UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

STATISTIČNO UREJANJE PODATKOV KRATKOROČNIH STATISTIK

Ljubljana, september 2016 RUDOLF SELJAK

IZJAVA O AVTORSTVU

Podpisani Rudolf Seljak, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Statistično urejanje podatkov kratkoročnih statistik, pripravljenega v sodelovanju s svetovalko doc. dr. Mojco Bavdaž

IZJAVLJAM

1. da sem predloženo delo pripravil samostojno;

2. da je tiskana oblika predloženega dela istovetna njegovi elektronski obliki;

3. da je besedilo predloženega dela jezikovno korektno in tehnično pripravljeno v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani, kar pomeni, da sem poskrbel, da so dela in mnenja drugih avtorjev oziroma avtoric, ki jih uporabljam oziroma navajam v besedilu, citirana oziroma povzeta v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani;

4. da se zavedam, da je plagiatorstvo – predstavljanje tujih del (v pisni ali grafični obliki) kot mojih lastnih – kaznivo po Kazenskem zakoniku Republike Slovenije;

5. da se zavedam posledic, ki bi jih na osnovi predloženega dela dokazano plagiatorstvo lahko predstavljalo za moj status na Ekonomski fakulteti Univerze v Ljubljani v skladu z relevantnim pravilnikom;

6. da sem pridobil vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v predloženem delu in jih v njem jasno označil;

7. da sem pri pripravi predloženega dela ravnal v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil soglasje etične komisije;

8. da soglašam, da se elektronska oblika predloženega dela uporabi za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;

9. da na Univerzo v Ljubljani neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve predloženega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja predloženega dela na voljo javnosti na svetovnem spletu preko Repozitorija Univerze v Ljubljani;

10. da hkrati z objavo predloženega dela dovoljujem objavo svojih osebnih podatkov, ki so navedeni v njem in v tej izjavi.

V Ljubljani, dne _____ Podpis študenta:_____

KAZALO

KAZALO TABEL

KAZALO SLIK

UVOD

Različne vrste podatkov predstavljajo osnovo za izvajanje statističnega raziskovanja. V idealnem svetu bi bili vsi ti podatki točni in brez napak. Ker pa, tako postopek pridobivanja kot tudi postopek obdelave podatkov, potekata v realnih okoliščinah in pogojih, so tako vhodni mikro podatki kot tudi izhodni statistični rezultati neizogibno »okuženi« z različnimi vrstami napak, ki vedno do določene mere izkrivljajo statistično sliko opazovanega pojava. Naloga vseh udeležencev v procesu izvedbe statističnega raziskovanja je zato, v okviru danih možnosti, čim več teh napak v podatkih zaznati in jih popraviti ter tako poskrbeti za čim bolj verodostojno sliko pojava (Arnež et al., 2012). Odkrivanje in odpravljanje napak v procesu izvedbe statističnega raziskovanja je še posebej pomembno v primeru raziskovanj s področja uradne statistike, saj so zahteve po točnosti in zanesljivosti teh statističnih rezultatov vedno zelo stroge. Dejstvo je tudi, da v primeru teh raziskovanj običajno obstaja precejšnja količina pomožnih podatkov in rezultatov, kar omogoča učinkovitejšo odpravo napak.

Urejanje podatkov prav gotovo predstavlja enega najbolj zahtevnih, če ne kar najbolj zahteven in drag del statističnega procesa. Nekoliko starejše študije ocenjujejo, da v primeru poslovnih raziskovanj postopki urejanja terjajo celo od 40 do 60 % celotnega proračuna raziskovanja (Granquist & Kovar, 1997). V zadnjem času študije izkazujejo nekoliko nižje, a še vedno visoke odstotke, nekje od 30 do 40 % (Black, 2009; Norberg, 2009). Ocenjuje se tudi, da je delež stroškov nekoliko nižji v primeru kratkoročnih raziskovanj kot v primeru letnih oziroma večletnih raziskovanjih (Norberg, 2009). V vsakem primeru obstaja splošen konsenz, da so stroški, ki jih statistične organizacije namenjajo postopkom urejanja, še vedno previsoki.

Ključen razlog, da so stroški urejanja tako visoki, je v tem, da velik del postopkov še vedno sloni na tako imenovanem ročnem urejanju. Izraz »ročno urejanje« označuje postopke, ki jih izvajajo za take postopke posebej usposobljene osebe in sloni predvsem na individualnem pristopu do preverjanja podatkov. Konkretno to pomeni, da podatke vsake enote, ki nam jih je sistem kontrol zaznal kot dvomljive, presojamo individualno, največkrat prek ponovnega kontakta z enoto (običajno preko telefona). Če v raziskovanju uporabljamo zgolj take, »ročne« postopke urejanja, je urejanje običajno zelo potrošno, tako s časovnega kot stroškovnega vidika. Zato ne čudi dejstvo, da je v zadnjih letih veliko aktivnosti in raziskovalnih dejavnosti povezanih z racionalizacijo in nadgradnjo postopkov ročnega urejanja. Vpeljava novih, učinkovitejših postopkov, ki bi vsaj delno sloneli na avtomatiziranih postopkih in bi v čim večji meri izkoriščali vse bolj zmogljivo računalniško opremo in tehnologijo, je osrednji cilj teh razvojnih aktivnosti (Seljak, 2012).

Eno od področij uradne statistike, kjer je uporaba hitrih in učinkovitih postopkov urejanja še posebej pomembna, je področje kratkoročnih poslovnih raziskovanj. Kratkoročna poslovna raziskovanja so raziskovanja, ki zagotavljajo statistične podatke za analizo kratkoročnega gibanja ponudbe in povpraševanja, proizvodnih dejavnikov in cen.

Učinkovito urejanje podatkov je v primeru kratkoročnih poslovnih raziskovanj še posebej velikega pomena. Čas med zaključkom zbiranja podatkov in objavo rezultatov je zelo kratek in ne dopušča podrobnega pregleda vseh podatkov in kontaktiranja velikega števila poročevalskih enot. Vpeljava postopkov selektivnega urejanja ter čim bolj avtomatiziranih postopkov je torej na tem področju še posebej ključnega pomena za zagotavljanje hitrih, vendar še vedno dovolj točnih rezultatov.

Namen magistrskega dela je tako s teoretskega kot s praktičnega vidika obravnavati področje statističnega urejanja podatkov kratkoročnih statističnih raziskovanj. Glavni cilj naloge je torej postaviti splošni teoretski in izvedbeni okvir, ki bo uporaben v vseh (oziroma vsaj v večini) kratkoročnih raziskovanj, in preizkusiti njegovo uporabnost na primeru izbranega konkretnega raziskovanja. V prvem delu naloge bomo tako najprej predstavili in opisali osnovne teoretske pojme in koncepte s področja statističnega urejanja podatkov. V nadaljevanju bomo nato predstavili področje kratkoročnih raziskovanj, predvsem glavne značilnosti teh raziskovanj, ki določajo tudi specifiko urejanja podatkov na tem področju. Osrednji del naloge bo namenjen opredelitvi teoretskega modela, ki bo upošteval vse razlikovalne značilnosti urejanja v primeru kratkoročnih raziskovanj. Veljavnost modela bo testirana na primeru kratkoročnih raziskovanj, ki jih trenutno izvaja Statistični urad Republike Slovenije (v nadaljevanju SURS). Teoretski model bomo primerjali s trenutno prakso urejanja v enem od kratkoročnih raziskovanj ter na podlagi teoretskih razmislekov in rezultatov empiričnih analiz predlagali izboljšave v trenutnem izvajanju postopkov. Preko implementacije modela na primeru konkretnega statističnega raziskovanja bomo preverjali tezo, da lahko s pomočjo dobro opredeljenega teoretskega modela pripomoremo k izboljšanju učinkovitosti postopkov urejanja. Zadnji del naloge bo posvečen analizi vplivov postopkov urejanja na posamezne dimenzije kakovosti, kot jih definira model ocenjevanja kakovosti, ki je v splošni uporabi v Evropskem statističnem sistemu (v nadaljevanju ESS).

1 STATISTIČNO UREJANJE PODATKOV

1.1 Statistično urejanje podatkov – osnovni pojmi in koncepti

Preden se posvetimo samemu urejanju podatkov, nekaj besed o sami predstavitvi podatkov, ki bodo glavni predmet naše obravnave. Za predstavitev podatkov bomo uporabili »klasičen« dvodimenzionalen matrični model, kjer bomo množico statističnih podatkov z n enotami in m spremenljivkami zapisali kot:

$$\mathbf{Y} = \begin{bmatrix} y_{11}, y_{12}, \cdots, y_{1m} \\ y_{21}, y_{22}, \cdots, y_{2m} \\ \cdots \\ y_{n1}, y_{n2}, \cdots, y_{nm} \end{bmatrix} \quad (1)$$

Nabor v raziskovanju opazovanih spremenljivk bomo predstavili kot vektor $\vec{Y} = (Y_1, Y_{2,\dots,} Y_m)$, kjer je vsaka komponenta Y_i slučajna spremenljivka z neko slučajno porazdelitvijo. Opazovane vrednosti $(y_{1i}, y_{2i,\dots,} y_{ni})$ so ena od možnih *n*-realizacij slučajne spremenljivke. Pozneje, ko bomo podrobneje obravnavali podatke v kratkoročnih raziskovanjih, bomo v ta predstavitveni model uvedli še tretjo, časovno dimenzijo.

Izraz urejanje podatkov v svojem osnovnem pomenu označuje vse postopke, s katerimi iščemo in odpravljamo napake v podatkih. Sodobni pristop razume proces urejanja podatkov v nekoliko širšem pomenu. Granquist (1995) tako opredeljuje naslednje glavne cilje urejanja:

- Poiskati in odpraviti tiste napake v podatkih, ki imajo nezanemarljiv vpliv na statistične rezultate, ter s tem zagotoviti konsistenten in popoln nabor individualnih mikro podatkov.
- Opredeliti izvore napak, kar posledično omogoča tudi vpeljavo izboljšav v statistični proces.
- Zagotoviti informacije za oceno kakovosti tako vhodnih mikro podatkov kot končnih statističnih rezultatov.

Ob presojanju učinkovitosti postopkov urejanja in ob načrtovanju za njihovo izboljšavo je predvsem pomembna delitev postopkov urejanja na **ročno** in **avtomatsko urejanje**. Izraz ročno urejanje tu označuje postopke, pri katerih je, predvsem v fazi izvajanja popravkov, še vedno prevladujoč in odločujoč človeški faktor. Pravilnost podatkov in potreba po njihovem popravljanju se presojata na ravni posamezne enote. Večinoma podatke preverjamo v ponovnem stiku s poročevalsko enoto ali pa popravljene vrednosti ocenimo na podlagi ekspertnih mnenj ustreznih izvajalcev. Nasprotno, pri postopkih avtomatskega urejanja vse faze izvajamo sistematsko, s pomočjo ustreznih računalniških programov oziroma aplikacij in brez ponovnega stika s poročevalsko enoto. V sodobnih praktičnih izvedbah urejanja podatkov se zelo pogosto uporablja ustrezna kombinacija obeh pristopov, ročnega in avtomatskega (Panekoek, Scholtus, & Van der Loo, 2013), zelo poredko pa zgolj ročno ali zgolj avtomatsko urejanje podatkov. Čeprav je tudi ročno urejanje še vedno pomemben dejavnik v praksi statističnih uradov, pa je z vidika strokovne zahtevnosti, predvsem pa z vidika koristnosti za proces racionalizacije procesov, neprimerno bolj zanimivo področje avtomatskega urejanja. Tako bo tudi magistrsko delo večji poudarek namenilo obravnavi področja avtomatskega urejanja.

Na začetku postopkov urejanja praviloma presojamo, ali je neki podatek ustrezen za naš statističen namen. Čeprav se opredelitev pojma ustreznosti podatka zdi na prvi pogled jasna in neproblematična, pa ob delu s konkretnimi podatki ta hitro izgubi svojo enopomenskost. Predvsem zaradi lažjih konceptualnih razmislekov je koristno, če ta pojem nekoliko podrobneje razčlenimo. Rivière (2002, str. 4) vpelje naslednja tri razlikovanja:

prava/napačna (angl. *right/wrong*), **pravilna/nepravilna** (angl. *correct/erroneous*) ter **sprejemljiva/dvomljiva** (angl. *acceptable/doubtfull*) vrednost podatka, z naslednjimi opredelitvami:

- Prava vrednost se nanaša na predpostavljeno vrednost v »objektivno obstoječem realnem svetu«. Tudi če predpostavimo, da taka prava vrednost sploh obstaja (čeprav je to nemalokrat vprašljivo), jo je v večini primerov zaradi različnih napak v procesu merjenja nemogoče točno izmeriti. Tako ostaja točna vrednost večinoma samo koncept na ravni teoretske predpostavke.
- Vrednost podatka je pravilna, če je kdo, ki bi zelo natanko poznal področje, tudi ob podrobnem preverjanju (tudi po morebitnem ponovnem stiku s poročevalsko enoto), ne bi spremenil.
- Vrednost podatka je sprejemljiva, če jo potrdi računalniški program za kontrolo podatkov.

Ker je v praktičnih izvedbah raziskovanj zelo težko ali celo nemogoče presojati, ali je nek podatek pravi/napačen, v nadaljevanju tega razlikovanja ne bomo več upoštevali, ampak bomo celotno razpravo zgradili na razmisleku o zadnjih dveh kategorizacijah.

Razlika med pravilnostjo in sprejemljivostjo podatka je torej predvsem v tem, da se pravilnost nanaša na presojo statistika (ali na podlagi ponovnega preverjanja ali preprosto na podlagi svojih izkušenj), sprejemljivost pa na formalno definiran in v računalniškem jeziku zapisan nabor logičnih kontrol. Iz razlike med sprejemljivostjo in pravilnostjo izhaja kategorizacija enot, ki jo prikazuje Tabela 1.

Tabela 1: Kategorizacija enot na podlagi sprejemljivosti/pravilnosti

	Dvomljive vrednosti	**Sprejemljive vrednosti**
Nepravilne vrednosti	a	b
Pravilne vrednosti	c	d

Vir: P. Riviere, General principles for data editing in business surveys and how to optimize it, 2002, str. 4.

Glede na opredeljene kategorije, lahko definiramo naslednje kazalnike, ki nam nakazujejo uspešnost in učinkovitost urejanja (Riviere, 2002, str. 4):

- **Stopnja zavrnitve** (angl. *failure rate*) je delež enot, za katere je vsaj ena od opredeljenih kontrol signalizirala napako:

$$f = \frac{a + c}{a + b + c + d} \tag{2}$$

- **Stopnja nepravilnosti** (angl. *error rate*) je delež enot z vsaj eno nepravilno vrednostjo:

$$e = \frac{a + b}{a + b + c + d} \tag{3}$$

- **Stopnja zaznanih napak** (angl. *hit rate*) je delež enot z dvomljivimi vrednostmi, od katerih je vsaj ena zares nepravilna:

$$h = \frac{a}{a + c} \tag{4}$$

Tako definirani kazalniki nam izkazujejo učinkovitost definiranega sistema kontrol podatkov in hkrati tudi učinkovitost postopkov, ki so vezani na ta sistem kontrol. Stopnja zaznanih napak je tako kazalnik uspešnosti odkrivanja napak, saj izkazuje, kolikšen delež izmed potencialno nepravilnih podatkov, ki jih je določil sistem logičnih kontrol podatkov, je resnično nepravilnih. Če je stopnja zaznanih napak nizka, to kaže predvsem na neučinkovit ali slabo opredeljen nabor logičnih kontrol podatkov. Če je stopnja zavrnitve visoka, stopnja zaznanih napak pa nizka, nam to nakazuje, da imamo sistem kontrol, ki kot nepravilne ali dvomljive označuje veliko število podatkov, ki so v resnici pravilni. V takem primeru je potem potrebno ponovno preveriti in po potrebi prenoviti sistem kontrol.

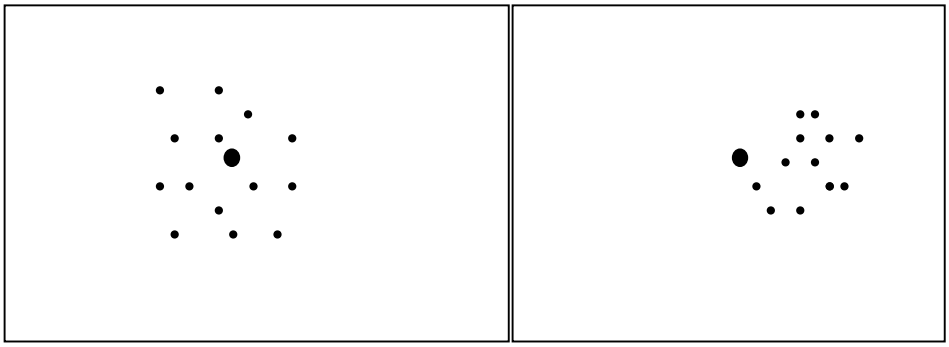
Pri obravnavi napak v podatkih, je zelo pomembna tudi kategorizacija na naslednji dve skupini napak:

- **Slučajne napake.** Napake, ki so posledica slučajnih vplivov v procesu pridobivanja podatkov. Tipičen primer slučajne napake je slučajna napaka pri elektronskem sporočanju podatkov (na primer napaka pri tipkanju). 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 (na primer napačno kodiranje). Možen vzrok za nastanek sistematične napake je tudi uporaba napačne merske enote (na primer evro namesto 1.000 evrov).

Razliko med slučajno in sistematično napako si najlažje predstavljamo, če predpostavimo, da isto meritev (pridobivanje podatka) mnogokrat ponovimo in nato rezultate meritev,

skupaj s pravo vrednostjo, predstavimo v dvodimenzionalnem grafikonu, kot prikazuje Slika 1.

Slika 1: Razlika med slučajno in sistematično napako statistične meritve



Odebeljena točka na sliki predstavlja pravo vrednost, manjše točke pa vrednosti, pridobljene pri ponovljenih meritvah. Kot vidimo, se meritve na levem delu slike − ta ponazarja slučajno napako − dokaj enakomerno »porazdelijo« okrog prave vrednosti. Povprečje teh meritev naj bi se približevalo pravi vrednosti. Nasprotno pa se pri sistematični napaki − ponazarja jo desni del slike − merjene vrednosti »gostijo« samo v enem delu; to pomeni, da se merjene vrednosti v povprečju lahko bistveno razlikujejo od prave vrednosti.

Uvedimo še formalni model statistične napake, s pomočjo katerega bomo zgoraj zapisano, opisno razliko med slučajno in sistematično napako, lahko tudi formalno opredelili. Naj bo Y_i slučajna spremenljivka, katere realizacije $\{y_{1i}, y_{2i}, \dots, y_{ni}\}$ so predmet opazovanja v statističnem raziskovanju. Zapišemo lahko:

$$Y_i = Y_i^0 + \varepsilon_i \tag{5}$$

kjer je $Y_i^0 = \lceil y_{1i}^0, y_{2i}^0, \dots, y_{ni}^0 \rceil^T$ vektor pravilnih vrednosti in $\varepsilon_i = \lceil \varepsilon_{1i}, \varepsilon_{2i}, \dots, \varepsilon_{ni} \rceil^T$ vektor napak. Za celoten nabor opazovanih spremenljivk $Y_1, Y_2, \dots, Y_m$ uporabimo matrični zapis:

$$\mathrm{Y} = \mathrm{Y}^0 + \mathrm{E}, \tag{6}$$

kjer je $\mathrm{Y}^0 = \begin{bmatrix} y_{11}^0, y_{12}^0, \dots, y_{1m}^0 \\ y_{21}^0, y_{21}^0, \dots, y_{2m}^0 \\ \dots \\ y_{n1}^0, y_{n2}^0, \dots, y_{nm}^0 \end{bmatrix}$ matrika pravilnih vrednosti in $\mathrm{E} = \begin{bmatrix} \varepsilon_{11}, \varepsilon_{12}, \dots, \varepsilon_{1m} \\ \varepsilon_{21}, \varepsilon_{22}, \dots, \varepsilon_{2m} \\ \dots \\ \varepsilon_{n1}, \varepsilon_{n2}, \dots, \varepsilon_{nm} \end{bmatrix}$ matrika napak. Za opazovano spremenljivko Y_i je napaka ε_i slučajna, če je njena pričakovana vrednost enaka 0, torej če velja $E(\varepsilon_i) = 0$. Če je pričakovana vrednost različna od 0, imamo opravka s sistematično napako. Za celoten nabor opazovanih podatkov lahko trdimo, da vsebuje zgolj slučajne napake, če velja $E(\mathrm{E}) = \lceil 0, 0, \dots, 0 \rceil^T$.

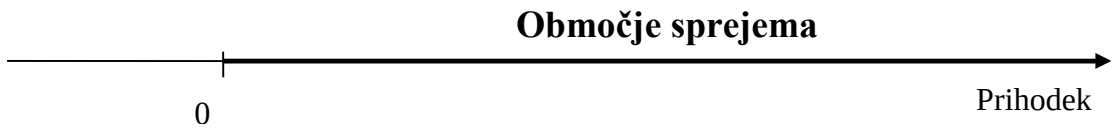
Osnovno orodje, na katerem slonijo vsi postopki urejanja podatkov, je smiseln in dosleden sistem kontrol, s katerim preverjamo smiselnost, ustreznost ter logično pravilnost podatkov. Kontrole lahko razdelimo v tri osnovne skupine:

- **Kontrola smiselnosti.** Pri podatku preverjamo, ali je vrednost v dovoljenem obsegu oziroma naboru. Če je denimo v poslovnem raziskovanju opazovana spremenljivka prihodek podjetja, mora le-ta biti nenegativen.
- **Kontrola doslednosti.** Preverjamo povezanost med več podatki, ki se nanašajo na isto opazovano enoto. Če je denimo v poslovnem raziskovanju Y_1 prihodek od prodaje na domačem trgu, Y_2 prihodek od prodaje na tujem trgu, Y_3 pa prihodek od prodaje, mora veljati $Y_1 + Y_2 = Y_3$.
- **Kontrola porazdelitve.** Pravilnost vrednosti spremenljivke ocenjujemo glede na porazdelitev vrednosti spremenljivk pri drugih enotah. Večinoma gre tu za iskanje izstopajočih vrednosti ali osamelcev.

Tretja skupina kontrol, kontrole porazdelitve, posega na področje kontrole podatkov na makro ravni in jo bomo nekoliko podrobneje obravnavali v razdelku 1.4, tu pa sedaj nekoliko podrobneje poglejmo, kaj nam določajo kontrole na mikro ravni, ki so v veliki večini kontrole iz prvih dveh skupin.

Sistem kontrol, ki jih določimo za namen kontrole podatkov, nam za vsako spremenljivko, ki v teh kontrolah nastopa, določa območje sprejema. Z območjem sprejema je podan obseg vrednosti določene spremenljivke, ki so sprejemljive glede na dane kontrole. Če so vrednosti vseh spremenljivk neke enote v območju sprejema, bo ta enota izpolnjevala vse pogoje, ki so podani z definiranimi kontrolami. Kontrole doslednosti, ki po definiciji vsebujejo le eno spremenljivko, določajo absolutno območje sprejema. V primeru iz definicije kontrole doslednosti kontrolo formalno zapišemo kot: $Prihodek \geq 0$. Grafično lahko v tem primeru območje sprejema prikažemo v enodimenzionalnem prostoru, kot prikazuje Slika 2:

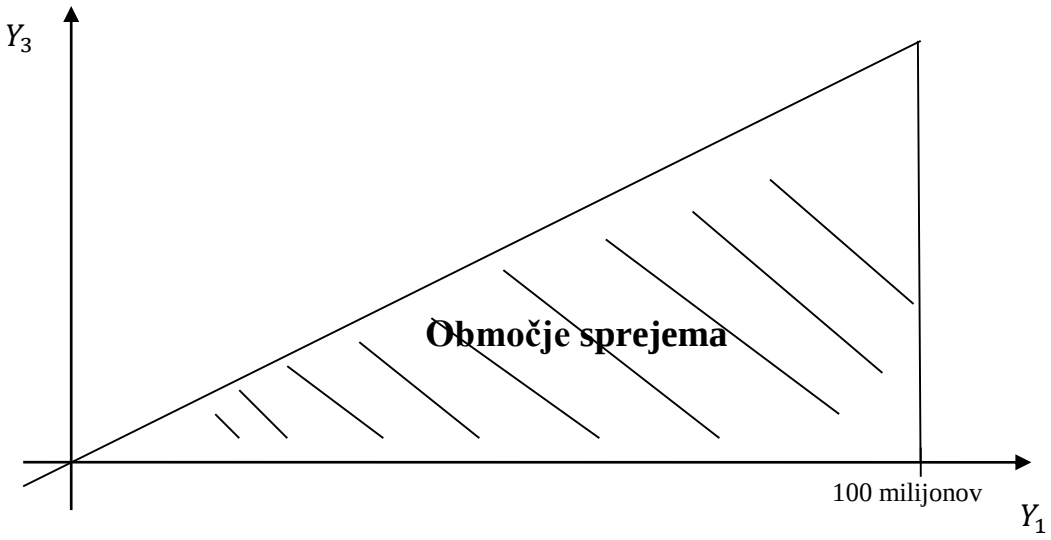
Slika 2: Območje sprejema za spremenljivko Prihodek



Pri kontrolah doslednosti, ki povezujejo več spremenljivk, je pripadajoče območje sprejema določeno pogojno glede na vrednosti drugih spremenljivk. Iz kontrole, ki smo jo navedli v primeru definicije kontrol doslednosti, lahko izpeljemo tudi naslednjo kontrolo: prihodek od prodaje na domačem trgu $\leq$ prihodek od prodaje ($Y_1 \leq Y_3$). Če dodamo še kontrolo smiselnosti $prihodek\ od\ prodaje \geq 0$ $(Y_1 \geq 0)$ ter (za izbrano skupino podjetij) še

dodatno kontrolo smiselnosti $prihodek\ od\ prodaje \leq 100\ milijonov\ evrov$ ($Y_1 \leq 100\ milijonov$), lahko območje sprejema za spremenljivko Y_1 določimo pogojno glede na vrednost spremenljivke Y_3, tako kot prikazuje Slika 3. V tem primeru moramo grafično predstavitev postaviti v dvodimenzionalen prostor.

Slika 3: Območje sprejema za dvodimenzionalno spremenljivko (Y_1, Y_3)



Območje sprejema za spremenljivki Y_1, Y_3 določajo tri kontrole, dve kontroli smiselnosti in ena kontrola doslednosti. Vsaka od teh kontrol določa svoje »podpodročje« sprejema, končno področje sprejema pa je presek vseh podpodročij.

Za postopke urejanja podatkov je pomembna še ena razdelitev napak, in sicer delitev na težke in lahke napake. Težke napake določajo kombinacije podatkov, pri katerih lahko z gotovostjo trdimo, da je vsaj eden od podatkov nepravilen. Če spet uporabimo gornji primer, lahko kot težko napako označimo primere enot, kjer ne velja, da je prihodek od prodaje na domačem trgu manjši ali enak prihodek od prodaje oziroma, da je prihodek od prodaje na domačem trgu večji kot prihodek od prodaje. Lahke napake določata kombinaciji, kjer lahko le z visoko verjetnostjo trdimo, da je vsaj eden od podatkov nepravilen. Če je na primer podjetje z enim zaposlenim poročalo letni prihodek 10 milijonov evrov, gre tu za dvomljivo vrednost pri eni od obeh spremenljivk, ni pa nujno, da gre za napako.

1.2 Selektivno urejanje podatkov

V prejšnji točki smo že omenili pojav v procesu urejanja podatkov, ko imamo ob kontroli podatkov visoko stopnjo zavrnitve in nizko stopnjo zaznanih napak. Taka situacija nam nakazuje, da imamo sistem kontrol, ki kot dvomljive označuje veliko število enot, ki so v resnici pravilni. Soočeni smo s problemom **pretiranega urejanja**. Pojem pretiranega

urejanja označuje prakso, ko izvajalci statističnih raziskovanj v želji čim bolj »prečistiti« zbrane podatke, postavijo preveč kontrol, s preveč strogo postavljenimi kriteriji. Posledica take prakse je obširen seznam enot, katerih vrednosti bi morali ponovno preveriti, kar vodi v zelo visoke stroške urejanja ali pa v neučinkovito urejanje, saj lahko zaradi omejenih sredstev in časovnih omejitev opravimo preverjanje na tej množici enot površno in tako ne odpravimo kakšne zares pomembne napake (Granquist 1995; Granquist & Kovar, 1997).

Eden od načinov za omejevanje vpliva pretiranega urejanja je **selektivno urejanje** (angl. *selective editing*). S selektivnim urejanjem lahko bistveno zmanjšamo količino in posledično stroške urejanja (Adolfsson et al., 2010). Selektivno urejanje označuje strategijo, pri kateri ročno urejanje, tu predvsem mislimo preverjanje podatkov preko ponovnega kontaktiranja, omejimo zgolj na napake, ki imajo nezanemarljiv vpliv na izkazane statistične rezultate (De Waal, Pannekoek, & Scholtus, 2011). Podatke pri enotah, za katere ocenimo, da odpravljanje napak ne bi bistveno vplivalo na rezultat, urejamo s pomočjo avtomatiziranih računalniško podprtih postopkov.

Celoten sistem selektivnega urejanja torej sloni na razdelitvi enot, pri katerih je sistem kontrol zaznal vsaj en dvomljiv podatek, na dva dela. Enote, za katere predpostavljamo, da bo odprava njihovih napak imela bistven vpliv na končne statistične rezultate, sestavljajo prvi del, preostali del enot z dvomljivimi podatki, pri katerih naj napake ne bi bistveno vplivale na rezultat(e), pa drugi del. Za delitev enot lahko uporabimo dva osnovna pristopa (De Waal et al., 2011). Prvi pristop imenujemo **vhodna razdelitev**, pri kateri določen del enot že vnaprej, pred začetkom zbiranja podatkov (običajno že pri samem izboru enot opazovanja), določimo kot prioritetne. Kriterij za razdelitev običajno sloni na podatkih prejšnjih izvedb raziskovanja ali na registrskih podatkih. V fazi urejanja nato prioritetne enote v primeru zaznane napake oziroma dvomljive vrednosti obravnavamo s pristopom ročnega urejanja, torej ponovno kontaktiramo enoto. Pri ostalih enotah napake odpravljamo z računalniško podprtimi postopki avtomatskega urejanja. Prednost takega pristopa je predvsem v že vnaprej poznani razdelitvi, ki omogoča začetek selektivnega urejanja v zelo zgodnji fazi statističnega procesa.

Drugi pristop imenujemo **izhodna razdelitev**. V tem primeru določitev prioritetnih enot poteka po tem, ko so vsi podatki (oziroma vsaj večina njih) že zbrani in lahko že dokaj natančno ocenimo ciljne statistične rezultate in vpliv posameznih enot, njihovih podatkov in zaznanih napak na te rezultate. Ta pristop tako temelji predvsem na analizi zbranih podatkov v trenutni izvedbi raziskovanja. Izhodna razdelitev enot v veliki večini primerov temelji na **funkciji pomembnosti** (angl. *score function*). Funkcija pomembnosti z matematičnim postopkom določa prioriteto enot v postopku urejanja (Latouche & Berthelot, 1992). Osnovna ideja funkcije pomembnosti je opredeliti kriterij, ki nam bo v podatkih zaznane napake razdelil na napake, ki imajo pomemben vpliv na statistične rezultate, in na tiste napake, katerih vpliv na statistične rezultate je zanemarljiv. V končni implementaciji sicer delitev poteka na ravni opazovane enote. Funkcija pomembnosti torej določa kriterije, ki

nam določajo enote, katerih (zaznane) napake imajo lahko pomemben vpliv na končne statistične rezultate.

Različne implementacije funkcije pomembnosti uporabljajo različne kriterije za določitev pomembnosti enot za urejanje, večinoma pa kriteriji funkcije temeljijo na naslednjih štirih osnovnih elementih:

- Velikost enote (glede na eno od ključnih opazovanih spremenljivk, na primer prihodek podjetja). Za določitev velikosti lahko uporabimo tekoče podatke, podatke prejšnjih izvedb ali podatke iz administrativnih virov.
- Število dvomljivih podatkov pri enoti.
- Relativna velikost dvomljivih podatkov pri enoti.
- Relativna pomembnost vsake od obravnavanih spremenljivk.

Da bi lahko uspešno kombinirali zgoraj naštete elemente, vpeljemo še dva koncepta: lokalno pomembnost in globalno pomembnost. Z elementom iz tretje alineje najprej za vsako od obravnavanih spremenljivk določimo lokalno pomembnost, to je pomembnost glede na določeno spremenljivko. V naslednjem koraku z ustreznim upoštevanjem elementov iz prve, druge in četrte alineje za vsako enoto določimo še globalno pomembnost, ki določa pomembnost celotne enote v postopku urejanja. Globalna pomembnost preko eksaktnih numeričnih vrednosti določa, katere enote »pošljemo« v ročno urejanje, z uporabo ponovnega kontakta poročevalskih enot.

Eden od pomembnih faktorjev, na katerem temelji lokalna funkcija pomembnosti, je razlika med poročano in pričakovano vrednostjo opazovane spremenljivke. **Pričakovana vrednost** je povprečna vrednost realiziranih vrednosti slučajne spremenljivke pri velikem številu poskusov. Ta teoretska opredelitev nam v primeru zbiranja podatkov ne pomaga veliko, saj imamo pač opravka z eno samo vrednostjo. Zato zapišimo za naš namen nekoliko poenostavljeno opredelitev: pričakovana vrednost je vrednost spremenljivke pri opazovani enoti, ki bi jo dobili, če bi namesto poročane vrednosti uporabili statistično oceno na podlagi vseh razpoložljivih pomožnih informacij. V periodičnih raziskovanjih lahko na primer za oceno pričakovane vrednosti spremenljivke Y_j pri enoti i uporabimo poročan podatek iz preteklega referenčnega obdobja. Ena od možnih ocen je tako

$$\widehat{y_{ij}(t)} = \bar{t} \cdot y_{ij}(t-1) \tag{7}$$

kjer je $\bar{t}$ povprečen trend izračunan s formulo

$$\bar{t} = \sum_{z \in R} (y_{zj}(t)/y_{zj}(t-1) \tag{8}$$

in je R množica vseh tistih enot, za katero imamo (čist, nedvomljiv) podatek za spremenljivko Y_j za obe referenčni obdobji.

