



Rudi Seljak

UREJANJE PODATKOV



Oktober 2019

Kazalo

1	SPLOŠNO O UREJANJU	3
2	KONTROLA PODATKOV	3
3	POPRAVKI PODATKOV	4
4	VSTAVLJANJE PODATKOV	4

1 SPLOŠNO O UREJANJU

V praksi je vsako pridobivanje podatkov in izračunavanje statističnih rezultatov »okuženo« z različnimi vrstami napak. Izraz urejanje podatkov označuje vse postopke, s katerimi:

- iščemo in odpravljamo napake v podatkih,
- pridobivamo informacije o kakovosti podatkov tako na mikro- kot tudi na makroravni,
- redno testiramo in izboljšujemo statistični proces.

Z vidika vpliva na statistične rezultate je pomembno ločevanje na naslednji vrsti napak:

- **slučajne napake**
Napake, ki so posledica slučajnih vplivov v procesu pridobivanja podatkov. Tipičen vzrok za slučajne napake je »zatipkanje« pri elektronskem sporočanju podatkov. Slučajne napake povečujejo varianco statističnih rezultatov, ne povzročajo pa pristranskosti ocen;
- **sistematične napake**
Napake, ki so posledica sistematičnih pomanjkljivosti v procesu pridobivanja podatkov. Do sistematičnih merskih napak pride običajno zaradi pomanjkljivosti pri merskem instrumentu, torej v našem primeru zaradi slabega vprašalnika, premalo usposobljenega ali motiviranega anketarja ali pa zaradi sistematične napake v postopku pretvorbe podatkov v računalniško obliko (npr. napačno kodiranje). Tipičen primer vzroka za nastanek sistematične napake je tudi uporaba napačne merske enote (npr. evro namesto 1000 evrov).

2 KONTROLA PODATKOV

Osnovno orodje, s katerim izvajamo statistično urejanje podatkov, je smiseln in dosleden sistem logičnih pravil, na podlagi katerih podatke »podvržemo« preverjanju, vnaprej opredeljenim kontrolam. Pravila logičnih kontrol delimo v tri osnovne skupine:

- **kontrola smiselnosti** (ang. »validity check«)
Za posamezno spremenljivko preverjamo, ali je vrednost v dovoljenem obsegu oziroma naboru
- **kontrola doslednosti** (ang. »consistency check«)
Preverjamo povezanost med vrednostmi več spremenljivk pri isti enoti.
- **kontrola porazdelitve** (ang. »distributional check«)
Pravilnost vrednosti spremenljivke ocenjujemo glede na porazdelitev vrednosti spremenljivk pri drugih enotah. Večinoma gre tu za iskanje izstopajočih vrednosti ali osamelcev.

Osamelec je vrednost, katerega vrednost se glede na opredeljena merila bistveno razlikuje od drugih vrednosti. Definicija osamelca ni absolutna in sam pojem je natančno določen šele z izbiro ustreznih kriterijev v posameznem raziskovanju. Kadar obravnavamo osamelce, je pomembno predvsem določiti:

- ali je vrednost osamelca napačen podatek ali pa gre za sicer izstopajoč, vendar pravilen podatek;
- ali je osamelec – če gre za pravilen podatek in če raziskovanje izvajamo na podlagi slučajnega vzorca – reprezentativen. Osamelec je reprezentativen, če v populaciji obstaja (vsaj približno) tolikšno število podatkov podobnega obsega, kot je faktor preračuna na populacijo (populacijska utež).

3 POPRAVKI PODATKOV

Popravke podatkov lahko na osnovni ravni razdelimo v dva sklopa: ročni ter avtomatski popravki podatkov. Z izrazom »ročni popravki« označujemo tiste popravke, pri katerih glavni in odločujoči dejavnik še vedno človeški dejavnik. Gre torej za popravke, ki jih izvajamo ob preverjanju podatkov preko ponovnega stika s poročevalsko enoto ali pa jih izvajamo na podlagi ekspertnih ocen ustreznih izvajalcev. Pri avtomatskih popravkih, v nasprotju z ročnimi popravki, izvajanja popravkov prepustimo računalniškim programom ali aplikacijam. V sodobnih praktičnih izvedbah urejanja podatkov se zelo pogosto uporablja ustrezna kombinacija obeh pristopov, ročnega in avtomatskega, zelo poredko pa zgolj ročno ali zgolj avtomatsko urejanje podatkov.

