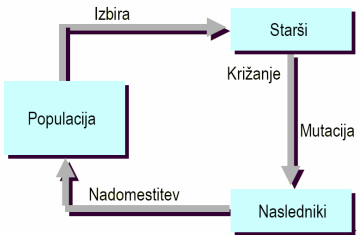
1. korak: V mrežo pošljemo dražljaj.
2. korak: Mreža poišče najbolj vzburjeni nevron v tekmovalnem sloju (središče), ki ima utežni vektor z najmanjšo razdaljo do dražljaja.
3. korak: Prilagodimo uteži najbolj vzburjenega nevrona, da se razdalja še zmanjša.
4. korak: Prilagodimo tudi nevrone v okolici. Bolj oddaljene od središča prilagodimo manj.

Postopek ponovimo za ves nabor učnih podatkov v naključnem vrstnem redu. To se imenuje učenje. Učenje je končano, ko smo z rezultati testiranja zadovoljni oziroma ko postane nevronska mreža stabilna.

1.3.5 Genetski algoritmi

Genetski algoritmi oziroma evolucijsko računanje posnemajo načela biološkega razvoja. Kot kaže Slika 7, nove elemente (naslednike), ki obdržijo določene lastnosti njihovih »staršev«, dobimo s selekcijo uspešnih staršev. Populacija je skupina osebkov, ki obstajajo hkrati. V vsakem koraku genetskega algoritma iz populacije staršev z genetskimi operatorji dobimo populacijo potomcev.

Slika 7: Evolucijski cikel genetskih algoritmov



Vir: B. Leskošek, Informacijske tehnologije za podporo odločanju: odločitveni in ekspertni sistemi, 2006.

Glavne aktivnosti v postopku genetskih algoritmov so: selekcija, križanje in mutacija (Liu & Motoda, 2005, str. 418–419):

- Selekcija (angl. *selection*) je izbiranje boljših elementov iz populacije oziroma elementov z določenimi lastnostmi.
- Križanje (angl. *crossover*) deluje na dveh osebkih populacije (starših) in nam vrne dva nova osebka (naslednika).
- Mutacija (angl. *mutation)* deluje na enem samem osebku tako, da ga naključno spremeni.

Genetski algoritmi se uporabljajo pri težkih optimizacijskih problemih, avtomatičnem programiranju, ekologiji in reševanju ekonomskih vprašanj.

1.4 VIZUALIZACIJA V PODATKOVNEM RUDARJENJU

Veliko podatkovnih baz vsebuje velike količine večdimenzionalnih podatkov, zaradi katerih je težko najti koristne informacije. Vizualizacija nam pomaga v procesu podatkovnega rudarjenja v dveh smereh. Prva nam lahko pripravi nazorno sliko rezultatov, ki jih dobimo z zapletenimi algoritmi. Druga je lahko uporabna za odkritje kompleksnih vzorcev v podatkih, ki jih ne zaznamo s predhodnimi metodami, ampak so lahko identificirani z vizualno predstavitvijo (Demšar, 2006, str. 13).

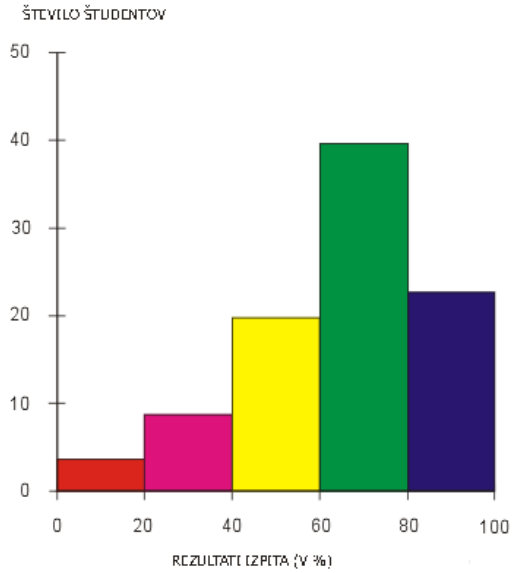
V teoriji obstajajo različne razdelitve tehnik vizualizacije. Grinstein in Ward (2002, str. 22) delita tehnike na geometrične in simbolične. Geometrične tehnike predstavljajo podatke kot raztreseni diagram, linije, površine ali volumne. Simbolično osredotočene tehnike predstavljajo nenumerične podatke z ikonami ali grafi. Poudarek pri teh je prikazati povezanost med podatkovnimi elementi. Bolj splošna razdelitev je glede na dimenzijo: enodimenzionalna, dvodimenzionalna ali večdimenzionalna predstavitev. Nekateri avtorji delijo vizualizacijske tehnike tudi na statične in dinamične (Demšar, 2006, str. 14).

Primeri najbolj pogosto uporabljenih tehnik vizualizacije so (Hoffman & Grinstein, 2002, str. 126):

- Standardni 2-D ali 3-D prikazi (stolpčni graf, linijski graf, histogram, tortni diagram); primer 2-D histograma je na Sliki 8 (na strani 17).
- Razpršeni diagram (angl. *scatterplot*)
- Matrika razpršenih prikazov (angl. *scatterplot matrix*)
- Analiza osnovnih komponent (angl. *principal component analysis*)
- Večrazsežno lestvičenje
- Kvantilni diagram (angl. *box plot*); primer kvantilnih diagramov je prikazan na Sliki 9 (na strani 17).
- Paralelne koordinate

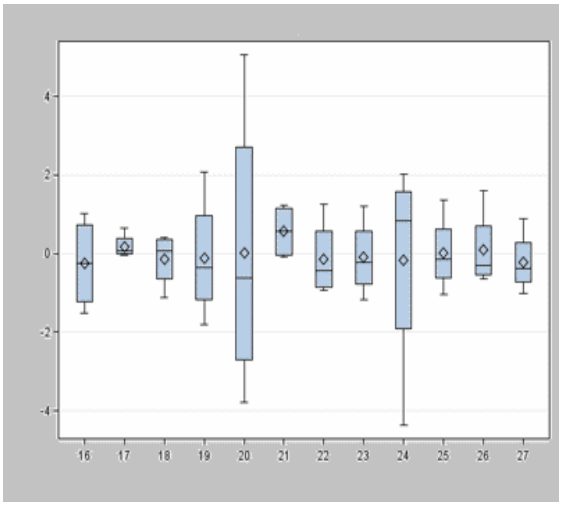
- Kohonenove mreže za razvrščanje v skupine – na Sliki 10 je prikazana Kohonenova mreža na primeru držav sveta; države so razdeljene v skupine glede na njihovo razvitost; svetlejša polja, ki niso pregrajena s temnimi, pomenijo podobne značilnosti.

Slika 8: Primer histograma



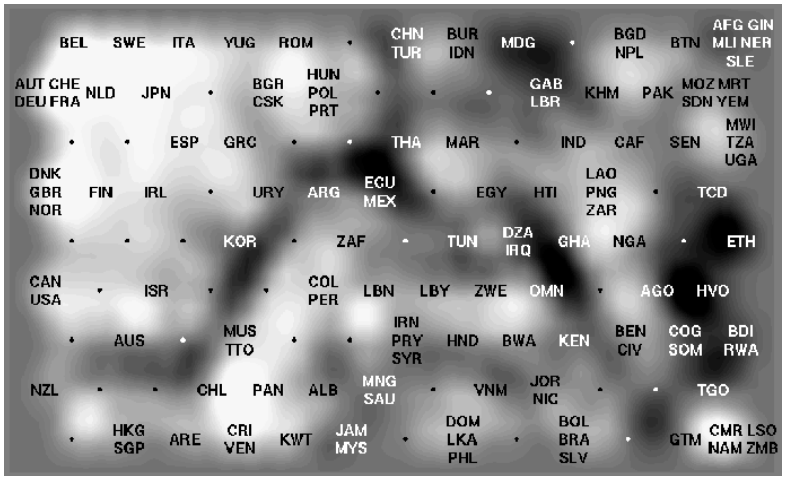
Vir: Search software quality, 2008.

Slika 9: Primer kvantilnega diagrama



Vir: SAS customer support, 2008.

Slika 10: Kohonenova mreža na primeru držav



Vir: A. Jakulin, Večrazsežno lestvičenje, 2008, str. 49.

2 GEOGRAFSKI INFORMACIJSKI SISTEMI

Lahko bi rekli, da je geografski informacijski sistem »pametna« karta, ki nam omogoča pridobivanje odgovorov na najrazličnejša vprašanja, na primer katera področja imajo primerno kvaliteto tal in so na prisojnih pobočjih, da bi jih lahko uporabili za vinograde; kako postaviti radijske oddajnike za optimalno pokritost prebivalstva in podobno. Ta sistem torej ne odgovarja zgolj na enostavna vprašanja glede pozicije, ampak kombinira najrazličnejše podatke – tako prostorske kot neprostorske (tematske). Zato so geografski podatki v današnjem času postali osnova za kvalitetno odločanje (Geografski informacijski sistemi – Rešitve za terenski zajem podatkov, 2008).

2.1 KAJ JE GEOGRAFSKI INFORMACIJSKI SISTEM?

V svetu se je vzporedno razvijalo več opredelitev geografskega informacijskega sistema, ki pa vsaka po svoje dejansko opisujejo isti sistem. Ena najbolj razširjenih definicij je:

Geografski informacijski sistem (v nadaljevanju GIS) je skupek programske in strojne opreme, postopkov, ki omogočajo urejanje, analiziranje, upravljanje, modeliranje in prikaz geografsko usmerjenih podatkov z namenom reševanja kompleksnih problemov planiranja in upravljanja virov (Tomlinson, 2003, str. 4).

Krajša definicija GIS pravi, da je to računalniško podprt sistem, ki omogoča uporabo podatkov, ki opisujejo realen prostor (Steinberg & Steinberg 2006, str. 6).