Na zelo splošni ravni lahko lokalno funkcijo pomembnosti za spremenljivko Y_j pri enoti i zapišemo kot produkt dveh osnovnih komponent (De Vaal, 2013): komponente vpliva (v) in komponente tveganja (t):

$$f_l(y_{ij}) = v(y_{ij}) \cdot t(y_{ij}) \tag{9}$$

Funkcija lokalne pomembnosti torej na tej splošni ravni združuje oceno tveganja za napako ter oceno vpliva napake na rezultat. Komponento tveganja lahko na primer ocenimo kot relativno razliko do pričakovane vrednosti:

$$v(y_{ij}) = \frac{|\hat{y}_{ij} - y_{ij}|}{\hat{y}_{ij}} \tag{10}$$

Veliko odstopanje od pričakovane vrednosti nakazuje večje tveganje za nepravilen podatek kot manjše odstopanje. Komponenta vpliva mora določati vpliv posamezne enote v statistični oceni. Če je naša ciljna statistična ocena populacijska vsota, je naravna izbira za komponento vpliva utežena vrednost $w_i \cdot \hat{y}_{ij}$, kjer je w_i[1] utež enote i, ki jo uporabimo pri izračunu statistične ocene. V tem primeru je nato lokalna pomembnost enote določena kot:

$$f_l(y_{ij}) = w_i \cdot |\hat{y}_{ij} - y_{ij}| \tag{11}$$

Tudi pri izračunu globalne funkcije pomembnosti obstaja več različnih načinov, kako globalno pomembnost izpeljati iz lokalnih pomembnosti posameznih vrednosti. Pomembno je, da lokalne pomembnosti najprej pretvorimo na primerljiv obseg vrednosti. V ta namen lahko na primer vsako lokalno pomembnost delimo z ocenjeno populacijsko vsoto spremenljivke (De Val, 2013) ali pa na primer delimo lokalno pomembnost s standardnim odklonom vseh pričakovanih vrednosti (Lawrence & McKenzie, 2000). Tako dobljeno prilagojeno lokalno pomembnost imenujemo standardizirana lokalna vrednost. Globalno pomembnost lahko iz standardiziranih lokalnih vrednosti dobimo tako, da le-te preprosto seštejemo (Latouche & Berthelot, 1992) ali pa namesto vsote vzamemo kako drugo agregatno funkcijo, na primer maksimum (Lawrence & McKenzie, 2000).

[1] Katero utež uporabljamo v dejanski izvedbi je odvisno od tega, v kateri fazi izvajamo postopek selektivnega urejanja. Če postopek izvajamo ob samem zbiranju podatkov, uporabimo vzorčno utež, saj druge uteži v tem času še nimamo na voljo. Če postopke izvajamo v poznejši fazi, uporabimo končno utež, ki vključuje tudi morebitno prilagoditev zaradi neodgovora in/ali kalibracijski faktor. Če podatki niso zbrani na podlagi slučajnega vzorca, je seveda vzorčna utež enaka 1.

Uspešna implementacija ene od verzij funkcije pomembnosti je torej osnoven predpogoj za delitev enot na podlagi pristopa izhodne razdelitve. Prednost pristopa izhodne razdelitve je, da v razdelitvi upošteva podatke tekočega raziskovanja. Predvsem je pomembno, da upošteva predviden vpliv zaznane napake na statistični rezultat. Pomanjkljivost pristopa je v tem, da je postopkovno precej bolj zahteven kot pristop vhodne razdelitve in ga je zato tudi precej težje vključiti v statistični proces. Selektivno urejanje na podlagi funkcije pomembnosti po navadi zahteva precejšnje preoblikovanje celotnega procesa urejanja. Predvsem pa pri uvajanju postopkov selektivnega urejanja zna predstavljati problem sprememba splošnega pogleda na postopek urejanja. Sprejetje stališča, da lahko v podatkih pustimo precejšnje število napak (če te nimajo bistvenega vpliva na rezultat) oziroma jih popravimo zgolj s postopki avtomatskega urejanja, za marsikaterega statistika še vedno predstavlja precejšen zalogaj (Lawrence & McKenzie, 2000).

Kljub nespornim prednostim uvajanja postopkov selektivnega urejanja, kjer je primarni cilj odpraviti napake, ki imajo nezanemarljiv vpliv na statistične rezultate, ima tak pristop tudi nekatere slabosti. Ena od teh je na primer dejstvo, da v primeru, ko se ne ukvarjamo z »manj vplivnimi« napakami oziroma zaradi teh napak ne vzpostavimo ponovnega stika s poročevalsko enoto, le-te ne dobijo povratnih informacij o napakah v procesu poročanja in je zato zmanjšana možnost, da v nadaljnjih poročanjih ne bi ponavljale napak.

1.3 Avtomatsko urejanje podatkov – Fellegi-Holt pristop

Kot smo že zapisali v uvodnem delu naloge, lahko na osnovni ravni postopke urejanja podatkov razdelimo na dve osnovni fazi: na fazo iskanja napak in na fazo odpravljanja napak. Prva faza, iskanje napak, je v veliki meri že računalniško avtomatizirana, druga, odpravljanje napak, pa še vedno poteka pretežno z bolj ali manj »ročnimi« postopki. Statistični uradi še vedno v veliki večini primerov, ko sistem logičnih kontrol zazna dvomljive podatke, uporabljajo preverjanje pravilnosti podatkov s ponovnim kontaktom poročevalske enote (običajno v obliki telefonskega preverjanja). Taka praksa zaradi svoje časovne zamudnosti in potrebnih človeških virov zelo zvišuje stroške raziskovanja. Da bi se ti stroški vsaj delno zmanjšali, so uradi vse bolj začeli uporabljati postopke avtomatskega urejanja podatkov. Pri teh postopkih računalniški programi ne skrbijo le za odkrivanje napak, ampak tudi za njihovo odpravljanje. V sodobnih izvedbah postopkov urejanja je v resnici najbolj pogosta ustrezna kombinacija postopkov ročnega in avtomatskega urejanja (Pannekoek et al., 2013; Norberg, 2009).

Ključni teoretski koncepti, na katerih sloni avtomatsko urejanje podatkov, so bili večinoma uvedeni v drugi polovici prejšnjega stoletja. Še posebej prelomen je bil članek iz leta 1976, v katerem je opisan sistematičen postopek, s katerim lahko nek zapis, ki ne zadošča vsaj enemu od vnaprej opredeljenih pogojev logičnih kontrol, z minimalnim številom popravkov, spremenimo v sprejemljiv zapis, torej zapis, ki zadošča vsem pogojem logičnih kontrol (Fellegi & Holt, 1976). Pristop, opisan v tem članku, ki ga bomo v nadaljevanju označevali

kot »Fellegi-Holt pristop« oziroma »Fellegi-Holt metoda« je kasneje postal osnova večine postopkov avtomatskega urejanja. Sama metoda je zasnovana na naslednjih treh osnovnih načelih:

- Popravki naj se izvedejo tako, da po opravljenih popravkih podatki zadoščajo pogojem vseh logičnih kontrol, hkrati pa je spremenjeno čim manjše število podatkov (spremenljivk) pri posamezni enoti. Cilj je imeti na koncu vse podatke sprejemljive, hkrati pa ohraniti čim večjo količino izvornih podatkov.
- Nove, popravljene (vstavljene) vrednosti naj izhajajo direktno iz pravil logičnih kontrol. S tem zagotovimo, da z na novo ocenjenimi podatki ne ustvarimo novih dvomljivih zapisov, torej zapisov, ki ne bi zadoščali vsem pravilom logičnih kontrol.
- Porazdelitev vrednosti spremenljivk po opravljenih popravkih naj v čim večji možni meri ohranja prvotna porazdelitev tistega dela podatkov, ki je brez zaznanih napak. To naj velja tako za enorazsežne porazdelitve posameznih spremenljivk, kot tudi za večrazsežne porazdelitve več spremenljivk.

Ker je eden od osnovnih principov metode pravilo, da vedno izvedemo minimalno število popravkov, potrebnih za to, da dobimo spet sprejemljiv zapis, pristop imenujemo tudi »pristop minimalne spremembe«. Pristop minimalne spremembe je skozi leta doživel mnoge prilagoditve in izvedbe, primerne za različne vrste podatkov, v prvotnem članku pa je bil postopek zastavljen kot metoda za urejanje popisnih podatkov, zato so bili tudi praktični primeri večinoma podani za kategorične spremenljivke. Je pa postopek zastavljen na tako abstraktni ravni, da ga je mogoče brez problema uporabiti tudi za primer številskih spremenljivk (Seljak & Špeh, 2004). V nadaljevanju bomo torej opisali osnovne korake v izvajanju metode, s tem da bomo izpustili zahtevnejše tehnično-matematične dele, kot so izreki in dokazi. Kljub temu pa moramo najprej vpeljati formalni-matematični zapis, saj celoten sistem temelji prav na zapisu spremenljivk in sistemu kontrol teh spremenljivk v eksaktni matematični obliki.

Predpostavimo, da so predmet opazovanja spremenljivke $Y_1, Y_2, \dots, Y_m$, katerih celoten obseg mogočih vrednosti predstavljajo množice $A_1, A_2, \dots, A_m$. Urejanje podatkov na mikro ravni temelji na sistemu kontrol, s katerimi preverjamo pravilnost posameznih zapisov. Te kontrole seveda lahko zapišemo na različne načine; pri tem lahko uporabljamo pogoje, povezane z različnimi logičnimi operatorji (in, ali, negacija, implikacija, ...). Izkaže se, da lahko z uporabo nekaj preprostih pravil, kontrole spremenimo v obliko, v kateri nastopa samo logični operator in ($\cap$). Vse podane kontrole torej lahko zapišemo kot množico kontrol v naslednji obliki:

$$(Y_1 \in A_1') \cap (Y_2 \in A_2') \dots \cap (Y_m \in A_m') = \text{Napaka}^2 \qquad (12)$$

[2] »Izid« logične kontrole bi v formalno-matematični obliki zapisali kot binarno spremenljivko, z dvema vrednostima: 1 – pravilna vrednost, 0 – napaka. Zaradi večje jasnosti v našem zapisu ohranjamo opisno obliko.

Tako obliko kontrole imenujemo normalna oblika kontrole (angl. *normal form of edit*). Množice $A'_1, A'_2, \dots, A'_2$ so podmnožice množic $A_1, A_2, \dots, A_m$ in v takem zapisu kontrole za posamezne spremenljivke določajo območje »nesprejemljivosti«. Ni težko videti, da območje sprejemljivosti za spremenljivko Y_i določa podmnožica $A_1 - A'_1$. Za spremenljivko Y_i, za katero je pripadajoča A'_i prava podmnožica množice A_i, pravimo, da eksplicitno nastopa v kontroli. Drugače povedano, tako območje nesprejemljivosti kot območje sprejemljivosti spremenljivke Y_i, ki ga določa obravnavana kontrola, sta pravi podmnožici celotnega nabora podatkov spremenljivke. Na splošno lahko v posamezni kontroli eksplicitno nastopajo vse spremenljivke, običajno pa eksplicitno nastopa le manjše število spremenljivk. Tako v enostavni kontroli smiselnosti $Y_1 \leq 0$, ki določa, da vrednost spremenljivke Y_1 ne sme biti negativna, eksplicitno nastopa samo ena spremenljivka. To kontrolo v normalni obliki zapišemo takole:

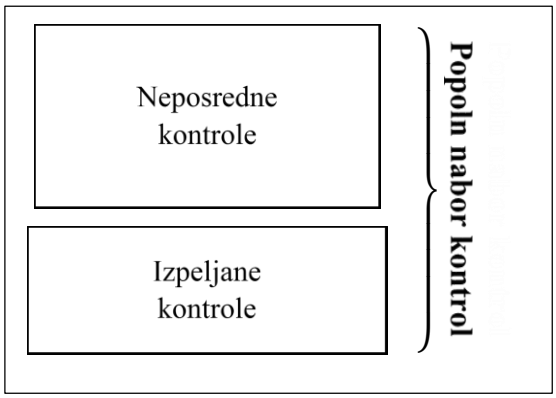
$$\big(Y_1 \in (-\infty, 0)\big) \cap (Y_2 \in A_2) \cap \dots \cap (Y_m \in A_\mathrm{m}) = \text{Napaka} \qquad (13)$$

Ker spremenljivke $Y_2, \dots, Y_m$ ne nastopajo eksplicitno v kontroli, njihove vrednosti na »izid« kontrole ne vplivajo in njihove vrednosti so lahko poljubne v celotnem naboru vrednosti.

Recimo, da je sistem opredeljenih kontrol tak, da nam omogoča zaznati vse nepravilne zapise. Kljub temu potrebujemo za izpeljavo nadaljnjih korakov v postopku še dodatne kontrole, imenovane izpeljane kontrole (angl. *implicit edits*). Gre za kontrole, ki jih lahko s posebnimi matematičnimi postopki izpeljemo iz začetnih, torej neposrednih kontrol (angl. *explicit edits*). Če imamo na primer kontroli $Y_1 < 100$ in $Y_1 > Y_2$, je izpeljana kontrola $Y_2 < 100$. Ta kontrola sicer ni potrebna za zaznavanje nepravilnih podatkov za Y_2, saj jih zaznamo že s prvima dvema kontrolama, vendar pa je potrebna za nadaljnje postopke v Fellegi-Holt metodi.

Če so neposredne kontrole zapisane v normalni obliki, je mogoče postopek pridobivanja izpeljanih kontrol prevesti tudi v avtomatiziran računalniški jezik; to pomeni, da je mogoče te kontrole izpeljati preko nekega avtomatiziranega postopka. Po izvedbi takega postopka imamo torej na voljo v normalni obliki tako nabor neposrednih kontrol kot tudi nabor izpeljanih kontrol. Tak seznam kontrol, prikazan na Slika 4, imenujemo **poln nabor kontrol** (angl. *complete set of edits*).

Slika 4: Poln nabor kontrol



Ko je določen poln nabor kontrol in ko je pri vsaki kontroli točno določeno, katere spremenljivke v njej eksplicitno nastopajo, vse zapise »podvržemo« kontrolam tega polnega sistema. Seznam vseh obravnavanih enot nam zdaj razpade na dva dela, na enote, ki so uspešno prestale vse kontrole (te enote v nadaljevanju imenujemo »čiste enote«), in na enote, ki najmanj ene kontrole niso prestale uspešno (te enote v nadaljevanju imenujemo »dvomljive enote«). Imamo torej seznam spremenljivk, za katere vemo, da ne ustrezajo vsaj enemu od pogojev kontrol, ne vemo pa, katere izmed spremenljivk so povzročile, da je zapis označen kot dvomljiv. Sledi korak v procesu, ki ga imenujemo tudi **lokalizacija napake** (angl. *error localization*). V tem koraku določimo − na podlagi védenja o tem, katerih kontrol enota ni prestala in katere spremenljivke v teh kontrolah eksplicitno nastopajo − minimalen seznam spremenljivk, ki jih moramo popraviti, da bo potem zapis prestal vse kontrole. Seveda ni zagotovila, da postopek določi res tiste spremenljivke, ki so povzročile napako. Lahko le zatrdimo, da je največja verjetnost, da je izmed vseh možnih naborov spremenljivk, povzročil napako izbrani nabor spremenljivk. To je pa tudi največ, kar lahko od takega sistema glede na dane informacije pričakujemo. Ta del postopka je, kar zadeva avtomatizacijo, še najpreprostejši, saj gre za enostavno iteracijo naslednjih korakov:

- Za vsako spremenljivko v zapisu iz nabora enot, ki so javile napako, preštej v koliko kontrolah, ki so javile napako, eksplicitno nastopa.
- Določi spremenljivko v zapisu, ki eksplicitno nastopa v največ kontrolah, ki so javile napako. Če je spremenljivk z enakim številom eksplicitnih nastopov v kontrolah več, določi eno izmed njih naključno. Ta spremenljivka je določena kot podatek, ki ga moramo popraviti.
- Iz nabora kontrol, ki so za obravnavano enoto javile napako, izloči vse tiste, v katerih eksplicitno nastopa spremenljivka, ki smo jo v prejšnjem koraku določili za popravek.
- Ponovi postopek na zmanjšanem naboru kontrol. Postopek ponavlja toliko časa, dokler ni nabor kontrol, ki so javile napako, prazen.

Ko smo določili, vrednosti katerih spremenljivk v zapisu z napakami moramo popraviti, moramo določiti še postopek za določitev novih vrednosti. Običajno uporabimo eno izmed poznanih metod za vstavljanje podatkov, pri tem pa moramo cel postopek postaviti tako, da bodo vse vstavljene vrednosti v območju sprejemljivosti, torej območju, ki bo zagotavljalo, da bo popravljeni zapis »prestal« vse logične kontrole. Teoretično je tako območje enostavno določljivo. Vrednost spremenljivke Y_i mora biti iz preseka $\mathrm{A}'_{i1} \cap \mathrm{A}'_{i2} \cap \ldots \cap \mathrm{A}'_{im}$, kjer so $\mathrm{A}'_{i1}, \mathrm{A}'_{i2}, \ldots, \mathrm{A}'_{im}$ podmnožice množice A_i, ki omejujejo nabor vrednosti za Y_i v polnem sistemu kontrol.

Kljub konceptualni enostavnosti opisanega postopka je v praksi postavitev splošnega sistema za določanje območja sprejema iz danega polnega sistema kontrol eden zahtevnejših delov celotnega postopka. To še posebej velja za številske spremenljivke. V tem primeru so logične kontrole večinoma podane v obliki linearnih enačb in neenačb in določitev območja sprejema je enakovredno reševanju sistema enačb in neenačb. Za splošno reševanje takih sistemov že obstajajo programske rešitve, ki večinoma uporabljajo tehnike linearnega programiranja, in ti postopki so vgrajeni tudi v nekatere statistične aplikacije, kot je denimo Banff (Kozak, 2005).

Če hočemo slediti načelu ohranitve robne in večrazsežne porazdelitve spremenljivk (ki jo določajo zapisi brez napak), je naravna izbira za postopek vstavljanja novih vrednosti **metoda notranjega darovalca** oziroma »hot-deck metoda« (Coutinho et al., 2013). Pri tej metodi v določeni podskupini izmed čistih enot najprej določimo tiste, katerih vrednosti so v pravem območju, izmed njih pa s slučajnim izborom ali s postopkom najbližjega soseda določimo darovano vrednost. Kadar moramo pri eni enoti popraviti vrednosti več spremenljivk, lahko uporabimo dva pristopa. Prvi pristop poteka tako, da vstavljamo spremenljivke zaporedoma, eno za drugo. S takim pristopom zagotavljamo le ohranjanje enodimenzionalnih robnih porazdelitev. Drug pristop poteka tako, da vstavljamo vse spremenljivke hkrati in ob tem upoštevamo povezave med njimi. Prvi pristop je gotovo enostavnejši, prednost drugega pa je, da nam zagotavlja tudi ohranjanje večrazsežnih porazdelitev.

Opisani postopek temelji na predpostavki, da so vsa pravila logičnih kontrol enakovredna. Kot smo navedli v razdelku 1.1, zelo pogosto kontrole delimo na težke kontrole in lahke kontrole. Težke kontrole so tako kontrole, ki zagotovo določajo napako v podatkih, lahke kontrole pa določajo samo dvomljive podatke, ni pa nujno, da gre res za napako. Lahko, da kljub temu, da kontrola javi napako, v zapisu nastopajo povsem veljavni podatki. Primeri takih kontrol so vse kontrole, ki omejujejo rast prihodka glede na preteklo obdobje. Lahko, recimo, postavimo pravilo, da je dvomljiv podatek, če se je prihodek glede na preteklo obdobje povišal za več kot 20 %: $Prihodek(t)/Prihodek(t-1) > 1{,}2$. V zgoraj opisanem pristopu vse kontrole obravnavamo kot težke kontrole. Postopek pa lahko prilagodimo tudi tako, da bo ustrezno upoštevana delitev kontrol na težke in lahke kontrole (Scholtus, 2013).

Čeprav je neizpodbitno dejstvo, da lahko z metodami avtomatskega urejanja bistveno racionaliziramo postopke urejanja podatkov in posledično zmanjšamo stroške raziskovanja, obstajajo tudi nekatere slabosti oziroma omejitve, ki se jih je pri postavitvi celotnega procesa urejanja potrebno zavedati. Te slabosti so predvsem:

- Podatki, popravljeni s postopki avtomatskega urejanja, niso pravilni, ampak zgolj sprejemljivi.
- Obstaja sindrom »črne škatle«. Postopki, ki poskrbijo za avtomatske popravke, so vsebinskim statistikom večinoma skriti in neznani. Zato je zelo pomembno, da postopki poleg samih izhodnih podatkov zagotavljajo tudi dodatne parapodatke, ki uporabnikom omogočajo vsaj delen vpogled v izvedene postopke. Taki parapodatki so: podrobna evidenca popravljenih podatkov (delež in število popravljenih podatkov za vsako spremenljivko), evidenca uporabljenih metod vstavljanja, vpliv postopkov avtomatskega urejanja na ključne statistične rezultate (Eurostat, 2007).
- V postopke avtomatskega urejanja vstopajo zgolj vhodni podatki ter sistem logičnih kontrol. Na podlagi teh vhodnih podatkov se izvedejo popravki. Sistem logičnih kontrol torej predstavlja ključen faktor pri izvedbi popravkov, zato je še posebej pomembno, da je ta sistem konsistenten in popoln.

1.4 Urejanje na makro ravni in problem osamelih vrednosti

Urejanje podatkov je kompleksen proces, ki lahko združuje več različnih postopkov, ki lahko potekajo v različnih korakih statističnega procesa. Na splošni ravni je koristno razlikovati predvsem urejanje v dveh fazah statističnega procesa. Urejanje v zgodnji fazi izvajanja raziskovanja, predvsem ob zbiranju ali vnosu podatkov (vhodno urejanje) ter urejanje v poznejši fazi izvajanja, predvsem pri izvajanju statistične obdelave in pripravi statističnih rezultatov (izhodno urejanje). Osnovna razlikovalna značilnost, ki določa specifiko postopkov vhodnega urejanja je ta, da imamo v tem primeru naenkrat dostop le do podatkov ene enote, za katero presojamo konsistentnost oziroma sprejemljivost podatkov. V tem primeru govorimo o urejanju na mikro ravni. Nasprotno pa imamo v primeru urejanja v poznejših fazah izvajanja raziskovanja na voljo že podatke vseh enot, zato lahko izvajamo tudi urejanje na makro ravni. Urejanje na makro ravni označuje vse postopke, s katerimi dvomljive podatke iščemo s pomočjo analize že agregiranih podatkov, uteženih vrednosti in statističnih porazdelitev.

Analiza že agregiranih podatkov se nanaša predvsem na postopke, pri katerih kontrole ustreznosti in smiselnosti namesto na vhodnih mikro podatkih izvajamo na izhodnih, statističnih agregatih, ki so končni rezultat našega raziskovanja (makro raven). Tudi v tem primeru obstaja več različnih pristopov preverjanja. Eden od pristopov je metoda postopne kontrole agregatov, pri kateri kombiniramo kontrolo na mikro in na makro ravni. Osnovna zamisel pristopa je v tem, da najprej preverjamo agregirane podatke za določeno domeno (na primer na ravni 4-mestne šifre Standardne klasifikacije dejavnosti), nato pa izvedemo

kontrolo na mikro ravni za vse enote iz skupin, ki so bile na makro ravni zaznane kot dvomljive. V tem primeru torej urejanje na makro ravni nastopa v procesu pred urejanjem na mikro ravni. Tak pristop je učinkovit predvsem v primeru periodičnih raziskovanj, kjer lahko preverjanje na makro ravni izvajamo na podlagi porazdelitev razmerij in razlik glede na rezultate predhodnega obdobja. V praktični izvedbi tega pristopa običajno razvrstimo skupine glede na razmerje ali razliko s preteklim obdobjem ter označimo skupine z najmanjšo in skupine z največjo vrednostjo. Drugače povedano: označimo in posledično izberemo za kontrolo na mikro ravni le enote iz repov porazdelitev, ali razmerij ali razlik.

Pri analizi porazdelitev podatkov je cilj postopkov predvsem zaznavanje osamelih vrednosti oziroma osamelcev. **Osamelec** (angl. *outlier*) je statistični podatek, katerega vrednost se glede na opredeljena merila bistveno razlikuje od drugih vrednosti. Taka definicija je sicer še zelo ohlapna, saj predvsem pri »opredeljenih merilih« dopušča zelo različne postopke določitve osamelcev. V resnici taka široka definicija izhaja iz narave osamelcev in iz dejstva, da je kriterije za določitev osamelca mogoče natančneje opredeliti šele v konkretnem statističnem raziskovanju in na podlagi realnih podatkov, s katerimi imamo opravka.

Osamelci imajo lahko zelo pomemben vpliv na statistične analize in končne statistične rezultate. Osborne in Overbay (2004) izpostavita predvsem naslednje možne negativne vplive v postopkih statistične analize:

- Osamelci zvišujejo variabilnost analiziranih podatkov in posledično zmanjšujejo moč statističnega sklepanja.
- Če osamele vrednosti niso porazdeljene slučajno, lahko izkrivijo sicer normalno porazdelitev podatkov.
- Osamelci imajo lahko močan vpliv na statistične ocene in lahko povzročijo tudi nezanemarljivo pristranskost teh ocen.

Osamelce lahko delimo v dve osnovni obliki (Barnett & Lewis, 1994):

- Osamelci, ki izhajajo iz napak v podatkih. Gre torej za dejansko nepravilen podatek.
- Osamelci, ki izhajajo iz variabilnosti podatkov. Gre torej za podatek, ki sicer odstopa od ostale porazdelitve, vendar je v resnici pravilen.

V postopkih statističnega urejanja podatkov je seveda predvsem pomembno, da zaznamo (in tudi ustrezno popravimo) osamelce iz prve skupine. Če osamelce obravnavamo v primeru podatkov, ki so izbrani na podlagi vzorčnega raziskovanja, pa je pomembna še ena presoja. In sicer presoja tega, ali je osamelec reprezentativen ali ne. Osamelec v vzorčnih podatkih je reprezentativen, če ima pravilno vrednost in (če vsaj približno) velja predpostavka, da v populaciji obstaja tako število s tako izstopajočimi vrednostmi, kot je vzorčna utež opazovane enote (Chambers, 1986). Če ocenimo, da je osamelec prava vrednost, vendar ne

reprezentativna (vrednost je tudi v populaciji osamela), taki enoti utež popravimo na vrednost 1. Pravimo, da smo enoto spremenili v samoreprezentativno enoto.

Obstaja veliko različnih pristopov in metod za določanje osamelcev (Tukey, 1977; Hidiroglou & Berthelot, 1986; Barnett & Lewis, 1994; De Waal et al., 2011), ki jih uporabljamo glede na specifiko podatkov statističnega raziskovanja. Večina teh metod temelji na podrobni analizi porazdelitve podatkov, predvsem srednjih vrednosti in mer variabilnosti[3]. V analizi je pomembna tudi opredelitev razdalje oziroma metrične funkcije, saj le-ta bistveno določa, katere vrednosti so od srednjih mer porazdelitve oddaljene toliko, da jih lahko označimo kot osamele vrednosti.

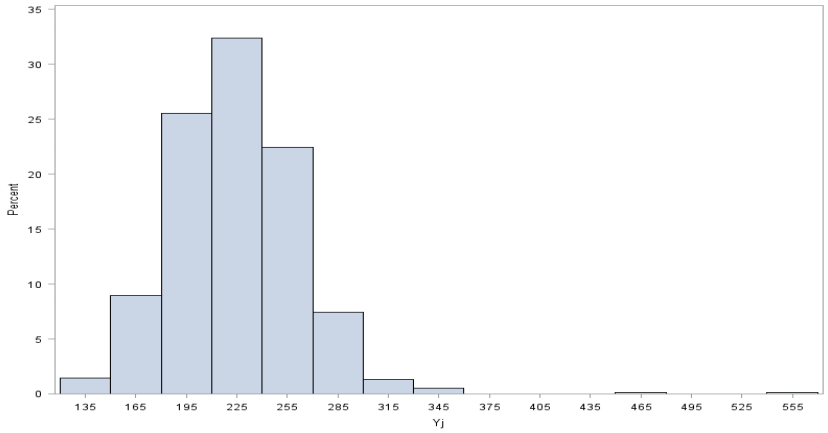
Osrednja tema naše naloge so periodična poslovna raziskovanja, zato v nadaljevanju nekoliko podrobneje poglejmo problem osamelcev v takih raziskovanjih. Najprej pa uvedimo nekoliko bolj formalen opis takih podatkov. Predpostavimo torej, da imamo raziskovanje, v katerem v rednih časovnih intervalih (mesecih, četrtletjih), v istih (ali pa vsaj delno istih) opazovanih enotah merimo vrednost številskih spremenljivk (na primer prihodek od prodaje, število zaposlenih). Podatke iz več časovnih obdobij bomo obravnavali kot en podatkovni set, kar pomeni, da v formalni opis podatkov uvedemo še dodatno, časovno dimenzijo. Če predpostavimo, da opazujemo T časovnih obdobij $t_1, t_2, \dots, t_T$, lahko z oznakami, ki smo jih uvedli v razdelku 1.1, podatke periodičnega raziskovanja zapišemo kot:

$$\mathbf{Y}(\mathrm{t}) = \begin{bmatrix} y_{11}(t), y_{12}(t), \dots, y_{1m}(t) \\ y_{21}(t), y_{22}(t), \dots, y_{2m}(t) \\ \dots \\ y_{n1}(t), y_{n2}(t), \dots, y_{nm}(t) \end{bmatrix}_{\{t=t_1, t_2, \dots, t_T\}} \tag{14}$$

Uvedeni model si lahko predstavljamo kot tridimenzionalni prostor, kjer prvo dimenzijo določajo opazovane enote, drugo dimenzijo opazovane spremenljivke in tretjo dimenzijo časovna obdobja. Ker se za zdaj omejujemo zgolj na univariatno analizo, bosta za določanje osamelcev v takih podatkih zanimivi predvsem dve enodimenzionalni robni porazdelitvi. Prva je porazdelitev podatkov izbrane spremenljivke Y_i v izbranem časovnem obdobju t_0, torej porazdelitev vrednosti $\{y_{ij}(t_0)\}_{i=1}^{n}$. To porazdelitev imenujemo tudi presečna porazdelitev, podatke, ki jo določajo, pa presečni podatki. Obstoj osamelcev v tej porazdelitvi lahko običajno razberemo že iz ustrezne grafične predstavitve podatkov, denimo iz histograma. Na Sliki 5 je kot primer predstavljena porazdelitev podatkov 1.000 enot[4], s »približno normalno« porazdelitvijo, ki pa vsebujejo dve precej izstopajoči osameli vrednosti.

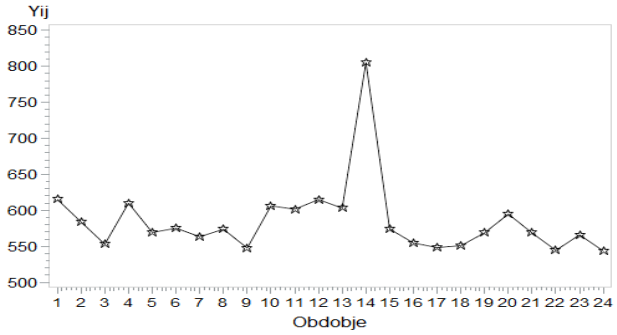
[3] Najbolj znan primer, ki je z leti postal nekakšno pravilo palca, je meja treh standardnih odklonov od srednje vrednosti (na primer aritmetične sredine).
[4] Podatki so umetni, tvorjeni zgolj za ponazoritev prej opisanega primera osamelcev.

Slika 5: Histogram porazdelitve presečnih podatkov



Druga je porazdelitev podatkov izbrane spremenljivke Y_i, pri izbrani enoti j, v različnih časovnih obdobjih, torej porazdelitev vrednosti $\{y_{ij}(t_z)\}_{z=1}^{T}$. Tako porazdelitev imenujemo tudi **longitudinalna porazdelitev**, podatke, ki jo določajo, pa longitudinalni podatki. Za prikaz longitudinalnih podatkov in za zaznavanje obstoja osamelcev v njih, je najprimernejši linijski diagram. Na Sliki 6 je kot primer predstavljena porazdelitev podatkov za spremenljivko Y_i pri enoti j za 24 zaporednih časovnih obdobij. Na sliki je jasno razvidna izstopajoča vrednost v 14. časovnem obdobju[5].

Slika 6: Linijski diagram longitudinalnih podatkov



V nadaljevanju bomo podrobneje predstavili metodo, ki kombinira zgoraj opisani porazdelitvi v novo enodimenzionalno porazdelitev, z analizo katere iščemo osamele vrednosti. Metoda je bila prvič opisana in predstavljena v Hidirogloujevem in Berthelotevem članku leta 1986 in jo bomo zato imenovali »Hidiroglou-Berthelot metoda« (v nadaljevanju H-B metoda).

Čeprav z metodo lahko preverjamo vrednosti več spremenljivk, postopek izvajamo za vsako spremenljivko posebej, zato v nadaljevanju obravnavamo le eno spremenljivko Y_i. Ker bo

[5] Tudi v tem primeru gre za popolnoma umetne podatke, tvorjene za ilustracijo primera.

obravnavana le ena spremenljivka, bomo za namen predstavitve metode opustili indeks spremenljivke, tako da bomo obravnavali spremenljivko Y. Sledeč zgoraj zapisanemu opisnemu modelu, zapišimo $Y(t) = (y_1(t), y_2(t), \dots, y_n(t))$ vrednosti opazovane spremenljivke pri n opazovanih enotah.

