REPUBLIKA SLOVENIJA
STATISTIČNI URAD

SURS Litostrojska cesta 54, 1000 Ljubljana

STATISTIČNA ZAŠČITA PODATKOV SPLOŠNA METODOLOŠKA POJASNILA

Opis

Statistična zaščita tabel

Statistična zaščita mikropodatkov

Sestavili:

Zadnjič osveženo

Opis

Doseganje strateških ciljev in izpolnjevanje nalog državne statistike sta med drugim odvisni od zaupanja opazovanih enot oz. dajalcev podatkov: oseb, gospodinjstev, podjetij, kmetijskih gospodarstev in drugih organizacij. To pomeni, da nam bodo svoje podatke zaupali, če bodo prepričani, da ravnamo tako, da v celoti zagotavljamo njihovo zaupnost (tj. varujemo njihovo identiteto) in zaupnost danih informacij, in da bodo pridobljeni podatki uporabljeni samo za statistični namen. S tem namenom jih tudi nenehno obveščamo o tem, jim razlagamo uporabljene postopke, in temu je namenjeno tudi to metodološko pojasnilo.

Del zaščite (varovanja) zaupnosti podatkov temelji na tem, da izkazani podatki ne omogočajo neposredne identifikacije (s pomočjo direktnih identifikatorjev) ali posredne identifikacije (na kateri koli drug način). K temu nas zavezujeta tako slovenska kot tudi evropska zakonodaja. Glede na to, da je poslanstvo državne statistike posredovati in objavljati statistične rezultate v čim širšem obsegu, vendar hkrati tako, da je tveganje razkritja informacij o enotah minimalno, so potrebne metode statistične zaščite, s katerimi dosežemo, da se spoštuje zakonodaja in obenem ohranja zadovoljiva oz. čim večja obvestilna (informacijska) vrednost posredovanih oziroma objavljenih statističnih rezultatov.

Ločimo dve vrsti metod za statistično zaščito podatkov:

- **metode, uporabljene na osnovnih mikropodatkih:**
  - *metode, ki spreminjajo podatke* (podatki naj bi se z uporabo take metode spremenili čim manj, vendar v tolikšni meri, da je tveganje razkritja najmanjše možno):
    - statistična zaščita mikropodatkov: metoda PRAM, dodajanje šuma, zaokroževanje, menjava podatkov;
    - statistična zaščita tabel: zaokroževanje, metoda CTA;
  - *metode, ki ne spreminjajo podatkov* (podatki imajo manjšo informacijsko vrednost, vendar so še vedno točni):

        - statistična zaščita mikropodatkov: metoda manjkajočih vrednosti, združevanje v razrede, kodiranje vrha, kodiranje dna, mikroagregacija, vzorčenje;
        - statistična zaščita tabel: metoda manjkajočih vrednosti, združevanje v razrede;
- **metode za generiranje sintetičnih podatkov** (podatke dobimo z upoštevanjem statističnega modela).

Statistična zaščita tabel

Tabela (preglednica) je podatkovna struktura, ki realizira matematično strukturo matrike in razvrsti podatke v celice po vrsticah in stolpcih. V njej so prikazani agregirani podatki (različne statistike), ki opisujejo združene podatke več enot. Vsako statistiko, ki jo nameravamo posredovati, si lahko predstavljamo kot tabelo dimenzije 1 x 1. Spremenljivke, ki napenjajo tabelo, se imenujejo pojasnjevalne spremenljivke. Število pojasnjevalnih spremenljivk določa število dimenzij tabele.
Pri postopku statistične zaščite tabel določimo pravila za določanje primarno občutljivih celic oz. vrednosti, ki so prikazane v celicah tabele. Od tipa tabele je odvisno, katera pravila bomo uporabili. Ločimo tri tipe tabel:

- **frekvenčne tabele** (v vsaki celici tabele je vpisano število enot, ki pripadajo tej celici);
- **vrednostne tabele** (v vsaki celici je vpisana vsota vrednosti določene spremenljivke enot, ki pripadajo tej celici);
- **druge tabele** (v celicah so vpisane vrednosti drugih tipov statistik – deležev, razmerij, indeksov in drugih), ki so običajno povezane z vrednostnimi in frekvenčnimi tabelami.

Pravila za določanje primarno občutljivih celic so:

- **prag** (če je število enot manjše od praga, potem je celica primarno občutljiva) – minimalna vrednost za prag naj bo 3;
- **pravilo dominantnosti** (če n največjih prispevkov prispeva več kot k % k vrednosti celice, potem je celica primarno občutljiva);
- **p%-pravilo** (če lahko enota z drugim največjim prispevkom oceni največji prispevek bolje kot na p % natančno, potem je celica primarno občutljiva).

Pri frekvenčnih tabelah upoštevamo le prag, medtem ko pri vrednostnih in drugih tabelah upoštevamo prag in p%-pravilo ali pa prag in pravilo dominantnosti. Pravila za določanje primarno občutljivih celic so zaupna in se ne smejo posredovati, saj bi z razkritjem parametrov zmanjšali zaščito enot.
Če se določene enote strinjajo (pridobimo njihovo pisno dovoljenje), da se njihovi individualni podatki objavijo, ali če je del podatkov pridobljen iz javnega vira, se lahko uporabi t. i. pravilo zahteve. To pomeni, da tem enotam dodelimo poseben status in tako zmanjšamo izgubo informacije v tabelah, v katerih se pojavljajo.
Z metodo za statistično zaščito tabel primarno občutljive celice vedno zaščitimo. Zaradi povezav znotraj tabele pa najpogosteje zaščitimo še dodatne celice (pravimo jim sekundarno zaščitene celice); če povezav ni, to ni potrebno. Pri metodi manjkajočih vrednosti zakrijemo primarno in sekundarno občutljive vrednosti tako, da jih ne objavimo, tj. jih ne vpišemo v celico, temveč namesto vrednosti vpišemo črko „z“. Izgubo informacije določa delež zakritih celic.
Pri statistični zaščiti tabel je treba paziti na hierarhije pojasnjevalnih spremenljivk, število dimenzij tabel in povezave med tabelami.