Steinberg & Steinberg (2006, str. 7) menita, da čeprav ni ene same definicije GIS, se strokovnjaki za te sisteme strinjajo v več glavnih načelih. Prvič, kot že omenjeno, GIS zahteva kombinacijo strojne in programske opreme. Drugič, GIS za delovanje potrebuje podatke, ki morajo imeti prostorsko ali lokacijsko komponento. Tretjič, GIS potrebuje strokovnjake, ki oblikujejo bazo podatkov in procesirajo zahtevane podatke. Večino nalog lahko opravi vsakdo, ki ima vsaj osnovno znanje računalništva. Čeprav so GIS-aplikacije od začetka geografskih uporabniških vmesnikov postale veliko preprostejše za uporabo, tudi danes zahtevajo od uporabnika, da ima znanje o zemljevidih in o analizi zemljevidov. Zadnje, morda celo najpomembnejše, GIS je sistem za analitike. GIS je namreč uporaben za raziskavo in predstavitev končnih rezultatov v obliki informacij, ki so poznane analitiku.

Vidmar, Zdešar, Herlec in Ferlan (2008, str. 1) menijo, da so značilnosti GIS naslednje:

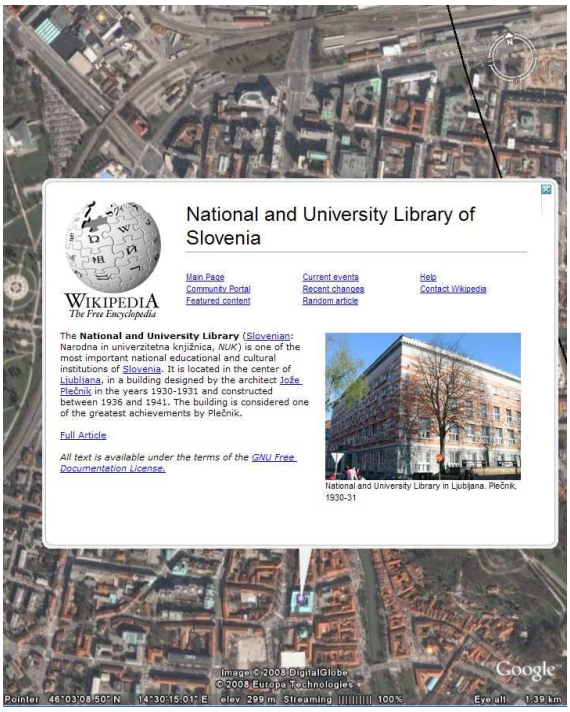
- Zajemanje prostorskih podatkov
- Urejanje, obdelava in pretvorba prostorskih podatkov
- Shranjevanje prostorskih podatkov
- Obnavljanje in spreminjanje prostorskih podatkov
- Manipulacija s podatki in izmenjava prostorskih podatkov
- Iskanje, povezovanje in predstavitev prostorskih podatkov
- Analize in kombinacije prostorskih podatkov

Kot raziskovalno orodje nas GIS oskrbi s priložnostmi za raziskovanje povezav v prostoru, velikokrat s podatki, ki prvotno niso namenjeni za raziskave v ekonomskih ali socialnih vedah (Tomlinson, 2003, str. 4).

V praksi lahko funkcije GIS predstavimo tudi samo z uporabo lista, svinčnika in ravnila in veliko ljudi še vedno uporablja ta pristop, ki pa seveda ni praktičen in učinkovit. Rezultati, ki nam jih GIS ponudi, so ponavadi v obliki zemljevida ali prostorskih informacij.

Ko omenimo prostorske informacije ali podatke, s tem mislimo, da so informacije povezane s specifično lokacijo, na primer naslovom. Slika 11 nam kaže primer, predstavljen v fotografiji nekega območja, kjer poskušamo prepoznati povezane podatke. Ti tabelarni podatki so povezani prek lokacije ali prostorske informacije (Steinberg & Steinberg, 2006, str. 8).

Slika 11: Fotografija območja in primer povezovanja podatkov



Vir: Google Earth, 2008

Pridobivanje kakovostnih GIS-podatkov je najzahtevnejši in najdražji del GIS. Pri tem lahko pridobivanje pomeni nakup ali izposojo podatkov ali pa zajem podatkov na terenu. Za pridobivanje geografskih podatkov je na voljo mnogo različnih tehnologij – daljinsko zaznavanje, aero snemanje, geodetsko snemanje in terenski zajem podatkov z GPS.

2.2 OBLIKA ZAPISA PODATKOV GEOGRAFSKIH INFORMACIJSKIH SISTEMOV

Ko opredelimo naše cilje raziskave GIS, je naslednji korak izbira podatkovnega modela, ki ga bomo uporabili. V večini primerov bo to odločitev namesto nas naredila programska rešitev. Če bomo uporabili dejansko okolje GIS, je pomembno vedeti, kateri program je zmožen operirati s temi podatki. To pomeni, da če nimamo znanja in možnosti dostopanja in odpiranja določenih podatkovnih formatov, nekateri široko dostopni in potencialno zelo uporabni nabori podatkov ne bodo imeli nikakršne vrednosti za uporabnika. Veliko najbolj priljubljenih programskih paketov ponuja podporo vseh širše poznanih podatkovnih formatov. Veliko spornih vprašanj (in potencialnih stroškov) je povezanih z enostavnim dostopom do nabora podatkov (Tomlinson, 2003, str. 107).

2.2.1 Prostorski zapisi podatkov

Prostorski podatki so informacije, povezane s specifično lokacijo, na primer populacija nekega mesta ali oseba, ki živi na določenem naslovu. V veliko primerih je zahteven del sestavljanja podatkov za GIS povezovanje informacij z lokacijo; gre za proces, znan kot geokodiranje. V posebnem naboru podatkov mora obstajati element, ki označi lokacijo. To je lahko poštna številka ali naslov ulice. Ta element znotraj podatkov identificira lokacijo, poznano kot geokoda (Steinberg & Steinberg, 2006, str. 21).

Obsežno razumevanje narave geografskih informacij je ključnega pomena v procesu zbiranja podatkov za uspešnost GIS v celoti. Nabori podatkov so razdeljeni na ekonomske podatke in tiste, ki so vezani na okolje.

Ekonomski podatki

Ekonomski podatki so lažje dostopni, saj jih ponavadi lahko dobimo pri državnih institucijah in so rezultat anketiranja. Te podatke uporabljajo tudi podjetja, ki združijo informacije iz drugih naborov podatkov, da dobijo podatke za marketinške potrebe. Ta sposobnost prepoznavanja določenih trgov, povezanih z geografskimi podatki, je poznana kot geodemografija in je eno najhitreje rastočih področij v geografskih informacijskih sistemih.

Okoljski podatki

Zbiranje in analiza informacij o okolju je ena gonilnih sil za razvoj GIS. Okoljski podatkovni seti so ponavadi veliki in zahtevajo veliko znanja.

Viri teh podatkov so:

- Obstoječe topografske karte
- Tematski zemljevidi, specifične geološke karte
- Podatki, zbrani prek satelitskih opazovanj

Okoljski podatki zelo pogosto vsebujejo meje med na primer tipi vegetacij, ki so lahko zabrisane in niso definirane z eno samo linijo (Steinberg & Steinberg, 2006, str. 127).

Ko prvič dobimo prostorske podatke, so ponavadi shranjeni v formatu, določenem za enega vodilnih programskih paketov GIS. Večina GIS-paketov je sposobna prebrati in prenesti veliko znanih formatov, čeprav je v nekaterih primerih treba pretvoriti podatke s posebnimi programi. Kot je prikazano v Tabeli 3 (na strani 22), je v uporabi veliko različnih programskih paketov, podatkovnih formatov in pripon datotek.

Tabela 3: Najbolj pogosto uporabljeni podatkovni formati za GIS

Ime	*Programski paket*	*Pripona datoteke*
Shapefile	ArcView, ArcGIS	*.shp, *.shx, *.dbf
Export File	ArcInfo, ArcGIS	*.e00, *.e01, in druge
MapInfo File	Map Info	*.mif, *.mid
Drawing Exchange File	AutoCAD	*.dxf, *.dwg
SDTS Files (standard za prostorske podatke)	National Institute of Standards and Technology	*.sdts, *.ddf
TIGER/Line Files	U.S. Census Bureau	trg#####.zip

Vir: GIS: Geographic Information Systems: Investigation Space and Place, Steinberg & Steinberg, 2006, str. 127

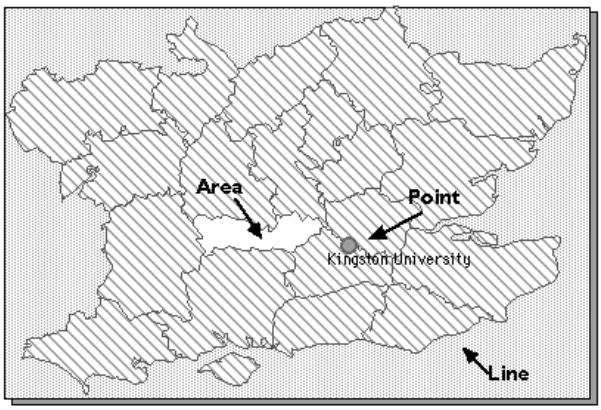
2.2.2 Podatkovni modeli geografskih informacijskih sistemov

Resničnost je preveč kompleksna tudi za najboljše GIS-aplikacije, zato je za predstavitev dejanskega stanja v prostorski bazi podatkov treba te podatke narediti lažje razumljive. Ta poenostavitev je podatkovni model. V podatkovnem modelu je dejansko stanje poenostavljeno v tri prostorske elemente, ki so uporabljeni za predstavitev realnosti.