Za namen nadaljnje predstavitve bomo sedaj nekoliko obrnili zapis in pri izbrani enoti j zapisali vektor opazovanih vrednostih v časovnih obdobjih $t_1, \dots, t_{c-1}, t_c$; $(Y_j(t_1), \dots, Y_j(t_{c-1}), Y_j(t_c))$. Obdobje t_c bomo imenovali tekoče obdobje, t_{c-1} pa predhodno obdobje. Ker preverjamo poročane vrednosti v tekočem obdobju t_c, nas zanimajo predvsem razmerja vrednosti $Y_j(t_c)$ glede na vrednosti predhodnega obdobja, torej razmerja $r_j = Y_j(t_c)/Y_j(t_{c-1})$. Ničelne vrednosti spremenljivke Y tako v tekočem kot v predhodnem obdobju iz analize izločimo in so obravnavane posebej. Predmet obravnave bo torej porazdelitev vrednosti $\{r_j\}$ pri neki podskupini celotne opazovane populacije s k elementi, torej porazdelitev za $\{r_1, r_2, \dots, r_k\}$.

Naj bo $R = \sum_{j=1}^{k} Y_j(t_c) / \sum_{j=1}^{k} Y_j(t_{c-1})$ razmerje za celotno skupino. Zapišemo lahko $R = \sum_{j=1}^{k} I_j r_j$, kjer je r_j razmerje pri posamezni enoti, $I_j = Y_j(t_{c-1c}) / \sum_{i=1}^{k} Y_i(t_{c-1})$ pa mera pomembnosti te enote. Vprašanje, ki ga razrešuje metoda, je, kako določiti enote, katerih vrednosti $\{r_j\}$ bistveno izstopajo iz porazdelitve in hkrati pomembno vplivajo na vrednost R. Najenostavnejša možnost bi bila proučevati kar porazdelitev $\{r_j\}$, vendar je zaradi razlogov, ki jih bomo pojasnili pozneje, ustrezneje te vrednosti najprej dvakrat transformirati.

Vrednosti $\{r_j\}$ se po definiciji nahajajo na intervalu $(0, \infty)$. Ker je običajno mediana take porazdelitve blizu 1, bo imela porazdelitev »rep« le na desni strani. Posledično bi pri obravnavi take porazdelitve zaznali ekstremne vrednosti le na desni strani, torej bi bile za nas problematične le enote, katerih vrednosti so se glede na predhodno obdobje močno povečale, ne pa tiste, katerih vrednosti so se močno zmanjšale. Zatorej najprej opravimo naslednjo transformacijo:

$$s_j = \begin{cases} 1 - r_M / r_j \,; \text{če } r_j < r_M \\ r_j / r_M - 1 \,; \text{če } r_j \geq r_M \end{cases} \tag{15}$$

kjer je r_M mediana vrednosti $\{r_j\}$. Ni težko videti, da je pol vrednosti $\{s_j\}$ manjših, pol pa večjih od 0, torej smo s transformacijo dobili simetrično porazdelitev okrog 0.
Ker so enako velike relativne spremembe r_j za vrednost skupnega razmerja R pomembnejše pri večjih vrednostih Y_j kot pri manjših, opravimo še naslednjo transformacijo:

$$E_j = s_j \{Max(Y_j(t_c), Y_j(t_{c-1}))\}^U, \tag{16}$$

kjer je U iz intervala $[0,1]$ in predstavlja vpliv velikosti absolutne vrednosti v transformaciji. Želimo namreč, da imajo enote z večjimi absolutnimi vrednostmi ($Y_j(t)$) tudi večjo pomembnost v analizi porazdelitve razmerij. Čim večji je U, tem večji je vpliv absolutne vrednosti v porazdelitvi $\{E_j\}$. Če je $U = 0$, je $\{E_j\}$=$\{s_j\}$ in ostanemo pri relativnih spremembah.

Osamelce določimo kot ekstremne vrednosti v porazdelitvi vrednosti $\{E_j\}$. V ta namen najprej določimo odklona:

$$d_{Q1} = Max(E_M - E_{Q1}, |A \cdot E_M|) \tag{17}$$

$$d_{Q3} = Max(E_{Q3} - E_M, |A \cdot E_M|), \tag{18}$$

kjer so E_{Q1}, E_M in E_{Q3} prvi kvartil, mediana in tretji kvartil v porazdelitvi $\{E_j\}$. A je izbrano število, s katerim preprečimo »čudne« rezultate takrat, kadar je vrednost $E_M - E_{Q1}$ ali $E_{Q3} - E_M$ zelo majhna, torej takrat, kadar so vrednosti $\{E_i\}$ močno zgoščene okoli mediane. Vrednost parametra A je običajno zelo majhna, na primer 0,05.

Kot ekstremne vrednosti (osamelce) označimo tiste zapise, pri katerih je vrednost E_j zunaj intervala $[E_M - C \cdot d_{Q1}, E_M + C \cdot d_{Q3}]$, kjer je C pozitivno celo število. Parameter C določa območje sprejema in s tem tudi »strogost« postopka za določitev osamelcev. Večja ko je vrednost za C, širše je območje sprejema in posledično manj osamelcev v porazdelitvi. Najustreznejšo vrednost za parameter C, kakor tudi za ostala parametra U in A, določimo za primer vsakega raziskovanja posebej. Pri tem je treba dobro analizirati morebitne že obstoječe zgodovinske podatke raziskovanja, če ti še ne obstajajo pa poiskati nek vir s sorodnimi podatki. Koristno je tudi pregledati prakse drugih izvajalcev, ki so to metodo že uporabili v sorodnih raziskovanjih (Mulry & Feldpausch, 2007; Hunt, Johnson, & King, 1999).

2 OCENA KAKOVOSTI STATISTIČNIH PRODUKTOV IN PROCESOV

2.1 Standardi za oceno kakovosti v okviru ESS

Spremljanje in ocenjevanje kakovosti statističnih rezultatov je pomemben del statističnega procesa, ki je v praksi statističnih uradov ter drugih organizacij, ki izračunavajo ter objavljajo statistične rezultate, prisoten že dolga desetletja. Tisto, kar na tem področju v zadnjem obdobju predstavlja novost, so nekoliko drugačni pristopi, predvsem pa:

- Širši, večdimenzionalni pogled na oceno kakovosti. Če se je še v drugi polovici prejšnjega stoletja kakovost statističnih rezultatov večinoma presojala skozi prizmo točnosti, je v zadnjem obdobju prevladal bolj celosten pristop, kjer se presoja več vidikov kakovosti (Biemer & Lyberg, 2003). Točnost s tem postane samo ena od dimenzij.
- Poenotenje oziroma harmonizacija različnih modelov presojanja kakovosti različnih statističnih organizacij. V postopku razvoja večdimenzionalnega pristopa k ocenjevanju kakovosti v statistiki je več različnih organizacij razvilo več različnih modelov z nekoliko drugače opredeljenimi dimenzijami (Brackstone, 1999; Carson, 2001; Eurostat 2003). V zadnjih letih je vse več aktivnosti v smeri poenotenja oziroma vsaj približevanja teh različnih modelov.
- Večji poudarek na (javnem) izkazovanju ocenjene kakovosti. Izvajalci uradne statistike vse bolj uveljavljajo prakso, da vsaj del informacij, ki izhajajo iz postopkov ocenjevanja kakovosti, dajejo na vpogled svojim uporabnikom in s tem povečujejo transparentnost in preglednost svojega dela ter objavljenih statističnih rezultatov.

Eurostat je pričel v sodelovanju z nacionalnimi statističnimi uradi svoj koncept spremljanja kakovosti razvijati sredi devetdesetih let prejšnjega stoletja. Pomemben mejnik v tem razvoju predstavlja leto 1998, ko je bila ustanovljena posebna delovna skupina, ki je v naslednjih letih pripravila nabor pomembnih metodoloških dokumentov, ki so postali temeljni kamen za harmonizirano spremljanje in poročanje o kakovosti statističnih rezultatov. Na podlagi teh dokumentov je bil opredeljen osnovni model ocenjevanja kakovosti v ESS. V začetni opredelitvi je model slonel na sedmih dimenzijah kakovosti, vendar je že kmalu prišlo do preoblikovanja v model s šestimi dimenzijami[6], ki je bil nato vrsto let splošno sprejet model ocenjevanja kakovosti. V nadaljevanju podajamo seznam ter kratek opis teh šestih dimenzij (Eurostat, 2009):

- **Ustreznost**. Dimenzija, s katero ugotavljamo, ali statistični rezultati zadovoljujejo potrebe oziroma zahteve uporabnikov. Presojamo predvsem naslednja dva vidika:

  – Ali so izkazani rezultati ustrezni, torej ustrezajo potrebam uporabnikov?
  – Ali so izkazani vsi statistični rezultati, ki jih uporabniki potrebujejo?

  Dimenzijo lahko še nadalje razdelimo na naslednje poddimenzije:

  - **Potrebe uporabnikov**. Opis, razvrstitev in klasifikacija uporabnikov ter opis njihovih potreb in zahtev v zvezi z obravnavanim statističnim raziskovanjem.
  - **Zadovoljstvo uporabnikov**. Informacije o tem, ali se redno spremlja in meri zadovoljstvo uporabnikov (anketa o zadovoljstvu uporabnikov) in analizira

[6] Popolnost, ki je bila na začetku samostojna dimenzija, je bila kasneje pridružena kot poddimenzija dimenziji Ustreznosti.

njihove rezultate. V primeru, da se redne analize izvajajo, je potrebno glavne rezultate teh analiz v poročilu o kakovosti tudi prikazati.

- **Popolnost statistik**. Informacija o tem, ali se izračunava in izkazuje vse statistike, ki so zahtevane. Informacija, ki je zanimiva predvsem za Eurostat, saj jim podaja informacijo o tem, kakšen delež statistik, ki so zahtevane glede na evropske uredbe in druge dogovore, države članice res zagotavljajo.

- **Točnost**. Točnost ocenjuje neujemanje med (v statističnem raziskovanju) ocenjeno in pravo, toda neznano populacijsko vrednostjo. Tako neujemanje imenujemo tudi statistična napaka. V okviru dimenzije je potrebno oceniti in izkazati vse različne vrste napak, ki vplivajo na točnost objavljenih rezultatov. Nekaj najbolj značilnih tipov statističnih napak je (Groves, 1989; Lessler, 1992):

  - **Vzorčna napaka.** Napaka, ki nastane zaradi dejstva, da v raziskovanju opazujemo le del populacije (vzorec).
  - **Napaka pokritja.** Napaka pokritja nastane zaradi neskladja med populacijo, ki je predmet našega raziskovanja, in vzorčnim okvirom, ki je dejanska danost dometa našega opazovanja.
  - **Napaka neodgovora.** Napaka, ki nastane kot posledica dejstva, da v fazi zbiranja podatkov nismo uspeli pridobiti (vseh) želenih podatkov o izbranih enotah opazovanja. Napake neodgovora lahko nastanejo zaradi zavračanja, nezmožnosti odgovarjanja, nekontaktiranja, …
  - **Merska napaka.** Napaka, ki nastane v procesu zbiranja podatkov. V procesu zbiranja smo sicer o opazovani enoti pridobili želeno informacijo, vendar je (iz različnih razlogov) netočna. Razlogi za mersko napako so lahko: nerazumevanje vprašanja, slabo postavljeno vprašanje anketarja, problem s priklicem informacije, …
  - **Napaka zaradi obdelave.** Napaka, ki nastane pri obdelavi podatkov, na primer pri vnosu, kodiranju ali urejanju podatkov.
  - **Napaka zaradi privzema modela.** Napaka, ki je posledica privzema neustreznega modela v postopkih kot so kalibracija uteži ali desezoniranje časovne vrste.

- **Pravočasnost in točnost objave.** Dimenzija izkazuje časovno uspešnost objave statističnih rezultatov in ima dve poddimenziji:

  - Pravočasnost meri časovni razmik med referenčnim obdobjem, na katero se podatki nanašajo, in datumom objave.
  - Točnost objave podatkov označuje skladnost med dejanskim in predhodno najavljenim datumom objave podatkov.

- **Dostopnost in jasnost.** Dimenzija, v okviru katere ocenjujemo konkretne fizične okoliščine, v katerih so podatki dostopni uporabniku (preko katerih komunikacijskih poti lahko uporabnik dostopa do statističnih podatkov). Jasnost se nanaša na podatkovno okolje, preko katerega uporabnik dostopa do informacij o vsebini podatkov, predvsem na razpoložljivost pojasnil in drugih metapodatkov.
- **Primerljivost.** Dimenzija, preko katere se presoja, v kolikšni meri je obravnavane rezultate možno primerjati z enakovrednimi rezultati v drugih državah (geografska primerljivost) ter v drugih časovnih obdobjih (časovna primerljivost).
- **Skladnost.** Dimenzija, s katero ugotavljamo skladnost statističnih rezultatov s primerljivimi statističnimi rezultati iz drugih virov in posledično primernost za nadaljnje povezovanje in agregacijo.

Kot sedma, »pridružena« dimenzija kakovosti, je v modelu ocenjevanja kakovosti v ESS dodana dimenzija **Stroški in obremenitev**. Pod to dimenzijo po eni strani ocenjujemo, kakšno obremenitev je raziskovanje predstavljalo za poročevalske enote, po drugi pa, kakšne stroške je povzročila na strani statistične organizacije. Ta dimenzija kakovosti sicer direktno ne izkazuje kakovosti, je pa z večino drugih dimenzij tesno povezana, nemalokrat pa raven dosežene kakovosti celo v veliki meri določa (Eurostat, 2014).

Dimenzije in poddimenzije kakovosti, kot smo opisali zgoraj, so osnova za skupni, standardni model presoje kakovosti v okviru ESS. Glavni trije cilji, ki jih želijo članice ESS doseči z vpeljavo standardnega modela ocenjevanja kakovosti, so (Eurostat, 2014):

- Vpeljati harmonizirano ocenjevanje in poročanje kakovosti med različnimi statističnimi področji.
- Vpeljati harmonizirano ocenjevanje in poročanje kakovosti za sorodna področja med različnimi statističnimi organizacijami, članicami ESS.
- Zagotoviti, da poročila vsebujejo vse potrebne informacije za identifikacijo in vpeljavo potrebnih izboljšav, ki bi vodile k višji kakovosti statističnih procesov in izdelkov.

2.2 Kompromisi med dimenzijami kakovosti

Dimenzije kakovosti so sicer v teoretskem modelu opredeljene kot ločeni, samostojni vidiki kakovosti, vendar pa se v praksi izvedbe statističnih raziskovanj te dimenzije pogosto prepletajo, vplivajo ena na drugo ter tako terjajo kompromise med ravnijo dosežene kakovosti v okviru posameznih dimenzij. Najbolj tipični kompromisi med dimenzijami kakovosti so (Eurostat, 2009, Holt, 1999):

- **Kompromis med točnostjo in pravočasnostjo.** Izboljšanje pravočasnosti (krajšanje časovnega intervala med referenčnim obdobjem in objavo rezultatov) običajno vpliva na manjšo točnost rezultatov.

- **Kompromis med točnostjo in ustreznostjo.** Želja, da bi v čim večji meri zadovoljili uporabnike, predvsem z objavljanjem rezultatov na zelo podrobni ravni, lahko vpliva na večjo verjetnost statističnih napak in s tem slabšo točnost rezultatov.
- **Kompromis med ustreznostjo in časovno primerljivostjo.** Spremembe, ki jih uvajamo v izvedbo statističnega raziskovanja z željo povečati ustreznost izkazanih rezultatov in boljše izpolnjevanje potreb uporabnikov statističnih rezultatov, lahko vplivajo na slabšo primerljivost v času.
- **Kompromis med točnostjo in časovno primerljivostjo.** Vpeljava izboljšav v metodologijo izvedbe statističnega raziskovanja, ki izboljšuje točnost rezultatov, lahko na drugi strani vpliva na slabšo primerljivost v času.
- **Kompromis med stroški in ostalimi dimenzijami.** Količina sredstev, ki jo imamo na voljo za izvedbo raziskovanja, nemalokrat določa ostale dimenzije kakovosti. Če na primer lahko kot način pridobivanja podatkov uporabimo terensko anketiranje, ki je stroškovno precej zahteven način, bomo v veliki večini primerov dobili bolj točne rezultate, kot pa če bomo na primer uporabili telefonsko anketiranje.

Z vidika področja, ki je glavna tema naše naloge, torej statističnega urejanja podatkov v kratkoročnih raziskovanjih, sta predvsem zanimiva dva kompromisa: kompromis med točnostjo in pravočasnostjo in kompromis med stroški in točnostjo. Prvi kompromis je v primeru kratkoročnih raziskovanj tako ali tako neprestano aktualen. Želja po vse bolj hitrih objavah je v jasnem nasprotju z željo po čim bolj točnih statističnih rezultatih že zato, ker tudi poslovni subjekti potrebujejo nekaj časa, da pridejo do končnih podatkov za opazovano obdobje. V nadaljevanju bomo zato razpravljali, kako lahko z vpeljavo učinkovitega modela urejanja pripomoremo k boljšemu ravnovesju med tema dvema dimenzijama kakovosti. Vzporedno s tem pa bomo vključili še razmislek, kako lahko z uporabo bolj učinkovitih postopkov zmanjšamo potrebna sredstva na strani urejanja, ki jih nato lahko razporedimo v druge faze izvajanja raziskovanja.

3 KRATKOROČNA POSLOVNA RAZISKOVANJA

3.1 Kratkoročne statistike v okviru Evropskega statističnega sistema

Statistična raziskovanja, ki imajo za enoto opazovanja poslovne subjekte (na primer podjetja ter njihove lokalne enote), imenujemo **poslovna raziskovanja**. Poglavitne značilnosti, ki določajo podatke poslovnih raziskovanj in jih hkrati razlikujejo od podatkov drugih področij (na primer podatkov oseb in gospodinjstev), so (Cox in Chinnappa, 1995; Snijkers & Bavdaž, 2011):

- Spremenljivke v podatkih so pretežno številske, med njimi v veliki meri prevladujejo zvezne, nenegativne.

- Porazdelitev spremenljivk je izrazito asimetrična. Podatki malega števila enot običajno prispevajo zelo visok delež v skupni vsoti.
- Podatki, ki jih enote poročajo, so v veliki meri pridobljeni iz računovodskih sistemov poslovnih subjektov.
- Večina statističnih raziskovanj je periodičnih, kar pomeni, da jih izvajamo v rednih časovnih intervalih. To z vidika kontrole in urejanja pomeni, da imamo pogosto za enoto na voljo zgodovinske podatke, ki jih lahko v teh postopkih ustrezno uporabimo.
- Obstaja veliko število administrativnih virov s podatki, ki imajo visoko korelacijo, s podatki, ki jih zbiramo v statističnem raziskovanju. Tudi te podatke lahko ustrezno uporabimo v postopkih urejanja.

Med poslovnimi raziskovanji so, predvsem z vidika hitrega dostopa do informacij o kratkoročnih ekonomskih gibanjih, zelo pomembna **kratkoročna raziskovanja** oziroma njihovi produkti, kratkoročne statistike. Kratkoročne statistike so, po opredelitvi uredbe Evropske komisije o kratkoročnih statističnih kazalcih (Svet Evropske unije, 1998; v nadaljevanju Uredba STS), statistični podatki, ki vsebujejo informacije (spremenljivke), potrebne za zagotavljanje enotne osnove za analizo kratkoročnega gibanja ponudbe in povpraševanja, proizvodnih dejavnikov in cen.

Po Uredbi STS področje kratkoročnih statistik pokriva naslednja ekonomska področja dejavnosti (na podlagi mednarodne klasifikacije dejavnosti Nace Rev. 2): industrija, gradbeništvo, trgovina na drobno in storitve (brez finančnih in zavarovalniških storitev). Pri opredelitvi metodologije izbora in opazovanja enot na različnih področjih sta pomembna predvsem naslednja koncepta:

- **Statistična enota oziroma enota opazovanja.** V okviru kratkoročnih statistik se kot enoti opazovanja uporabljata predvsem naslednji statistični enoti:

  - Podjetje (angl. *Enterprise*). Podjetje je najmanjša kombinacija pravnih enot, ki ima kot organizacijska enota za izdelavo proizvodov ali ponudbo storitev pri svojem odločanju določeno stopnjo samostojnosti, predvsem za razporejanje svojih tekočih poslovnih sredstev. Podjetje lahko opravlja eno ali več dejavnosti, in to na eni ali več lokacijah (Svet ES, 1993).
  - Enota enovrstne dejavnosti (angl. *Kind of Activity Unit – KAU,* v nadaljevanju EED). EED združuje vse dele podjetja, ki prispevajo k izvajanju dejavnosti, definirane na ravni razreda (štiri-mestne številke) po NACE Rev.1[7], in ustreza enemu ali več operativnim oddelkom podjetja. Informacijski sistem podjetja mora biti sposoben, da za vsako EED opredeli ali izračuna vsaj vrednost proizvodnje, vmesno porabo,

[7] Ker je veljavna osnovna uredba, se le-ta sklicuje na takrat veljavno klasifikacijo dejavnosti NACE Rev. 1. Sedaj veljavna klasifikacija dejavnosti je Nace Rev. 2.

stroške delovne sile, operativne presežke in število zaposlenih ter ustvarjanje bruto osnovnih sredstev (Uredba Sveta št. 696/93).

V trenutni uredbi je za področji industrije in gradbeništva kot enota opazovanja določena EED, za področji trgovine na drobno in storitev pa podjetje. V trenutno še nastajajoči novi krovni uredbi za poslovne statistike (angl. *Framework Regulation Integrating Business Statistics – FRIBS)* so predvidene določene spremembe tudi na tem področju.

- **Glavna dejavnost opazovane enote.** Glavna dejavnost enote je gospodarska dejavnost, ki največ prispeva k dodani vrednosti enote. V preprostem primeru, ko enota izvaja zgolj eno gospodarsko dejavnost, je glavna dejavnost takšne enote določena s šifro Standardne klasifikacije dejavnosti, ki opisuje to dejavnost. Če enota izvaja več gospodarskih dejavnosti, se glavna dejavnost določi na podlagi dodane vrednosti teh dejavnosti s pomočjo metode »od zgoraj navzdol«. Metoda »od zgoraj navzdol« sledi hierarhičnemu načelu; dodeljena dejavnost enote na najnižji ravni klasifikacije mora biti skladna z razvrstitvijo enote na višjih ravneh. V praktični izvedbi določimo dejavnost enote na podlagi podatkov o dodanih vrednostih enote na različnih ravneh klasifikacije, skozi naslednje korake (Statistični urad Republike Slovenije, 2008):

  - za enoto določimo področje, ki ima pri enoti največji delež dodane vrednosti,
  - v okviru tega področja določimo oddelek, ki ima največji delež dodane vrednosti v tem področju,
  - v okviru tega oddelka določimo skupino, ki ima največji delež dodane vrednosti v tem oddelku,
  - v okviru te skupine določimo razred, ki ima največji delež dodane vrednosti v tej skupini.

Zaradi zelo pomembnega časovnega vidika in zelo pomembne dimenzije pravočasnosti je urejanje podatkov v primeru kratkoročnih statistik še posebej pomembno. Raziskovanja, ki zagotavljajo podatke za kratkoročne statistike, so glede na nekatere značilnosti toliko različne od preostalih poslovnih raziskovanj, da postopki urejanja v tem primeru zahtevajo precej drugačen pristop in načrt. Glavne razlikovalne značilnosti kratkoročnih raziskovanj, ki zagotavljajo kratkoročne statistike, so:

- Običajno imamo pri zbiranju podatkov opraviti z majhnim številom (vhodnih) spremenljivk.
- Pravočasnost je poglavitna zahteva, ki jo s kratkoročnimi statistikami želimo izpolniti. V klasičnem kompromisu med pravočasnostjo in točnostjo se nemalokrat daje prednost pravočasnosti.
- Prvi objavljeni rezultati, ki zadovoljujejo zahtevo po pravočasnosti, so običajno pozneje (enkrat ali večkrat) revidirani zaradi izboljšanja točnosti objavljenih rezultatov.

- Zgodovinski podatki so zelo močan pomožen vir za fazo urejanja podatkov in nemalokrat določajo postopke urejanja.

4 OPREDELITEV TEORETSKEGA MODELA UREJANJA V KRATKOROČNIH RAZISKOVANJIH

4.1 Izkušnje in prakse drugih statističnih uradov

4.1.1 Uvodna pojasnila

Ker postopki urejanja podatkov običajno predstavljajo zelo »potrošen« del celotnega procesa, se v večini uradov v zadnjih letih trudijo te postopke izboljšati v smeri večje učinkovitosti in posledično manjše potrošnje. Široka paleta raznolikih praks, ki jih uradi uporabljajo pri urejanju svojih podatkov, je predstavljena tudi v znanstvenih člankih, člankih s konferenc in v drugih strokovnih prispevkih. Ker je fokus naše naloge na urejanju podatkov kratkoročnih raziskovanj, v nadaljevanju podajamo tri primere praks, ki se direktno nanašajo na to področje in ki so bili opisani v člankih oziroma prispevkih s konferenc. Eden od kriterijev, ki smo ga upoštevali, je tudi možen doprinos praks k opredelitvi splošnega modela urejanja, ki ga bomo predstavili v nadaljevanju.

4.1.2 Zvezni statistični urad Republike Nemčije (Destatis)

Kot vse več evropskih statističnih uradov, tudi nemški statistični urad kratkoročne statistike izračunava na podlagi podatkov pridobljenih iz administrativnega vira, in sicer podatkov davčnega urada, ki so prvotno namenjeni obračunu davka na dodano vrednost (v nadaljevanju podatki DDV). Wein (2009) podaja opis reševanja enega od problemov, ki zelo pogosto nastopa v primeru uporabe administrativnih podatkov v kratkoročnih raziskovanjih. Ti podatki namreč zelo pogosto ne pokrivajo celotne ciljne populacije (Vlag, 2012; Marolt & Seljak, 2006), kar posledično vodi do problema manjkajočih podatkov na delu enot. Za razliko od manjkajočih podatkov pri neodgovoru, za te enote že vnaprej vemo, da od njih ne bomo pridobili podatka, saj tu manjkajoči podatek izhaja iz narave podatkovnega vira. Kot navaja Wein, je bilo potrebno v njihovem primeru med 30 in 40 % celotnega prihodka oceniti z metodami vstavljanja. Zato so razvili učinkovitejši postopek vstavljanja, ki ga lahko povzamemo v naslednjih točkah:

- V postopek so vedno vključeni podatki za preteklih 24 mesecev.
- Na tistih zgodovinskih podatkih, ki že imajo končne rezultate (več kot 6 mesecev stari podatki), so uporabili 9 različnih metod vstavljanja in ugotavljali, pri kateri metodi je bila najmanjša razlika (vsota absolutnih razlik po mesecih) med vstavljeno vrednostjo in kasneje res poročano vrednostjo.

- Za tekoči mesec je bila nato uporabljena (prej ugotovljena) najbolj učinkovita metoda. Če je bila povprečna razlika manjša od polovice ocenjene vrednosti, so vstavljeno vrednost popravili za prej ugotovljeno razliko (za povprečno razliko po mesecih).
- Uporabljene metode lahko razdelimo v tri skupine:

  - Metode za podjetja, katerih prihodek izkazuje sezonske vzorce (te metode ocenjujejo prihodek v tekočem obdobju na podlagi prihodka izpred enega leta).
  - Metode za podjetja, katerih prihodek je določen s prihodkom predhodnega meseca (te metode ocenjujejo prihodek v tekočem obdobju na podlagi prihodka izpred enega meseca).
  - Metode za enote, katerih prihodek je slučajen - neodvisen od predhodnih mesecev (te metode temeljijo na mediani/povprečju prihodka v preteklih mesecih).

4.1.3 Nizozemski statistični urad (Centraal Bureau voor de Statistiek)

Tudi ta urad uporablja podatke DDV kot glavni vir za oceno indeksov prihodka na mesečni ravni. Podatke »klasičnega raziskovanja«, zbranega na podlagi statističnega vprašalnika, uporabljajo samo še v primerih skupin podjetij in za nekatere dejavnosti, kjer je ocenjeno, da tovrstni administrativni podatki niso ustrezen vir. Hoogland (2009) in Vlag (2012) poročata o poglavitnih pomanjkljivostih podatkov DDV, ki povzročijo, da mora biti znatna količina teh podatkov podvržena statističnemu urejanju:

- Samo del podatkov (za večje enote) je na voljo na mesečni ravni. Ostali podatki so na voljo na četrtletni ali na letni ravni. Podjetja, ki so pod določenim pragom letnega prihodka, so celo oproščena poročanja.
- V podatkih DDV je precej merskih napak. Metodološka opredelitev opazovanega pojava (na primer prihodek) ni nujno povsem v skladu s podatkom iz administrativnega vira. Vse te napake običajno pomenijo precejšnjo količino dela, ki ga je potrebno opraviti v fazi statističnega urejanja podatkov.

  - Identifikacija enot v podatkih DDV je lahko različna od identifikacije enot v raziskovanju. Povezovanje enot iz dveh različnih virov je lahko vir napak in manjkajočih podatkov.
  - Enote poročanja v sistemu DDV niso vedno enake enoti opazovanja v statističnem raziskovanju. Pretvorba na ustrezne statistične enote lahko doprinese še k dodatni merski napaki.

Hoogland (2006) opiše postopek, s katerim poskušajo odpraviti čim večjo količino napak, ki izhajajo iz prej opisanih pomanjkljivosti. V nadaljevanju podajamo kratek opis postopkov:

- V prvi fazi preverjanja podatkov se išče dvomljive vzorce poročanja v podatkih DDV. V primeru analize četrtletnih podatkov sta prevladujoča predvsem naslednja dva dvomljiva vzorca poročanja:

  – (0,0,0,x), $x > 0$. Enota je neničelne podatke poročala samo v zadnjem četrtletju. Tak vzorec poročanja lahko pomeni, da je enota v zadnjem četrtletju v bistvu poročala letni prihodek. V takem primeru je potrebno v fazi urejanja ta prihodek z ustrezno metodo »porazdeliti« med vsa štiri četrtletja.
  – (x,x,x,x), $x \neq 0$. Enota je v vseh štirih četrtletjih poročala enak prihodek. Za nekatere dejavnosti je to povsem smiseln vzorec poročanja (na primer oddajanje nepremičnin v najem).
  – Podatke, ki so bili v prvi fazi označeni kot dvomljivi, se popravi s postopki statističnega urejanja. Del enot se uredi ročno, del avtomatsko. Ročno urejanje se uporablja predvsem pri enotah, katerih podatki so bili zbrani z vprašalniki, vse dvomljive podatke DDV pa se popravi zgolj z uporabo postopkov avtomatskega urejanja. Za bolj učinkovit sistem urejanja je vpeljan sistem, ki kombinira pristop selektivnega urejanja in urejanja na makro ravni. Postopek je iterativen in je sestavljen iz ponavljajočih se podpostopkov, ki izmenično kombinirajo mikro in makro raven:

    - Opravi se določanje dvomljivih vrednosti na mikro ravni in prva faza avtomatskih popravkov.
    - Izračunajo se ciljni statistični agregati. Izvede se analiza in kontrola na makro ravni. V ta namen se vse domene grupirajo v »sorodne« skupine. V okviru vsake skupine se izračuna razlika med največjo in najmanjšo vrednostjo in če ta razlika preseže predpisani prag, je celica označena kot dvomljiva.
    - Dvomljive celice gredo v nadaljnje »ročno« preverjanje. Po potrebi se izvršijo dodatni popravki podatkov.
    - Ponovi se kontrola agregatov in zaznavanje dvomljivih vrednosti.

4.1.4 Statistični urad Švedske (Statistics Sweden)

Ta urad je eden vodilnih v razvoju in uvajanju novih in naprednih metod urejanja podatkov. Predvsem je njihov doprinos zelo pomemben na področju racionalizacije postopkov urejanja podatkov in uvajanju selektivnega urejanja (Granquist, 1995, Hedlin 2003, Norberg, 2009, Norberg et al., 2014). V zadnjih letih so razvili tudi odprtokodno programsko aplikacijo za izvajanje selektivnega urejanja – SELEKT, ki je na voljo tudi ostalim statističnim uradom. Norberg, Lindgren in Tongur (2014, str. 2-9) opisujejo uporabo tega programskega orodja za postopke selektivnega urejanja v 11 raziskovanjih. Kot navajajo avtorji, je uvedba postopkov selektivnega urejanja bistveno zmanjšala količino podatkov za urejanje in posledično tudi bistveno zmanjšala stroške raziskovanja. Za kratkoročno raziskovanje o plačah v privatnem sektorju poročajo o 12-odstotnem zmanjšanju enot za urejanje, za

kratkoročno raziskovanju o zaposlenih pa 22–27-odstotno zmanjšanje enot za urejanje (Norberg et al., 2014, str. 5-6).

