Univerza v Ljubljani

EKONOMSKA FAKULTETA

Viri podatkov

9. srečanje:

Statistična zaščita podatkov

Mojca Bavdaž (mojca.bavdaz@ef.uni-lj.si)

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Osnovni pojmi

Zaupnost (*angl.* confidentiality)

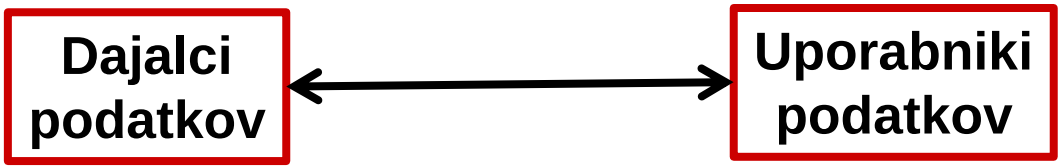
Zaupnost je status podatkov, za katerega sta se dogovorila oseba/organizacija, ki je podatke priskrbela, in organizacijo, ki je te podatke dobila, opisuje pa stopnjo nudene zaščite. (Dalenius,1988)

Zasebnost (*angl.* privacy)

Pravica odločati, katere informacije o sebi bomo delili z drugimi. (Fellegi, 1972)

Razkritje (*angl.* disclosure)

Ko oseba/organizacija iz objavljenih podatkov izve o drugi osebi/organizaciji nekaj novega, česar brez teh podatkov ne bi vedela.



Statistična zaščita podatkov

Statistična zaščita podatkov
(*angl.* statistical disclosure control/limitation)

⇒ Uporaba metod za znižanje tveganja razkritja.

Področja uporabe:

- Uradna statistika (zaupanje respondentov)
- Zdravstveni podatki
- Elektronsko poslovanje
- …

Vrste rezultatov, ki naj se jih ščiti:

- (Statične) statistične tabele
- Dinamične poizvedbe po bazah
- Mikropodatki *(Netflix!)*

Tipologija podatkov glede na stopnjo razkritja

Mikro podatki

- neanonimizirani mikro podatki (data enclaves),
- anonimizirani mikro podatki,
- statistično zaščiteni mikro podatki,
- public use microdata/files

Agregirani podatki

- v tabelah/bazah
- v grafičnih prikazih

Vrste spremenljivk glede na vlogo v statistični zaščiti podatkov

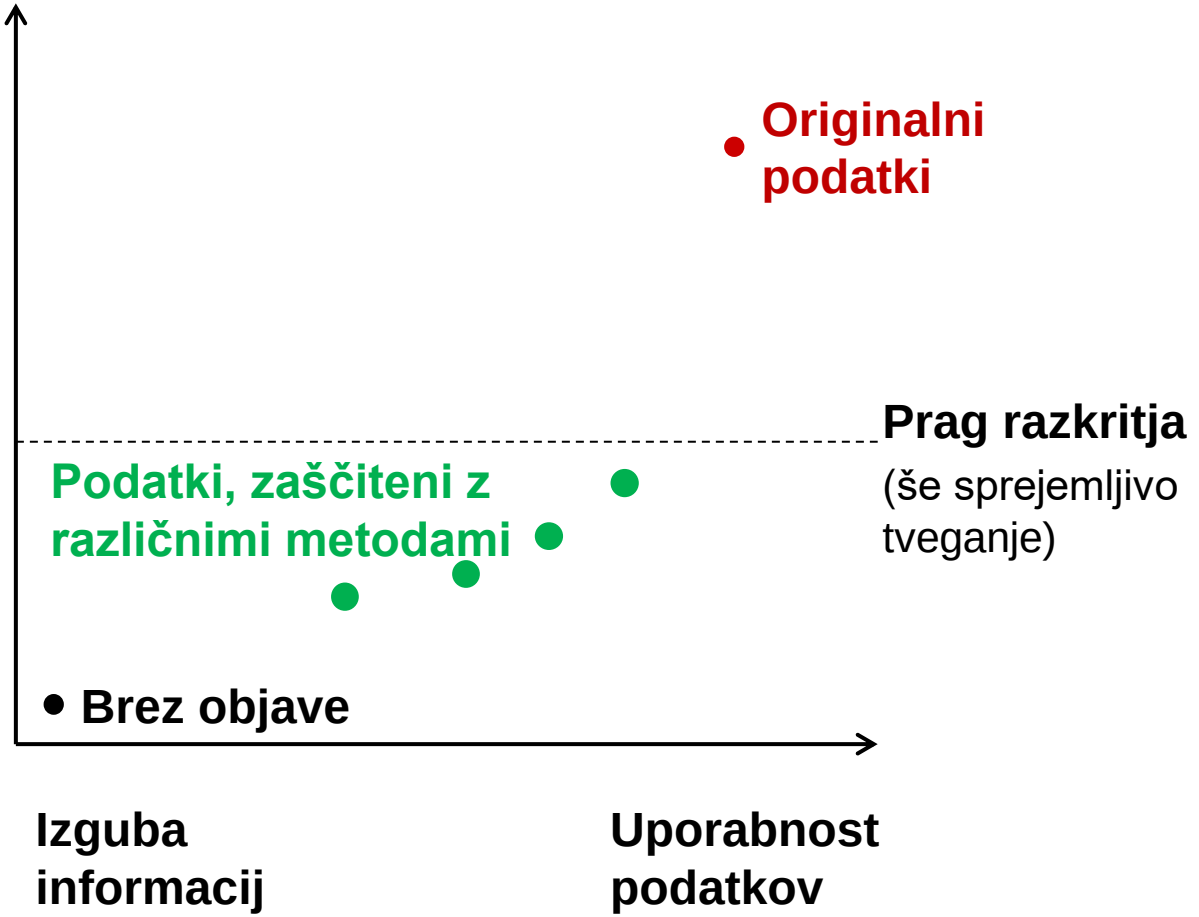
- Identifikatorji
  - spremenljivke, ki natančno določijo enoto
  - npr. EMŠO, davčna številka, matična številka podjetja
  - večinoma kategorialne spremenljivke (nominalke)
- Kvazi identifikatorji ali ključne spremenljivke
  - spremenljivke, ki pri določitvi enote dopuščajo dvom, a njihovo povezovanje lahko določi enoto
  - npr. ime, naslov, spol, starost, telefonska številka
- Zaupne spremenljivke
  - spremenljivke z občutljivo vsebino
  - nacionalnost, dohodek, zdravstveno stanje, veroizpoved
- Neobčutljive spremenljivke