Kot kaže Slika 12, so ti trije elementi (Zagavec, 2008):

- Točkovni elementi (angl. *point*)
- Črta oziroma linijski elementi (meje, prelomi) (angl. *line*)
- Površina oziroma ploskovni elementi (angl. *area*)

Slika 12: Prostorski elementi



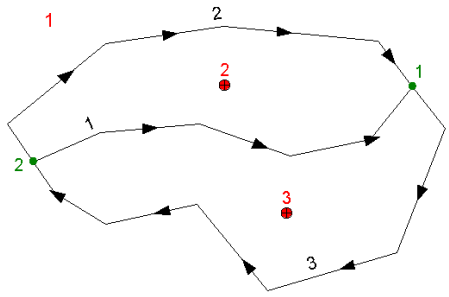
Vir: R. Tomlinson, Thinking about GIS, 2003, str. 36

V GIS sta v uporabi dva različna podatkovna modela: vektorski in rastrski.

Vektorski podatkovni model

Uporaba vektorskega načina vnosa in obdelave podatkov med GIS prevladuje. V matematiki je vektor posplošeno definiran kot usmerjena kvantiteta, določena z izhodiščem, velikostjo in smerjo. V GIS in računalniški grafiki vektor večinoma ni tako strogo definiran. V GIS vektor enostavno pomeni linijski segment neke podatkovne strukture, kateremu je dodan spisek opisnih atributov (Zagavec, 2008). Na Sliki 13 je prikazan primer vektorskega podatkovnega modela.

Slika 13: Poenostavljen prikaz vektorskega podatkovnega modela

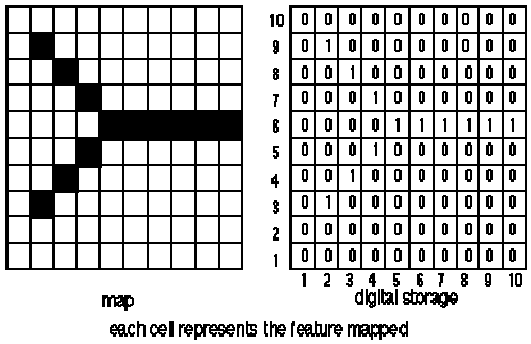


Vir: GIS: Geographic Information Systems: Investigation Space and Place, Steinberg & Steinberg, 2006, str. 67

Rastrski podatkovni model

Rastrski podatkovni model je ena izmed dveh glavnih oblik notranjega podatkovnega sistema, ki se uporablja v GIS. Rastrske podatke si najlažje predstavljamo kot funkcionalno povezano dvodimenzionalno pravokotno mrežo celic ali slikovnih elementov (pikslov), kjer je vsaki celici prirejena ena številka. Ta številka dejansko pomeni vrednost nekega parametra, položaj te številke v mreži pa geografski položaj njene vrednosti relativno glede na ostale vrednosti v mreži. Velikost celice določa pravokotno območje v naravi, za katerega velja vrednost celice. Celica v satelitski rastrski sliki ima lahko, recimo, velikost 10 metrov, kar pomeni, da predstavlja področje v naravni velikosti 10 x 10 metrov. Bistveno za rastrske podatke je, da jih lahko zelo učinkovito in nazorno prikažemo grafično. Na Sliki 14 (na strani 24) je prikazan poenostavljen primer rastrskega podatkovnega modela.

Slika 14: Poenostavljen prikaz rastrskega podatkovnega modela



Vir: GIS: Geographic Information Systems: Investigation Space and Place, Steinberg & Steinberg, 2006, str. 68

Med prednostmi rastrske metode je njena zmožnost sprejemanja podatkov neposredno iz sistema daljinskega zaznavanja ter prikaz prehodnih podatkov. Rastrski sistem se lahko relativno temeljito shranjuje, kar postavlja stvarne meje pokritemu področju ali ločljivosti ali obojemu (Steinberg & Steinberg, 2006, str. 23).

2.3 IZHODNI PODATKI

Končna faza geografskih informacijskih sistemov je output, ki za večino pomeni zemljevid. Zemljevid, narejen v GIS, je privlačen in zelo nazorno pokaže končne rezultate. Kot je znano, slika pove tisoč besed. Mnogo GIS-aplikacij nam ne prikaže rezultata, ki ga pričakujemo. Barve na zemljevidu so ponavadi izbrane naključno, zato so si med seboj preveč podobne ali nesmiselne. Na primer vijolična vodna površina in modro kopno.

Treba je paziti torej na barve in simbole (Steinberg & Steinberg, 2006, str. 118):

- Prva pogosta napaka je izbira preveč podobnih barv. S tem dobimo zemljevid, iz katerega ne vidimo rezultatov.
- Druga pogosta napaka je to, da na enem zemljevidu predstavimo preveč informacij.

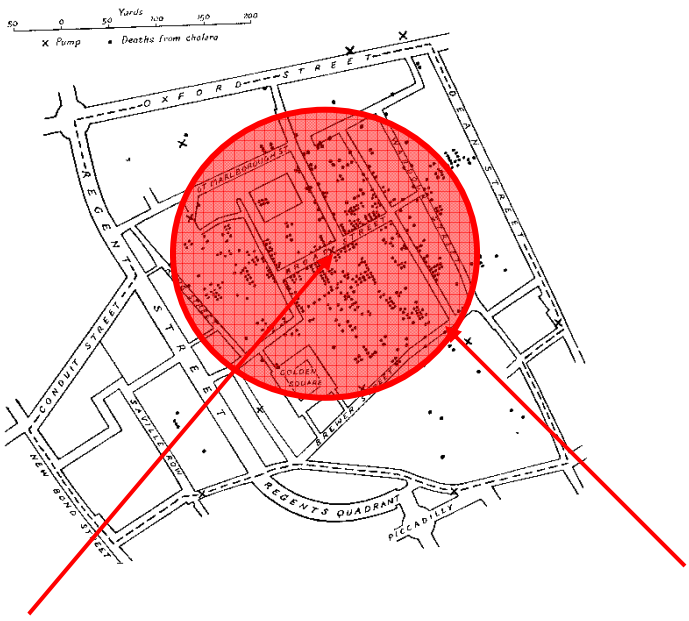
3 PODATKOVNO RUDARJENJE V GEOGRAFSKIH INFORMACIJSKIH SISTEMIH

Geografsko podatkovno rudarjenje je proces odkrivanja zanimivih in nepoznanih vzorcev v geografskih bazah podatkov. Veliko število prostorskih baz podatkov je vzrok za naraščajoče zanimanje za podatkovno rudarjenje in iskanje uporabnih geografskih vzorcev in povezav (Shekhar & Vatsavi, 2003, str. 520).

3.1 UPORABA PODATKOVNEGA RUDARJENJA V GEOGRAFSKIH INFORMACIJSKIH SISTEMIH

Geografsko podatkovno rudarjenje vsebuje kombinacijo dveh ključnih programskih orodij: geografske informacijske sisteme (GIS) in sisteme za podatkovno rudarjenje. GIS in podatkovno rudarjenje sta tehnologiji, ki se lahko združita za pridobivanje močnih, ponavadi trženjskih informacij iz množice podatkov (Duke, 2001). Klasičen primer uporabe podatkovnega rudarjenja v geografskih informacijskih sistemih je prikazan na Sliki 15, ki prikazuje odkritje okuženega zajetja vode pri smrtih zaradi kolere. Pike na zemljevidu prikazujejo lokacije umrlih zaradi kolere, križci pa zajetja vode. Iz slike je razvidno, da je skupina umrlih locirana v ulicah okrog nekega zajetja vode, in tako so ugotovili, katero zajetje vode je bilo okuženo.

Slika 15: Klasičen primer geografskega podatkovnega rudarjenja: smrti zaradi kolere v Londonu leta 1854



Skupina umrlih zaradi kolere

Okuženo zajetje pitne vode

Vir: S. Shekhar & R. R. Vatsavi, Mining Geospatial Data, 2003, str. 521

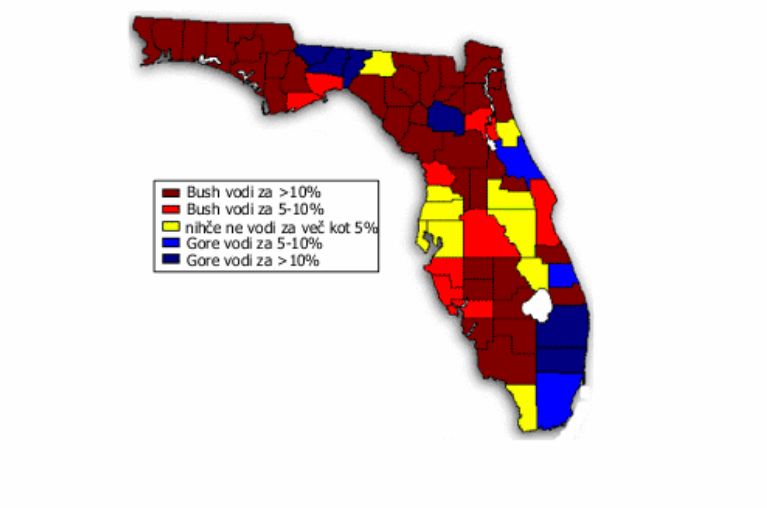
Kot že omenjeno, so geografski elementi ponavadi v obliki točk (na primer določen kupec), črt (na primer ceste) ali površin (na primer občine). Če so podatki pripeti k določenem zemljevidu, elementi lahko predstavljajo zelo detajlne informacije. GIS nam omogoča, da povežemo tabele glede na neposredno bližino tako, kot da bi povezali ključe relacijske baze podatkov. S tem lahko odgovorimo na vprašanja, kot so:

- Kakšen je povprečen prihodek kupca v radiu enega kilometra od določene lokacije?
- Kje živijo naši najboljši kupci?