4.1.5 Statistični urad Kanade (Statistics Canada)

Za konec poglejmo še primer implementacije metode za iskanja osamelcev. Belcher (2003, str. 27-29) opiše metodo, ki jo v ta namen uporabljajo na kanadskem statističnem uradu. Metodo sicer uporabljajo za iskanje osamelcev v letnih finančnih podatkih, za nas pa je zanimiva, ker predstavlja uporabno posplošitev H-B metode, ki smo jo opisali v razdelku 1.4. H-B metoda je sicer zelo primerna za iskanje osamelcev v izrazito asimetričnih porazdelitvah (kot so običajno porazdelitve podatkov poslovnih raziskovanj), saj upošteva velikost enote in s tem relativno pomembnost enote v porazdelitvi. Osnovna oblika H-B metode podaja interval sprejemljivosti za transformirano porazdelitev razmerij $\{E_i\}$. Zgornja in spodnja meja intervala zaupanja, ki sta povsem ustrezna za računsko določanje osamelih vrednosti, pa nimata nobene informativne vrednosti za vsebinske statistike. Zato je koristno pretvoriti te meje nazaj v interval sprejemljivosti za $y_i(t)$, glede na vrednost $y_i(t-1)$. Če je SM_E, spodnja meja področja sprejemljivost za $\{E_i\}$, določena kot $SM_E = E_M - C \cdot d_{Q1}$, lahko izpeljemo spodnjo mejo za $y_i(t)$ kot:

$$SM_{y_i(t)} = \frac{r_M \cdot xy_i(t-1)}{1 - \frac{SM_E}{[y_i(t-1)]^U}} \tag{19}$$

Pri določanju analognega izraza za zgornjo mejo ($ZM_{y_i(t)}$) lahko analitično izpeljemo zgolj naslednjo zvezo:

$$\left(\frac{\frac{ZM_{y_i(t)}}{y_i(t-1)}}{r_M} - 1 \right) \cdot \left(ZM_{y_i(t)} \right)^U = ZM_E \tag{20}$$

Izkaže se, da iz te zveze analitično ni mogoče izpeljati eksplicitnega izraza za $ZM_{y_i(t)}$, lahko pa enačbo rešimo s pomočjo numeričnih metod, denimo z uporabo Newtonove iteracijske metode za iskanje ničel nelinearne funkcije (Bohte, 1993). Če iskano zgornjo mejo $ZM_{x_i(t)}$ na kratko označimo kot x, lahko iteracijsko formulo po Newtonovi metodi zapišemo kot:

$$x_{n+1} = x_n - (x_n^{U+1} - y_i(t-1) \cdot r_M \cdot x_n^U - ZM_E \cdot y_i(t-1) \cdot \frac{r_M}{U+1} \cdot x_n^U - U \cdot y_i(t-1) \cdot r_M \cdot x_n^{U-1}) \tag{21}$$

Kot prvi približek za x_0 lahko vzamemo vrednost $y_i(t-1)$. Z določitvijo spodnje in zgornje meje sprejemljivosti za vsako posamezno enoto lahko nato precej lažje izvajamo postopke urejanja (na primer avtomatske popravke) na mikro ravni.

4.2 Teoretski model urejanja

4.2.1 Uvodna pojasnila

S teoretskim modelom urejanja podatkov kratkoročnih poslovnih raziskovanj bomo skušali v čim večji meri upoštevati vse ključne faktorje, ki na urejanje vplivajo, določajo njegovo izvedbo in izhodne rezultate urejanja. Ti ključni faktorji so predvsem:

- Podatke raziskovanja pridobimo iz dveh različnih virov (»klasično statistično raziskovanje« in administrativni vir), ki jih je v postopkih urejanja potrebno obravnavati ločeno.
- Predvsem administrativni podatki imajo lahko različno periodiko poročanja (na primer četrtletno, potrebujemo pa mesečne podatke), kar povzroči, da moramo za del enot, ki so že vnaprej določene, podatke oceniti.
- Zelo pomembna je longitudinalna komponenta podatkov, zato morajo tako logične kontrole kot tudi postopki za izvajanje popravkov upoštevati to komponento.
- Rezultati raziskovanja se (za isto referenčno obdobje) izračunavajo in objavljajo večkrat.
- Zaradi izrazito asimetrične porazdelitve podatkov imajo nekatere enote precej večji vpliv na statistične rezultate urejanja kot druge. Te enote je tudi v postopkih urejanja treba obravnavati posebej (selektivno urejanje na podlagi funkcije pomembnosti).
- Celoten postopek nastanka napak v podatkih, njihovo odkrivanje in odpravljanje, poganja sklop mehanizmov, od katerih je del v rokah in pod kontrolo izvajalcev raziskovanj, del pa ostaja izvajalcem skrit in ga lahko samo (do določene mere) modeliramo.

4.2.2 Mikro podatki

Mikro podatki, ki so pridobljeni v okviru izvedbe statističnega raziskovanja in iz katerega so v končni fazi izvedeni statistični rezultati, so osnova za postopke urejanja. Zato moramo najprej opredeliti formalno predstavitev teh podatkov. Podatke na mikro ravni bomo opisovali s trodimenzionalno matriko dimenzije $n \times m \times T$, ki smo jo uvedli v razdelku 1.4.

Naj bo $\mathbf{Y}^0$ tridimenzionalna matrika pravilnih vrednosti m spremenljivk, pri n enotah v T zaporednih referenčnih obdobjih. Nadalje naj bo $\widehat{\mathbf{Y}}^z$ matrika prvotno zbranih podatkov, $\widehat{\mathbf{Y}}^{p_1}$, $\widehat{\mathbf{Y}}^{p_2}$, …, $\widehat{\mathbf{Y}}^{p_k}$ pa matrike podatkov, ki so pripravljeni za posamezne objave statističnih rezultatov oziroma statistik. Predpostavka je torej, da smo rezultate, ocenjene na podlagi podatkov matrike $\mathbf{Y}$, objavili k- krat. Ker je povsem realna predpostavka, da vse matrike podatkov vsebujejo napake, lahko v splošnem trdimo, da se vse matrike, iz katerih smo pripravili posamezne objave, razlikujejo od $\mathbf{Y}^0$. Razlike posameznih matrik do matrike pravih vrednosti imenujemo matrike napake in jih označimo kot E^{p_1}, E^{p_2}, …, E^{p_k}. Velja torej:

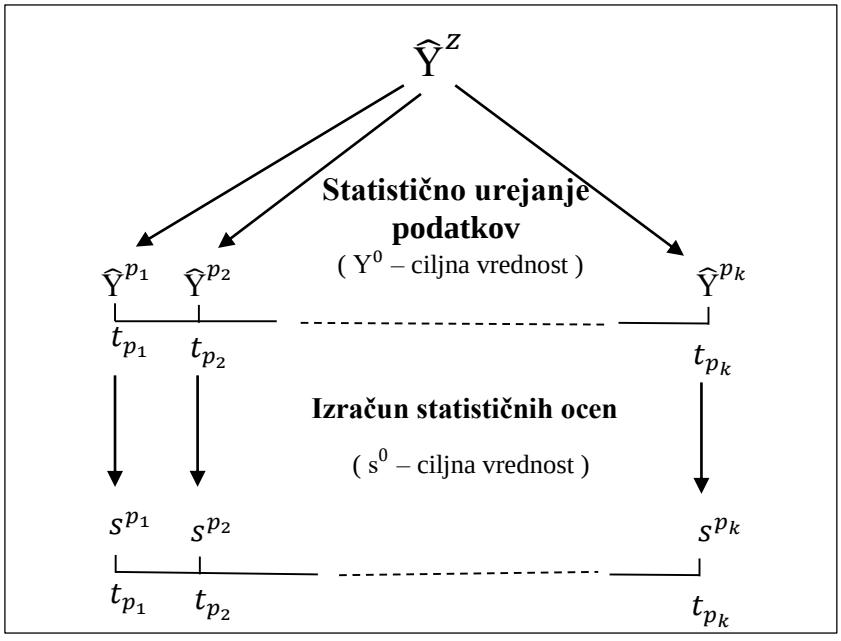
$$\mathrm{E}^{p_i} = \hat{\mathbf{Y}}^{p_i} - \mathbf{Y}^0 \tag{22}$$

Čeprav običajno na podlagi podatkov statističnega raziskovanja objavimo več statistik, tu predpostavimo zgolj eno ciljno statistiko s, ki jo iz podatkov v matriki mikro podatkov izračunamo s pomočjo agregatne funkcije $ag(.)$, torej velja $s = ag(\mathbf{Y})$. Ker matrika $\mathbf{Y}$ vsebuje podatke več referenčnih obdobij, lahko v splošnem predpostavimo, da lahko za izračun statistike s uporabimo podatke različnih referenčnih obdobij. Primer take statistike je (časovni) indeks na različna časovna obdobja. Na primer, indeks tekočega obdobja glede na predhodno obdobje računamo iz podatkov dveh zaporednih časovnih obdobij in lahko zapišemo: $I_{t_0/t_0-1} = ag(\mathbf{Y}(t)|_{t\in\{t_0,t_0-1\}})$.

Če v zgornjem zapisu zamenjamo splošno matriko podatkov s prej definiranimi različnimi verzijami podatkov, je $s^0 = ag(\mathbf{Y}^0)$, prava vrednost ciljne statistike s, $s^{p_i} = ag(\mathbf{Y}^{p_i})$ pa ocena, pripravljena za i-to objavo statistike $s(t)$. Razlike $s^{p_i} - s^0$ predstavljajo statistične napake posameznih objavljenih statistik. Naš cilj je, da so te napake z vsako objavo manjše.

Na Sliki 7 je podan shematski prikaz različnih verzij podatkov, ki jih v različnih časovnih točkah uporabimo za izračun in izkazovanje statističnih rezultatov.

Slika 7: Shematski prikaz modelne predstavitve podatkov

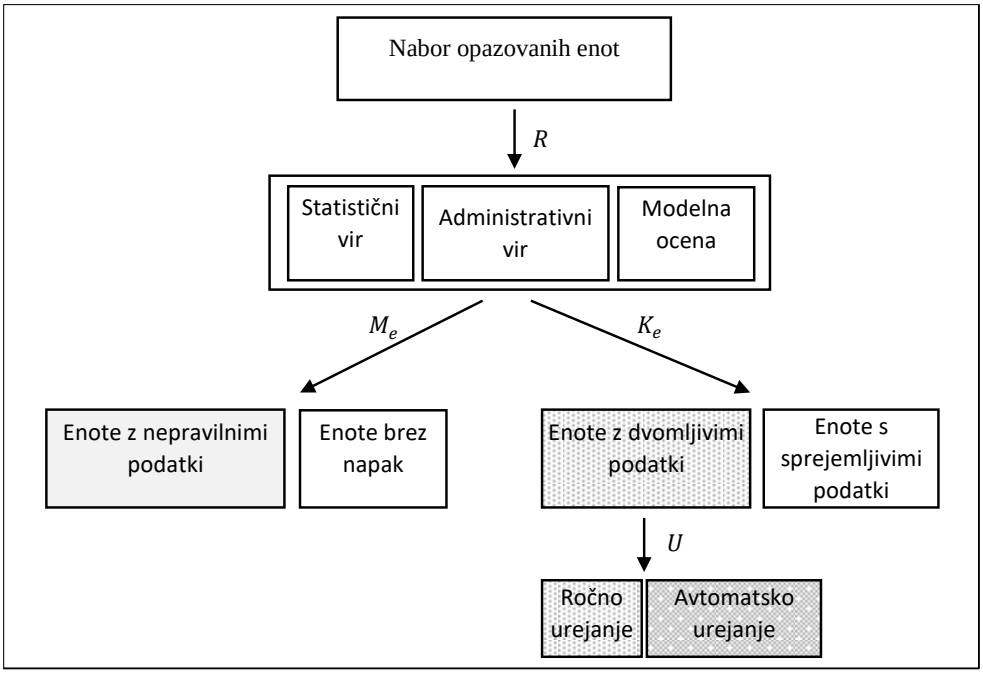


4.2.3 Modelni mehanizmi

V naslednjem koraku bomo definirali sklop mehanizmov, preko katerih se izvajajo posamezni koraki v postopku statističnega urejanja podatkov. Ena skupina teh mehanizmov

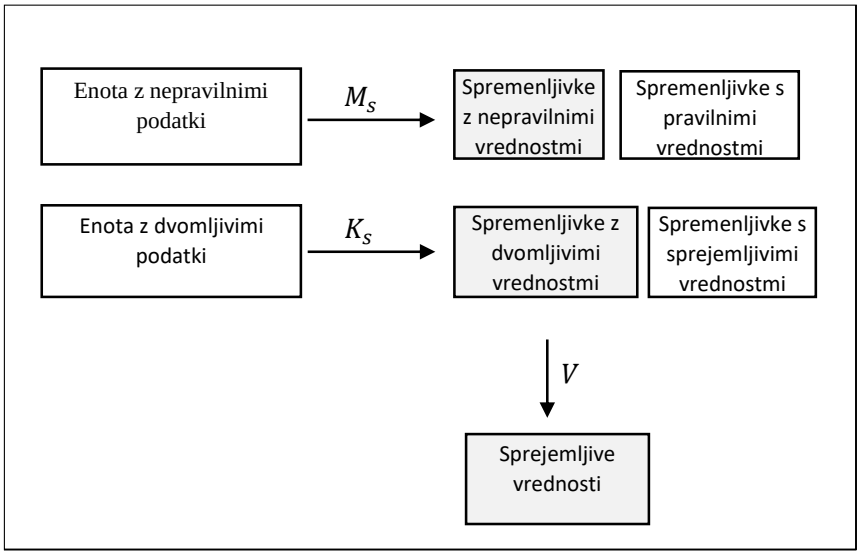
deluje na ravni opazovane enote, druga skupina pa na ravni opazovane spremenljivke. Slika 8 prikazuje delovanje mehanizmov na ravni enote.

Slika 8: Mehanizmi na ravni statistične enote



Slika 9 prikazuje delovanje mehanizmov na ravni spremenljivke.

Slika 9: Mehanizmi na ravni statistične spremenljivke



Prvi mehanizem sicer določa postopek, ki neposredno ne spada v okvir postopkov urejanja podatkov, ima pa na urejanje pomemben vpliv. Gre za mehanizem, ki celoten nabor opazovanih enot razdeli glede na uporabljeni podatkovni vir. Pri tem poudarjamo, da

mehanizem, ki ga opisujemo, ne vključuje postopka izbora enot opazovanja, saj tega dela procesa v naši nalogi ne obravnavamo. Zaradi tega je tudi naš model neodvisen od načina izbora enot opazovanja (slučajni vzorec, zajem s pragom, popis…).

Predpostavka, s katero je ta prvi mehanizem pogojen, je torej predpostavka obstoja več podatkovnih virov za podatke raziskovanja. V splošnem lahko uporaba različnih podatkovnih virov razdeli nabor podatkov na ravni enot ali na ravni spremenljivk ali pa celo na ravni obeh. V prvem primeru so različni viri uporabljeni za različne enote, v drugem primeru za različne spremenljivke, v tretjem primeru pa uporabljamo kombinacijo obeh pristopov. Zelo redek je primer delitve glede na časovno komponento, kar bi pomenilo, da za eno referenčno obdobje uporabljamo en vir, za drugo pa drugega. S tako situacijo smo soočeni zgolj v primeru prehoda na nov vir, kar pa ni tema naše naloge. Ker je v primeru kratkoročnih raziskovanj prevladujoč prvi pristop, torej uporaba različnih virov za različne enote, bomo v našem modelu predpostavili zgolj tako porazdelitev različnih podatkovnih virov. Glede na konkretne prakse, predstavljene v prejšnjem razdelku, bomo predpostavili obstoj treh podatkovnih virov:

- **Statistični vir.** Podatke zberemo s »klasičnim« statističnim raziskovanjem, z uporabo anketnega vprašalnika.
- **Administrativni vir.** Podatki so pridobljeni iz zbirk podatkov nekega administrativnega organa (na primer Finančni urad Republike Slovenije). Prvotno ti podatki niso bili zbrani v statistični namen.
- **Statistična ocena.** Podatki, ki smo jih ocenili z uporabo ustreznega statističnega modela. Pri tem poudarimo, da tu ne gre za podatke, za katere smo predvidevali, da jih bomo zbrali iz enega od prej naštetih virov, vendar nam zaradi različnih razlogov tega ni uspelo. Tu gre za enote, za katere že vnaprej vemo, da za določeno periodo podatkov iz administrativnega vira ne bomo dobili, lahko jih pa s pomočjo nekih pomožnih podatkov ocenimo. Tak primer so enote, ki so zavezane k poročanju podatkov DDV zgolj na četrtletni in ne na mesečni ravni.

Nabor n enot torej razdelimo v tri podskupine: n_{sr} enot, katerih podatke pridobivamo iz statističnega raziskovanja; n_{ad} enot, katerih podatke pridobivamo iz administrativnega vira; n_{mo} enot, katerih podatke ocenimo z modelno oceno. Velja $n = n_{sr} + n_{ad} + n_{mo}$ in $Y_{(n,m,t)} = (Y_{(n_{sr},m,t)}, Y_{(n_{ad},m,t)}, Y_{(n_{mo},m,t)})^T$. Funkcijo, ki razdeli celoten nabor enot v tri skupine, označimo R:

$$n \xrightarrow{R} (n_{sr}, n_{ad}, n_{mo}) \tag{23}$$

Poglejmo še nekoliko podrobneje delovanje mehanizma R. Izhodni rezultat, torej delitev v tri podskupine, določajo poleg števila vseh opazovanih enot (n) še naslednji vhodni parametri:

- Porazdelitev vrednosti opazovanih spremenljivk $Y_1, Y_2, \dots, Y_m$. Zaradi lažje izvedljivosti običajno izberemo samo eno ključno spremenljivko (Y_k), katere porazdelitev analiziramo. Prav tako v tem primeru zanemarimo longitudinalni vidik in obravnavamo zgolj eno referenčno obdobje izbrane ključne spremenljivke. To referenčno obdobje je lahko konkretno obdobje opazovanja (na primer zadnji mesec, za katerega so ob izboru enot opazovanja podatki na voljo), ali pa povsem fiktivno obdobje (na primer povprečje preteklega leta). Kot vhodni parameter mehanizma torej vzamemo enodimenzionalno porazdelitev F_{Y_k}, kjer je F kumulativna porazdelitvena funkcija izbrane ključne spremenljivke.
- Ciljni delež »pokritosti« ključne spremenljivke z vrednostmi pri enotah, katerih vrednosti pridobimo iz statističnega vira (d_{sr}). Število enot, katerih podatke pridobivamo iz statističnega raziskovanja, je nato direktna funkcija obeh parametrov ter ciljnega števila opazovanih enot:

$$n_{sr} = f(F_{Y_k}, n, d_{sr}) \tag{24}$$

- Delež enot, za katere podatke pridobivamo iz administrativnega, in je administrativen podatek res razpoložljiv na ravni ciljne referenčne časovne periode (d_{ad}). Ta delež je sicer določen z naravo administrativnih podatkov in ga lahko pred pričetkom izvajanja raziskovanja zgolj ocenimo.

Mehanizem R lahko torej v obliki funkcijske enačbe zapišemo kot:

$$(n_{sr}, n_{ad}, n_{mo}) = R(n, F_{Y_k}, d_{sr}, d_{ad}) \tag{25}$$

Zgoraj opisana razdelitev enot poteka pred samim pričetkom izvajanja raziskovanja in lahko trdimo, da je mehanizem razdelitve (R) v rokah in pod kontrolo izvajalcev raziskovanja. Drugače je z naslednjim mehanizmom, ki ga bomo predpostavili. Predpostavimo namreč, da obstaja mehanizem, ki v fazi pridobivanja podatkov razdeli pridobljene podatke na pravilne in na nepravilne podatke. Nadalje predpostavimo dve ravni tega mehanizma. Prvi mehanizem deluje na ravni enot, torej razdeli celoten nabor opazovanih enot na dva dela: na enote, ki imajo vsaj en nepravilen podatek, v vsaj enem časovnem obdobju (n_n), in na enote, za katere so vsi podatki pravilni (n_p). Velja $n = n_n + n_p$ in $Y_{(n,m,t)} = (Y_{(n_n,m,t)}, Y_{(n_p,m,t)})^T$. Ta mehanizem razdelitve bomo označevali kot funkcijo M_e:

$$n \xrightarrow{M_e} (n_n, n_p) \tag{26}$$

Za razliko od mehanizma R, pri katerem smo lahko dokaj natančno opredelili parametre, ki določajo njegovo delovanje, je v tem primeru ta naloga malo težja. Večina parametrov bo

temeljila zgolj na predpostavkah, ki jih bomo poskušali potrditi kasneje v empiričnem delu naloge. Delovanje mehanizma M_e torej določajo naslednji parametri:

- Način pridobivanja podatkov. Pri tem, tako kot prej, predpostavimo obstoj treh podatkovnih virov, katerih porazdelitev med enote opazovanja je določena z mehanizmom R. Način pridobivanja podatkov pri enoti i označimo z np_i.
- Velikost enote. Predpostavka je, da velikost enote (glede na izbrano ključno spremenljivko) pomembno vpliva na verjetnost, da bo neka enota vsebovala nepravilne podatke. Predpostavimo, da velikost enote določa ista ključna spremenljivka Y_k, katere porazdelitev določa, katere enote vključimo v statistično raziskovanje. Predpostavimo še, da se velikost enote skozi referenčna obdobja ne spreminja in lahko uporabimo kar vrednost ključne spremenljivke, ki je bila določena ob postopku izbora enot opazovanja. Vrednost ključne spremenljivke pri enoti i označimo z y_{ki}^0.
- Vključenost v raziskovanje. Predpostavka je, da je pri enotah, ki so že dlje časa vključene v raziskovanje, manjša verjetnost za nastanek napak. Število obdobij, v katerih je bila enota i do referenčnega obdobja t že vključena v raziskovanje, označimo z $r_i^v(t)$.

Funkcijo M_e na enoti opazovanja bomo sedaj opredelili kot binarno funkcijo, ki ima vrednost 1, če ima enota kak nepravilen podatek, in vrednost 0, če nima nobenega nepravilnega podatka. Nato lahko zapišemo:

$$n_n = \sum_{t=1}^{T} \sum_{i=1}^{n} Me(y_i(t)) = \sum_{t=1}^{T} \sum_{i=1}^{n} Me(np_i, y_{ki}^0, r_i^v(t)) \quad (27)$$

$$n_p = n - n_n \quad (28)$$

Druga raven, na kateri deluje razdelitev podatkov na pravilne in nepravilne, je raven spremenljivke. Mehanizem na tej ravni deluje samo na množici n_n enot (iz poljubnega referenčnega obdobja), ki imajo kakšen nepravilen podatek. Naj bo torej $y_i(t) = (y_{i1}(t), y_{i2}(t), \dots, y_{im}(t))$ enota i v časovnem obdobju t, za katero je vsaj eden od m pridobljenih podatkov nepravilen. Mehanizem M_s pri vsaki izmed teh enot razdeli m opazovanih spremenljivk na dva dela: m_n spremenljivk z nepravilnimi in m_n spremenljivk s pravilnimi podatki:

$$m \xrightarrow{M_s} (m_n, m_p) \quad (29)$$

Tudi pri mehanizmu M_s gre za »skriti mehanizem«, kar pomeni, da o parametrih, ki mehanizem določajo, lahko le bolj ali manj uspešno predpostavljamo. V našem modelu bomo predpostavili zgolj en, precej splošen parameter, to je kompleksnost vprašanja, s katerim pridobimo podatek (vrednost spremenljivke). Ta parameter, ki ga bomo imenovali utež kompleksnosti in ga označevali kot w_j^C, torej pripišemo vsaki spremenljivki Y_j. Utež

kompleksnosti lahko interpretiramo kot verjetnost, da bomo pri določeni spremenljivki pridobili nepravilen podatek. Ta verjetnost je v resnici pogojna glede na enoto opazovanja, vendar bomo v našem modelu privzeli poenostavitev, da je ta verjetnost (in s tem tudi utež) neodvisna od enote opazovanja, kar pomeni, da bo utež kompleksnosti opredeljena samo na ravni spremenljivke. Če tudi v tem primeru M_s opredelimo kot binarno funkcijo, bo to funkcija, ki bo definirana na spremenljivki pri enoti iz nabora n_n enot z nepravilnimi vrednostmi. Vrednost funkcije bo enaka 1, če bo vrednost spremenljivke nepravilna, sicer pa bo njena vrednost 0.

$$m_n = \sum_{j=1}^{m} Ms(y_{ij}(t)) = \sum_{j=1}^{m} Ms(w_j^C) \tag{30}$$

$$m_p = m - m_n \tag{31}$$

Mehanizma, ki smo ju opisali zgoraj, delujeta v fazi pridobivanja podatkov. Naslednja dva mehanizma pa delujeta v naslednji fazi, to je v fazi statistične obdelave oziroma bolj natančno v fazi urejanja podatkov. Prvi mehanizem bo razdelil nabor n opazovanih enot v podmnožico n_d dvomljivih in tej podmnožici komplementarno podmnožico n_a sprejemljivih enot. Kot smo že navedli v razdelku 1.1, to razdelitev določa sistem logičnih kontrol, kar pomeni, da je popolnoma v rokah izvajalca raziskovanja. Mehanizem, ki opravlja to delitev, bomo označevali kot funkcijo K_e:

$$n \xrightarrow{K_e} (n_d, n_a) \tag{32}$$

Edini parameter, ki poleg samih podatkov določa delovanje mehanizma K_e, je nabor za raziskovanje opredeljenih logičnih kontrol $\mathcal{K}$. Povsem splošno obliko logičnih kontrol smo zapisali v razdelku 1.3, za primer kratkoročnih raziskovanj pa bomo ta splošni zapis nekoliko poenostavili. Glede na prakse tujih uradov, pa tudi glede na prakso SURS, sta v teh raziskovanjih prevladujoča dva tipa kontrol:

- Kontrole, ki jih lahko zapišemo v obliki linearne enačbe ali neenačbe, torej oblike:

$$a_0 + \sum_{j=1}^{m} \sum_{t=1}^{T} a_j(t) \cdot Y_j(t) \ \Delta \ 0 \tag{33}$$

  kjer je Δ eden od operatorjev $\{=, <, >, \leq, \geq\}$. Pri tem je pomembno poudariti, da so vsi koeficienti $\{a_j(t)\}$ konstante, od samih podatkov neodvisni koeficienti. To množico kontrol bomo imenovali »linearne kontrole« in jo označili s $\mathcal{K}_L$. V množico $\mathcal{K}_L$ bomo uvrščali tudi vse kontrole, ki jih iz več linearnih (ne)enačb tvorimo z uporabo logičnih operatorjev $\{\wedge, \vee, \neg\}$.

- Kontrole, ki izhajajo iz porazdelitve funkcije g več longitudinalnih komponent izbrane spremenljivke Y_0, $g(Y_0)$. Funkcija g je običajno razmerje vrednosti spremenljivke v dveh časovno sosednih obdobjih, lahko pa tudi kaka druga enostavna funkcija, recimo razlika vrednosti dveh sosednih obdobij. Porazdelitev vrednosti funkcije g pri različnih enotah naj določa kumulativna porazdelitvena funkcija G_{Y_0}. Z ustrezno transformacijo $\mathcal{L}$ porazdelitve G_{Y_0} in s podanim vektorjem parametrov $p = (p_1, p_2, \dots, p_r)$ lahko nato določimo zgornjo in spodnjo mejo območja sprejemljivosti v porazdelitvi G_{Y_0}: $\mathcal{L}(G_{Y_0}) = (s_{G_{Y_0}}, s_{G_{Y_0}})$. Kot za primer HB metode pokaže Belcher (2003), je mogoče potem določiti obratno transformacijo, ki dane meje sprejemljivosti porazdelitve pretvori v meje sprejemljivosti za vsako posamezno enoto. Ta obratna informacija, ki jo bomo označili $\mathcal{L}^{-1}$, torej pretvori vektor $(s_{G_{Y_0}}, s_{G_{Y_0}})$ v $(n$ x 2 x T) matriko spodnjih in zgornjih mej:

$$\mathcal{L}^{-1}\left(\left(s_{G_{Y_0}}, s_{G_{Y_0}}\right)\right) = \begin{bmatrix} s_{10}(t), & z_{10}(t) \\ s_{20}(t), & z_{20}(t) \\ \dots & \dots \\ s_{n0}(t), & z_{n0}(t) \end{bmatrix} \tag{34}$$

Na podlagi te matrike lahko nato zapišemo množico kontrol oblike:

$$y_{i0}(t) < s_{i0}(t) \tag{35}$$

$$y_{i0}(t) > z_{i0}(t) \tag{36}$$

Kontrole take oblike bomo imenovali »kontrole longitudinalne porazdelitve« in označili $\mathcal{K}_{LP}$. V končni obliki ima tudi ta skupina kontrol linearno obliko, vendar s prejšnjima dvema skupinama obstaja bistvena razlika. Koeficienti v neenačbi namreč niso konstantni za celoten set kontroliranih podatkov, ampak so določeni za vsako enoto posebej. Vsak par spodnje in zgornje meje, je torej funkcija porazdelitve G_{Y_0} in same vrednosti spremenljivke: $(s_{i0}(t), z_{i0}(t)) = f(G_{Y_0}, y_{i0}(t))$. V realizaciji te skupine kontrol, pri kateri je funkcija g razmerje dveh longitudinalnih vrednosti, torej v splošnem $g(Y_0(t)) = Y_0(t)/Y_0(t-p)$, se pojavi problem ničelnih vrednosti v imenovalcu. V primeru, ko je $Y_0(t-p) = 0$, razmerij ne moremo vključiti v analizirano porazdelitev, lahko pa razlika med $Y_0(t), Y_0(t-p)$ tudi v tem primeru predstavlja dvomljiv podatek. Na primer, če je enota poročala ničelni podatek v preteklem mesecu, v tem mesecu pa zelo visok podatek, je to lahko signal za morebitno napako v poročanju. Zaradi tega moramo vključiti še dodatno kontrolo, ki bo take primere označila kot dvomljive. V tem primeru je za določitev mej sprejemljivosti smiselno upoštevati samo vrednosti spremenljivke Y_0 v tekočem obdobju. Najenostavneje rešitev je, da kot mejo uporabimo enega od momentov porazdelitve spremenljivke $Y_0(t)$ (na primer standardni odklon). Kontrolo potem lahko zapišemo v obliki:

$$Y_0(t-p) = 0 \wedge Y_0(t) > m(F_{Y_0(t)}) \tag{37}$$

kjer je $F_{Y_0(t)}$ kumulativna funkcija porazdelitve za $Y_0(t)$ in m izbrani moment. Množico kontrol longitudinalne porazdelitve torej v splošnem lahko zapišemo kot množico kontrol oblik (35), (36) in (37):

$$\mathcal{K}_{LP} = \{y_{i0}(t) < s_{i0}(t), y_{i0}(t) > z_{i0}(t), y_{i0}(t-p) = 0 \wedge y_{i0}(t) > m(F_{Y_0(t)})\} \quad (38)$$

Pogosto se pri validaciji podatkov kratkoročnih raziskovanj uporabljajo tudi kontrole, ki jih z uporabo logičnih operatorjev tvorimo iz linearnih kontrol s fiksnimi koeficienti in kontrol longitudinalne porazdelitve. Tipičen primer je, ko kontrolo longitudinalne porazdelitve omejimo samo na podmnožico opazovanih podatkov, na primer samo na tiste podatke, pri katerih je izbrana spremenljivka pozitivna. Po dogovoru bomo tudi te kontrole uvrščali v množico $\mathcal{K}_{LP}$.

Zapišemo torej lahko:

$$\mathcal{K} = \mathcal{K}_L \cup \mathcal{K}_{LP} \quad (39)$$

Mehanizem K_e deluje na ravni enot. Za vsako dvomljivo enoto nato obstaja še en mehanizem, ki deluje na ravni spremenljivke. Ta mehanizem nabor spremenljivk pri dvomljivi enoti razdeli v podmnožico m_a spremenljivk s sprejemljivimi vrednostmi in m_d spremenljivk z dvomljivimi vrednostmi. Ta mehanizem, ki deluje na naboru n_d dvomljivih enot, bomo označevali s funkcijo K_s:

$$m \xrightarrow{K_s} (m_d, m_a) \quad (40)$$

Če v okviru raziskovanj izvajamo samo postopke ročnega urejanja, je ta mehanizem implementiran v okviru ponovnega preverjanja podatkov, ki ga je teoretsko težko opisati, pa tudi z vidika obravnavane teme ni posebno zanimiv. Zato se na tem mestu omejimo na opis mehanizma v primeru izvajanja avtomatskih popravkov. Določanje dvomljivih spremenljivk v enoti z dvomljivimi podatki v veliki večini primerov temelji na eni od izvedb Fellegi-Holt pristopa oziroma na prvem delu postopka, ki smo ga v razdelku 1.3 imenovali lokalizacija napake. Kateri parametri določajo izvedbo mehanizma lokalizacije napake, je odvisno od posamezne izvedbe postopka, osnovni postopek in nekatere v strokovnih prispevkih predstavljene posplošitve (na primer De Waal et al., 2011, Pannekoek et al., 2013, Eurostat, 2007) izpostavljajo predvsem naslednja dva parametra, ki ju bomo vključili v naš model:

- Število kontrol, v katerih je določena spremenljivka eksplicitno vključena in so signalizirale napako. Naj bo $n_e^{ij}(t)$ to število za spremenljivko j pri neki enoti i v časovnem obdobju t.

- Utež zanesljivosti spremenljivke. Spremenljivki, za katero je manjša verjetnost nepravilnega poročanja, ima večjo utež zanesljivosti. Naj bo w_j^z utež zanesljivosti za spremenljivko j. Predpostavka je, da je ta utež časovno nespremenljiva.

Po analogiji z mehanizmom M_s tudi tu opredelimo mehanizem K_s kot binarno funkcijo, definirano na naboru dvomljivih enot. Vrednost funkcije bo enaka 1, če bo vrednost spremenljivke dvomljiva, sicer pa bo njena vrednost 0.

$$m_d = \sum_{j=1}^{m} Ks(y_{ij}(t)) = \sum_{j=1}^{m} Ks(n_e^{ij}(t), w_j^z) \tag{41}$$

$$m_a = m - m_d \tag{42}$$

Cilj učinkovitega sistema urejanja podatkov je, da se izhodni rezultati mehanizmov Ke in Ks čimbolj ujemajo z izhodnimi rezultati mehanizmov Me in Ms ob čim manjših stroških in čim manjšem številu sprememb, ki jih izvedemo na podatkih. To lahko dosežemo tudi s primerno uporabo naslednjega mehanizma.