Popravke lahko delimo tudi na individualne ter sistematske popravke. Individualni popravki so popravki, ki jih izvajamo samo pri določeni enoti. Sistematski popravki so popravki, kjer z uporabo determinističnega pravila, določen popravek izvedemo na celotnem sklopu enot, ki zadoščajo nekemu pogoju. V primeru ročnih popravkov v veliki meri prevladujejo individualni popravki, v primeru avtomatskih popravkov pa v veliki meri sistematski popravki.

V postopkih izvajanja popravkov je zelo pomembno zagotoviti sledljivost sprememb. Zagotavljanje sledljivosti pomeni, da smo po koncu izvajanja popravkov sposobni točno določiti, kateri podatki in za koliko so se spremenili. V konkretni izvedbi to pomeni, da popravljeni podatek nikoli ne »prekrije« originalnega podatka, ampak se vedno tvori nova verzija podatkov. Na SURS še višjo stopnjo sledljivosti zagotavljamo s tako imenovanimi statusi spremenljivk. Status spremenljivke je procesni metapodatek, ki je pridružen vsaki spremenljivki in v kateri je zabeležena vsaka sprememba podatka ter postopek oziroma metoda s katero je bila sprememba izvedena.

4 VSTAVLJANJE PODATKOV

Izraz vstavljanje podatkov (ang. »data imputation«) označuje vse postopke, pri katerih manjkajoče ali v procesu urejanja podatkov zaznane napačne vrednosti nadomestimo s statističnimi ocenami. Postopek vstavljanja podatkov torej uporabljamo vedno, kadar manjkajoče ali napačne vrednosti nadomeščamo z računalniškimi postopki. Obstaja veliko različnih metod vstavljanja, na SURS pa najpogosteje uporabljamo naslednje:

- **Metoda logičnega vstavljanja.** Pri tej metodi vstavimo vrednost, ki logično sledi iz podatkov, ki nam jih je za obravnavano enoto uspelo pridobiti.
- **Metoda povprečne vrednosti.** Vstavljena vrednost je povprečna vrednost podatkov tistih enot, katerih podatke imamo. Povprečne vrednosti običajno ne izračunavamo iz celotnega nabora podatkov, ampak iz podatkov za določeno celico vstavljanja, v kateri je enota, za katero vstavljamo podatek.
- **Metoda notranjega darovalca.** Pri metodi notranjega darovalca manjkajočo ali napačno vrednost nadomestimo z eno od vrednosti pri enotah, za katere imamo podatek in za katere je vrednost »potrjena« kot ustrezna. Enoto, pri kateri vstavljamo podatek, imenujemo prejemnik (ang. »recipient«), enoto, od katere smo podatek prevzeli, pa darovalec (ang. »donor«).
- **Metoda uporabe zgodovinskih podatkov.** Metoda je uporabna pri periodičnih raziskovanjih, pri katerih isto enoto vključimo v raziskovanje v več zaporednih opazovanih obdobjih. Vstavljanje z uporabo zgodovinskih podatkov lahko uporabimo pri enoti, ki sicer za tekoče obdobje ni sporočila podatka, sporočila pa ga je za preteklo ali za katero od preteklih obdobj. Podatek iz preteklih izvedb (zgodovinski podatek) nato uporabimo za oceno vstavljene vrednosti v tekočem obdobju. Način uporabe zgodovinskega podatka za oceno manjkajočega podatka v tekočem obdobju (kako, na kakšen način uporabimo zgodovinski podatek) določa eno od izvedb metode.