- Katere konkurenčne prednosti imajo naši kupci v določeni občini?

GIS lahko neposredno odgovori na takšna in podobna vprašanja. Demografska analiza podatkov je pogosto uporabljena za analizo trga za potencialno lokacijo, širše trženjske regije ali individualne kupce. Na žalost te detajlne informacije, kljub njihovi kvaliteti, pogosto niso ključne pri končnih odločitvah. Na Sliki 16 je prikazan primer geografskega podatkovnega rudarjenja pri rezultatih predvolilnih raziskav leta 2000 na Floridi. GIS je uporabljen v tem primeru za prikaz, kateri kandidat vodi na določenem območju.

Slika 16: Primer geografskega podatkovnega rudarjenja: rezultati predvolilnih raziskav na Floridi



Vir: P. Duke, Geospatial Data Mining for Market Intelligence, 2001

Naslednji logični korak za nadaljevanje GIS-analize je vključiti možnost analiziranja in združitve velikega števila spremenljivk v eno napoved ali oceno. To je prednost usmerjenega podatkovnega rudarjenja in razlog za analizo podatkov s kombinacijo GIS in podatkovnega rudarjenja.

GIS lahko združuje zgodovinske podatke o kupcu ali prodaji skupaj z demografskimi, poslovnimi in marketinškimi podatki. Te baze podatkov so idealne za pripravo napovednih modelov za nove lokacije, potencialne kupce in podobne probleme (Duke, 2001).

3.2 PRISTOPI K PODATKOVNEMU RUDARJENJU V GEOGRAFSKIH INFORMACIJSKIH SISTEMIH

Glavni namen komercialnih orodij za podatkovno rudarjenje je analiziranje in identifikacija potrošniških vzorcev, podatkov, zbranih prek internetnih strani, in podobnih podatkov. Pridobivanje zanimivih vzorcev in geografskih podatkov je bolj zahtevno kot

iskanje vzorcev v tradicionalnih numeričnih podatkih, zaradi kompleksnosti tipov prostorskih podatkov, povezav in algoritmov (Shekhar & Vatsavi, 2003, str. 520).

Pridobivanje znanja iz prostorskih podatkov torej zahteva posebne pristope. Obstajata dve vrsti pristopov k geografskem podatkovnem rudarjenju:

- Uporabiti nove prostorsko usmerjene algoritme, ki so narejeni posebej za preiskovanje prostorskih podatkov.
- Posebej najti prostorske povezave in značilnosti v fazi obdelave podatkov in nato uporabiti klasične tehnike podatkovnega rudarjenja.

Primeri prostorsko orientiranih algoritmov so prostorska regresija, prostorska asociativna pravila, prostorsko razvrščanje v skupine in prostorsko usmerjeno ugotavljanje odstopajočih vrednosti.

Primeri prostorskih značilnosti in klasičnih tehnik podatkovnega rudarjenja vsebujejo uporabo algoritmov podatkovnega rudarjenja na prostorskem modelu. Na primer opis prostorskih odvisnosti prek pravil in apliciranja teh pravil na drevo odločanja in metode SOM na prostorski model (Demšar, 2006, str. 23).

3.3 PRIMER UPORABE PODATKOVNEGA RUDARJENJA V GEOGRAFSKIH INFORMACIJSKIH SISTEMIH

V primeru sem preizkušala uporabo podatkovnega rudarjenja v GIS. Poudarek je bil na neusmerjenem podatkovnem rudarjenju, zanimalo me je predvsem področje razvrščanja v skupine in uporaba metode SOM (angl. *self-organizing maps*). Primer je izmišljen, vendar sem uporabila realne podatke o gospodinjstvih v Lizboni.

3.3.1 Opredelitev problema

V tem primeru je bil namen geodemografsko razdeliti mesto Lizbona. Želela sem pridobiti nekaj znanja o geodemografski tipologiji Lizbone, predvsem pa sem hotela preizkusiti orodja in spoznati metode razvrščanja v skupine v praksi. Najprej sem morala pregledati in bolje spoznati podatke in si postaviti nekaj vprašanj, na katera sem s podatkovnim rudarjenjem poskušala odgovoriti. Eno izmed možnih vprašanj bi bilo, ali obstaja vzorec med porazdelitvijo starejših prebivalcev v Lizboni. Predvidevala sem, da bom našla vzorec, da nepremičnine največkrat najamejo brezposelni in mladi, ki iščejo prvo zaposlitev. Zanimal me je tudi vzorec med starostjo prebivalstva in spolom. Prišla sem do zanimivih ugotovitev.

3.3.2 Podatki in programska orodja

Za dosego rezultatov sem potrebovala dva tipa podatkov (geografske in socio-ekonomske) in tri različne programe (MS Excel, ArcView in SOM_PAK).

Geografske podatke je vsebovala datoteka Shape (angl. *Shapefile*) mesta Lizbone. Datoteka vsebuje kode območij (četrti) v Lizboni. Shapefile je format, ki se uporablja v programu ArcView. Je standard za izmenjavo prostorskih podatkov med uporabniki geografskih informacijskih sistemov. Sestoji iz treh datotek: iz glavne datoteke (.shp), v kateri je shranjena geometrija objektov; indeksne datoteke (.shx), kjer so shranjeni indeksi za hitrejše poizvedbe po prostorskih podatkih, in datoteke z atributi (.dbf), kjer se shranjujejo atributi objektov.

Kot **ekonomske podatke** sem imela na voljo podatke o gospodinjstvih v Lizboni, ki so bili zbrani z anketiranjem. Takšni podatki so sicer zelo uporabni, saj lahko pričakujemo, da bomo našli določene vzorce med prebivalstvom, ampak morda nekoliko preveč kompleksni, saj zlahka spregledamo pomembne informacije. Podatki, ki sem jih uporabila, so bili v tabelarni obliki. Stolpci predstavljajo določene spremenljivke (starost, zaposlitev in drugo), vrstice pa določene enote (četrti) v Lizboni. Podatki vsebujejo 882 zapisov (vrstic) in ti se ujemajo s številom enot. Vsaka enota vsebuje informacije o 600 ljudeh in 250 gospodinjstvih. Uporabila sem povprečne vrednosti za vsako spremenljivko v določeni enoti. Podatki so bili normalizirani že pred uporabo.

Glede na vprašanja, na katera sem želela odgovoriti, sem se odločila, da bom zaradi lažje analize izbrala le 13 spremenljivk (v %), ki so:

- Najete nepremičnine
- Lastne nepremičnine
- Moški, stari med 14 in 19 let
- Ženske, stare med 14 in 19 let
- Moški, stari med 20 in 24 let
- Ženske, stare med 20 in 24 let
- Moški, stari med 25 in 65 let
- Ženske, stare med 25 in 65 let
- Moški, stari nad 65 let
- Ženske, stare nad 65 let
- Iskalci prve zaposlitve
- Nezaposleni
- Zaposleni

Ti podatki so ključni za razvoj geodemografske tipologije, vsak zapis v socio-ekonomskih podatkih se ujema z zapisom geografskih podatkov v datoteki Shape (kode četrti). Vsak zapis v datoteki Shape lahko združimo z določeno spremenljivko v tej tekstovni datoteki.

Ti podatki so pripravljeni za uporabo metode SOM in s tem na procesiranje s programom SOM_PAK, ki lahko odpre le tekstovno datoteko. Pri teh podatkih sem opazila dve stvari: da spremenljivke nimajo imen in da imam 40 stolpcev, obstaja pa le 39 spremenljivk. Ugotovila sem, da zadnji stolpec oziroma spremenljivka predstavlja kodo območja, ki identificira vsak zapis in pomeni ključ oziroma nekakšen skupni imenovalec in jo lahko združim z datoteko Shape, da bom lahko dobila kot rezultat zemljevid Lizbone.

Uporabila sem tri različne **programe**. Za analizo in »predprocesiranje« podatkov sem uporabila Microsoft Excel, saj je pregleden in enostaven za uporabo. Za geografsko vizualizacijo rezultatov sem uporabila ArcView. ArcView je programsko orodje GIS in nam omogoča vizualizacijo, preučevanje in analize prostorskih podatkov. Program SOM_PAK sem uporabila kot podporo metodi SOM za prikaz razvrščanja v skupine. Za uporabo programa SOM_PAK je potrebna datoteka .bat (razlaga v prilogi 1). Ker je metoda SOM nevronska mreža, je program SOM_PAK razdeljen na več delov: učenje algoritma, merjenje napak in vizualizacijo. Parametre vrednosti, ki nam vrnejo različne rezultate v razvrščanju v skupine, lahko spremenimo. Program se zažene in kot rezultat dobimo U-matriko. [2]

3.3.3 Postopek

1. Kot sem že omenila, sem najprej **pregledala podatke** in si postavila vprašanja, na katera sem želela odgovoriti:

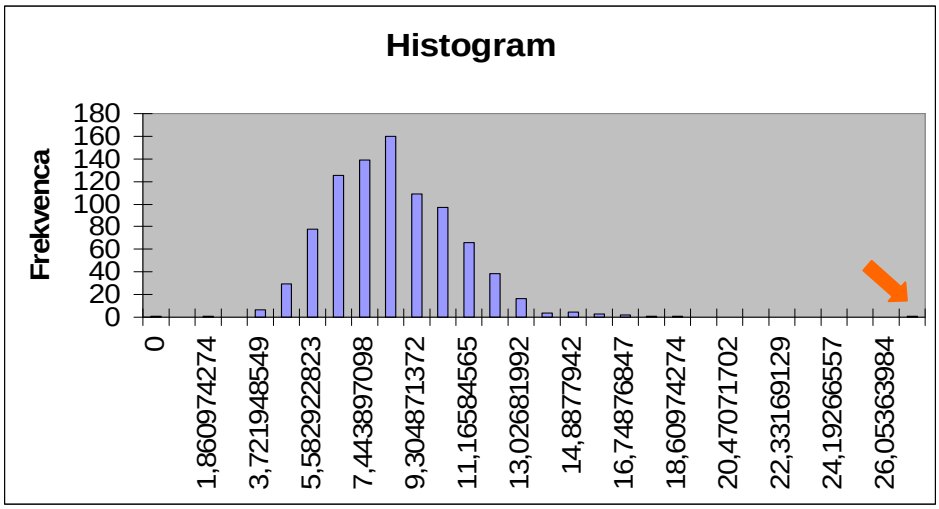
- Na katerih območjih živi starejše prebivalstvo?
- Kakšne značilnosti ima prebivalstvo, ki živi na elitnih lokacijah?
- Ali obstaja vzorec med nezaposlenimi in najetimi nepremičninami?