Gre za še en mehanizem, ki deluje na naboru dvomljivih enot, in sicer mehanizem, ki za vsako od dvomljivih enot določi, ali jo bomo obravnavali s postopki ročnega ali s postopki avtomatskega urejanja. Mehanizem torej razdeli n_d enot na n_{dr} enot, ki gredo v ročno, in n_{da} enot, ki gredo v avtomatsko urejanje. Ta mehanizem, katerega implementacijo smo v razdelku 1.2 imenovali funkcijo pomembnosti, bomo označevali s funkcijo U:

$$n_d \xrightarrow{U} (n_{dr}, n_{da}) \tag{43}$$

Nekaj podrobnosti o delovanju mehanizma U smo navedli že v razdelku 1.2, tu sedaj podajamo še malo bolj formaliziran opis. Kot je bilo navedeno v 1.2, lahko v mehanizem vključimo celo vrsto parametrov, za naš namen pa bomo predpostavili nekoliko poenostavljen mehanizem, katerega implementacijo bomo tudi predstavili v naslednjem poglavju. Predpostavka tu bo, da delovanje funkcije pomembnosti določata dva parametra: nabor pričakovanih vrednosti in utež enote. Nadalje še predpostavimo, da mehanizem deluje zgolj na enotah v eni časovni točki, kar je tudi povsem realna predpostavka v resnični implementaciji mehanizma. Podatke iz preteklih obdobij sicer uporabimo pri določitvi pričakovanih vrednosti, samo razdelitev enot pa izvedemo samo za nabor podatkov tekočega obdobja.

Naj bosta torej $\mathbf{Y}^p$ in $\widehat{\mathbf{Y}}^p$ po vrsti dvodimenzionalni matriki, dimenzije (n_d, m) poročanih in pričakovanih vrednosti za nabor dvomljivih enot, w pa vektor uteži za n_d dvomljivih enot. Vpeljimo najprej funkcijo $d(\mathrm{w}\cdot\mathbf{Y}^p, \mathrm{w}\cdot\widehat{\mathbf{Y}}^p)$, ki določa razdaljo med uteženimi poročanimi in uteženimi pričakovanimi vrednostmi. Ta razdalja je lahko, na primer, vsota absolutnih razlik

med uteženo poročano in uteženo pričakovano vrednostjo pri vsaki od obravnavanih spremenljivk. Formalno gledano je d funkcija, ki slika v prostor $(\mathbb{R}_0^+)^{n_d}$, saj vsaki enoti priredi neko nenegativno realno število. Mehanizem U nato razdeli dvomljive enote na nabor tistih, ki jih obravnavamo z ročnim, in nabor tistih, ki jih obravnavamo z avtomatskim urejanjem, na podlagi vrednosti dobljenega vektorja dimenzije n_d. Zapišemo torej lahko:

$$U = U(d\left(\mathrm{w}\cdot\mathbf{Y}^p,\mathrm{w}\cdot\widehat{\mathbf{Y}}^p\right)) \tag{44}$$

Na koncu definirajmo še en mehanizem, in sicer mehanizem, ki pri enotah z neustreznimi vrednostmi, vsaki od m_d vrednosti, ki smo jih prej določili kot nepravilne, določimo novo vrednost. Ta mehanizem bomo označevali s funkcijo V:

$$\left(y_{..1}, y_{..2}, \dots, y_{..m_d}\right) \xrightarrow{V} \left(y'_{..1}, y'_{..2}, \dots, y'_{..m_d}\right) \tag{45}$$

Pri podrobnejši specifikaciji mehanizma V se bomo naslonili na drugi in tretji princip Fellegi-Holt pristopa. Na kratko: postopek vstavljanja naj upošteva pravila logičnih kontrol; vstavljene vrednosti naj v čim večji možni meri ohranjajo porazdelitev podatkov brez napak. Tudi v tem primeru se bomo omejili na opis mehanizma v primerih izvajanja avtomatskih popravkov, saj je postopek določanja nove vrednosti z vidika teoretskega opisa zanimiv le v tem primeru. Postopki ročnega urejanja, ki večinoma temeljijo na kognitivnih procesih posameznika in direktno določijo pravilno vrednost, niso predmet obravnave te naloge. Naj bo torej $y_{ij}(t)$ podatek pri enoti i iz nabora n_{dr} dvomljivih enot, ki jih obravnavamo s postopki avtomatskega urejanja, za spremenljivko Y_j iz nabora m_d dvomljivih spremenljivk, v izbranem časovnem obdobju t. Naj bo nadalje $\mathbf{Y}$ tridimenzionalna matrika opazovanih podatkov, $\mathcal{K}$ za te podatke opredeljen nabor logičnih kontrol, $\mathbf{Y}^S$ pa pod-matrika z naborom podatkov vseh sprejemljivih podatkov. Novo ocenjeno vrednost $\widehat{y_{ij}}(t)$, potem lahko zapišemo kot:

$$\widehat{y_{ij}}(t) = V(y_{ij}(t), \mathcal{K}, F_{\mathbf{Y}}s) \tag{46}$$

kjer je $F_{\mathbf{Y}}s$ porazdelitvena funkcija za $\mathbf{Y}^S$.

Kot primer konkretne implementacije funkcije V si poglejmo zelo enostaven primer podatkov z eno spremenljivko Y_0, v naboru kontrol pa imamo samo dve spremenljivki tipa (35) in (36), torej $\mathcal{K} = \{y_{i0}(t) < s_{i0}(t), y_{i0}(t) > z_{i0}(t)\}$, meje v teh kontrolah pa so določene glede na funkcijo g, kjer velja $g = g(Y_0(t), Y_0(t-1))$. Naj bo $y_{i0}(t)$ dvomljiva vrednost, ki bi jo radi na novo ocenili. Postopek določitve nove vrednosti lahko potem poteka v naslednjih korakih:

- Za nabor enot s sprejemljivimi podatki S, izračunamo množico vrednosti $\{r_i\}$, kjer je $\{r_i\} = \{g(y_{i0}(t)\}_{i \in S}$.
- Iz nabora S slučajno izberemo enoto d. To enoto imenujemo **darovalec**.
- Novo ocenjeno vrednost potem izračunamo kot:

$$\widehat{y_{i0}}(t) = g^{-1}(y_{i0}(t-1), g\big(y_{d0}(t-1), y_{d0}(t)\big)) \tag{47}$$

V primeru, da za funkcijo g vzamemo razmerje $g\big(Y_0(t), Y_0(t-1)\big) = Y_0(t)/Y_0(t-1)$, je ocenjena vrednost enaka vrednosti iz preteklega obdobja, pomnoženi s koeficientom rasti pri darovalcu:

$$\widehat{y_{i0}}(t) = y_{i0}(t-1) \cdot (y_{d0}(t)/y_{d0}(t-1)) \tag{48}$$

V tem primeru imenujemo postopek izračuna nove vrednosti »metoda razmerja darovalca« (angl. *donor ratio imputation*).

5 EMPIRIČNA ANALIZA

5.1 Kratkoročni raziskovanji na SURS

5.1.1 Uvodna pojasnila

Empirična analiza bo izvedena na podatkih dveh kratkoročnih raziskovanj, ki jih izvaja SURS: Prihodek od prodaje in vrednosti zalog v industriji (v nadaljevanju IND-PN/M) in Mesečno statistično raziskovanje o trgovini na drobno, trgovini z motornimi vozili in popravilih motornih vozil (v nadaljevanju TRG/M). V nadaljevanju najprej podajamo nekaj osnovnih informacij o vsakem raziskovanju.

5.1.2 Raziskovanje IND-PN/M

5.1.2.1 Splošni podatki o raziskovanju

Glavni namen raziskovanja je mesečna ocena gibanja prihodka od prodaje ter gibanja vrednosti zalog v industriji. Pomemben kazalnik, ki ga ocenjujemo na podlagi tega raziskovanja, je tudi indeks industrijske proizvodnje. Spremljanje mesečnega gibanja indeksa industrijske proizvodnje je pomembno za hitro ugotavljanju sprememb v gospodarskem razvoju države (Češek Vozel, 2014). Enota opazovanja je po uredbi (Uredba Sveta št. 1165/98) enota enovrstne dejavnosti. Uredba tudi določa rok za pošiljanje podatkov Eurostatu, prve podatke raziskovanja (indeks industrijske proizvodnje) je potrebno poslati 45 dni od konca referenčnega obdobja. Podatke izbrane enote poročajo preko spletnega

vprašalnika, le manjši del enot (okrog 7 %) podatke še vedno sporoča po pošti preko papirnega vprašalnika.

5.1.2.2 Izbor enot opazovanja

Za izbor enot opazovanja je uporabljen postopek zajema s pragom (Benedetti, Bee, & Espa, 2010), pri čemer sta kot osnova za določitev praga upoštevani spremenljivki število zaposlenih oseb in letni prihodek od prodaje. V nabor enot opazovanja se tako vključijo vse enote, ki imajo vsaj 18 zaposlenih oseb, če so bila v raziskovanje vključena že v preteklem letu, oziroma 22 zaposlenih oseb, če so v raziskovanje vključena na novo. V nekaterih dejavnostih, kjer enote, izbrane po tem kriteriju, ne dosegajo želenega praga 75 % deleža zaposlenih, se vključi še ustrezno število dodatnih enot z manjšim številom zaposlenih oseb (Češek Vozel, 2014). Prav tako se v izbor vključi vse enote, ki presegajo prag letnega prihodka od prodaje, ki pa se lahko skozi leta nekoliko spreminja. Izbor enot torej poteka s podobnim postopkom, kot je na Sliki 10 prikazan za raziskovanje TRG/M, le da v tem primeru uporabimo samo prvi korak v postopku izbora. V zajem je vključenih približno 8 % enot ciljne populacije oziroma približno 77 % števila zaposlenih oseb v populaciji (Češek Vozel, 2014).

5.1.2.3 Statistična obdelava podatkov

Statistična obdelava podatkov se izvaja skozi naslednje korake (Češek Vozel, 2014):

- kontrola ter urejanje podatkov,
- ocena manjkajočih vrednosti,
- deflacija,
- izračun indeksov.

V naslednjih razdelkih bomo podrobneje obravnavali predvsem prva dva koraka obdelave.

5.1.2.4 Uporabljeni podatki

Za empirično analizo bomo uporabili mikro podatke za obdobje dveh let, in sicer od januarja 2014 do decembra 2015. Uporabljeni bodo naslednji podatki:

- Originalni podatki, kot so jih poročale opazovane enote. Te podatke bomo imenovali »surovi podatki«.
- Podatki, ki so bili v procesu izvajanja raziskovanja urejeni s postopki ročnega urejanja. Te podatke bomo imenovali »urejeni podatki«.
- Podatki o enotah opazovanja na letni ravni. Uporabljeni bodo podatki o številu zaposlenih ter o dejavnosti (po Standardni klasifikaciji dejavnosti) opazovane enote.

V okviru raziskovanja se sicer opazuje pet spremenljivk, vendar bomo mi v našo analizo vključili le tri. Podatkovna matrika bo v tem primeru torej izgledala takole:

$$\mathbf{Y}(\mathrm{t}) = \begin{bmatrix} y_{11}(t), y_{12}(t), y_{13}(t) \\ y_{21}(t), y_{22}(t), y_{23}(t) \\ \cdots \\ y_{n1}(t), y_{n2}(t), y_{n3}(t) \end{bmatrix}_{\{t=t_1, t_2, \ldots, t_{24}\}} \tag{49}$$

Analizirane opazovane spremenljivke so:

Y_1…Prihodek od prodaje na domačem trgu
Y_2…Prihodek od prodaje na tujem trgu
Y_3…Vrednost zalog

Podobno kot v primeru raziskovanja TRG/M (glej razdelek 5.1.2), bomo tudi tu privzeli predpostavko o enotnem, nespremenljivem številu enot opazovanja. V našem primeru je za analizirano obdobje $n \cong 1.900$.

5.1.3 Raziskovanje TRG/M

5.1.3.1 Splošni podatki o raziskovanju

Namen mesečnega statističnega raziskovanja TRG/M je spremljanje gibanja prihodka od prodaje v dejavnostih, navedenih v nazivu raziskovanja. V raziskovanju torej spremljamo zgolj eno spremenljivko, Prihodek od prodaje. Podatki o prihodku se nanašajo na celo podjetje, tudi na morebitne stranske dejavnosti, ki ne spadajo v ciljne dejavnosti (Lunder & Seljak, 2010).

Ena glavnih značilnosti raziskovanja je dejstvo, da so podatki o prihodku od prodaje zbrani iz dveh različnih virov. Za večja, za raziskovanje bolj pomembna podjetja, podatke pridobimo z anketo. Enote, vključene v to raziskovanje, podatke večinsko poročajo preko spletne aplikacije. Možnost spletnega poročanja je bila uvedena v začetku leta 2014, pred tem so vse poročevalske enote podatke poročale po pošti, na papirnem vprašalniku. Manjši del enot (približno 6 %) tudi še zdaj poroča podatke s papirnim vprašalnikom. Enote, za katere so podatki pridobljeni z anketnim vprašalnikom, bomo v nadaljevanju imenovali »terenske enote«. Za manjša podjetja se podatek o prihodku od prodaje izpelje iz podatkov, ki jih SURS pridobi od Finančne uprave Republike Slovenije (v nadaljevanju FURS) in so prvotno namenjeni obračunu davka na dodano vrednost (»podatki DDV«). SURS je v okviru izvajanja kratkoročnih statistik začel uporabljati podatke DDV leta 2006 (Marolt & Seljak, 2006), za raziskovanje TRG/M pa leta 2008 (Seljak, 2008). Enote, za katere se podatek o prihodku od prodaje izpelje iz administrativnih podatkov, bomo v nadaljevanju imenovali »DDV enote«. Dejstvo, da so podatki raziskovanja pridobljeni iz dveh različnih virov, tudi

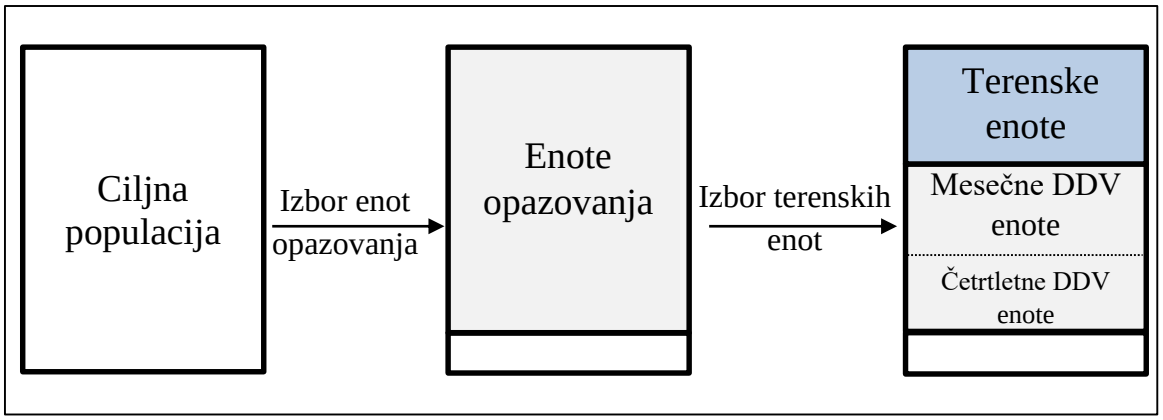
v veliki meri določa postopek izbora enot opazovanj ter postopek urejanja podatkov, ki ju bomo nekoliko podrobneje predstavili v nadaljevanju.

Še ena pomembna značilnost raziskovanja TRG/M je dejstvo, da so po uredbi (Uredba Sveta št. 1165/98) države članice zavezane poročati Eurostatu rezultate tega raziskovanja že v roku 30 dni po koncu referenčnega obdobja. Ker SURS pridobi podatke DDV šele 45 dni po koncu referenčnega obdobja, so ti prvi rezultati, poslani na Eurostat in tudi objavljeni na spletni strani SURS, ocenjeni zgolj na podlagi podatkov enot, vključenih v statistično raziskovanje.

5.1.3.2 Izbor enot opazovanja

Postopek izbora enot temelji na zajemu s pragom. Pri tem v postopku določamo dva praga. Prvi prag nam določi nabor vseh enot opazovanja, drugi prag pa nabor terenskih enot, ki jih vključimo v statistično raziskovanje (Seljak & Zaletel, 2007). Na tem mestu ne bomo podrobneje opisovali kriterijev, ki določajo oba praga, navedimo le, da je ključna spremenljivka v postopku določitve pragov prihodek od prodaje v preteklem letu, kot pomožna spremenljivka pa se uporabi tudi število zaposlenih konec preteklega leta. V nabor opazovanih enot je izbranih približno 30 % enot ciljne populacije, za poročanje v statističnem raziskovanju pa približno 8 % opazovanih enot. Prihodek od prodaje pri enotah, ki so izbrane kot enote opazovanja, predstavlja približno 97 % prihodka v celotni populaciji, prihodek od prodaje terenskih enot pa približno 80 % prihodka enot opazovanja (SURS, 2008). Med enotami DDV je del takih, ki podatke posredujejo le s četrtletno periodiko. Mesečne podatke teh enot je potrebno oceniti z modelno oceno. Delež teh enot sicer v času izbora enot še ni točno poznan, ga pa lahko precej dobro ocenimo. Število »četrtletnih enot DDV« predstavlja približno 12 % vseh enot opazovanja, prihodek od prodaj teh enot pa približno 0,5 % vsega prihodka od prodaje. Postopek izbora z dvema pragoma prikazuje Slika 10.

Slika 10: Postopek izbora enot opazovanja in terenskih enot



5.1.3.3 Statistična obdelava podatkov

Postopek statistične obdelave podatkov poteka v več korakih (Lunder & Seljak, 2010):

- kontrola ter urejanje podatkov pri terenskih enotah,
- kontrola in urejanje podatkov DDV enot,
- združevanje podatkov iz obeh virov,
- ocena manjkajočih podatkov,
- deflacija,
- izračun indeksov.

Postopke statistične obdelave, ki jih lahko uvrstimo v področje urejanja podatkov, bomo podrobneje razdelali v naslednjem razdelku.

5.1.3.4 Uporabljeni podatki

V okviru empirične analize bomo uporabili mikro podatke raziskovanja TRG/M za 24 mesecev, in sicer za obdobje januar 2014 do december 2015. Uporabljeni bodo naslednji podatki:

- Originalni terenski podatki, kot so jih poročale opazovane enote. Poročana je zgolj ena spremenljivka, prihodek od prodaje. Te podatke bomo v nadaljevanju imenovali »surovi terenski podatki«.
- Urejeni terenski podatki, ki so bili v okviru izvajanja raziskovanja urejeni s postopki ročnega urejanja. Te podatke bomo v nadaljevanju imenovali »urejeni terenski podatki«.
- Podatki DDV, kot jih je SURS pridobil od FURS. Ker s stališča SURS gre še za neurejene podatke, bomo te podatke v nadaljevanju imenovali »surovi podatki DDV«. SURS sicer pridobi vse postavke, ki jih podjetja FURS poročajo v namen obračuna DDV, vendar bo kot podatek uporabljena samo ocena prihodka od prodaje, ki je iz teh postavk izpeljana.
- Podatki o vseh enotah, ki so določene kot enote opazovanja v raziskovanju TRG/M. Ker se izbor izvaja enkrat letno, so ti podatki vezani zgolj na leto kot referenčno obdobje. Iz tega nabora bomo za naše analize uporabili podatke o letnem prihodku o prodaji[8], podatek o številu zaposlenih, podatek o glavni dejavnosti (po Standardni klasifikaciji dejavnosti) enote ter podatek o tem, kateri podatkovni vir je bil pri enoti uporabljen (terensko raziskovanje, podatki DDV).

Ker v okviru tega raziskovanja opazujemo in obdelujemo zgolj eno spremenljivko, je naša podatkovna matrika v tem primeru dvodimenzionalna:

[8] Vir za letni prihodek je pri enotah, ki so bile vključene v nabor terenskih enot preteklega leta, poročani podatek v preteklem letu, za ostale enote pa letni prihodek ocenimo iz podatkov DDV preteklega leta.

$$\mathbf{Y}(\mathrm{t}) = \begin{bmatrix} y_{11}(t) \\ y_{21}(t) \\ \cdots \\ y_{n1}(t) \end{bmatrix}_{\{t=t_1, t_2, \ldots, t_{24}\}} \tag{50}$$

Število opazovanih enot se skozi leta spreminja, ker se spreminja sama populacija, nekoliko pa se to število spreminja celo skozi mesece, saj se enote, ki postanejo neustrezne (na primer prenehajo poslovati ali zamenjajo glavno dejavnost) iz raziskovanja izločajo. Za naš model bomo kljub temu predpostavili, da je n konstantno število, in sicer si nabor opazovanih enot predstavljajmo kot unijo vseh enot, ki so bile v kateremkoli mesecu vključene kot opazovana enota. V podatkovni matriki bo enota, ki v resnici v raziskovanju v nekem referenčnem obdobju ni bila opazovana, imela manjkajočo vrednost opazovane spremenljivke. Za obdobje, izbrano za našo empirično analizo, je $n \cong 2900$.

5.2 Implementacija modelnih mehanizmov

5.2.1 Izbor enot opazovanja

5.2.1.1 Raziskovanje IND-PN/M

Ker so v tem raziskovanju za vse enote opazovanja podatki zbrani s terenskim statističnim raziskovanjem, je tu mehanizem R, s katerim delimo enote glede na različen vir podatkov, povsem trivialen. Velja namreč:

$$(n_{sr}, n_{ad}, n_{mo}) = (1 \cdot n, 0 \cdot n, 0 \cdot n) \tag{51}$$

5.2.1.2 Raziskovanje TRG/M

Z oznakami, uvedenimi v razdelku 4.2.3, lahko zapišemo:

$$(n_{sr}, n_{ad}, n_{mo}) = R\big(n, F_{Y_k}, d_{sr}, d_{ad}\big) \tag{52}$$

kjer je n število opazovanih enot, F_{Y_k} kumulativna porazdelitvena funkcija ocene letnega prihodka od prodaje opazovanih enot, d_{sr} delež terenskih enot in d_{ad} delež DDV enot. S podatki, navedenimi v razdelku 5.1.3, lahko nadalje zapišemo:

$$(n_{sr}, n_{ad}, n_{mo}) \approx (0{,}08 \cdot n; 0{,}8 \cdot n; 0{,}12 \cdot n) \tag{53}$$

$$F_{Y_k}(n_{sr}) \approx 0{,}8 \tag{54}$$

$$F_{Y_k}(n_{sr} + n_{ad}) \approx 0{,}995 \tag{55}$$

Predvsem pomemben je ciljni delež prihodka od prodaje, ki ga prispevajo terenske enote, saj ta ciljni delež v bistvu določa celotno razdelitev enot opazovanja v tri skupine. Poleg tega je ta ciljni delež edini, ki je povsem v rokah izvajalcev raziskovanja. Zato še posebej zapišimo:

$$n_{sr} \approx F_{Y_k}^{-1}(0{,}8) \tag{56}$$

5.2.2 Validacija podatkov

V okviru validacije podatkov bomo pod drobnogled vzeli mehanizme, ki določajo razdelitev podatkov v kategorije pravilnosti/nepravilnosti ter sprejemljivosti/dvomljivosti, kot smo jih opredelili v razdelku 1.1. Z oznakami, ki smo jih uvedli v razdelku 4.2.3, gre za mehanizme, konkretizirane s funkcijami Me, Ms, Ke ter Ks.

5.2.2.1 Raziskovanje IND-PN/M

Poleg mikro podatkov raziskovanja, opisanih v razdelku 5.1.2, bomo v naši analizi uporabili še nabor logičnih kontrol. V tem razdelku bomo uporabili nabor kontrol, ki jih na SURS uporabljajo v trenutni izvedbi raziskovanja. Pri tem omenimo, da bomo v analizo vključili le tisto podmnožico logičnih kontrol, ki se nanaša direktno na vsebino analiziranih spremenljivk (Tabela 2). Da bi model ostal preglednejši, smo izpustili nekatere kontrole, ki se nanašajo na pravilnost statusa poročanja enote (odgovor, neodgovor, neustrezna enota) in manjkajoče podatke (neodgovor enote in neodgovor spremenljivke). Vse kontrole so si vsebinsko zelo podobne, in sicer preverjajo spremembo poročanega prihodka glede na preteklo obdobje. Kot dvomljive so zaznane enote, katerih razmerje s podatkom iz preteklega meseca je preseglo 4 pri vrednostih do 20.000 evrov, oziroma 1,5 pri poročanih vrednostih nad 20.000 evrov. Podobno so zaznana dvomljiva zmanjšanja poročanega prihodka, v tem primeru z mejnima razmerjema 0,2 in 0,5.

Tabela 2: Logične kontrole raziskovanja IND-PN/M

Oznaka kontrole	Pravilo kontrole
LK1	$Y1(t)+Y2(t)+Y3+ Y3(t-1) < 0$
LK2	$Y1(t) \geq 0 \wedge Y1(t-1)>0 \wedge Y1 \leq 20.000 \wedge Y1 > 4 * Y1(t-1)$
LK3	$Y1(t) \geq 0 \wedge Y1(t-1)>0 \wedge Y1 > 20.000 \wedge Y1 > 1,5 * Y1(t-1)$
LK4	$Y2(t) \geq 0 \wedge Y2(t-1)>0 \wedge Y2 \leq 20.000 \wedge Y2>4 * Y2(t-1)$
LK5	$Y2(t) \geq 0 \wedge Y2(t-1)>0 \wedge Y2 > 20.000 \wedge Y2 > 1,5 * Y2(t-1)$
LK6	$Y3(t) \geq 0 \wedge Y3(t-1)>0 \wedge Y3 \leq 20.000 \wedge Y3>4 * Y3(t-1)$
LK7	$Y3(t) \geq 0 \wedge Y3(t-1)>0 \wedge Y3 > 20.000 \wedge Y3 > 1,5 * Y3(t-1)$
LK8	$Y1(t) \geq 0 \wedge Y1(t-1)>0 \wedge Y1 \leq 20.000 \wedge Y1<0,2 * Y1(t-1)$
LK9	$Y1(t) \geq 0 \wedge Y1(t-1)>0 \wedge Y1 > 20.000 \wedge Y1 <0,5 * Y1(t-1)$
LK10	$Y2(t) \geq 0 \wedge Y2(t-1)>0 \wedge Y2 \leq 20.000 \wedge Y2<0,2 * Y2(t-1)$
LK11	$Y2(t) \geq 0 \wedge Y2(t-1)>0 \wedge Y2 > 20.000 \wedge Y2 <0,5 * Y2(t-1)$
LK12	$Y3(t) \geq 0 \wedge Y3(t-1)>0 \wedge Y3 \leq 20.000 \wedge Y3<0,2 * Y3(t-1)$
LK13	$Y3(t) \geq 0 \wedge Y3(t-1)>0 \wedge Y3 > 20.000 \wedge Y3 <0,5 * Y3(t-1)$

Vir: SURS, Seznam logičnih kontrol raziskovanja IND-PN/M, b.l.

Čeprav velika večina kontrol, ki jih predstavlja Tabela 2 (LK2-LK13), preverja longitudinalno dimenzijo podatkov, so to kontrole s fiksnimi koeficienti, ki so neodvisni od dejansko opazovane porazdelitve longitudinalnih podatkov. Zato vse te kontrole, tako kot kontrolo LK1, uvrščamo v skupino kontrol linearne neenačbe. Simulacijo prilagoditve trenutnih kontrol na kontrole longitudinalne porazdelitve bomo predstavili v nadaljevanju.

Sledi analiza rezultatov pravilnosti in sprejemljivosti surovih podatkov raziskovanja glede na zgoraj opredeljene kontrole. Ker so bili pri vseh podatkih, ki niso zadoščali pogojem vsaj ene od logičnih kontrol, uporabljeni postopki ročnega urejanja, lahko jemljemo urejene podatke za pravilne podatke. Osnovni pogoj, da tako analizo sploh lahko izvedemo, je, da imamo na razpolago izvorne poročane podatke. V primeru raziskovanja IND-PN/M so nam bili taki surovi podatki na razpolago, vendar ne za povsem vse poročane podatke. Za enote, ki so poročale podatke v papirni obliki in zelo pozno, zaradi samega postopka vnosa podatkov surovi podatki niso na voljo. Ker gre le za manjši delež enot, to ne bi smelo bistveno vplivati na rezultate analize. Tabela 3 prikazuje število zaznanih napak po posameznih kontrolah. Ker bi bil prikaz po vseh kontrolah in za vse mesece nekoliko nepregleden, podajamo le zbirni prikaz za obe opazovani leti.

Tabela 3: Število napak glede na logične kontrole

Kontrola	2014	2015
LK1	153	145
LK2	130	127
LK3	2.341	2.251
LK4	156	169
LK5	1.960	1.979
LK6	25	28
LK7	597	607
LK8	458	433
LK9	693	722
LK10	811	765
LK11	710	679
LK12	204	183
LK13	172	182

Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2015, b.l.

Kot vidimo, izrazito izstopa število zaznanih napak pri kontrolah LK3 in LK5. Pri teh kontrolah gre za kontrolo rasti glede na pretekli mesec pri večjih enotah. Veliko število zaznanih napak kaže na najbrž nekoliko prestrogo postavljeno mejo pri teh kontrolah, je pa tudi posledica že prej omenjenega problema fiksnih koeficientov.

V nadaljevanju podajamo analizo učinkovitosti predstavljenega nabora kontrol. Pri analizi smo najprej na podatkih enot, vključenih v analizo (enote, za katere imamo surove podatke), opravili naslednji dve delitvi:

- Vse analizirane enote razdelimo na tiste, pri katerih je vsaj ena logična kontrola zaznala napako (dvomljive enote), in enote, kjer zaznanih napak ni bilo (sprejemljive enote).
- Vse dvomljive enote razdelimo na tiste, pri katerih je bil popravljen vsaj en podatek (enote z nepravilnimi podatki), in enote, pri katerih ni bil popravljen noben podatek (enote s pravilnimi podatki).

Glede na to razdelitev, lahko nato izračunamo stopnjo zavrnitve in stopnjo zaznanih napak, kot smo jih definirali v razdelku 1.1. Tabela 4 prikazuje rezultate analize za vse mesece let 2014 in 2015.

Tabela 4: Analiza validacije podatkov raziskovanja IND-PN/M

Leto	Mesec	Vse analizirane enote	Enote z dvomljivimi podatki	Enote z dvomljivimi podatki z vsaj enim popravkom	Stopnja zavrnitve (%)	Stopnja zaznanih napak (%)
2014	1	1.450	760	97	52,4	12,8
2014	2	1.432	515	80	36,0	15,5
2014	3	1.438	582	77	40,5	13,2
2014	4	1.462	491	68	33,6	13,8
2014	5	1.435	451	58	31,4	12,9
2014	6	1.471	499	55	33,9	11,0
2014	7	1.452	519	69	35,7	13,3
2014	8	1.439	580	63	40,3	10,9
2014	9	1.441	763	80	52,9	10,5
2014	10	1.430	448	52	31,3	11,6
2014	11	1.374	464	58	33,8	12,5
2014	12	1.399	533	79	38,1	14,8
2015	1	1.343	669	65	49,8	9,7
2015	2	1.426	498	78	34,9	15,7
2015	3	1.425	603	73	42,3	12,1
2015	4	1.449	475	55	32,8	11,6
2015	5	1.420	481	56	33,9	11,6
2015	6	1.428	501	68	35,1	13,6
2015	7	1.429	495	64	34,6	12,9
2015	8	1.433	593	52	41,4	8,8
2015	9	1.435	745	52	51,9	7,0
2015	10	1.428	483	54	33,8	11,2
2015	11	1.387	455	42	32,8	9,2
2015	12	1.403	554	53	39,5	9,6

Vir:SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2015.

Kar je hitro razvidno iz Tabele 4, je precej visoka stopnja zavrnitve po eni in nizka stopnja zaznanih napak po drugi strani. O možnih izboljšavah postopka, s katerimi bi »izboljšali« ti dve stopnji, bomo razpravljali v naslednjem razdelku.

Z oznakami, uvedenimi v razdelku 4.2.3, in z oceno povprečnih deležev enot, pri katerih smo popravljali podatke, lahko zapišemo:

$$M_e(n) = \left(n_n, n_p\right) \approx (0{,}05 \cdot n; 0{,}95 \cdot n) \tag{57}$$

Kot lahko hitro izračunamo iz podatkov Tabeli 4, je približno pri 5% enot popravljen vsaj en podatek. Ta delež smo v naši implementaciji mehanizma M_e uporabili kot približek za delež enot z nepravilnimi podatki. Pri tem omenimo, da je ta podatek verjetno nekoliko podcenjen, saj v izračunu nismo upoštevali kategorije b iz Tabele 1, torej enot, ki jih logične

kontrole niso zaznale kot enote z dvomljivimi podatki, vsebujejo pa nepravilne podatke. Ta delež pač, glede na razpoložljive podatke, ostane skrit. Glede na stopnjo zavrnitve in posledično visokega števila dvomljivih enot, pa lahko predpostavimo, da je delež enot v tej kategoriji zanemarljiv.

V razdelku 4.2.3 smo predpostavili tri parametre, ki določajo delovanje mehanizma M_e. Ker je v primeru tega raziskovanja uporabljen samo en podatkovni vir, ostaneta dva parametra: velikost enote ter število obdobij vključenosti v raziskovanja. Ker to presega okvir naše naloge, na tem mestu ne bomo podrobneje analizirali pravilnosti te predpostavke, ampak bomo zgolj nakazali vpliv/ne-vpliv odvisnosti mehanizma od omenjenih parametrov. V Tabeli 5 je prikazan delež enot s popravljenimi podatki glede na velikostni razred. Velikostne razrede smo definirali na podlagi podatka o številu zaposlenih s standardnima mejama 50 in 250 zaposlenih. Zaradi boljše preglednosti, prikazujemo samo deleže na ravni leta.