Hundepool, A. et al. (2010). *Handbook on Statistical Disclosure Control. Version 1.2.* ESSNet SDC. Najdeno na http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf

Kompromisi

Tveganje razkritja

- enostavnost sklepa
- posledice (ne)točnosti sklepa za “vohljača”
- posledice teh sklepov za organizacijo/ponudnika podatkov



Prirejeno po Duncan, G.T. , Elliot, M., & Salazar-Gonzalez, J.-J. (2011). *Statistical Confidentiality. Principles and Practice.* Springer.

Metode statistične zaščite

- Metode so lahko namenjene:
  - mikropodatkom / tabelam / obojemu
  - kategorialnim spr. / zveznim spr./ obojim
- Glede na pristop ločimo:
  - deterministične metode
  - verjetnostne metode
- Glede na rezultat ločimo:
  - metode, ki ustvarijo sintetične podatke, tako da ohranijo določene statistične značilnosti originalnih podatkov
  - metode, ki zakrivajo originalne podatke:
    - zakrivanje z vnosom motenj *(angl.* perturbative masking)
    - zakrivanje brez vnosa motenj (*angl.* non-perturbative masking; podatkov ne spreminjajo, ampak jih izpuščajo, prekodirajo, agregirajo, vzorčijo ipd.)

Mikropodatki: zakrivanje z vnosom motenj

Metoda	Zvezne spremenljivke	Kategorialne spremenljivke	
Dodajanje šuma	X		⇒
Zaokroževanje	X		
Ponovno vzorčenje	X		
Mikroagregacija	X	(X)	⇒
Menjava rangov/podatkov	X	X	⇒
PRAM		X	⇒

Dodajanje šuma

- cilj: zaščita pred povezovanjem z zunanjimi podatki z dodajanjem stohastičnega šuma
- različne metode dodajanja šuma (več šuma na osamelce, ohranjanje povprečij, ohranjanje korelacij)

Mikroagregacija

- cilj: vse enote porazdeliti po skupinah in znotraj vsake skupine posamične vrednosti spremenljivke nadomestiti z aritmetično sredino ali kako drugo vrednostjo
- upoštevanje načela k-anonymity, minimalnega števila enot v skupini, ki še zagotavlja anonimnost (običajno k=3)

Menjava rangov/podatkov

- osnovna ideja: zamenjati vrednosti neke spremenljivke na delu enot
- postopek pri upoštevanju rangov: razvrstitev po velikosti določene spremenljivke, zamenjava podatkov znotraj določenega intervala, ponovitev postopka na naslednji spremenljivki. Tak pristop se je izkazal kot dober kompromis