2. Sledila je **priprava podatkov**: analiziranje podatkov z namenom zaslediti odstopajoče vrednosti. Če so podatki čisti, se ta korak lahko izpusti, vendar so v realnosti podatki zmeraj nečisti, zato navadno prav ta korak vzame največ časa. Odstopajoče vrednosti lahko odkrijemo tudi v U-matriki, zato je možno to fazo tudi preskočiti in najprej analizirati rezultate metode SOM ter se pozneje vrniti nazaj. V tem primeru sem se odločila, da bom najprej uredila podatke in nato nadaljevala s postopkom.

[2] več o U-matriki je opisano v poglavju 1.3.3

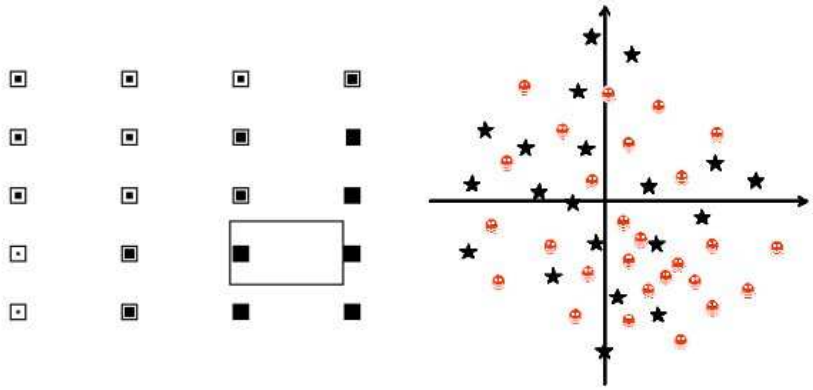
Predvidevala sem, da geografskih podatkov ni treba urediti, spreminjanje teh podatkov brez določenega geografskega znanja pa tudi ne bi bilo možno. Zato sem pregledala le socioekonomske podatke (datoteka .txt). Zaradi večje preglednosti sem podatke odprla v Excelu. Najprej sem izbrisala dva zapisa (vrstici), saj nista vsebovala numeričnih vrednosti in sta bila s tega vidika nekoristna. Ker drugih nepravilnosti v podatkih ni bilo mogoče opaziti, sem naredila histograme za vsako spremenljivko posebej. Približno polovica spremenljivk je vsebovala odstopajoče vrednosti. Nekatere odstopajoče vrednosti sem izločila, saj metoda SOM ni občutljiva na manjkajoče vrednosti. Druge, po mojem mnenju bolj pomembne odstopajoče vrednosti, sem nadomestila s povprečno vrednostjo spremenljivke. Na spodnjem grafu je primer odstopajoče vrednosti za spremenljivko moški 20–24 let. Kot odstopajočo vrednost lahko vidimo vrednost 26 odstotkov, kar bi pomenilo, da na določenem območju živi večina oziroma 26 odstotkov moških, starih od 20 do 24 let. Sicer je možno, da gre za pravilno vrednost, saj je lahko na tem območju študentsko naselje.

Slika 17: Histogram za spremenljivko moški 20-24 let



3. Naslednji korak je **uporaba metod,** in sicer kot je bilo že prej omenjeno, metode SOM. Metoda SOM oziroma samoorganizirajoče mreže (angl. *self-organizing maps*) je vizualizacijsko in analitično orodje za večdimenzionalne podatke. Je nevronska mreža, ki oponaša delovanje človeških možganov. V tem primeru je nevronska mreža uporabljena za razvrščanje v skupine. Prek treniranja mreže in učenja se mreža specializira za identifikacijo določenih vzorcev, ki jih želimo najti. Cilj metode SOM je dobiti določeno število nevronov, ki so organizirani glede na vzorce za klasifikacijo (vhodni vzorci oziroma angl. *input patterns*). Zelo pomembna značilnost SOM je razvoj urejene mreže, v kateri imajo bližnji nevroni podobne značilnosti. Je enoslojna nevronska mreža, kjer so nevroni postavljeni v večdimenzionalnem prostoru. V tem prostoru lahko definiramo izhodni prostor (angl. *output space*). Slika 18 na strani 31 nam kaže osnovno SOM-arhitekturo (Bação in Lobo, 2006b, str. 5).

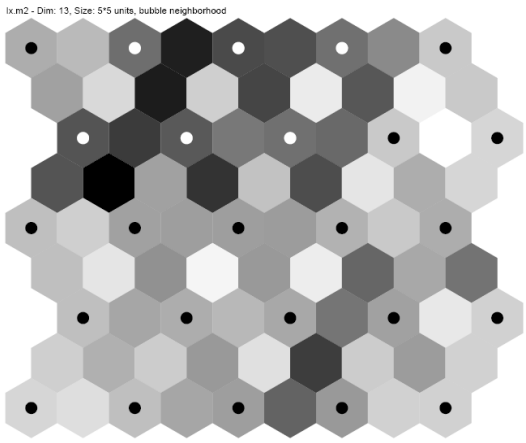
Slika 18: Osnovna SOM-arhitektura, na levi strani izhodni prostor (output space), na desni strani primer vhodnega prostora



Vir: F. Bação & V. Lobo, The self-organizing map (SOM), 2006, str. 6

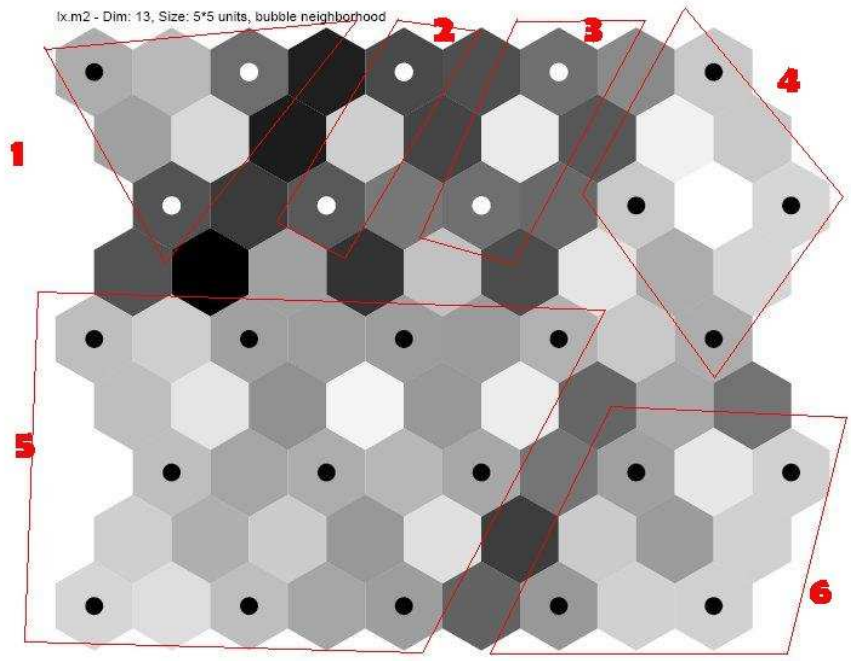
Izhodni podatki metode SOM so v obliki U-matrike. Prvotni cilj je bil dobiti mrežo 5*5, tako, da bi iz vseh 882 zapisov o enotah (četrtih) v Lizboni dobila 25 skupin. Velikost mreže se ponavadi spreminja glede na želeno natančnost pri končnem rezultatu. V U-matriki bele oziroma črne pike predstavljajo skupine. Med belimi in črnimi pikami ni razlike, ampak so različne le zaradi večje preglednosti. Polja brez pik predstavljajo razdalje med sosednjimi nevroni. Če so nevroni daleč narazen, lahko vidimo to kot meje med skupinami. V U-matriki so razdalje med nevroni prikazane kot sivine. Temnejša polja pomenijo večje razlike, svetlejša polja pa skupne značilnosti. Torej če so razlike med sosednjimi nevroni majhne, potem ti nevroni predstavljajo skupino s podobnimi značilnostmi. S preizkušanjem različnih vrednosti parametrov v .bat datoteki in primerjanjem rezultatov sem kot rezultat metode SOM dobila jasno sliko U-matrike, ki je prikazana na Sliki 19. Če na tej stopnji opazimo, da podatki vsebujejo vrednosti, ki se razlikujejo od preostalih, se moramo vrniti na začetek in urediti podatke.

Slika 19: Rezultat metode SOM: U-matrika



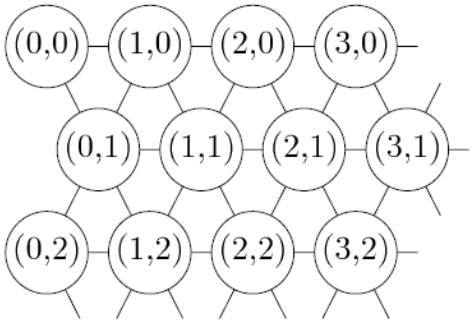
Ker pa je 25 skupin preveč za analizo glede na število spremenljivk in ker ima prebivalstvo v določenih četrtih podobne lastnosti, sem število skupin zmanjšala s 25 na 6 (kot je prikazano na Sliki 20). Nevroni, ki so povezani s polji svetle barve, morajo biti povezani v isti skupini (angl. *cluster*). V literaturi se omenjajo primerjave s svetlimi polji in dolinami ter temnejšimi polji in gorami.