Tabela 5: Delež enot z nepravilnimi podatki po velikostnih razredih

Leto	**Velikostni razred**	**Delež enot s popravljenimi podatki (%)**
2014	Mala podjetja (število zaposlenih < 50)	4,3
2014	Srednja podjetja (50 ≤ število zaposlenih < 250)	6,1
2014	Velika podjetja (število zaposlenih ≥ 250)	4,1
2015	Mala podjetja (število zaposlenih < 50)	3,7
2015	Srednja podjetja (50 ≤ število zaposlenih < 250)	5,0
2015	Velika podjetja (število zaposlenih ≥ 250)	5,1

Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2014, b.l.

Prikazani podatki ne nakazujejo vpliva velikosti na verjetnost za nepravilno poročan podatek, je pa res, da bi to vprašanje zahtevalo povsem samostojno analizo in podrobnejši vpogled v poročane in popravljene podatke.

Drug parameter, katerega vpliv bi radi presodili, je število obdobij vključenosti. Ker za analizo tega podatka nismo imeli na voljo, smo uporabili »pomožen podatek«, in sicer podatek o tem, ali je v referenčnem letu enota prvič vključena v raziskovanje (»nove enote«), ali je bila vključena tudi v preteklem letu (»stare enote«). Tabela 6 prikazuje deleže za vse mesece leta 2014, saj je zanimivo tudi gibanje deleža od prve vključitve naprej.

Tabela 6: Delež enot z nepravilnimi podatki glede na vključenost v raziskovanje

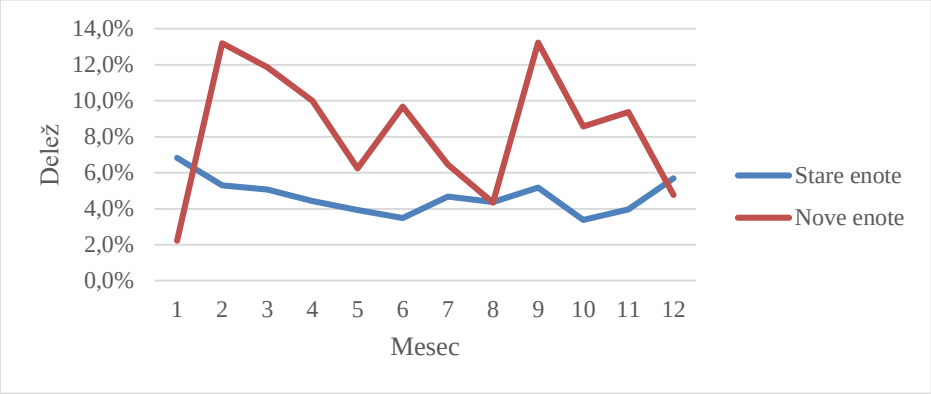
Leto	Mesec	Nova enota (1-DA, 0-NE)	Delež enot s popravljenimi podatki (%)
2014	1	0	6,8
2014	1	1	2,2
2014	2	0	5,3
2014	2	1	13,2
2014	3	0	5,1
2014	3	1	11,9
2014	4	0	4,4
2014	4	1	10,0
2014	5	0	3,9
2014	5	1	6,3
2014	6	0	3,5
2014	6	1	9,7
2014	7	0	4,7
2014	7	1	6,5
2014	8	0	4,4
2014	8	1	4,3
2014	9	0	5,2
2014	9	1	13,2
2014	10	0	3,4
2014	10	1	8,6
2014	11	0	4,0
2014	11	1	9,4
2014	12	0	5,7
2014	12	1	4,8

Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2014, b.l.

V večini mesecev je opazna precejšnja razlika v deležu enot s popravljenimi podatki v obeh kategorijah. Delež popravkov pri novo-vključenih enotah je izrazito manjši zgolj v prvem mesec, torej prav v mesecu, ko enote prvič poročajo podatke. Razlog za to je precej preprost. Ob prvem poročanju ni možno izvajati kontrol za iskanje longitudinalnih osamelcev, kar pomeni, da ostaja samo ena kontrola doslednosti. Kontrola longitudinalne komponente podatkov se torej lahko začne izvajati šele z drugim mesecem poročanja.

Za bolj nazoren prikaz gibanja deležev skozi čas na Slika 11 podajamo še linijski diagram.

Slika 11: Delež enot s popravki v letu 2014



Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2014, b.l.

Mehanizem M_s deluje samo na naboru enot, ki imajo vsaj en nepravilen podatek, torej na približno 5 % vseh opazovanih enot, in deli vse opazovane spremenljivke na tiste z nepravilnimi in tiste s pravilnimi podatki. Delovanje mehanizma določa en tridimenzionalni parameter, ki za vsako spremenljivko določa verjetnost nepravilnega podatka: (w_1, w_2, w_3). Ta parameter smo v razdelku 4.2.3 imenovali utež kompleksnosti.

Tabela 7 prikazuje rezultat empirične analize, s katero smo za vsako spremenljivko raziskovanja IND-PN/M ugotavljali število ter delež nepravilnih podatkov v naboru enot z vsaj enim nepravilnim podatkom.

Tabela 7: Število in delež popravkov po spremenljivkah

Leto	Mesec	Število enot s popravki	Število popravkov po spremenljivkah			Delež popravkov po spremenljivkah (%)		
			Y1	Y2	Y3	Y1	Y2	Y3
2014	1	97	50	31	54	51,5	32,0	55,7
2014	2	80	43	28	40	53,8	35,0	50,0
2014	3	77	40	27	34	51,9	35,1	44,2
2014	4	68	35	23	36	51,5	33,8	52,9
2014	5	58	31	18	27	53,4	31,0	46,6
2014	6	55	27	22	26	49,1	40,0	47,3
2014	7	69	35	23	36	50,7	33,3	52,2
2014	8	63	30	19	32	47,6	30,2	50,8
2014	9	80	41	26	37	51,3	32,5	46,3
2014	10	52	29	17	24	55,8	32,7	46,2
2014	11	58	31	18	26	53,4	31,0	44,8
2014	12	79	38	29	38	48,1	36,7	48,1
2015	1	65	26	17	43	40,0	26,2	66,2
2015	2	78	28	28	45	35,9	35,9	57,7
2015	3	73	35	24	39	47,9	32,9	53,4
2015	4	55	22	19	34	40,0	34,5	61,8
2015	5	56	22	23	33	39,3	41,1	58,9
2015	6	68	32	31	33	47,1	45,6	48,5
2015	7	64	24	27	37	37,5	42,2	57,8
2015	8	52	27	20	28	51,9	38,5	53,8
2015	9	52	20	13	29	38,5	25,0	55,8
2015	10	54	22	19	30	40,7	35,2	55,6
2015	11	42	23	20	23	54,8	47,6	54,8
2015	12	53	26	23	29	49,1	43,4	54,7

Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2015, b.l.

Če obravnavamo mehanizem M_s kot slučajno spremenljivko, deleže, ki jih prikazuje Tabela 7, pa kot realizacijo te slučajne spremenljivke, lahko ocenimo uteži kompleksnosti kot pričakovano vrednost te slučajne spremenljivke oziroma kot aritmetično sredino vseh predstavljenih deležev:

$$\begin{aligned} w_1 &\cong 0{,}48 \\ w_2 &\cong 0{,}35 \\ w_3 &\cong 0{,}53 \end{aligned} \tag{58}$$

Če upoštevamo še ocenjen delež enot z vsaj enim nepravilnim podatkom, lahko za vsako spremenljivko izračunamo oceno nepravilnega poročanja:

$$
\begin{aligned}
P\left(\hat{Y}_1^Z(t) - Y_1^0(t) \neq 0\right) &= P(M_s(y_{i1}(t)) = 1 | M_e(y_i(t)) = 1) \cong 0{,}024 \\
P\left(\hat{Y}_2^Z(t) - Y_2^0(t) \neq 0\right) &= P(M_s(y_{i2}(t)) = 1 | M_e(y_i(t)) = 1) \cong 0{,}018 \\
P\left(\hat{Y}_3^Z(t) - Y_3^0(t) \neq 0\right) &= P(M_s(y_{i3}(t)) = 1 | M_e(y_i(t)) = 1) \cong 0{,}027
\end{aligned} \tag{59}
$$

Predpostavka, da so zgoraj zapisane verjetnosti konstantne skozi čas, ni povsem realna, saj Tabela 7 izkazuje kar precejšnjo variabilnost deležev skozi čas, vendar bo za namene našega modela ta predpostavka povsem sprejemljiva poenostavitev.

Naslednja dva mehanizma, to je mehanizma K_e in K_s, smo v resnici vključili že v zgoraj predstavljene analize. Rezultati, predstavljeni v Tabelah 3 in 4, so namreč dobljeni na podlagi implementacije (tudi) teh dveh mehanizmov. Glede na podatke v Tabeli 4, lahko za rezultat delovanja mehanizma K_e zapišemo oceno:

$$
K_e(n) = (n_d, n_a) \approx (0{,}39 \cdot n, 0{,}61 \cdot n) \tag{60}
$$

Ker se v rednem postopku izvajanja raziskovanja ne izvaja postopkov avtomatskega urejanja, tudi ni izvedena lokalizacija napake. Za oceno izhodnih parametrov mehanizma K_s bomo izhajali kar iz podatkov, ki jih prikazuje Tabela 3. Za enote, za katere je napako signalizirala kontrola LK1, v kateri eksplicitno nastopajo vse tri spremenljivke, bomo kot dvomljive smatrali vse tri spremenljivke. V vseh drugih kontrolah nastopa samo po ena spremenljivka, zato lokalizacija napake tu izhaja iz same kontrole. Na podlagi tega pristopa lahko izpeljemo oceno:

$$
K_s(Y_1, Y_2, Y_3) = \begin{bmatrix} m_d(Y_1),\ m_a(Y_1) \\ m_d(Y_2),\ m_a(Y_2) \\ m_d(Y_3),\ m_a(Y_3) \end{bmatrix} = \begin{bmatrix} 0{,}57 \cdot n_d, 0{,}43 \cdot n_d \\ 0{,}57 \cdot n_d, 0{,}43 \cdot n_d \\ 0{,}17 \cdot n_d, 0{,}83 \cdot n_d \end{bmatrix} \tag{61}
$$

Če združimo oceni izhodnih parametrov mehanizmov K_e, K_s, lahko za vsako spremenljivko ocenimo še verjetnost (p_n), da bo zaradi podatkov te spremenljivke signalizirana ena od napak:

$$
\begin{aligned}
p_n(Y_1) &= 0{,}57 \cdot 0{,}39 = 0{,}22 \\
p_n(Y_2) &= 0{,}57 \cdot 0{,}39 = 0{,}22 \\
p_n(Y_3) &= 0{,}17 \cdot 0{,}39 = 0{,}07
\end{aligned} \tag{62}
$$

5.2.2.2 Raziskovanje TRG/M

Ker je v primeru raziskovanja TRG/M opazovana le ena spremenljivka, je nabor logičnih kontrol omejen zgolj na longitudinalno kontrolo podatkov. Tudi v tem primeru ne bomo obravnavali manjkajočih podatkov, saj je za reševanje problema neodgovora enote predviden povsem ločen postopek. Nabor logičnih kontrol, ki se trenutno uporabljajo v okviru validacije podatkov, je sicer različen za primer terenskih enot in za primer DDV enot,

pa tudi uporaba kontrol je v obeh primerih nekoliko drugačna, zato vsako od teh dveh skupin prikazujemo in analiziramo ločeno. Tabela 8 tako prikazuje kontrole, ki se uporabljajo samo za terenske enote.

Tabela 8: Logične kontrole za terenske enote raziskovanja TRG/M

Oznaka kontrole	Pravilo kontrole
LK1	$Y1(t) \geq 0 \wedge Y1(t-1) > 0 \wedge Y1(t) < 0.6 * Y1(t-1)$
LK2	$Y1(t) \geq 0 \wedge Y1(t-1) > 0 \wedge Y1(t) > 1.8 * Y1(t-1)$
LK3	$Y1(t) \geq 0 \wedge Y1(t-12) > 0 \wedge Y1(t) < 0.8 * Y1(t-12)$
LK4	$Y1(t) \geq 0 \wedge Y1(t-12) > 0 \wedge Y1(t) > 1.8 * Y1(t-12)$

Vir: SURS, Seznam logičnih kontrol raziskovanja TRG/M, b.l.

Vse kontrole preverjajo longitudinalni vidik poročanih podatkov pri posameznih enotah. Tudi tu, podobno kot pri kontrolah, predstavljenih v Tabeli 2, imamo kontrole s fiksnimi koeficienti, zato tudi vse te kontrole uvrščamo v skupino kontrol linearne neenačbe.

Tabela 9 prikazuje število zaznanih napak glede na logične kontrole, zbirno za vsako leto.

Tabela 9: Število napak pri terenskih enotah glede na logične kontrole

Kontrola	2014	2015
LK1	175	157
LK2	130	95
LK3	467	359
LK4	107	142

Vir: SURS, Mikro podatki raziskovanja TRG/M, januar 2014 - december 2015, b.l.

Tabela 10 prikazuje rezultate validacije podatkov terenskih enot, kjer so uporabljeni postopki ročnega urejanja. Za vsak dvomljiv podatek torej vemo, ali je bil kasneje določen kot nepravilen ali pravilen. Podobno kot v primeru raziskovanja IND-PN/M, lahko torej tudi tukaj poleg stopnje zavrnitve ocenimo tudi stopnjo zaznanih napak.

Tabela 10: Analiza validacije podatkov terenskih enot raziskovanja TRG/M

Leto	Mesec	Analizirane enote	Enote z dvomljivimi podatki	Enote z dvomljivimi podatki z vsaj enim popravkom	Stopnja zavrnitve (%)	Stopnja zaznanih napak (%)
2014	1	270	69	20	25,6	29,0
2014	2	270	59	13	21,9	22,0
2014	3	270	67	28	24,8	41,8
2014	4	270	55	16	20,4	29,1
2014	5	270	56	18	20,7	32,1
2014	6	270	68	24	25,2	35,3
2014	7	270	52	18	19,3	34,6
2014	8	270	64	17	23,7	26,6
2014	9	270	61	18	22,6	29,5
2014	10	270	62	22	23,0	35,5
2014	11	270	56	19	20,7	33,9
2014	12	270	64	21	23,7	32,8
2015	1	281	82	31	29,2	37,8
2015	2	281	58	19	20,6	32,8
2015	3	281	61	21	21,7	34,4
2015	4	281	48	16	17,1	33,3
2015	5	281	52	22	18,5	42,3
2015	6	281	51	19	18,1	37,3
2015	7	281	47	18	16,7	38,3
2015	8	281	43	17	15,3	39,5
2015	9	281	60	25	21,4	41,7
2015	10	281	46	20	16,4	43,5
2015	11	281	48	18	17,1	37,5
2015	12	281	51	18	18,1	35,3

Vir: SURS, Mikro podatki raziskovanja TRG/M, januar 2014 - december 2015, b.l.

V primerjavi z raziskovanjem IND-PN/M je v tem primeru precej nižja stopnja zavrnitve, precej višja pa stopnja zaznanih napak. To je pričakovano, saj tu analiziramo zgolj podmnožico po prihodku od prodaje največjih (terenskih) enot, pri katerih so manj pogosta izrazita nihanja v časovni vrsti poročanih podatkov posamezne enote.

Tabela 11 prikazuje kontrole, ki jih uporabljamo za kontrolo DDV podatkov.

Tabela 11: Logične kontrole za DDV enote raziskovanja TRG/M

Oznaka kontrole	Pravilo kontrole
LK5	Y1(t-1) > 0 ∧ Y1(t) > HB_PRET_Z ∧ Y1(t+1) >0 ∧ Y1> HB_PRIH_Z
LK6	¬ (Y1(t-1) >0) ∧ Y1(t+1) >0 ∧ Y1> HB_PRIH_Z
LK7	Y1(t-1) >0 ∧ Y1(t) > HB_PRET_Z ∧ ¬ t(Y1(t+1) >0)
LK8	Y1(t-1) >0 ∧ Y1(t) < HB_PRET_S ∧ Y1(t+1) >0 ∧ Y1(t) < HB_PRIH_S
LK9	¬ (Y1(t-1)>0) ∧ Y1(t+1) >0 ∧ Y1(t) < HB_PRIH_Z
LK10	Y1(t-1) >0 ∧ Y1(t) < HB_PRET_S ∧ ¬ (Y1(t+1) >0)

Vir: SURS, Seznam logičnih kontrol raziskovanja TRG/M, interno gradivo.

Te kontrole potrebujejo nekaj dodatnih pojasnil. Gre za kontrole longitudinalne porazdelitve, kjer se meje sprejemljivosti podatka določajo glede na porazdelitve poročanih podatkov. Na primer, kontrola LK5 določa meje sprejemljivosti glede na razmerje poročanega podatka v obdobju t glede na poročan podatek v preteklem obdobju $t-1$ ter glede na poročan podatek v prihodnjem obdobju $t+1$. Podatek je dvomljiv samo, če je tako razmerje s preteklim obdobjem, kot tudi razmerje s prihodnjim obdobjem, zaznano kot osamelec v porazdelitvi razmerij. Na ta način, seveda, lahko preverjamo samo podatke znotraj časovne vrste poročanj, za katere imamo tako pretekli kot tudi prihodnji podatek.

Vsi parametri v kontrolah, ki se začenjajo s predpono HB, so parametri, ki izhajajo iz implementacije HB metode, in smo jih iz mej sprejemljivosti porazdelitve funkcije $g(Y_1)$ dobili s transformacijo $\mathcal{L}^{-1}$ (glej razdelek 4.2.2). V resnici v kontrolah sprejemljivost podatka določamo glede na dve longitudinalni funkciji spremenljivke Y_1:

- V prvem primeru proučujemo porazdelitev razmerij vrednosti spremenljivke v tekočem in preteklem obdobju: $g_1(Y_1) = Y_1(t)/Y_1(t-1)$. Iz analize porazdelitve te funkcije izhajata spodnja in zgornja mejo intervala sprejemljivosti, ki ju lahko zapišemo kot $(\text{HB\_PRET\_S}, \text{HB\_PRET\_Z}) = \mathcal{L}^{-1}(g_1(Y_1))$.
- V drugem primeru proučujemo porazdelitev razmerij vrednosti spremenljivke v tekočem in prihodnjem obdobju: $g_2(Y_1) = Y_1(t)/Y_1(t+1)$. Na prvi pogled se zdi, da taka analiza v praksi ni mogoča, saj ob času pridobivanja podatkov za obdobje t, še nimamo na voljo podatkov za $t+1$, vendar moramo upoštevati, da vedno naenkrat analiziramo podatke za T referenčnih obdobij. Spodnjo in zgornjo mejo intervala sprejemljivosti v tem primeru zapišemo kot $(\text{HB\_PRIH\_S}, \text{HB\_PRIH\_Z}) = \mathcal{L}^{-1}(g_2(Y_1))$.

S tako definiranimi kontrolami v podatkih v resnici iščemo primere longitudinalnih osamelcev, kot jih prikazuje Slika 6. Za vsa »notranja« obdobja je vrednost spremenljivke osamelec, če je izrazit skok tako glede na preteklo kot tudi glede na prihodnje obdobje. Samo ob »robnih obdobjih« (prvo in zadnje analizirano obdobje) podatek opredelimo kot osamelec, če je izrazit skok samo v eni smeri. Vse tri možne primere osamelcev prikazujemo na Sliki 12.

Slika 12: Različni primeri longitudinalnih osamelcev

Ker so pri podatkih DDV enot uporabljeni postopki avtomatskega urejanja, nimamo podatka o pravilnosti/nepravilnosti dvomljivih podatkov, zato lahko v tem primeru izračunamo zgolj stopnjo zavrnitve, ko jo prikazuje Tabela 12.

Tabela 12: Analiza validacije podatkov DDV enot raziskovanja TRG/M

Leto	Mesec	DDV enote	Enote z dvomljivimi podatki	Stopnja zavrnitve (%)
2014	1	2.132	19	0,89
2014	2	2.131	10	0,47
2014	3	2.127	9	0,42
2014	4	2.157	9	0,42
2014	5	2.155	6	0,28
2014	6	2.152	11	0,51
2014	7	2.157	14	0,65
2014	8	2.150	11	0,51
2014	9	2.142	6	0,28
2014	10	2.151	7	0,33
2014	11	2.142	13	0,61
2014	12	2.125	12	0,56
2015	1	2.106	19	0,90
2015	2	2.153	13	0,60
2015	3	2.151	7	0,33
2015	4	2.188	9	0,41
2015	5	2.181	9	0,41
2015	6	2.175	8	0,37
2015	7	2.184	8	0,37
2015	8	2.178	10	0,46
2015	9	2.173	4	0,18
2015	10	2.179	5	0,23
2015	11	2.162	13	0,60
2015	12	2.151	25	1,16

Vir: SURS, Mikro podatki raziskovanja TRG/M, januar 2014 - december 2015, b.l.

Kot vidimo, imamo tu precej nižjo stopnjo zavrnitev kot v primeru terenskih enot, vendar je to razumljivo, saj so v tem primeru vsi dvomljivi podatki obravnavani kot nepravilni in

popravljeni s postopki vstavljanja. Tako je razumljivo, da so tu uporabljeni precej bolj »konservativni« parametri HB metode, ki v končni fazi za vsak podatek določajo mejo sprejemljivih vrednosti.

5.2.3 Selektivno urejanje podatkov

V okviru tega razdelka bomo proučevali v primeru enega raziskovanja že obstoječo oziroma v primeru drugega raziskovanja možno implementacijo mehanizmov U in V, ki v povezavi z že prej opisanima mehanizmoma K_e in K_s predstavljata osnovo za izvajanje postopkov selektivnega urejanja podatkov, ki smo jih s teoretske plati predstavili v razdelku 1.2.

5.2.3.1 Raziskovanje IND-PN/M

V raziskovanju trenutno ne potekajo postopki selektivnega urejanja, zato bomo take postopke za namene tega dela simulirali. V prenovljen postopek bomo vključili postopke, ki smo jih s teoretske plati razdelali v poglavju 1, njihov modelni opis pa predstavili v razdelku 4.2.3. Prenovljen postopek bomo izvedli v treh glavnih korakih:

- Opredelitev prenovljenega nabora logičnih kontrol, ki temelji na določanju osamelih vrednosti na podlagi porazdelitve razmerij.
- Selektivno urejanje podatkov na principu izhodne razdelitve. Vsaki enoti bo glede na poročane podatke izračunana funkcija pomembnosti, na podlagi vrednosti te funkcije pa nato izvedena delitev enot na enote, ki jih obravnavamo s postopki ročnega urejanja, in enote, katerih vrednosti popravljamo s postopki avtomatskega urejanja.
- Določitev novih vrednosti. Enotam, katere smo določili za ročno urejanje, »čiste podatke« samo prepišemo, ostalim enotam pa ocenimo nove vrednosti z na novo postavljenim postopkom avtomatskega urejanja.

V naboru logičnih kontrol, ki se trenutno uporabljajo v raziskovanju, so samo linearne kontrole. V prenovljenem naboru kontrol bomo vse linearne kontrole, ki zaznavajo dvomljive vrednosti v longitudinalni komponenti izbrane spremenljivki, nadomestili s kontrolami longitudinalne porazdelitve, torej kontrolami, ki jih zapišemo v oblikah (35), (36) ali (37). Prenovljen nabor kontrol prikazuje Tabela 13.

Tabela 13: Prenovljen sistem logičnih kontrol pri raziskovanju IND-PN/M

Oznaka kontrole	Pravilo kontrole
LK2	Y1(t)+Y2(t)+Y3+ Y3(t-1) < 0
LK3	Y1(t-1)>0 ∧ Y1(t) >_HB_PRET_Z_Y1 ∧ Y1(t+1)>0 ∧ Y1(t) >_HB_PRIH_Z_Y1
LK4	Y1(t-1)=Null ∧ Y1(t+1)>0 ∧ Y1(t) >_HB_PRIH_Z_Y1
LK5	Y1(t-1)>0 ∧ Y1(t) >_HB_PRET_Z_Y1 ∧ Y1(t+1)=Null*
LK6	Y2(t-1)>0 ∧ Y2(t) >_HB_PRET_Z_Y2 ∧ Y2(t+1)>0 ∧ Y2(t) >_HB_PRIH_Z_Y2
LK7	Y2(t-1)=Null ∧ Y2(t+1)>0 ∧ Y2(t) >_HB_PRIH_Z_Y2
LK8	Y2(t-1)>0 ∧ Y2(t) >_HB_PRET_Z_Y2 ∧ Y2(t+1)=Null
LK9	Y3(t-1)>0 ∧ Y3(t) >_HB_PRET_Z_Y3 ∧ Y3(t+1)>0 ∧ Y3(t) >_HB_PRIH_Z_Y3
LK10	Y3(t-1)=Null ∧ Y3(t+1)>0 ∧ Y3(t) >_HB_PRIH_Z_Y3
LK11	Y3(t-1)>0 AND Y3(t)> _HB_PRET_Z_Y3 ∧ Y3(t+1)=Null
LK12	Y1(t-1)>0 ∧ Y1(t) <_HB_PRET_S_Y1 ∧ Y1(t+1)>0 ∧ Y1(t) <_HB_PRIH_S_Y1
LK13	Y1(t-1)=Null ∧ Y1(t+1)>0 ∧ Y1(t) <_HB_PRIH_S_Y1
LK14	Y1(t-1)>0 ∧ Y1(t) <_HB_PRET_S_Y1 ∧ Y1(t+1)=Null
LK15	Y2(t-1)>0 ∧ Y2(t) <_HB_PRET_S_Y2 ∧ Y2(t+1)>0 ∧ Y2(t) <_HB_PRIH_S_Y2
LK16	Y2(t-1)=Null ∧ Y2(t+1)>0 ∧ Y2(t) <_HB_PRIH_S_Y2
LK17	Y2(t-1)>0 ∧ Y2(t) <_HB_PRET_S_Y2 ∧ Y2(t+1)=Null
LK18	Y3(t-1)>0 ∧ Y3(t) <_HB_PRET_S_Y3 ∧ Y3(t+1)>0 ∧ Y3(t) <_HB_PRIH_S_Y3
LK19	Y3(t-1)=Null ∧ Y3(t+1)>0 ∧ Y3(t) <_HB_PRIH_S_Y3
LK20	Y3(t-1)>0 ∧ Y3(t) <_HB_PRET_S_Y3 ∧ Y3(t+1)=Null
LK21	¬ (Y1(t-1) >0) ∧ ¬ (Y1(t+1) >0) ∧ Y1(t) >Y1_m0
LK22	¬ (Y2(t-1) >0) ∧ ¬ (Y2(t+1) >0) ∧ Y2(t) >Y2_m0
LK23	¬ (Y3(t-1) >0) ∧ ¬ (Y3(t+1) >0) ∧ Y3(t) >Y3_m0

* Oznaka Null je oznaka za manjkajočo vrednost

Kontrola LK1 ostaja enaka prvi linearni kontroli iz trenutnega nabora kontrol. Kontrole LK3-LK20 so kontrole longitudinalne porazdelitve, katerih parametri izhajajo iz HB metode. V fazi simulacije postopkov smo preizkusili več različnih parametrov za HB metodo, rezultati pa bodo prikazani za vrednosti parametrov: $C = 10$, $U = 0{,}7$, $A = 0{,}05$. Ta del kontrol nadomešča kontrole LK2-LK13 iz Tabele Tabela *2*, linearne kontrole s fiksnimi koeficienti, s katerimi se v trenutni izvedbi zaznava izstopajoče vrednosti glede na preteklo oziroma prihodnje obdobje.

Kontrole LK21-LK23 so kontrole, ki smo jih v sistem uvedli na novo. Gre za kontrole oblike (37), kjer za primere, za katere nimamo pozitivnega podatka v preteklem in prihodnjem obdobju, iščemo vrednosti spremenljivk nad določeno mejo, ki izhaja iz porazdelitev spremenljivke v tekočem obdobju. V našem primeru $m(F_{Y_0(t)})$ iz splošnega zapisa kontrole (37) določimo s pomočjo pomožnih spremenljivk $\{z^{Y_j}\}_{j=1,..,3}$, ki jo za obravnavano spremenljivko Y_j, v obdobju t, izračunamo za vsako enoto i kot $z_i^{Y_j}(t) = \frac{y_{ij}(t)}{z_i}$, kjer je z_i število zaposlenih pri enoti i, ki je bilo določeno ob izboru enot opazovanja ob začetku leta.

Gre torej za povprečno vrednost spremenljivke na zaposleno osebo. Mejo Y_j_m0 nato izračunamo kot:

$$Y_{ij}(t)\_m0 = z_i \cdot (\bar{z}^{Y_j}(t) + 0{,}5 \cdot \mathrm{std}\left(z^{Y_j}(t)\right)) \tag{63}$$

kjer je $\bar{z}^{Y_j}(t)$ povprečna vrednost in $\mathrm{std}\left(z^{Y_j}(t)\right)$ standardni odklon porazdelitve pomožne spremenljivke z^{Y_j} v obdobju t.

Prenovljen nabor določa novo realizacijo mehanizma K_e, ki razdeli enote na sprejemljive in dvomljive. Ti dve podmnožici bomo označili z $\mathcal{U}_a$ in $\mathcal{U}_d$. Tabela 14 prikazuje število enot, ki jih uvrščamo v množico $\mathcal{U}_d$, torej enot, za katere je novi nabor logičnih kontrol signaliziral vsaj eno napako.

Tabela 14: Validacija podatkov raziskovanja IND-PN/M z novim naborom logičnih kontrol

Leto	Mesec	Vse analizirane enote	Enote z dvomljivimi podatki	Stopnja zavrnitve (%)
2014	1	1.450	78	5,4
2014	2	1.432	75	5,2
2014	3	1.438	80	5,6
2014	4	1.462	85	5,8
2014	5	1.435	70	4,9
2014	6	1.471	74	5,0
2014	7	1.452	78	5,4
2014	8	1.439	107	7,4
2014	9	1.441	70	4,9
2014	10	1.430	62	4,3
2014	11	1.374	63	4,6
2014	12	1.399	82	5,9
2015	1	1.343	64	4,8
2015	2	1.426	69	4,8
2015	3	1.425	97	6,8
2015	4	1.449	88	6,1
2015	5	1.420	76	5,4
2015	6	1.428	79	5,5
2015	7	1.429	77	5,4
2015	8	1.433	80	5,6
2015	9	1.435	64	4,5
2015	10	1.428	72	5,0
2015	11	1.387	72	5,2
2015	12	1.401	178	12,7

Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2015, b.l.

Glede na povprečno stopnjo zavrnitve, lahko v tem primeru zapišemo:

$$K_e(n) = (n_d, n_a) \approx (0{,}06 \cdot n, 0{,}94 \cdot n) \tag{64}$$

Kot hitro vidimo, je v tem primeru glede na trenutno uporabljan nabor kontrol, izrazito nižja stopnja zavrnitve, torej posledično izrazito manjše število enot, za katere je potrebno preveriti podatke. Kakšen vpliv ima ta nižja stopnja zavrnitve na rezultate, bomo nakazali v nadaljevanju.

V naslednjem koraku bomo izvedli ključni korak v postopku selektivnega urejanja, to je delitev enot na tiste, za katere bomo uporabili ročne, in tiste, za katere bomo uporabili avtomatske popravke. Gre torej za implementacijo mehanizma U. V ta namen pri enotah, pri katerih je bila signalizirana vsaj ena napaka, za vsako spremenljivko določimo njeno pričakovano vrednost. V postopku najprej za izbrano spremenljivko Y_j celoten nabor n_d dvomljivih vrednosti razdelimo v dve podmnožici. Prvo podmnožico sestavljajo vse dvomljive enote, za katere imamo pozitivno vrednost spremenljivke iz preteklega obdobja, torej velja $Y_j(t-1) > 0$. Druga podmnožica je tej podmnožici komplementarna množica vseh enot, pri katerih nimamo vrednosti iz preteklega obdobja, ali pa je le-ta enaka 0. Postopek določitve pričakovane vrednosti bo sedaj različen glede na to, v katero podmnožico spada enota.