Post-randomizacija

(*angl.* Post Randomization Method, PRAM)

namerna napačna klasifikacija

cilj: spremeniti del vrednosti kategorialne spremenljivke na osnovi vnaprej določenih verjetnosti prehoda iz ene v drugo kategorijo

Mikropodatki: zakrivanje brez vnosa motenj

Metoda	Zvezne spremenljivke	Kategorialne spremenljivke	
Prekodiranje	X	X	⇒
Vzorčenje		X	
Delno izpuščanje		X	⇒

Prekodiranje *(angl*. recoding):

- cilj: zmanjšati število možnih vrednosti spremenljivke;
- slabost: čeprav je problem navadno v specifični kombinaciji (npr. redek poklic+manjši kraj), se prekodiranje izvede na vseh podatkih (zato imenovano tudi globalno prekodiranje)
- kategorialne spremenljivke: združevanje več kategorij v eno, manj informativno
- zvezne spremenljivke: iz zveznih v diskretne vrednosti (manj običajno)
- prekodiranje najnižjih/najvišjih vrednosti (*angl.* bottom/top coding): združevanje vrednosti pod/nad pragom

Delno izpuščanje (*angl.* local suppression)

- cilj: izpustiti določene vrednosti iz spremenljivke, ki bi omogočila razkritje
- običajno se je potrebno odločiti, katera v kombinaciji vrednosti kategorialnih spremenljivk naj bo manjkajoča (npr. kraj ali poklic)

Tveganje razkritja v tabelah

Frekvenčne tabele

Vrednostne tabele (*angl.* magnitude tables) $\Rightarrow$ vsote

Povezane tabele

Primer problematične frekvenčne tabele

Zakonski stan	Polni delovni čas	Krajši delovni čas	Skupaj
Poročen	6	0	6
Razvezan	5	1	6
Samski	2	2	4
Skupaj	13	3	16

Hundepool, A. et al. (2010). *Handbook on Statistical Disclosure Control. Version 1.2.* ESSNet SDC. Najdeno na http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf

Katera polja so problematična?

Določanje občutljivosti polj v tabeli

Pravila	Polje v tabeli je občutljivo, ko…
Pravilo najmanjše frekvence	je frekvenca manjša od vnaprej definiranega minimuma *n* (običajno *n*=3)
(*n, k*) - pravilo dominantnosti	je vsota *n* največjih vrednosti večja od *k%* vsote v polju
Pravilo *p%*	je vsota v polju minus dve največji vrednosti manjša od *p%* največje vrednosti

Hundepool, A. et al. (2010). *Handbook on Statistical Disclosure Control. Version 1.2.* ESSNet SDC. Najdeno na http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf

Metode za zaščito tabel

- Predhodna zaščita mikropodatkov
- Preoblikovanje tabele
  - Združevanje kategorij, uporaba višje hierarhične ravni
  - Uporaba praga, najmanjše frekvence
- Spremembe po pripravi tabele
  - Uporaba manjkajočih vrednosti
    - Primarna
    - Sekundarna
  - Zaokrožitev vrednosti po vseh poljih
  - Motnja polja, npr. ‘barnardisation’, ki vsako polje z neničelno vrednostjo spremeni za +1, 0 ali -1 v skladu z verjetnostjo

- *Paziti na povezane tabele, nerazkrivanje pravil za zaščito, število dimenzij v tabeli*

Primer objavljenih pravil

EHIS Wave 1 variables: ANONYMISATION RULES

Povezava na primer

http://ec.europa.eu/eurostat/documents/203647/203710/EHIS_wave_1_anonymisation_rules.pdf

Uporaba manjkajočih vrednosti

	Regija				
	A	B	C	D	Skupaj
Squash	58	47	36	89	230
Fitness	71	124	24	31	250
Odbojka	92	157	59	28	454
Ostalo	800	934	651	742	3127
Skupaj	1021	1262	770	890	4061

Primarna

	Regija				
	A	B	C	D	Skupaj
Squash	58	X	36	89	230
Fitness	71	124	24	31	250
Odbojka	92	157	59	X	454
Ostalo	800	934	651	742	3127
Skupaj	1021	1262	770	890	4061

Sekundarna

	Regija				
	A	B	C	D	Skupaj
Squash					
Fitness					
Odbojka					
Ostalo					
Skupaj					4061

Vir: Jukić, N., SURS