Slika 20: U-matrika z zmanjšanim številom skupin



Da ugotovimo, za kateri nevron (skupino) gre, pa moramo upoštevati številčenje v programu SOM_PAK (poleg U-matrike program generira tudi datoteko s temi podatki), ki je določeno tako, kot kaže Slika 21:

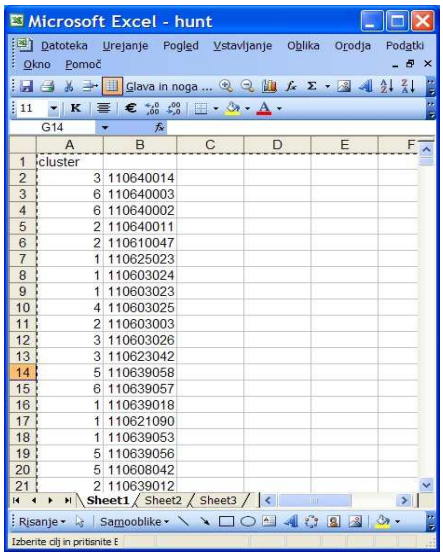
Slika 21: Številčenje skupin v U-matriki



Vir: K. Kaski & T. Kohonen, Exploratory data analysis by the self-organizing map, 1995, str. 501

Da bi določili skupine v podatkih in ne le na sliki, moramo združiti socioekonomske podatke o gospodinjstvih in datoteko, ki vsebuje podatke o številčenju. Ta del je zelo enostaven. Odpremo obe datoteki v Excelu in jih brez sortiranja ali spreminjanja skopiramo skupaj. Da bi definirali skupine (kot je označeno na Sliki 21), moramo najprej združiti prva dva stolpca v datoteki, ki vsebujeta številčne oznake skupin. Na primer tako bo vsak zapis, ki ima 1 in 4 v prvih dveh stolpcih, klasificiran kot skupina 14. Če se odločimo, da sta na primer skupini 01 in 11 podobni, jih združimo v isto skupino. Nato podatke shranimo kot .dbf datoteko (datoteka z atributi), da bodo te združene skupine lahko zaznane v programu ArcView. Ker nas nato zanimajo le številke skupin in kode četrti, lahko ostale stolpce izbrišemo. Nato dobimo tabelo z dvema stolpcema, kot prikazuje Slika 22.

Slika 22: Skupine in kode četrti



	A	B
1	cluster	
2	3	110640014
3	6	110640003
4	6	110640002
5	2	110640011
6	2	110610047
7	1	110625023
8	1	110603024
9	1	110603023
10	4	110603025
11	2	110603003
12	3	110603026
13	3	110623042
14	5	110639058
15	6	110639057
16	1	110639018
17	1	110621090
18	1	110639053
19	5	110639056
20	5	110608042
21	2	110639012

4. Sledi **merjenje rezultatov**. Da bi ugotovili, kako se razlikujejo rezultati, nam SOM_PAK generira tudi datoteko s povprečnimi vrednostmi vsake skupine in vsake spremenljivke. Če pregledamo te podatke, lažje razumemo določene lastnosti vsakega območja v mestu. V Tabeli 4 so povprečni rezultati za vsako od šestih skupin. Iz teh vrednosti sem dobila graf, ki je na Sliki 23 (na strani 35), ki nam kaže vsako skupino glede na spremenljivke.

Tabela 4: Povprečne vrednosti izbranih spremenljivk za določene skupine

Skupina/spremenljivka (%)	Najeta nepremičnina	Lastna nepremičnina
1.	15,96	79,57
2.	37,91	57,39
3.	53,16	42,20
4.	56,59	38,85
5.	57,13	38,04
6.	89,49	8,39

Skupina/spremenljivka (%)	m 14–19	ž 14–19	m 20–24	ž 20–24
1.	6,25	5,18	16,42	79,11
2.	5,15	4,13	41,77	53,24
3.	5,20	3,95	53,83	41,70
4.	5,88	4,48	56,31	39,13
5.	6,36	5,32	58,21	37,00
6.	8,97	7,68	90,32	7,63

Skupina/spremenljivka (%)	m 25 - 64	ž 25 - 64	m 65+	ž 65+
1.	6,12	5,03	17,92	77,65
2.	5,12	4,12	44,55	50,35
3.	5,19	3,79	53,38	42,02
4.	5,97	4,60	55,76	39,64
5.	6,33	5,26	59,46	35,84
6.	9,21	8,01	91,35	6,67

Skupina/spremenljivka (%)	Prva zaposlitev	Nezaposleni	Zaposleni
1.	19,87	75,63	6,02
2.	46,18	49,09	5,22
3.	52,88	42,22	5,13
4.	55,17	40,19	5,95
5.	60,74	34,57	6,36
6.	92,64	5,78	9,96

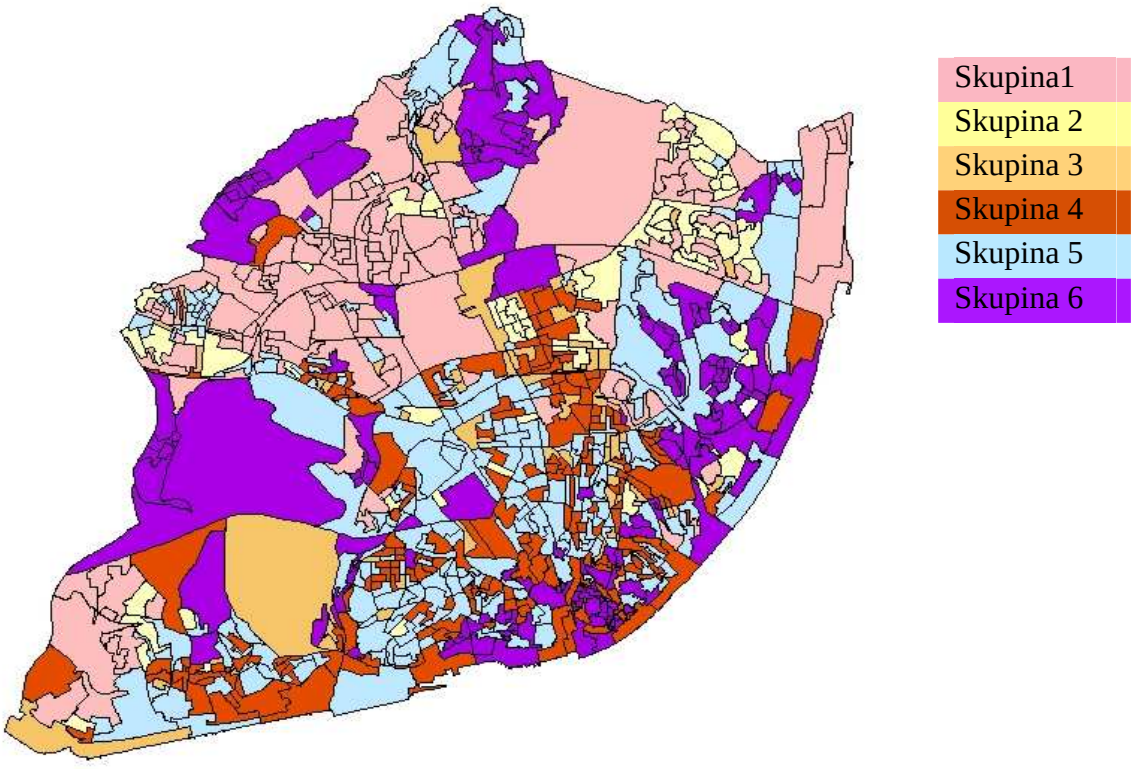
Slika 23: Povprečne vrednost spremenljivk za vsako skupino

Značilnosti posameznih skupin so torej razvidne iz tabele in grafa:

- Skupina 1: najvišji odstotek zaposlenih srednjih let
- Skupina 2: starejše prebivalstvo, zaposleni ter lastniška stanovanja
- Skupina 3: prevladuje starejše prebivalstvo z lastniškimi stanovanji, mladih je zelo malo
- Skupina 4: starejše prebivalstvo, visok odstotek najetih stanovanj
- Skupina 5: vse spremenljivke imajo povprečne vrednosti, zato nam ta skupina ne daje nobenih posebnih informacij
- Skupina 6: večinoma najeta stanovanja, največji odstotek tistih, ki iščejo prvo zaposlitev ter nezaposleni

5. **Predstavitev rezultatov**. Da bi dobili zemljevid Lizbone (kot nam kaže Slika 24) in s tem združene podatke o gospodinjstvih in geografske podatke, moramo združiti datoteko z atributi (datoteka s končnico .dbf) in geografske podatke (Shapefile datoteka s končnico .shp). Tako združimo dve plasti podatkov v eno. Program ArcView je zelo enostaven in nam omogoča, da to naredimo z enim klikom. Lahko izberemo tudi različne barve, tako da so rezultati bolje vidni.

Slika 24: Prikaz rezultatov – zemljevid Lizbone, razdeljen na 6 skupin



3.3.4 Ugotovitve

Ob pregledu zemljevidov posameznih skupin sem dobila pričakovane rezultate.[3] Kot sem predvidevala, sem našla vzorec med najetimi nepremičninami in brezposelnimi. Prav tako nam rezultati pokažejo, da mladi, ki iščejo prvo zaposlitev, živijo v najetih stanovanjih. Kot pričakovano, so zaposleni ponavadi lastniki nepremičnin.