Za izračun pričakovane vrednosti pri enotah iz prve podmnožice na množici vseh n_a sprejemljivih enot najprej izračunamo naslednji nabor razmerij vrednosti tekočega in preteklega obdobja:

$$\mathcal{R}_a = \{Y_{ij}(t)/Y_{ij}(t-1)\,; Y_{ij}(t) > 0 \ \wedge\ Y_{ij}(t-1) > 0\}_{i=1,\dots,n_a} = \{r_{ij}\}_{i=1,\dots,n_a} \tag{65}$$

V naslednjem koraku nato definiramo funkcijo D, ki vsaki dvomljivi enoti določi darovalca iz nabora enot, ki tvorijo množico $\mathcal{R}_a$. Darovalec bo v našem primeru enota, ki ima z enoto, za katero računamo pričakovano vrednost, najbolj podoben podatek o številu zaposlenih oseb. Če je takih enot več, izmed njih darovalca izberemo s slučajnim izborom.

$$\begin{gathered} D\colon \mathcal{U}_d \rightarrow \mathcal{R}_a \\ D(y_{ij}(t), z_i) = r_{dj} \end{gathered} \tag{66}$$

Pričakovano vrednost za $y_{ij}(t)$ nato izračunamo kot:

$$\hat{y}_{ij}(t) = y_{ij}(t-1) \cdot r_{dj} \tag{67}$$

Za drugo podmnožico enot, za katere nimamo pozitivnega podatka spremenljivke iz preteklega obdobja, pričakovano vrednost izračunamo s pomočjo pomožne spremenljivke o

številu zaposlenih oseb. Naj bo tako kot prej $\bar{z}^{Y_j}(t)$ povprečna vrednost spremenljivke Y_j na eno zaposleno osebo pri vseh sprejemljivih enotah. Pričakovano vrednost nato izračunamo kot:

$$\hat{y}_{ij}(t) = \bar{z}^{Y_j}(t) \cdot z_i \tag{68}$$

Ko smo za vse spremenljivke pri dvomljivih enotah izračunali pričakovane vrednosti, lahko na podlagi teh vrednosti in primerno izbrane razdalje, definiramo ustrezno implementacijo mehanizma U oziroma funkcije pomembnosti. Najprej pri vseh dvomljivih enotah, za vsako od obravnavanih spremenljivk, izračunamo vrednost lokalne funkcije pomembnosti:

$$l_{ij} = \left|\hat{y}_{ij}(t) - y_{ij}(t)\right| \tag{69}$$

Nato za vsako enoto izračunamo še vrednost globalne funkcije pomembnosti:

$$g_i = \sum_{j=1}^{3} \frac{l_{ij}}{\mathrm{std}(Y_j)} \tag{70}$$

Razdelitev enot na podmnožico, namenjeno ročnemu oziroma avtomatskemu urejanju, nato izvedemo na podlagi kumulativne funkcije porazdelitve vrednosti globalne funkcije pomembnosti. V ročno urejanje bomo uvrstili toliko enot z najvišjo vrednostjo g_i, da bomo pokrili ciljni delež p_r. Implementacijo mehanizma U lahko torej formalno zapišemo kot:

$$U = U(F_g, p_r) \tag{71}$$

Za namene naše simulacije smo izbrali ciljni delež $p_r = 0{,}95$. S postopki ročnega urejanja bomo torej urejali toliko največjih enot glede na globalno funkcijo pomembnosti, da bomo s temi enotami pokrili 95 % kumulativne porazdelitve F_g. Rezultat delitve enot preko mehanizma U po zgoraj opisanem postopku prikazuje Tabela 15.

Tabela 15: Razdelitev enot glede na kriterij ročnega in avtomatskega urejanja

Leto	Mesec	Dvomljive enote	Ročno urejanje - število	Avtomatsko urejanje - število	Ročno urejanje – delež (%)	Avtomatsko urejanje – delež (%)
2014	1	78	33	45	42,3	57,7
2014	2	75	18	57	24,0	76,0
2014	3	80	41	39	51,3	48,8
2014	4	85	35	50	41,2	58,8
2014	5	70	32	38	45,7	54,3
2014	6	74	32	42	43,2	56,8
2014	7	78	32	46	41,0	59,0
2014	8	107	46	61	43,0	57,0
2014	9	70	26	44	37,1	62,9
2014	10	62	33	29	53,2	46,8
2014	11	63	31	32	49,2	50,8
2014	12	82	40	42	48,8	51,2
2015	1	64	18	46	28,1	71,9
2015	2	69	33	36	47,8	52,2
2015	3	97	42	55	43,3	56,7
2015	4	88	32	56	36,4	63,6
2015	5	76	35	41	46,1	53,9
2015	6	79	32	47	40,5	59,5
2015	7	77	45	32	58,4	41,6
2015	8	80	45	35	56,3	43,8
2015	9	64	28	36	43,8	56,3
2015	10	72	19	53	26,4	73,6
2015	11	72	37	35	51,4	48,6
2015	12	178	87	91	48,9	51,1

Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2015, b.l.

Glede na povprečne vrednosti deležev v tabeli, lahko zapišemo oceno:

$$U(n) = (n_{dr}, n_{da}) \approx (0{,}44 \cdot n_d, 0{,}56 \cdot n_d) \tag{72}$$

Nadaljnje postopke urejanja sedaj izvajamo ločeno za vsako od prej določenih podmnožic. Pri enotah, ki smo jih določili za ročno urejanje, končne podatke preprosto prepišemo iz tabele urejenih podatkov po trenutnem postopku urejanja. Ker se v trenutnih postopkih izvaja samo ročno urejanje, predpostavimo, da je za to podmnožico rezultat urejanja po prenovljenem postopku enak rezultatom urejanja po trenutnem postopku. Za enote, za katere smo predpostavili avtomatske popravke, pa je postopek določitve popravljenih vrednosti treba na novo opredeliti. Za ta, zadnji del postopka, bomo izvedli še dva koraka. V prvem bomo implementirali mehanizem K_s, v drugem pa mehanizem V.

Z mehanizmom K_s oziroma z lokalizacijo napake pri vsaki izmed n_{da} enot, ki jih urejamo avtomatsko, določimo, pri kateri izmed treh spremenljivk bomo spremenili vrednost, katere vrednosti pa bomo pustili nespremenjene. Pri implementaciji tega koraka smo natanko sledili lokalizaciji napake po Fellegi-Holt pristopu, ki smo ga opisali v razdelku 1.3 in sloni na številu kontrol, v katere je spremenljivka eksplicitno vključena in pri katerih je bila signalizirana napaka. Tabela 16 prikazuje končni rezultat tega postopka, zbirno po letih, skupaj z lokalizacijo napak pri enotah, ki smo jih popravljali ročno.

Tabela 16: Rezultati lokalizacije napak

		Število			**Delež (%)**		
Leto	**Način**	**Y1**	**Y2**	**Y3**	**Y1**	**Y2**	**Y3**
2014	Ročno	128	84	65	32,1	21,1	16,3
2014	Avtomatsko	227	235	147	43,2	44,8	28,0
2015	Ročno	90	62	78	19,9	13,7	17,2
2015	Avtomatsko	245	197	183	43,5	35,0	32,5

Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2015, b.l.

Pri enotah, ki smo jih urejali avtomatsko, moramo v simulaciji izvedbe selektivne urejanja, izvesti še en korak, in sicer implementacijo mehanizma V. Pri spremenljivkah, ki smo jih določili v postopku lokalizacije napake, moramo določiti novo, popravljeno vrednost. V ta namen bomo uporabili kar pričakovane vrednosti spremenljivk, ki smo jih določili za namene računanja funkcije pomembnosti. Če je torej $y_{ij}(t)$ vrednost spremenljivke, ki jo moramo popraviti, lahko enostavno zapišemo:

$$U(y_{ij}(t)) = \hat{y}_{ij}(t) \tag{73}$$

Z določitvijo novih vrednosti pri dvomljivih spremenljivkah je postopek selektivnega urejanja končan. Na koncu poglejmo še kakšen vpliv ima spremenjen postopek urejanja na statistične rezultate. Pogledali bomo vpliv na eno ključno statistiko, in sicer na indeks prihodka od prodaje, tekoči mesec glede na pretekli mesec. Indeks bomo izračunali kot enostaven agregatni indeks:

$$I_{t/t-1} = \frac{\sum_{i=1}^{n}(y_{i1}(t) + y_{i2}(t))}{\sum_{i=1}^{n}(y_{i1}(t-1) + y_{i2}(t-1))} \tag{74}$$

Tabela 17 za leti 2014 in 2015 prikazuje dve indeksni vrsti. Prva je izračunana iz mikro podatkov redne obdelave, torej samo s postopki ročnega urejanja. Druga je izračunana iz podatkov, ki smo jih dobili po prej opisani simulaciji selektivnega urejanja. Pri tem poudarjamo, da se indeksi, ki smo jih dobili iz podatkov redne obdelave, ne ujemajo povsem z objavljenimi indeksi, to pa iz dveh razlogov: v naših analizah nismo imeli na razpolago povsem vseh mikro podatkov; glede na postopek v rednem raziskovanju smo uporabil

nekoliko poenostavljen postopek izračuna z navadnim agregatnim indeksom. Sicer pa same vrednosti indeksa niti niso pomembne. Pomembna je razlika med obema indeksnima vrstama.

Tabela 17: Primerjava indeksnih vrst, dobljenih z različnima postopkoma urejanja

Leto	Mesec	Indeks - trenutno urejanje	Indeks - selektivno urejanje	Absolutna razlika
2014	2	98,9	98,5	0,4
2014	3	112,8	111,3	1,5
2014	4	95,6	95,4	0,2
2014	5	95,4	95,6	0,2
2014	6	107,6	107,4	0,2
2014	7	101,1	102,3	1,2
2014	8	78,3	79,0	0,7
2014	9	132,0	130,2	1,8
2014	10	99,7	99,2	0,5
2014	11	94,5	94,7	0,2
2014	12	93,6	93,5	0,1
2015	1	100,8	101,2	0,4
2015	2	103,2	102,8	0,4
2015	3	114,7	114,7	0,0
2015	4	89,1	90,6	1,5
2015	5	102,0	100,4	1,6
2015	6	108,6	108,6	0,0
2015	7	95,2	95,2	0,0
2015	8	79,0	81,1	2,1
2015	9	132,7	129,3	3,4
2015	10	98,1	98,0	0,1
2015	11	97,3	97,3	0,0
2015	12	88,9	89,0	0,1

Vir: SURS, Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2015, b.l.

Kot vidimo, smo s prenovljenimi postopki dobili skoraj enako indeksno vrsto, s tem, da smo bistveno zmanjšali število popravkov. Nekoliko podrobnejša analiza teh rezultatov bo podana v zadnjem poglavju.

5.2.3.2 Raziskovanje TRG/M

V trenutni praksi izvajanja raziskovanja že poteka nekoliko poenostavljena različica selektivnega urejanja podatkov. Razdelitev enot na tiste, pri katerih se izvajajo postopki ročnega, in tiste, pri katerih se izvajajo postopki avtomatskega urejanja, poteka po pristopu vhodne razdelitve. V resnici to vhodno razdelitev določa že sama narava podatkovnih virov. Pri DDV podatkih ni možno izvajati postopkov ročnega urejanja, saj SURS nima pristojnosti

preverjanja teh podatkov pri opazovanih enotah in so zato postopki avtomatskega urejanja edina izvedljiva možnost. V tem primeru torej že pred izvajanjem raziskovanja vemo, katere enote bomo obravnavali s postopki ročnega, in katere s postopki avtomatskega urejanja. S podatki, navedenimi v razdelkih 5.1.3 in 5.2.2, lahko ocenimo:

$$U(n) = (n_{dr}, n_{da}) \approx (0{,}75 \cdot n_d, 0{,}25 \cdot n_d) \tag{75}$$

Ker glede na naravo raziskovanja ni realno, da bi pri tem raziskovanju lahko uvedli selektivno urejanje na podlagi izhodne razdelitve, bomo simulirali zgolj spremembo v opredelitvi logičnih kontrol za terenske enote. V trenutni izvedbi namreč v tem naboru nastopajo samo linearne kontrole s fiksnimi koeficienti. S simulacijo bomo nakazali, kakšno spremembo v delovanju mehanizma K_e bi povzročili, če bi te kontrole nadomestili s kontrolami longitudinalne porazdelitve. Ker hočemo imeti nove kontrole, ki bi bile vsebinsko čim bližje že obstoječim kontrolam, smo v nov nabor vključili zgolj dve kontroli longitudinalne porazdelitve, ki nadomeščata kontrole LK1-LK4 v Tabeli 8. Ti dve kontroli prikazuje Tabela 18.

Tabela 18: Nove kontrole za validacijo terenskih enot raziskovanja TRG/M

Oznaka kontrole	**Pravilo kontrole**
LK1	Y1(t-1)>0 ∧ Y1(t)>0 ∧ Y1(t)> HB_PRET_Z
LK2	Y1(t-1)>0 ∧ Y1(t)>0 ∧ Y1(t)< HB_PRET_S

HB_PRET_S in HB_PRET_Z sta tudi tukaj, tako kot prej, spodnja in zgornja meja porazdelitve razmerij s podatki preteklega meseca, dobljeni s HB metodo. Pri uporabi HB metode smo simulirali različne vrednosti parametrov in nato primerjali izhodne rezultate kontrol z rezultati trenutnih kontrol, ki smo jih predstavili v razdelku 5.2.2. Na tem mestu (Tabela 19 in Tabela 20) podajamo rezultate primerjalne analize le za dva različna nabora vrednosti parametrov. V primerjavo smo vključile enote, ki so bile zaznane kot dvomljive v rednem izvajanju kontrol, in jih analizirali glede na to, ali so dvomljive tudi po novem sistemu ter glede na to, ali jim je bila pozneje v postopkih ročnega urejanja spremenjena vrednost. Iz analize smo izpustili tiste enote, ki so bile zaznane kot dvomljive po novem naboru kontrol, ne pa po starem. Za te enote namreč ne vemo, ali imajo pravilno ali nepravilno vrednost spremenljivke, in jih ne moremo nadalje kategorizirati.

- Vrednosti HB parametrov v prenovljenih tabelah: $C = 0{,}5$; $U = 0{,}3$; $A = 0{,}01$.

Tabela 19: Primerjalna analiza validacije pri prvem naboru HB parametrov

		Stare kontrole			
		Število enot		Prihodek od prodaje	
		Dvomljive/ Pravilne (%)	Dvomljive/ Popravljene (%)	Dvomljive/ Pravilne (%)	Dvomljive/ Popravljene (%)
Nove kontrole	Sprejemljive	49,2	22,6	37,3	16,7
	Dvomljive	18,8	9,3	25,1	20,9

Vir: SURS, Mikro podatki raziskovanja TRG/M, januar 2014 - december 2015, b.l.

Vidimo, da je kar precejšnja razlika v kategorizaciji enot. Na primer, približno 50% enot, ki so bile prej zaznane kot dvomljive, ne pa popravljene, zdaj niso zaznane kot dvomljive. Kategorija, ki je s stališča učinkovitosti novih kontrol najbolj problematična, so enote, katerih vrednost spremenljivke je bila popravljena, torej jo lahko jemljemo kot nepravilno, ni pa bila zaznana kot dvomljiva v novem naboru kontrol. Teh enot je v tem primeru nekaj več kot 20%, njihov prihodek pa v naboru analiziranih enot predstavlja malo več kot 15%.

- Vrednosti HB parametrov v prenovljenih tabelah: $C = 0{,}5$; $U = 0{,}7$; $A = 0{,}01$

Tabela 20: Primerjalna analiza validacije pri drugem naboru HB parametrov

		Stare kontrole			
		Število enot		Prihodek od prodaje	
		Dvomljive/ Pravilne (%)	Dvomljive/ Popravljene (%)	Dvomljive/ Pravilne (%)	Dvomljive/ Popravljene (%)
Nove kontrole	Sprejemljive	50,2	23,5	29,6	12,5
	Dvomljive	17,9	8,5	32,8	25,1

Vir: SURS, Mikro podatki raziskovanja TRG/M, januar 2014 - december 2015, b.l.

Vidimo, da se v tem primeru delež enot v »problematični« kategoriji ni bistveno spremenil, vendar pa je precej nižji delež prihodka pri teh enotah. To je posledica tega, da smo zvišali vrednost parametra U in s tem v zaznavanju dvomljivih podatkov dali večjo težo večjim enotam.

Na koncu poglejmo še, kakšen je vpliv spremenjenega nabora kontrol na statistični rezultat. Tudi v tem primeru bomo kot analizirano statistiko vzeli indeks tekoči mesec glede na pretekli mesec. V primeru prenovljenih kontrol smo popravili samo tiste dvomljive enote, ki so bile popravljene tudi po prejšnjem sistemu kontrol. Vse ostale vrednosti, torej tiste, ki so bile zaznane kot pravilne po starem sistemu, kot tudi tiste, ki po novem sistemu sploh niso bile zaznane kot dvomljive, smo pustili nespremenjene. Primerjavo obeh indeksnih vrst prikazuje Tabela 21. Ob tem poudarjamo, da tu sami indeksi nimajo nikakršnega

vsebinskega pomena, saj so izračunani zgolj na podlagi podatkov terenskih enot. Pomembna je primerjava dveh indeksnih vrst.

Tabela 21: Vpliv prenovljenega sistema kontrol na indeksno vrsto

Leto	Mesec	Indeks-redna obdelava	Indeks - nove kontrole	Absolutna razlika
2014	2	91,0	90,8	0,2
2014	3	116,5	116,6	0,1
2014	4	104,1	104,3	0,2
2014	5	94,4	94,4	0,0
2014	6	103,6	103,6	0,0
2014	7	102,8	102,8	0,0
2014	8	92,8	93,1	0,3
2014	9	107,4	106,8	0,6
2014	10	104,2	104,2	0,0
2014	11	90,9	91,1	0,2
2014	12	108,5	108,4	0,1
2015	1	85,5	85,8	0,3
2015	2	95,9	95,6	0,3
2015	3	116,8	116,3	0,5
2015	4	100,9	101,3	0,4
2015	5	99,6	99,6	0,0
2015	6	104,4	104,2	0,2
2015	7	96,8	97,2	0,4
2015	8	93,2	93,2	0,0
2015	9	107,0	106,6	0,4
2015	10	102,5	102,9	0,4
2015	11	94,5	94,3	0,2
2015	12	111,0	110,9	0,1

Vir: SURS, Mikro podatki raziskovanja TRG/M, januar 2014 - december 2015, b.l.

Kot vidimo, ima prenovljen sistem kontrol zelo majhen vpliv na indeks. To pomeni, da smo tudi z novim naborom kontrol popravili vse pomembne napake, ki vplivajo na končni rezultat. Ta sklep je zelo pomemben s stališča možnosti izboljšanja učinkovitosti urejanja, saj je pri prenovljenem sistemu kontrol precej manjše število dvomljivih enot. Za leto 2014 je tako na primer nov sistem signaliziral 400 dvomljivih enot, medtem ko jih je stari nabor kontrol 755.

5.3 Vpliv urejanja na dimenzije kakovosti

Urejanje podatkov ima pomemben vpliv na dimenzije kakovosti, ki smo jih opredelili v razdelku 2.1. Postopki urejanja, s katerimi smo se ukvarjali v nalogi, neposredno vplivajo

predvsem na točnost, pravočasnost in časovno primerljivost, zelo pomembni pa so tudi za »pridruženo« dimenzijo, ki izkazuje stroške in obremenitve statističnega raziskovanja. Podrobna analiza vplivov urejanja na dimenzije kakovosti bi zahtevala samostojno raziskavo in presega okvire naše naloge. Na tem mestu zato podajamo zgolj kratek oris te problematike.

5.3.1 Točnost

V bistvu je osnovni namen vseh postopkov urejanja izboljšati točnost končnih rezultatov. Je pa prav pri zagotavljanju točnosti pomembno upoštevati naslednje dejstvo. Popolna točnost, kot ujemanje med pravo in ocenjeno vrednostjo statistike, je zgolj cilj, ki je v veliki večini izvedb statističnih raziskovanj nedosegljiv. Že zaradi dejstva (če odmislimo vse druge razloge za ne-točnost), da bomo v zelo redkih primerih odkrili in odpravili vse napake v podatkih. Zato je v povezavi s točnostjo in z urejanjem vedno potrebno sprejeti neko tolerančno mejo točnosti, ki jo že vnaprej predpostavimo. V naši simulaciji selektivnega urejanja pri raziskovanju IND-PN/M smo tako tolerančno mejo v bistvu posredno določili s postavitvijo ciljnega deleža funkcije pomembnosti. Poglejmo zelo grobo oceno, kakšen vpliv ima ciljni delež na točnost izhodnega rezultata. Naj bo T vsota prihodka od prodaje vseh opazovanih enot v izbranem referenčnem obdobju t, torej $T = \sum_{i=1}^{n}(y_{i1}(t) + y_{i2}(t))$. Naj bo nadalje T_a vsota pri sprejemljivih in T_d vsota pri dvomljivih enotah. Prihodek pri dvomljivih enotah lahko nato delimo na prihodek pri enotah, ki gredo v ročno urejanje ($T_{d,r}$), in prihodek pri enotah, ki gredo v avtomatsko urejanje ($T_{d,a}$). Prihodek, pri enotah, ki gredo v avtomatsko urejanje, na koncu lahko razdelimo na tiste izmed teh enot, ki imajo pravilno vrednost ($T_{d,a}^{p}$), in tiste, ki imajo nepravilno vrednost ($T_{d,a}^{n}$). Celotno vsoto torej lahko zapišemo kot:

$$T = T_a + T_d = T_a + T_{d,r} + T_{d,a} = T_a + T_{d,r} + T_{d,a}^{p} + T_{d,a}^{n} \tag{76}$$

Če predpostavimo zdaj, da je prihodek enot enakomerno porazdeljen v vsaki izmed zgornjih kategorij in lahko deleže po številu, ki smo jih ocenili v empirični analizi, »prenesemo« na prihodek, lahko zapišemo oceno:

$$\begin{aligned} T &= 0{,}6 \cdot T + 0{,}4 \cdot T = 0{,}6 \cdot T + 0{,}4 \cdot (0{,}95 \cdot T + 0{,}05 \cdot T) = \\ &= 0{,}6 \cdot T + 0{,}38 \cdot T + 0{,}02 \cdot (0{,}88 \cdot T + 0{,}12 \cdot T) = \\ &= 0{,}6 \cdot T + 0{,}38 \cdot T + 0{,}0176 \cdot T + 0{,}0024 \cdot T \end{aligned} \tag{77}$$

Zadnja komponenta v vsoti predstavlja delež prihodka pri enotah, ki so nepravilne in jih ne urejamo ročno. Iz zgornjih izračunov izhaja ocena deleža 0,24%. Povprečna absolutna razlika indeksov, izražena v indeksnih točkah, ki jo prikazuje Tabela 17, je 0,7 točke, kar je nekoliko večje odstopanje, vendar še vedno zelo blizu, glede na to, da gre za zelo grobo oceno.

5.3.2 Pravočasnost in časovna primerljivost

Pravočasnost je zelo pomembna dimenzija v primeru kratkoročnih raziskovanj. Statistični uradi se nenehno soočajo z željami in zahtevami po statističnih rezultatih, ki bi bili na voljo zelo hitro po koncu referenčnega obdobja. Primer take zahteve je raziskovanje TRG/M, katerega rezultati morajo biti po uredbi posredovani Eurostatu v roku 30 dni po koncu referenčnega meseca. Dejstvo je, da je pravočasnost v kratkoročnih statistikah v tesni povezanosti s točnostjo, obe pa sta v veliki meri določeni z učinkovitostjo postopkov urejanja podatkov. To soodvisnost prikazuje Slika 7, kjer je prikazan osnovni pristop k izračunu in izkazovanju statističnih ocen. Prvo oceno statističnega rezultata (s^{p_1}) objavimo zelo hitro, nato pa ta rezultat večkrat revidiramo, tako da upoštevamo dodatne podatke in izvajamo dodatno urejanje podatkov. Vse do objave končne ocene (s^{p_k}). Tisto, kar bi izvajalci statističnega raziskovanja v tem primeru morali zagotavljati, je:

- Vsaka poznejša objava je bližje pravi vrednosti, torej:

$$|s^{p_1} - s^0| > |s^{p_2} - s^0| > \cdots > |s^{p_k} - s^0| \tag{78}$$

- Število začasnih rezultatov ne sme biti preveliko, saj s tem ustvarjamo preveliko število časovnih vrst za opis istega pojava in s tem vplivamo na slabšo časovno primerljivost rezultatov. V raziskovanju TRG/M so, na primer, objavljeni rezultati začasni 10 mesecev in se v tem obdobju lahko spreminjajo, potem pa postanejo končni.

5.3.3 Stroški in obremenitve

Kot smo že večkrat omenili, imajo postopki urejanja velik vpliv predvsem na stroške raziskovanja, vplivajo pa lahko tudi na (poleg pridobivanja podatkov) dodatno obremenitev poročevalskih enot zaradi preverjanja podatkov. Tisto, kar smo že pokazali v razdelku 5.2.3, je, da lahko s pametnim načrtovanjem, predvsem pa z vključitvijo selektivnega in avtomatskega urejanja, bistveno znižamo stroške ter obremenitve. Na tem mestu podajamo še zelo grobo kvantitativno oceno prihranka za primer simulacije uvedbe prenovljenih postopkov v raziskovanje IND-PN/M. V ta namen uvedimo poenostavljen stroškovni model urejanja podatkov na letni ravni. Fiksni stroški urejanja naj znašajo 1.000 enot, variabilni del na enoto urejanja pa 1 enoto v primeru ročnega in 0,1 enote v primeru avtomatskega urejanja. Če sedaj upoštevamo podatke za leto 2014 iz Tabel 4 in 14, dobimo oceno, prikazano v Tabeli Tabela *22*:

Tabela 22: Ocena zmanjšanja stroškov z novimi postopki urejanja

	Trenutno urejanje	**Prenovljeni postopek**	
	Ročno urejanje	**Ročno urejanje**	**Avtomatsko urejanje**
Število enot	6.605	399	525
Fiksni stroški	1.000	1.000	
Variabilni stroški	6.605	399	52,5
Stroški- skupaj	7.605	1.451,5	

Vidimo, da so se stroški zmanjšali kar za dobrih 80 %. Večina tega zmanjšanja niti ni na račun uvedbe selektivnega urejanja, ampak na račun prenovljenega nabora kontrol, ki ima precej nižjo stopnjo zavrnitve. Omeniti je potrebno še, da nismo upoštevali stroškov, ki jih potrebujemo za razvoj in uvedbo novih metod, ki prav gotovo niso zanemarljivi. Po drugi strani pa gre v tem primeru za enkratni vložek statističnega urada, ki se, glede na predstavljeno oceno, kmalu povrne.

6 RAZPRAVA S PREDLOGI IZBOLJŠAV

6.1 Razprava

Postopek urejanja podatkov, s katerim iščemo in odpravljamo napake v podatkih, je eden najbolj zahtevnih delov statističnega procesa, saj nemalokrat zahteva velik vložek, tako glede na porabljen čas kot tudi glede na delež celotnih sredstev, namenjenih raziskovanju. Po drugi strani je urejanje zelo pomemben del statističnega procesa s stališča kakovosti končnih rezultatov. Vsak izvajalec želi v končni fazi procesa, pri pripravi statističnih agregatov, uporabljati prečiščene in konsistentne podatke ter tako zagotoviti čim bolj točne in zanesljive rezultate. Podatki, ki ne vsebujejo napak in v največji možni meri izkazujejo merjeni pojav, so končen, čeprav zelo težko dosegljiv cilj. Vsa ta dejstva so torej vzrok, da statistični uradi namenjajo veliko svojega časa in veliko svojih sredstev razvoju in vpeljavi novih, naprednejših postopkov urejanja podatkov, s katerimi bi te postopke racionalizirali in jih naredili čim bolj učinkovite.

Kljub temu, da se torej v zadnjem času vpeljuje cela vrsta novih metod in postopkov, predvsem na področju avtomatskega urejanja podatkov, pa še vedno manjkajo celovitejši teoretski modeli, ki bi vključevali vse sodobne izsledke in nove pristope ter bi ponujali trdno teoretsko osnovo za vpeljavo novih postopkov urejanja v proces statističnih raziskovanj uradne statistike. V nalogi smo skušali prispevati kamenček v tem nepopolnem mozaiku, in sicer smo predstavili teoretski model statističnega urejanja v primeru kratkoročnih poslovnih raziskovanj, katerega ustreznost je bila preizkušena tudi na primeru konkretnih statističnih raziskovanj, ki jih izvaja SURS. Celoten model temelji na dveh osnovnih elementih: na naboru mikro podatkov ter na sklopu mehanizmov, ki vsak na svoj način spreminjajo vhodne

mikro podatke. V nadaljevanju podajamo še nekaj dodatnih razmišljanj o vsakem od teh dveh elementov.

Mikro podatke smo v osnovni obliki predstavili s tridimenzionalno matriko. Čeprav gre tu za povsem formalen opis, je vidik tridimenzionalnosti za razumevanje postopkov urejanja v kratkoročnih raziskovanjih zelo pomemben, če ne celo bistven. Pomembno je, da podatkov ne gledamo več samo v eni časovni točki, ali mogoče v povezavi še z enim izbranim časovnim obdobjem, ampak v T časovnih točkah naenkrat. Primer takega pogleda je opredelitev več različnih vrst longitudinalnih osamelcev, ki jih grafično prikazuje Slika 12. V praksi smo ta pristop implementirali s kontrolami, ki jih prikazuje Tabela 11. Tisto, kar tak tridimenzionalen pogled sploh naredi smiseln, je dejstvo, da v primeru kratkoročnih raziskovanj, rezultate za isto referenčno obdobje objavljamo večkrat. Drugače povedano, rezultati za T obdobij so začasni, kar pomeni, da vedno preverjamo in obdelujemo podatke za T časovnih obdobij naenkrat. To značilnost podatkov in rezultatov kratkoročnih raziskovanj prikazuje Slika 7. Če bi v modelni opredelitvi mikro podatkov upoštevali še vidik začasnosti podatkov, bi v bistvu morali podatke predstaviti s štiridimenzionalno matriko, kjer bi četrto dimenzijo predstavljala verzija mikro podatkov p_i, na podlagi katerih je izpeljana statistika s^{p_i}. Ker bi bili štiridimenzionalni podatki z vidika preglednosti in jasnosti precej zahtevni, smo četrto dimenzijo v večini naših izpeljav izpuščali. Nanjo smo se sklicevali zgolj v zvezi s kompromisi med dimenzijami kakovosti v razdelku 5.3.

Modelne mehanizme lahko v grobem razdelimo v dve skupini. Mehanizme iz prve skupine imenujemo »načrtovani mehanizmi«, saj jih vzpostavljajo in kontrolirajo izvajalci raziskovanja. Mehanizmi druge skupine, ki jih imenujemo »skriti mehanizmi«, so mehanizmi, ki se vzpostavljajo zunaj samega izvajanja raziskovanja in na katere izvajalci raziskovanja nimajo neposrednega vpliva. Za vsak mehanizem v modelu smo opredelili množico parametrov, ki določajo delovanje in tudi izhodne rezultate mehanizma. V primeru načrtovanih mehanizmov so ti parametri predmet načrtovanja in implementacije, medtem ko so v primeru skritih mehanizmov stvar modeliranja na podlagi čim bolj realnih modelnih predpostavk. Celoten sklop mehanizmov z vhodnimi parametri prikazuje Slika 13.

Parametrizacija skritega mehanizma M_e, ki smo jo predstavili v razdelku 4.2.3, je temeljila na informacijah o enotah opazovanja, ki smo jih imeli na voljo za izvajanje empirične analize. Izmed teh informacij smo kot parametre, ki določajo delovanje mehanizma, izbrali velikost enote in število zaporednih poročanj. Empirična analiza ni potrdila odvisnost delovanja od prvega parametra in je le delno potrdila odvisnost od drugega. Dejstvo je, da obstajajo dejavniki, ki delovanje mehanizma bolj značilno določajo (na primer izkušenost osebe, ki posreduje podatke; kakovost in popolnost informacijskega sistema podjetja), vendar tega zaradi pomanjkanja informacij nismo uspeli preizkusiti, zato lahko te dejavnike za zdaj samo hipotetično predpostavimo. Če bodo uradi take informacije v prihodnosti bolj sistematično zbirali in evidentirali, bo mogoče opis delovanja mehanizma M_e precej izboljšati.

Načrtovane mehanizme določajo trije osnovni elementi: vhodni mikro podatki, parametri, ki opredeljujejo delovanje mehanizma in izhodni mikro podatki, ki so rezultat delovanja mehanizma. Cilj načrtovalcev in izvajalcev postopkov urejanja je taka parametrizacija načrtovanih mehanizmov, ki bi v čim večji meri »izničila« delovanje skritih mehanizmov M_e in M_s ob čim manjšem vložku v smislu porabljenega časa in sredstev. Mehanizmi, ki so v tem pogledu bistveni, so mehanizmi K_e, K_s, U in V, zato podajamo še nekaj dodatnih razmislekov o teh mehanizmih.

Mehanizem K_e je določen z naborom logičnih kontrol, ki predstavljajo osnovno orodje v izvajanju postopkov urejanja podatkov. V primeru longitudinalnih raziskovanj je ključna kontrola časovne (longitudinalne) dimenzije podatkov. Te kontrole so v praktičnih izvedbah še vedno pretežno linearne kontrole (33), kar pomeni, da so koeficienti, ki določajo mejo sprejemljivosti, fiksni in vnaprej določeni (glej Tabeli 2 in 8). Kot smo pokazali z našo empirično študijo, lahko s prehodom na kontrole longitudinalne porazdelitve bistveno izboljšamo učinkovitost urejanja. Meje sprejemljivosti so v tem primeru določene ob vsaki izvedbi raziskovanja, glede na porazdelitve pridobljenih podatkov. Pri uporabi longitudinalnih kontrol izpostavljamo še en pomemben vidik. Pomembno je, da so te kontrole opredeljene tako, da je celostno upoštevana časovna komponenta podatkov. Ko presojamo sprejemljivost nekega podatka, je potrebno upoštevati umeščenost tega podatka v serijo zaporednih časovnih poročanj. Drugače povedano, pri presoji sprejemljivosti podatka upoštevamo tako pretekle kot tudi prihodnje podatke. Implementacijo takega pristopa predstavljajo kontrole, ki jih prikazujeta Tabela 11 in Tabela 13.

Mehanizem U je orodje, katerega delovanje je ključno za optimizacijo postopkov urejanja z vidika racionalizacije vloženih sredstev ob pogoju ohranitve zadostne točnosti izhodnih rezultatov. Vzpostavitev in učinkovito delovanje tega mehanizma sta zato bistvena elementa pri modernizaciji postopkov urejanja. Tisto, kar ta mehanizem vzpostavlja, je drugačen pogled na pomembnost enot za urejanje. Pomembnost enote ni več funkcija zgolj atributov enote, ki jih poznamo pred začetkom pridobivanja podatkov, ampak funkcija teh atributov v kombinaciji s podatki, ki jih enota poroča. Na primer, enota, ki jo v izboru podatkov obravnavamo kot majhno enoto (na primer po številu zaposlenih ali letnem prihodku od prodaje), lahko poroča podatek, ki bistveno odstopa od pričakovane vrednosti in ga moramo zato obravnavati z višjo prioriteto. V empirični analizi smo pokazali, da lahko z vzpostavitvijo mehanizma U bistveno zmanjšamo količino potrebnega urejanja brez bistvenega vpliva na točnost izhodnih statistik. Kratkoročna raziskovanja so za vzpostavitev tega mehanizma še posebej primerna, saj imamo v poročanih podatkih preteklih obdobij zelo močan vir za določitev pričakovane vrednosti. V izvajanju empirične analize smo se tudi prepričali, da je za učinkovitost tega mehanizma zelo pomembna določitev ustreznih parametrov, pri tem mislimo predvsem na funkcijo pomembnosti in ciljni delež funkcije pomembnosti. Zato je pred implementacijo teh postopkov potrebno izvesti podrobne in temeljite analize različnih parametrizacij mehanizma. Omenimo še, da vzpostavitev mehanizma U pomeni precejšnjo spremembo celotnega statističnega procesa, saj se bistveno

spremeni tok podatkov. Uvedba mehanizma in vseh s to uvedbo povezanih postopkov zato predstavlja za statistične urade precejšen izziv.