Statistična zaščita mikropodatkov

Mikropodatke (individualne podatke) je treba zaščititi zaradi velikega števila spremenljivk, ki so v datoteki mikropodatkov, saj so nekatere kombinacije spremenljivk redke in zato je tveganje, da se razkrije identiteta enote, veliko.
V procesu statistične zaščite mikropodatkov moramo vedeti, kateremu tipu uporabnikov bodo zaščitene datoteke mikropodatkov namenjene. Če je taka datoteka namenjena raziskovalcem, potem bo izguba informacije manjša, saj enote zaščitimo le pred nenamernim razkritjem. V takem primeru raziskovalec podpiše tudi pogodbo s SURS-om, s katero se zaveže k varovanju podatkov, ki jih bo prejel. Če je datoteka namenjena javnosti (v to skupino štejemo tudi študente), potem bo izguba informacije večja, saj enote zaščitimo tudi pred namernim razkritjem. V takem primeru je datoteka na voljo na spletni strani ponudnika podatkov, ne da bi bilo treba podpisati pogodbo.

Pri procesu statistične zaščite mikropodatkov moramo najprej izbrati statistike, katerih vrednosti bomo skušali čim manj spremeniti.
V naslednjem koraku je treba določiti 3 nabore spremenljivk:

- **identifikacijske spremenljivke:** spremenljivke, ki omogočajo razkritje enote:
    - *direktni identifikatorji:* spremenljivke, ki enolično določajo enote (npr. davčna številka, matična številka podjetja), omogočajo neposredno identifikacijo;
    - *indirektni identifikatorji:* spremenljivke, katerih kombinacije lahko privedejo do razkritja enote (npr. ime, naslov, spol, starost, regija), omogočajo posredno identifikacijo;
- **zaupne izhodne spremenljivke:** spremenljivke, ki nosijo občutljivo informacijo o enoti (npr. prihodek, vera, politično prepričanje, zdravstveno stanje …);
- **nezaupne izhodne spremenljivke:** druge spremenljivke.

Redke kombinacije preostalih identifikacijskih spremenljivk so tiste, ki jih je treba zaščititi, saj obstaja velika verjetnost, da se enoto, na katero se ta redka kombinacija nanaša, lahko prepozna.
Nato določimo sprejemljivo tveganje razkritja (verjetnost prepoznave pri točno določenem scenariju). Za to imamo na voljo več metod. Omenili bomo dve največkrat uporabljeni:

- na podlagi števila enot, ki posamezni kombinaciji identifikacijskih spremenljivk pripadajo (ta metoda je uporabna, če imamo v datoteki celotno populacijo);
- z zahtevnejšo matematično metodo, pri kateri določimo statistični model za populacijo na podlagi vzorčnih podatkov, ki so v datoteki.

Določimo število enot, ki jih uporabnik lahko potencialno še vedno prepozna. Zavedati se moramo, da kljub statistični zaščiti tveganje razkritja ni nikoli ničelno.
V naslednjem koraku določimo metode statistične zaščite in jih uporabimo na podatkih. Direktne identifikatorje in določene indirektne identifikatorje, tj. take, katerih kombinacija z veliko verjetnostjo enolično določa enoto (npr. ime, priimek, naslov), moramo odstraniti. Preostali indirektni identifikatorji so spremenljivke, ki jih v podatkih obdržimo, saj so pomembne za statistično analizo (npr. starost, spol, statistična regija). Te spremenljivke imenujemo ključne spremenljivke. Izberemo metode, ki jih bomo uporabili na ključnih spremenljivkah, da bo verjetnost posredne identifikacije sprejemljiva. Metode so

izbrane glede na namen podatkov (npr. datoteka za javnost, datoteka za raziskovalce). Zaščitimo lahko tudi zaupne izhodne spremenljivke. Nekatere metode so uporabne le za zvezne spremenljivke (npr. dodajanje šuma), druge le za znakovne spremenljivke (npr. metoda manjkajočih vrednosti); obstajajo pa tudi metode, ki so primerne za obe vrsti spremenljivk (npr. združevanje v razrede). Metode uporabimo v taki meri, da zmanjšamo tveganje razkritja enot v zaščiteni datoteki do sprejemljive (dopustne) stopnje. Uporabniku zaščitenih podatkov pojasnimo, katere metode smo uporabili na podatkih, saj se mora zavedati, da so določeni podatki spremenjeni (zmoteni) in da se vrednosti v zaščiteni datoteki razlikujejo od vrednosti v originalni datoteki. Ne smemo pa razkriti parametrov metod (npr. nabor indirektnih identifikatorjev, praga pri metodi manjkajočih vrednosti), saj bi z razkritjem parametrov zmanjšali zaščito enot.

Sestavili:

Andrejka Smukavec, Manca Golmajer

Zadnjič osveženo

5. 10. 2016