Tiste skupine z večjim odstotkom zaposlenih živijo bolj na obrobju mesta. Na elitnih lokacijah (Belem, Campolide) živi prebivalstvo, ki spada v prvo skupino, in starejše prebivalstvo z lastniškimi stanovanji. V strogem centru živi revnejše starejše prebivalstvo in najemniki.

SKLEP

V diplomski nalogi sta predstavljeni in na primeru preizkušeni dve tehnologiji, obe pomembni vsaka zase, združeni pa imata bistveno vlogo v poslovnem svetu. Podatkovno rudarjenje postaja vedno pomembnejši del poslovnega odločanja. Geografski informacijski sistemi zagotavljajo obvladovanje prostorskih geografskih podatkov in možnost pridobitve pomembnih informacij, ki jih drugače ne moremo pridobiti. Združitev informacij, pridobljenih s podatkovnim rudarjenjem, in informacij, pridobljenih z geografskimi informacijskimi sistemi, nam pomaga k uspešnim poslovnim odločitvam.

V primeru sem ugotovila, da so pri podatkovnem rudarjenju v geografskih informacijskih sistemih zelo pomembne vizualizacijske tehnike. Strukture podatkov ne bi bilo možno opisati brez ustrezne integracije vizualizacije in podatkovnega rudarjenja. Ugotovila sem tudi, da so tehnike podatkovnega rudarjenja, predvsem razvrščanje v skupine z metodo SOM, uporabne na več področjih. Metodo lahko uporabimo za razvrščanje v skupine pri socioekonomskih podatkih in te podatke nato združimo z geografskimi in rezultate predstavimo na zemljevidu. Vzorci, ki sem jih dobila pri podatkovnem rudarjenju, bi bili uporabni pri trženjskih raziskavah.

Pri primeru sem imela največ dela s pripravo podatkov. Če podatki niso čisti, lahko dobimo napačne ali neuporabne rezultate. Veliko časa sem namenila tudi za spoznavanje tehnik, metod in uporabljenih orodij. Metoda SOM je sicer zapleteno orodje, vendar z uporabo programov, ki olajšajo delo, je identificirati povezave in vzorce v rezultatih zelo preprosto.

[3] zemljevidi posameznih skupin so prikazani v Prilogi 2.

Ne le v trženju, ampak tudi na drugih področjih lahko z geografskim podatkovnim rudarjenjem napovemo izid ali poiščemo vzorce, ki nam bodo prinesli določeno znanje. Področja, na katerih bi še bilo geografsko podatkovno rudarjenje uporabno, so prometne konice, migracijski vzorci, vremenske napovedi, seljenje živali, širjenje bolezni in okoljske spremembe ter podatki s prostorsko komponento, ki jih je težko ali nemogoče analizirati s tradicionalnimi metodami podatkovnega rudarjenja.

V več pogledih sta znanost in tehnika naprednejša od podjetij. Veliko strokovnjakov v podjetjih že uporablja pristope, podobne omenjenim. Trenutni procesi so na žalost zelo nepovezani, pogosto zahtevajo ročno prenašanje podatkov med različnimi programi. To omejuje uporabo tehnologij tistim, ki imajo znanje in izkušnje z upravljanjem baz podatkov, geografskimi informacijskimi sistemi, s podatkovno analizo, napovedovanjem, programiranjem in, najpomembnejše, določenimi poslovnimi aplikacijami. Večja povezanost podatkov in tehnologij bi procese poenostavila in zagotovila ustrezne informacije. To bi omogočilo podjetjem osredotočenje na poslovne priložnosti.

LITERATURA IN VIRI

1. Arshan, H. (2008). *Tools for Decision Analysis*. Najdeno 10. julija 2008 na spletnem naslovu: home.ubalt.edu/ntsbarsh/opre640a/RiskTree.gif.
2. Bação, F. (2008). *Predmet Data Mining, prosojnice predavanj*. Lizbona: Universidade Nova de Lisboa.
3. Bação, F . & Lobo, V. (2005). *Fundamentals of Data Preparation and pre-processing.* Najdeno 4. novembra 2007 na spletnem naslovu http://www.isegi.unl.pt/ensi no/ docentes/fbacao/data_mining_erasmus.htm
4. Bação, F. & Lobo, V. (2006a). *Introduction to Unsupervised Classification.* Najdeno 4. novembra 2007 na spletnem naslovu http://www.isegi.unl.pt/ensino/docen tes/fbacao/data_mining_erasmus.htm.
5. Bação, F. & Lobo, V. (2006b). *The Self-Organizing Map (SOM).* Najdeno 4. novembra 2007 na spletnem naslovu http://www.isegi.unl.pt/ ensino/docentes/fbacao/da ta_mining_erasmus.htm.
6. Batagelj, T. (2002). *Nevronske mreže.* Najdeno 3. junija 2008 na spletnem naslovu: www.drustvo-informatika.si/dogodki/dsi2002/prispeliReferati/batagelj.doc.
7. Berry, M. & Linoff, G. (1998). *Mastering Data Mining: the Art and Science of Customer Relationship Management*. Indianapolis: Wiley.
8. Berry, M. & Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. (2st ed.) Indianapolis: Wiley.
9. Demšar, U. (2006). *Data Mining of Geospatial Data: Combining Visual and Automatic Method.* Stockholm: KTH, Kungliga Tehniska högskolan.
10. Duke, P. (2001). Geospatial Data Mining for Market Intelligence. *TDAN*. Najdeno 22. februarja 2008 na spletnem naslovu http://www.tdan.com/view-articles/4921.
11. *Geografski informacijski sistemi – Rešitve za terenski zajem podatkov.* Najdeno 14. aprila 2008 na spletnem naslovu http://www.geoservis.si/uporabno/gis/gis.htm.
12. *Google Earth.* Najdeno 15. januarja na spletnem naslovu: 2008 na http://earth.google.com.
13. Grinstein, G. & Ward, M. (2002). Introduction to Data Visualization. V U., Fayyad, G., Grinstein,. & A., Wierse, *Information Visualization in Data Mining and Knowledge Discovery* (str. 21–45). San Francisco: Morgan Kaufmann Publishers.
14. Grobelnik, M. (1999). *Data Mining s Clementine*. Najdeno 14. aprila 2008 na spletnem naslovu: www-ai.ijs.si/MarkoGrobelnik/awamida99/clementine.doc.
15. Hoffman, P. & Grinstein, G. (2002). Multidimensional Information Visualizations for Data Mining. V U., Fayyad, G., Grinstein,. & A., Wierse, *Information Visualization in Data ining and Knowledge Discovery* (str. 125–128*).* San Francisco: Morgan Kaufmann Publishers.

16. *I slovar, slovar informatike.* Slovensko društvo Informatika. Najdeno na spletnem naslovu http://www.islovar.org.
17. Jakulin, A. (2008). *Večrazsežno lestvičenje*. Najdeno na spletnem naslovu www.stat.columbia.edu/~jakulin/Predavanja/MDS.ppt.
18. Jaklič, J. (2006). *Predmet Tehnologija za podporo odločanju, prosojnice predavanj*. Ljubljana: Ekonomska fakulteta.
19. Kaski, K. & Kohonen, T. (1995). Exploratory Data Analysis by the Self-Organizing Map. V *Proceedings of the Third International Conference on Neural Networks and Capital Markets* (str. 498–507). London
20. Košmelj, K. & Breskvar, L. (2006). *Metode za razvrščanje enot v skupine; osnove in primer.* Najdeno 4. junija 2008 na spletnem naslovu http://aas.bf.uni-lj.si/september2006/11kosmelj.pdf.
21. Kovačič, A., Jaklič, J., Štemberger, M. & Groznik, A. (2004). *Prenova in informatizacija poslovanj*a. Ljubljana: Ekonomska fakulteta.
22. Leskošek, B. (2006). *Informacijske tehnologije za podporo odločanju: odločitveni in ekspertni sistemi.* Najdeno 15. maja na spletnem naslovu http://www.fsp.uni-lj.si/bojan/.
23. Liu, H. & Motoda, H. (2003). Feature Extraction, Selection and Construction. V Ye, N. (2003). *The handbook of data mining* (str. 409–422*).* Mahwah, New Jersey: Lawrence Erlbam Associates.
24. Mertik, M. & Jevtić, S. (2001). *Posojilo – da ali ne (Uporaba inteligentnih sistemov v bančništvu).* Najdeno 5. maja 2008 na spletnem naslovu http://www.drustvo-informatika.si/dogodki/arhiv/dsi2001/sekcija_i/mertik.doc
25. Rhodes, P. (2005). Discovering new relationships. A brief overview of data mining and knowledge discovery. V U., Fayyad, G., Grinstein,. & A., Wierse, *Information visualization in data mining and knowledge discovery* (str. 277–297). San Francisco: Morgan Kaufmann Publishers.
26. *SAS customer support – [spletna stran podjetja SAS]*. Najdeno 5. julja 2008 na spletnem naslovu: http://support.sas.com/rnd/base/topics/statgraph/proctemplate/a002650843.htm.
27. *Search Software Quality – spletna stran za pomoč pri razvijanju in upravljanju kvalitetne programske opreme.* Najdeno 10. julija 2008 na spletnem naslovu: http://searchsoftwarequality.techtarget.com/sDefinition//0,,sid92_gci330729,00.html.
28. Shekhar, S. & Vatsavi, R. R. (2003). Mining Geospatial Data. V N., Ye. (2003). *The Handbook of Data Mining* (str. 520–547). Mahwah, New Jersey: Lawrence Erlbam Associates.
29. Si, J. & Nelson, B.J. (2003). Artificial Neural Network Models for Data Mining. V N., Ye. (2003). *The handbook of data mining* (str. 41–65). Mahwah, New Jersey: Lawrence Erlbam Associates.