Postopke, ki jih poganja mehanizem U, dopolnjujejo postopki, ki smo jih modelirali z mehanizmom V. Gre za oceno novih, popravljenih vrednosti pri enotah in spremenljivkah, ki so jih prejšnji mehanizmi določili kot nepravilne, in so določeni, da bodo ocenjeni z avtomatskimi popravki. V bistvu gre za ustrezno uporabo tako imenovanih metod vstavljanja (angl. *imputation methods*), ki jih v statističnih raziskovanjih največ uporabljamo za oceno manjkajočih vrednosti, v našem primeru pa bodo vstavljene vrednosti nadomestile nepravilne podatke. Pri načrtovanju mehanizma V sta zelo pomembni dve zahtevi, ki naj jim zadoščajo podatki, ki so rezultat delovanja mehanizma. Prva zahteva je, da so vse nove vrednosti v območju sprejema, kot smo ga definirali v razdelku 1.1. Drugi pogoj je, da z novimi vrednostmi ne spremenimo bistveno porazdelitev (predvsem longitudinalnih), ki jih določajo že pred urejanjem sprejemljivi podatki. Demonstracija postopka, s katerim zagotavljamo obe zahtevi, je za primer podatkov raziskovanja IND-PN/M podana v okviru empirične analize v razdelku 5.2.3.

Dejstvo je, da je predvsem zahtevo po sprejemljivosti težje izpolniti, če je v obravnavanem naboru podatkov veliko spremenljivk, predvsem pa, če je med njimi veliko povezav, ki jih določa sistem logičnih kontrol. S tega vidika so kratkoročna longitudinalna raziskovanja manj zahtevna (v primerjavi z letnimi presečnimi), saj imamo v teh raziskovanjih večinoma opraviti z majhnim številom spremenljivk, pri katerih je predvsem pomembna longitudinalna komponenta, ni pa toliko povezav med posameznimi spremenljivkami v istem časovnem obdobju.

Glavni rezultat te naloge je teoretski model urejanja v primeru kratkoročnih raziskovanj. Na primeru konkretnih raziskovanj smo v empiričnem delu pokazali, da ko obstoječe prakse postavimo nasproti teoretskemu modelu, lahko iz tega izpeljemo precej koristnih napotkov za izboljšave. Predloge teh izboljšav podajamo v naslednjem razdelku.

6.2 Predlogi izboljšav

6.2.1 Raziskovanje IND-PN/M

- Vse kontrole, ki so namenjene zaznavanjem odstopajočih longitudinalnih vrednosti posamezne spremenljivke (LK2-LK13), in so sedaj linearne kontrole s fiksnimi koeficienti, bi bilo smiselno nadomestiti s kontrolami longitudinalne porazdelitve. Z uvedbo kontrol, ki bi temeljile na vsakokratni porazdelitvi longitudinalnih podatkov, bi lahko znižali stopnjo zavrnitve in zvišali stopnjo zaznanih napak. Pri tem bi bilo, seveda, z dodatnimi analizami potrebno izbrati metodo ter najbolj ustrezno parametrizacijo metode. HB metoda, ki smo jo uporabili v naših analizah, je samo eden od možnih pristopov.

- V nabor kontrol bi bilo koristno vključiti tudi kontrole longitudinalne porazdelitve, ki pri enotah, ki nimajo (pozitivnih) podatkov iz preteklega obdobja, določajo osamelce samo iz porazdelitve v tekočem obdobju (kontrole oblike (37)). Kontrole te oblike, ki smo jih predlagali v naši simulaciji (LK21-LK23), so samo ena od možnosti implementacije.
- Glavni rezultat empiričnega dela naše naloge je jasna indikacija o velikih potencialih, ki jih imajo postopki selektivnega urejanja za racionalizacijo postopkov urejanja. Kot smo pokazali v razdelku 5.2.3, lahko bistveno zmanjšamo število enot, ki jih moramo ročno preveriti in urediti, brez bistvenega vpliva na točnost končnega rezultata. Rezultati simulacije postopkov selektivnega urejanja sicer potrebujejo še nekaj dodatnih pojasnil:

  – Primerjalna analiza statističnih rezultatov po dveh različnih postopkih urejanja je bila opravljena samo za raven »Industrija skupaj«. SURS objavlja rezultate raziskovanja tudi na nižjih ravneh, in sicer na ravni področja in oddelka Standardne klasifikacije dejavnosti 2008 ter na ravni namenskih skupin (Češek-Vozel, 2013). Pred uvedbo postopkov selektivnega urejanja bi bilo potrebno opraviti primerjalno analizo še na teh področjih in po potrebi prilagoditi parametrizacijo.
  – Za razdelitev enot na tiste, ki jih urejamo ročno, in tiste, ki jih urejamo avtomatsko, smo upoštevali ciljni delež $p_r = 0{,}95$. Enote, katerih podatke naj bi se urejalo ročno, torej »pokrijejo« 95 % vrednosti funkcije pomembnosti. V resnici je ta delež zelo visok in bi ga lahko tudi nekoliko znižali, pa bi bili obe indeksni vrsti še vedno dovolj blizu. Ena od možnih alternativ je, da ta delež, ki smo ga uporabili za delitev enot na ravni celotnega nabora podatkov, nekoliko znižamo, hkrati pa uvedemo tudi delitev enot na nižjih ravneh, na primer na ravni oddelka Standardne klasifikacije dejavnosti.
  – Posredna predpostavka primerjalne analize je, da so rezultati, dobljeni iz ročno urejenih podatkov, točni oziroma da ti mikro podatki ne vsebujejo več napak. Ta predpostavka je seveda nerealna, saj se zavedamo, da nikoli ne odkrijemo vseh napak v podatkih. Pred uvedbo postopkov selektivnega urejanja bi bilo zato potrebno izvesti še dodatno analizo, kjer bi se nekaj časa vzporedno preverjali dvomljivi podatki, ki bi jih signalizirala oba nabora kontrol. Tako bi lahko vsaj približno ocenili delež enot, ki so po trenutnem postopku sprejemljive, v resnici pa so nepravilne.

- Čeprav načini zbiranja podatkov niso neposredna tema te naloge, vseeno dodajamo še priporočilo s tega področja. Po analogiji z raziskovanjem TRG/M, bi bilo tudi v tem primeru vredno raziskati možnosti, da bi za del enot pridobili oceno iz podatkov DDV v kombinaciji z modelno oceno. V tem primeru bi lahko prag zajema tudi nekoliko spustili in s tem zmanjšali pristranskost zaradi zajema.

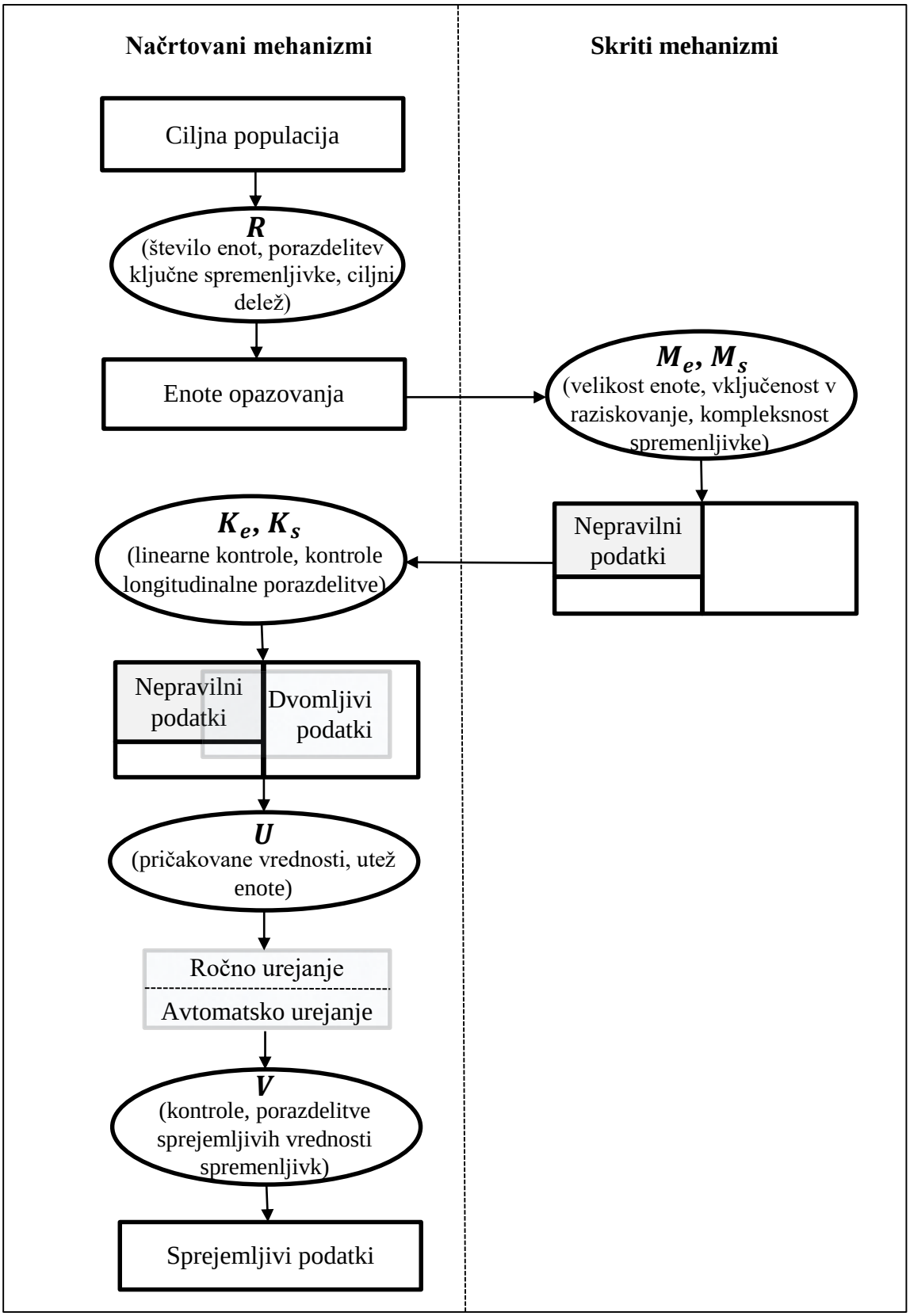
6.2.2 Raziskovanje TRG/M

- Vse kontrole za terenske enote so trenutno linearne kontrole s fiksnimi koeficienti. Smiselno bi bilo razmisliti o nadomestitvi teh kontrol s kontrolami longitudinalne porazdelitve. Kot smo pokazali v razdelku 5.2.3, bi s tako prenovo kontrol lahko precej

zmanjšali stopnjo zavrnitve in količino ročnega urejanja podatkov. Prav tako smo tudi pokazali, da bi bilo ob uvedbi kontrol longitudinalne porazdelitvi potrebno analizirati različne parametrizacije metode za iskanje osamelih vrednosti. Nevarnost je namreč, da bi z neustrezno opredeljenimi parametri prišli do velikega števila sprejemljivih enot, ki bi bile v resnici nepravilne.

- Trenuten nabor kontrol za DDV enote ima precej nizko stopnjo zavrnitve. Možno je sicer, da to izvira iz nizke stopnje napak, vendar je v resnici precej bolj realna možnost, da parametri, ki se trenutno uporabljajo, rezultirajo v preširokem območju sprejema oziroma preozkem področju zavrnitve. S prilagoditvijo parametrov bi dobili ožje področje sprejema, kar bi najbrž pripomoglo k bolj učinkovitemu urejanju podatkov DDV enot.
- Trenuten nabor kontrol za DDV enote temelji na porazdelitvi razmerij podatkov tekočega in preteklega meseca. Ker imajo podatki o prihodku od prodaje v trgovini na drobno izrazit sezonski vpliv, bi bilo po analogiji kontrol za terenske enote najbrž smiselno nabor dopolniti še s kontrolami, ki bi izhajale iz porazdelitve podatkov tekočega meseca glede na podatke istega meseca preteklega leta, torej porazdelitve množice $\{r_i\}_{i=1,\dots,n} = \{y_i(t)/y_i(t-12)\}_{i=1,\dots,n}$.
- Tako v kontrolah za terenske enote kot tudi v kontrolah za DDV enote, bi bilo potrebno dodati še kontrolo oblike (37), ki bi izhajala iz analize porazdelitve podatkov tekočega obdobja. Ta kontrola bi bila pomembna predvsem za zaznavanje osamelih vrednosti pri enotah, ki v raziskovanju poročajo prvič.

Slika 13: Mehanizmi urejanja podatkov



SKLEP

Statistično urejanje podatkov označuje vse postopke, s katerimi iščemo in odpravljamo napake v podatkih. Končni namen postopkov je izboljšati kakovost vhodnih podatkov, posledično kakovost izhodnih statistik, na dolgi rok pa tudi kakovost statističnih procesov, če izsledke urejanja uporabimo za vpeljavo izboljšav. Iz vsega tega sledi, da gre za zelo pomemben del statističnega procesa, kateremu izvajalci statističnih raziskovanj posvečajo zelo veliko pozornosti. Razvojne aktivnosti na tem področju gredo predvsem v smeri večje racionalizacije postopkov, s katerimi bi zmanjšali sicer zelo visoke vložke, tako v smislu stroškov kot tudi porabljenega časa, ki jih urejanje zahteva. Navedena dejstva še posebej veljajo za področje uradne statistike, kjer so zahteve po visoki kakovosti rezultatov, vzporedno z vse večjimi pritiski na zmanjšanje stroškov, še posebej izrazite.

V nalogi smo se posvetili problemu urejanja podatkov na enem od zelo pomembnih področij uradne statistike, na področju kratkoročnih raziskovanj. Kratkoročna raziskovanja so mesečna ali četrtletna raziskovanja, s katerimi zaznavamo kratkoročna gospodarska gibanja. Kratkoročna raziskovanja imajo precej specifičnih značilnosti, ki jih je potrebno upoštevati v razmislekih o postopkih urejanja, predvsem pa je potreben specifičen pristop pri načrtovanju novih, modernejših postopkov urejanja. Osrednji cilj naloge je bil postaviti in preizkusiti teoretski model, ki bi na abstraktni, hkrati pa tudi uporabni ravni, opisal urejanje podatkov v kratkoročnih raziskovanjih. V prvem delu naloge smo tako predstavili splošne pojme in koncepte s področja urejanja ter predstavili nekatere prakse tujih statističnih uradov. V drugem delu smo opredelili teoretski model ter ga v empiričnim delu preizkusili na primeru konkretnih podatkov dveh raziskovanj, ki jih izvaja SURS. Postopke urejanje podatkov smo v zadnjem delu razdelali še skozi prizmo dimenzij kakovosti standardnega modela ocenjevanja, ki se uporablja v ESS. Ključni izsledki naloge so:

- Glavni značilnosti kratkoročnih poslovnih raziskovanj, ki temeljno določata postopke urejanja, sta časovna dimenzija podatkov in izrazito asimetrična porazdelitev podatkov, ki jih urejamo.

  - Časovna komponenta izhaja iz dejstva, da iste enote v raziskovanju opazujemo daljše časovno obdobje. Za večino enot imamo zato na voljo daljšo časovno vrsto poročanih podatkov.
  - Asimetrična porazdelitev pomeni, da so z vidika vpliva na statistični rezultat podatki nekaterih enot precej bolj pomembni kot podatki drugih. Ta lastnost je osnova za uspešno izvajanje postopkov selektivnega urejanja.

- Dodatni faktor, ki tudi v precejšnji meri določa postopke urejanja, je uporaba več podatkovnih virov. Vse več uradov namreč v svojih kratkoročnih raziskovanjih kombinira podatke, pridobljene z anketnim vprašalnikom, s podatki iz administrativnih virov. Ker za administrativne vire uradi običajno nimajo pravice preverjati veljavnost

dvomljivih podatkov direktno pri poročevalskih enotah, za te podatke odpade možnost klasičnega ročnega urejanja.

- Postopke urejanja lahko opišemo z zaporedjem parametriziranih mehanizmov, ki delujejo na množici opazovanih podatkov in od katerih jih je del načrtovan, del pa skrit.
- Parametre modela lahko precej dobro ocenimo, če imamo na voljo nabor uporabljenih logičnih kontrol, mikro podatke pred in po urejanju, in če je bil vsaj del podatkov urejen s postopki ročnega urejanja, ki naj bi vodili do pravilnih vrednosti.
- Z implementacijo ustrezno parametriziranih načrtovanih mehanizmov lahko bistveno izboljšamo postopke urejanja, predvsem v smislu racionalizacije in učinkovitosti.
- Empirična analiza je za konkretni raziskovanji pokazala, da so možne izboljšave predvsem pri boljši opredelitvi nabora logičnih kontrol ter pri uvedbi postopkov selektivnega urejanja.

Dejstvo je, da urejanje podatkov združuje pod svojim okriljem raznolik nabor teorij in praks, zaradi česar je zelo težko v okviru enega teoretskega modela zajeti vse vidike in dejavnike. V naš model smo vključili tiste parametre, za katere smo presodili, da so ključni za opis urejanja podatkov v kratkoročnih raziskovanjih. Delno je izbor parametrov določala tudi dosegljivost podatkov za empirično analizo, s katero smo testirali veljavnost in uporabnost modela. V vsakem primeru se je potrebno zavedati, da je model z uporabo dodatnih testnih podatkov in z izvedbo dodatnih še bolj poglobljenih analiz mogoče še precej izboljšati in ga narediti še bolj uporabnega. Prihodnje raziskovalno delo na tem področju bi se moralo osredotočiti predvsem na preizkus modela v dodatnih primerih konkretnih raziskovanj. Predvsem boljša opredelitev in parametrizacija skritih mehanizmov bi prispevala k bistveni izboljšavi modela. Prav tako ostaja izziv bolj natančen teoretski opis soodvisnosti postopkov urejanja in dimenzij kakovosti.

LITERATURA IN VIRI

1. Adolfsson, C., Arvidson, G., Gidlund, P., Norberg, A., & Nordberg, L. (2010). *Development and Implementation of Selective Data Editing at Statistics Sweden.* Prispevek predstavljen na konferenci European Conference on Quality in Official Statistics, Helsinki, Finland, 3-6 May 2010. Najdeno 16. julija 2016 na spletnem naslovu https://q2010.stat.fi/media//presentations/Norberg_et_all__Statistics_Sweden_slutversion.pdf
2. Arnež, M., Belak, E., Blažič, P., Blejec, Z., Dolenc, D., Garvas, T., Glinšek, E., Jokić, N., Katnič, N., Klasinc, S., Kleindienst, K., Lisec, M., Nikić, B., Novak, T., Noč Razinger, M., Ostrež, T., Repotočnik, Z., Rutar, K., Seljak, R., Smrekar, T., Smukavec, A., Urbajs, V., Vrabič Kek, B., Šegan, V., Šnuderl, K., Špeh, T.,& Žavbi, M. (2012). *Smernice za zagotavljanje kakovosti*. Ljubljana: Statistični urad Republike Slovenije.
3. Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.
4. Belcher, R. (2003). ”Application of the Hidiroglou-Berthelot Method of Outlier Detection for Periodic Business Surveys”. *SSC Annual Meeting. Proceedings of the Survey Methods Section* (str. 25-30). Halifax, Nova Scotia: Statistical Society of Canada.
5. Benedetti, R., Bee, M., & Espa, G. (2010). A Framework for Cut-off Sampling in Business Survey Design, *Journal of Official Statistics*, *26*(4), 651–671.
6. Biemer, P. P., & Lyberg, L. (2003). *Introduction to Survey Quality*. New York: John Wiley & Sons.
7. Black, O. (2009). *Improving validation for business surveys - the Eden Project.* Članek predstavljen na konferenci European Establishment Statistics Workshop, Stockholm, Sweden, 7-9 September 2009. Najdeno 16. julija 2016 na spletnem naslovu http://enbes.wikispaces.com/Black+2009+-+Improving+validation+...
8. Bohte, Z. (1993). *Numerično reševanje nelinearnih enačb*. Ljubljana: DMFA.
9. Brackstone G. (1999). Managing Data Quality in a Statistical Agency. *Survey Methodology, 25*(2), 139-149.
10. Carson C. S. (2001, februar). Toward a Framework for Assessing Data Quality. IMF working paper. Najdeno 9. junija 2016 na spletnem naslovu https://www.imf.org/external/pubs/ft/wp/2001/wp0125.pdf
11. Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association, 81*(396),1063–1069.
12. Cox, B. G., & Chinnappa, B. N. (1995). Unique Features of Business Surveys. V B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (ur.), *Business Survey Methods* (str. 1–17). New York: John Wiley & Sons.
13. Coutinho, W., De Waal, T., & Shlomo, N. (2013). Calibrated Hot-Deck Donor Imputation Subject to Edit Restrictions. *Journal of Official Statistics, 29*(2), 299–321.
14. Češek Vozel, N. (2014, b.d.). Standardno poročilo o kakovosti za raziskovanje Prihodek od prodaje in vrednosti zalog v industriji. Statistični urad Republike Slovenije. Najdeno 9. junija 2016 na spletnem naslovu:

http://www.stat.si/StatWeb/Common/PrikaziDokument.ashx?IdDatoteke=7946

15. De Waal, T., J. Pannekoek, & S. Scholtus (2011). *Handbook of statistical data editing and imputation*. New Jersey: John Wiley & Sons.
16. Eurostat (2003): *Definition of Quality in Statistics. Working Group "Assessment of quality in statistics", Sixth meeting, Luxembourg, 2-3 October 2003.* Najdeno 9. junija 2016 na spletnem naslovu http://ec.europa.eu/eurostat/documents/64157/4373735/02-ESS-quality-definition.pdf/f0fdc8d8-6a9b-48e8-a636-9a34d073410f
17. Eurostat (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys.* Metodološki priročnik, pripravljen v okviru projekta EDIMBUS. Najdeno 16. julija 2016 na spletnem naslovu http://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf/6e51b229-8628-422d-8c4c-7ede411e107f
18. Eurostat (2009). *ESS Standard for Quality Reports*, Luxembourg: Office for Official Publications of the European Communities
19. Fellegi, I. P., & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association 71*, 17- 35.
20. Granquist L. (1995). Improving the traditional editing process. In *Business Survey Methods*, pages 385–401, John Wiley and Sons.
21. Granquist, L., & Kovar, J. (1997). Editing of Survey Data: How much is enough? V L. Lyberg, P Biemer M. Collins, E. Leeuw, C. Dippo, N. Schwarz, D. Trewin (ur.) *Survey Measurement and Process* Quality (str. 415-436). New York: Wiley.
22. Groves, R. M. (1989). *Survey Errors and Survey Costs*. New Jersey: John Wiley and Sons.
23. Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics 19*(2), 177–199.
24. Hidiroglou, M.A., & Berthelot, J.M. (1986). Statistical Editing and Imputation for periodic business surveys, *Survey Methodology*, *12*(1), 73-83.
25. Hoogland, J. (2009), *Detection of potential influential errors in VAT turnover data used for short-term statistics.* Prispevek predstavljen na konferenci Work Session on Statistical Data Editing, Neuchâtel, Switzerland, 5-7 October 2009. Najdeno 16. julija 2016 na spletnem naslovu http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2009/wp.18.e.pdf
26. Hoogland, J., Van Bemmel, K., & De Wolf, P. (2011). *Editing of Mixed Source Data for Turnover Statistics*. Prispevek predstavljen na konferenci Work Session on Statistical Data Editing, Ljubljana, Slovenia, 9-11 May 2011. Najdeno 16. julija 2016 na spletnem naslovu http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2011/wp.11.e.pdf

27. Holt T., Jones, T. (1999) Quality work and conflicting quality objectives. *Proceedings of the 84th DGINS Conference* (str. 15–24). Luxembourg: Office for Official Publications of the European Communities.
28. Hunt, J. W., Johnson, J. S., & King, K. S. (1999). "Detecting Outliers in the Monthly Retail Trade Survey Using the Hidiroglou-Berthelot Method". *JSM Proceedings, Survery Research Methods Section* (str. 539-543). Alexandria: American Statistical Association. Najdeno 16. julija 2016 na spletnem naslovu http://www.amstat.org/sections/srms/Proceedings/papers/1999_093.pdf
29. Kozak, R. (2005), "The Banff System for Automated Editing and Imputation", *SSC Annual Meeting. Proceedings of the Survey Methods Section* (str. 2-10). Saskatoon, Saskatchewan: Statistical Society of Canada.
30. Lawrence, D., & McKenzie, R. (2000). The General Application of Significance Editing. *Journal of Official Statistics*, *16*(3), 243–253.
31. Latouche, M., & Berthelot, J. M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics 8*(3), 389-400.
32. Lessler, J. T., Kalsbeek W. D. (1992). Nonsampling Error in Surveys. New York: John Wiley & Sons,
33. Lunder, D., & Seljak, R. (2010*). Standardno poročilo o kakovosti za raziskovanje Mesečno statistično raziskovanje o trgovini Na drobno, trgovini z motornimi vozili in Popravilih motornih vozil.* Ljubljana: Statistični urad Republike Slovenije. Najdeno 9. junija 2016 na spletnem naslovu: http://www.stat.si/StatWeb/Common/PrikaziDokument.ashx?IdDatoteke=884
34. Marolt, K., & Seljak, R. (2006).Uporaba administrativnih virov kot sredstvo za hitro in učinkovito zagotavljanje kratk*oročn*ih statistik – indeksi prihodka v trgovini na debelo, *Zbornik referatov 16. mednarodnega statističnega posvetovanja Statistični dnevi 2006: Merjenje razvojne vloge in učinkovitosti javnega sektorja in politik* (str. 339-350). Radenci: Statistični urad Republike Slovenije.
35. Mulry, M. H., & Feldpausch, R. M. (2007). *Investigation of Treatment of Influential Values.* Prispevek predstavljen na konferenci ICES-III, June 18-21, 2007, Montreal, Quebec, Canada. Najdeno 16. julija 2016 na spletnem naslovu https://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000229.PDF
36. Norberg, A. (2009). *Editing at Statistics Sweden – Yesterday, today and tomorrow*. Prispevek predstavljen na konferenci European Establishment Statistics Workshop, Stockholm, Sweden, 7-9 September 2009. Najdeno 16. julija 2016 na spletnem naslovu http://enbes.wikispaces.com/Norberg+2009+-+Editing+at+Statistics+Sweden+-+...
37. Norberg, A., Lindgren, K., & Tongur, C. (2014). *Experiences from Selective Editing at Statistics Sweden*. Prispevek predstavljen na konferenci Work Session on Statistical Data Editing. Paris, France, 28-30 April 2014. Najdeno 16. julija 2016 na spletnem naslovu http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_1_Sweden_Norberg.pdf

38. Osborne, J.W., Overbay, A. (2004): The power of outliers (and why researchers should always check for them). Practical Assessment, Research & Evaluation, 9(6). Najdeno 1. julija 2016 na spletnem naslovu http://PAREonline.net/getvn.asp?v=9&n=6
39. Pannekoek, J., & De Waal, T. (2005), Automatic Edit and Imputation for Business Surveys: the Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics*, *21*(2), 257-286.
40. Pannekoek, J., Scholtus, S., & Van der Loo, M. (2013). Automated and Manual Data Editing: A View on Process Design and Methodology, *Journal of Official Statistics*, *29*(4), 511–537.
41. Riviere, P. (2002). *General principles for data editing in business surveys and how to optimize it.* Prispevek predstavljen na konferenci UNECE Work Session on Statistical Data Editing, Helsinki, Finland, 27-29 May 2002. Najdeno 16. julija 2016 na spletnem naslovu http://www.unece.org/fileadmin/DAM/stats/documents/2002/05/sde/16.e.pdf
42. Seljak, R., & Špeh, T. (2004). *Automatic editing system for two short-term business surveys*. Prispevek predstavljen na konferenci Work Session on Statistical Data Editing, Ottawa, Canada, 16-18 May 2005. Najdeno 16. julija 2016 na spletnem naslovu http://www.unece.org/fileadmin/DAM/stats/documents/2005/05/sde/wp.43.e.pdf
43. Scholtus, S. (2013). Automatic editing with hard and soft edits. *Survey methodology 39*(1), 59-89.
44. Seljak, R., & Zaletel, M. (2007). *Tax data as a means for the essential reduction of the short-term surveys response burden*. Prispevek predstavljen na konferenci ICES-III, 18-21 June, 2007, Montreal, Quebec, Canada. Najdeno 16. julija 2016 na spletnem naslovu https://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000189.PDF
45. Seljak, R. (2008). *Combining the tax and the survey data for the purposes of the short-term statistics production*. Prispevek predstavljen na konferenci ESSnet-ISAD workshop, Vienna, 29-30 May, 2008. Najdeno 16. julija 2016 na spletnem naslovu http://ec.europa.eu/eurostat/documents/3888793/5845197/KS-RA-09-005-EN-TOC.PDF/7b642a49-416c-46ba-9bb0-eecbd68efc83?version=1.0
46. Seljak, R. (2012). *Statistično urejanje podatkov*. Ljubljana: Statistični urad Republike Slovenije.
47. Snijkers, G., & Bavdaž, M. (2011). Business surveys. V: Lovrić, M. (ur.), *International encyclopedia of statistical science* (str. 191-194). Berlin, Heidelberg: Springer.
48. Statistični urad Republike Slovenije. (2008). *Increase of efficiency of data collection and dissemination of retail trade turnover indices. Final report for grant agreement: 44402.2006.004-2006.342* (neobjavljeno interno gradivo). Ljubljana: Statistični urad Republike Slovenije.
49. Statistični urad Republike Slovenije (b.l.). *Mikro podatki raziskovanja IND-PN/M, januar 2014 - december 2015* (neobjavljeni interni podatki). Ljubljana: Statistični urad Republike Slovenije.

50. Statistični urad Republike Slovenije (b.l.). *Mikro podatki raziskovanja TRG/M, januar 2014 - december 2015* (neobjavljeni interni podatki). Ljubljana: Statistični urad Republike Slovenije.
51. Statistični urad Republike Slovenije (b.l.). *Seznam logičnih kontrol raziskovanja IND-PN/M* (neobjavljeno interno gradivo). Ljubljana: Statistični urad Republike Slovenije.
52. Statistični urad Republike Slovenije (b.l.). *Seznam logičnih kontrol raziskovanja TRG/M* (neobjavljeno interno gradivo). Ljubljana: Statistični urad Republike Slovenije.
53. Svet Evropske unije (1998, 5. junij). *Uredba o kratkoročnih statističnih kazalcih.* Uradni list Evropske unije L 162, 05/06/1998 str. 1 –15.
54. Svet Evropske unije (1993, 30. marec). *Uredba o statističnih enotah za opazovanje in analizo gospodarstva v Skupnosti.* Uradni list Evropske unije L 76, 30.3.1993, str. 1–11, spremenjena z Uredbo Evropskega parlamenta in Sveta (ES) št. 1158/2005, Uradni list Evropske unije L 191, 22.7.2005, str. 1-15.
55. Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesely.
56. Vlag, P. (2012). *Imputing Missing Values When Using Administrative Data for Short-Term Enterprise Statistics.* Prispevek predstavljen na konferenci Work Session on Statistical Data Editing, Oslo, Norway, 24 - 26 September 2012. Najdeno 16. julija 2016 na spletnem naslovu http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/23_Netherlands.pdf
57. Wein, E. (2009). *Automatic imputation for short term statistics.* Prispevek predstavljen na konferenci Work Session on Statistical Data Editing, Neuchâtel, Switzerland, 5-7 October 2009. Najdeno 16. julija 2016 na spletnem naslovu http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2009/wp.2.e.pdf

PRILOGE

KAZALO PRILOG

Priloga 1: Seznam uporabljenih kratic

DDV	Davek na dodano vrednost
EED	Enota enovrstne dejavnosti
ESS	Evropski Statistični sistem
FURS	Finančna uprava Republike Slovenije
SURS	Statistični urad Republike Slovenije

Priloga 2: Slovensko-angleški slovar uporabljenih izrazov

Slovenski izraz	**Angleški izraz**
Dvomljiva vrednost	Doubtful value
Enota enovrstne dejavnosti	Kind of Activity Unit
Funkcija pomembnosti	Score function
Izpeljane kontrole	Implicit edits
Krovna uredba za poslovne statistike	Framework Regulation Integrating Business Statistics
Lokalizacija napake	Error localisation
Metoda razmerja darovalca	Donor ratio imputation
Metode vstavljanja	Imputation methods
Napačna vrednost	Wrong value
Neposredne kontrole	Explicit edits
Nepravilna vrednost	Erroneous value
Normalna oblika kontrole	Normal form of edit
Osamelec	Outlier
Podjetje	Enterprise
Poln nabor kontrol	Complete set of edits
Prava vrednost	Right value
Pravilna vrednost	Correct value
Selektivno urejanje	Selective editing
Sprejemljiva vrednost	Acceptable value
Stopnja nepravilnosti	Error rate
Stopnja zavrnitve	Failure rate
Stopnja zaznanih napak	Hit rate