30. Steinberg, S. & Steinberg, S. L. (2006). *GIS: Geographic Information Systems: Investigation Space and Place.* Thousand Oaks: Sage publications.
31. Thearling, K. *An Introduction to Data Mining: Discovering Hidden Value in Your Data Warehouse.* Najdeno 5. maja 2008 na spletnem naslovu http://www.thearling.com/text/dmwhite/dmwhite.htm.
32. Tomlinson, R. (2003). *Thinking about GIS.* Redlans, California: ESRI Press
33. Vidmar, G., Zdešar, A., Herlec, U. & Ferlan, M. *Zajem in vizualizacija geoloških podatkov za GIS.* Najdeno 14. aprila 2008 na spletnem naslovu http://mfc-2.si/pdf/zajem_in_vizualizacija_geoloskih_podatkov_za_GIS.pdf.
34. *Wikipedia – spletna enciklopedija.* Najdeno na spletnem naslovu: http://sl.wikipedia.org/.
35. Zagavec, D. *Osnove geografskih informacijskih sistemov.* Najdeno 6. maja 2008 na spletnem naslovu http://users.volja.net/damijanz/GIS/GIS_podat.gif.
36. Zidar, M. (2005). *Podatkovno rudarjenje in KXEN analitično orodje.* Ljubljana: Ekonomska fakulteta.

PRILOGE

Priloga 1: RAZLAGA DATOTEKE .BAT

```
randinit -xdim 5 -ydim 5 -din tutorial1.txt -cout lx.m0 -
topol hexa -neigh bubble
vsom -din tutorial1.txt -cin lx.m0 -cout lx.m1 -rlen 30000 -
alpha 0.8 -radius 5
vsom -din tutorial1.txt -cin lx.m1 -cout lx.m2 -rlen 50000 -
alpha 0.1 -radius 3
umat -cin lx.m2 -ps 1 > output.ps
vcal -din tutorial1.txt -cin lx.m2 -cout lx.m3
visual -din tutorial1.txt -cin lx.m3 -dout lx.m4
```

datoteka s podatki — **datoteka SOM** — **velikost U-matrike**

```
randinit -din ex.dat -cout ex.map -xdim 3 -ydim 4
         -topol rect -neigh bubble -rand 1
```

```
vsom -din tutorial1.txt -cin lx.m0 -cout lx.m1 -rlen 30000 -
alpha 0.8 -radius 5
```

število ponovitev (prvo učenje mreže)

```
vsom -din tutorial1.txt -cin lx.m1 -cout lx.m2 -rlen 50000 -
alpha 0.1 -radius 3
```

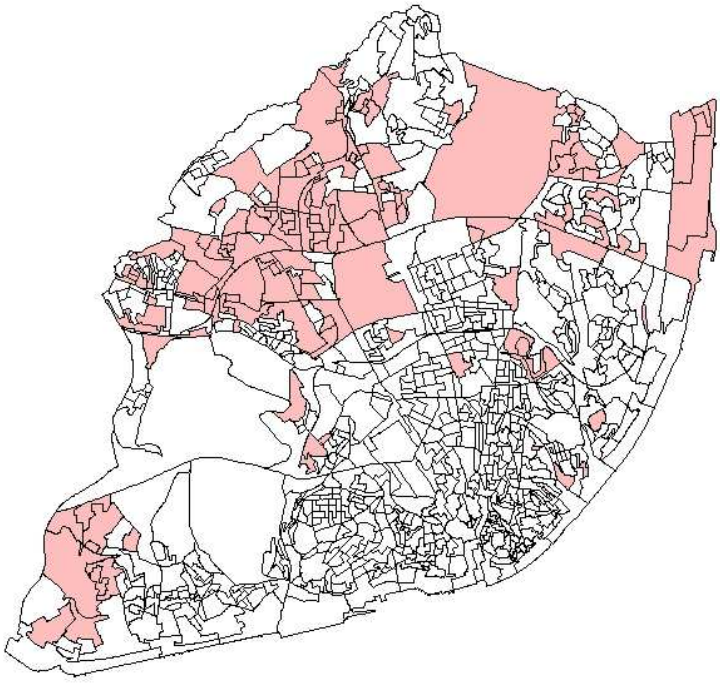
število ponovitev (drugo učenje mreže)

datoteka, ki vsebuje končni rezultat: U-matriko

```
umat -cin lx.m2 -ps 1 > output.ps
```

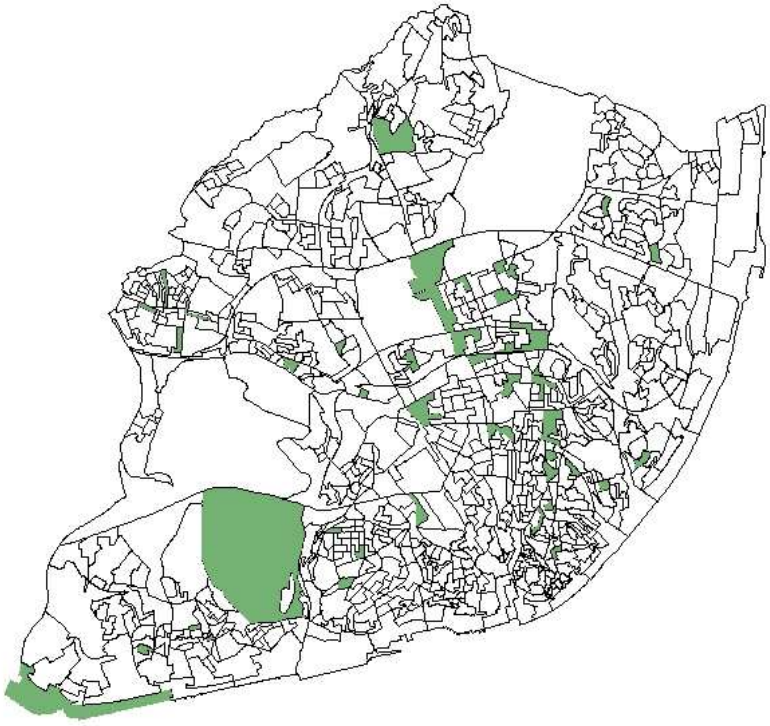
Priloga 2: REZULTATI PRIMERA 1: ZEMLJEVIDI ZA VSAKO SKUPINO POSEBEJ

Skupina 1:



Skupina 2:

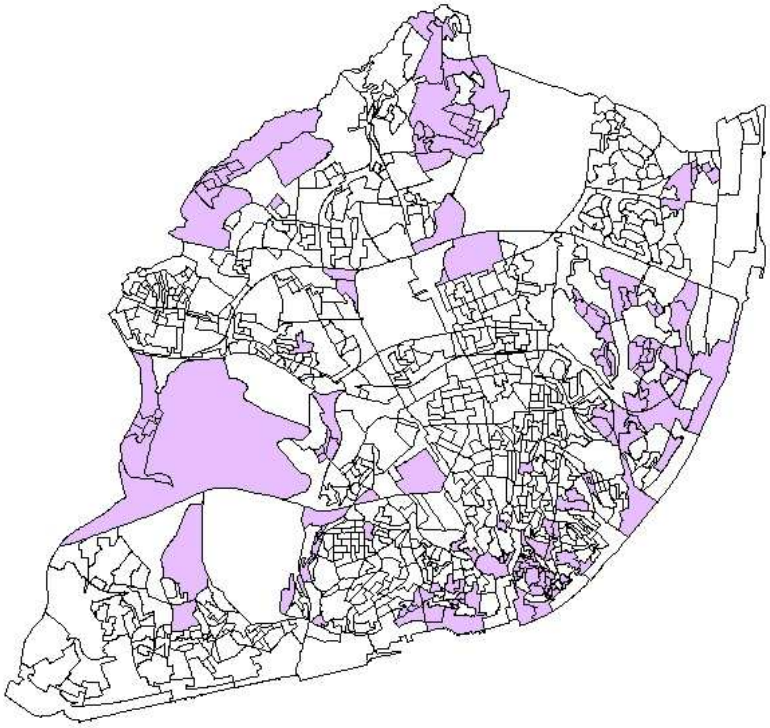
Skupina 3:



Skupina 4:

Skupina 5:

Skupina 6:



Priloga 3: SEZNAM UPORABLJENIH TUJIH IZRAZOV

Tuji izraz	**slovenski prevod**
self- organizing maps	samoorganizirajoče mreže
data mining	podatkovno rudarjenje
knowledge discovery from databases	odkrivanje znanja iz podatkov
data preparation	priprava podatkov
outlier	odstopajoča vrednost
decision trees	odločitvena drevesa
if-then	če-potem
support	podpora
confidence	zaupanje
improvement	izboljšanje
clustering	razvrščanje v skupine
selection	selekcija
crossover	križanje
mutation	mutacija
scatterplot	razpršeni diagram
scatterplot matrix	matrika razpršenih prikazov
principal component analysis	analiza osnovnih komponent
box plot	kvantilni diagram
point	točka
line	linija
area	površina
shapefile	datoteka shape

Priloga 4: SEZNAM UPORABLJENIH KRATIC

Kratica	Angleško	Kratka razlaga pojma
KDD	*knowledge discovery from databases*	Pridobivanje znanja iz podatkov
GIS	*geographic information systems*	Sistem za urejanje in upravljanje prostorsko orientiranih podatkov
GPS	*global positioning system*	Satelitski navigacijski sistem, ki se uporablja za določanje natančnega položaja in časa kjerkoli na Zemlji
SQL	*structured query language*	Programski jezik, namenjen delu s podatkovnimi bazami
OLAP	*on-line analytical processing*	Sprotna analitična obdelava podatkov
SOM	*self-organizing maps*	Samoorganizirajoče mreže, ki delujejo na principu nevronskih mrež