KOMBINIMI STATISTIKOR PËR INTEGRIMIN E TË DHËNAVE ME BURIME TË NDRYSHME

EVA XHAVA, INSTITUTI I STATISTIKAVE
exhava@instat.gov.al

Abstrakti

Në krijimin e bazave të të dhënave që përdoren shpesh nga analistët dhe statisticienët, disa skedarë të dhënash janë kombinuar nëpërmjet teknikave të kombinimit statistikor për të pasuruar skedarin e të dhënave marrëse. Ky proces kërkon ruajtjen e supozimit të pavarësisë me kusht (CIA)[1] , i cili mund të çojë në gabime serioze në lidhjet ndërmjet variablave. Në studimin për këtë artikull, janë kryer metoda të kombinimit statistikor. Rezultatet janë bazuar në të dhëna reale nga skedari i Taksës mbi Vlerën e Shtuar (TVSH) dhe skedari i Anketës Tremujore pranë Ndërmarrjeve (ATN), dhe për përpunimin e tyre është përdorur gjuha e programimit R. Bazuar në një rast studimi, janë kryer disa analiza të parakushteve dhe të rezultateve të metodave të kombinimit statistikor dhe janë përmendur përfitimet e përdorimit të informacioneve ndihmëse. Është konfirmuar se supozimi i pavarësisë së kushtëzuar (CIA) mund të jetë një kushtëzim serioz që mund të kalohet në sajë të përdorimit të informacioneve ndihmëse të përshtatshme. Metodat "hot deck"[2] u panë si më të preferuara se metodat e tjera të kryera në këtë rast studimi. Ky studim u motivua nga nevoja për të analizuar së bashku të dhënat e shitjeve neto të mbledhura nga burime të ndryshme dhe në këtë rast, nga skedari i TVSH-së dhe ATN-së.

FJALË KYÇE:

Kombinimi statistikor; Informacionet ndihmëse; Supozimi i pavarësisë së kushtëzuar

1 Conditinal Independence Assumption

2 Hot deck është një metodë imputimi që merret me të dhëna që mungojnë në të cilën çdo vlerë e munguar është e zëvendësuar me një vlerë të vrojtuar në një anketë nga një njësi e ngjashme.

1. HYRJE

Kombinimi statistikor është një metodë e bazuar në modele për të ofruar informacion të bashkuar mbi variabla dhe indikatorë të mbledhur nëpërmjet burimeve të shumëfishta[3] (anketa të dizajnuara nga e njëjta popullatë). Përfitimet potenciale të kësaj mënyre kanë të bëjnë me mundësinë për të nxitur përdorimin komplementar dhe potencialin analitik të burimeve të të dhënave ekzistuese. Në këtë mënyrë, kombinimi statistikor mund të jetë një mjet për të rritur efiçiencën e përdorimit të të dhënave ekzistuese. Më shpesh, qëllimi i kombinimit është të shtojë përdorimin e informacionit, por teknikat e kombinimit janë përdorur gjithashtu për bashkimin e vlerësuesve të vrojtuar në anketa të ndryshme dhe për të përmirësuar saktësinë e këtyre vlerësuesve nëpërmjet integrimit me anketa më të mëdha.

Janë dy mënyra kryesore, në terma të output-it, që mund të merren nëpërmjet kombinimit:

a) Mënyra makro i referohet identifikimit të ndonjë strukture që përshkruan lidhjet ndërmjet variablave jo të vrojtuar së bashku, të tilla si shpërndarjet e bashkuara, shpërndarjet margjinale ose matricat e korrelacionit (D'Orazio, 2006).

b) Mënyra mikro i referohet krijimit të një skedari të dhënash mikro të plotë, ku të dhëna mbi të gjithë variablat janë të disponueshme. Kjo arrihet nëpërmjet gjenerimit të një bashkësie të dhënash të re, nga dy bashkësi të dhënash që janë të bazuara në një bashkësi variablash të përbashkët ndërmjet dy "rekordesh sintetike mikro". Në praktikë, procedurat e kombinimit mund të shihen si një problem imputimi të variablave target nga një anketë dhuruese në një marrëse. X dhe Y janë variablat e anketës dhuruese ndërsa anketa marrëse përmban variablat X dhe Z. Y, Z janë mbledhur nëpërmjet dy kampioneve të ndryshme nga e njëjta popullatë; variablat X janë mbledhur në të dy kampionet dhe janë lidhur me Y dhe Z. Lidhja ndërmjet këtyre variablave të përbashkët me variablat e vrojtuar vetëm në një nga bashkësitë e të dhënave - bashkësia e të dhënave dhuruese - do të eksplorohet dhe përdoret për të imputuar në bashkësinë tjetër të të dhënave – bashkësia e të dhënave marrëse – variablat jo të vrojtuar drejtpërdrejt. Kështu që një bashkësi të dhënash sintetike gjenerohet me informacion të plotë mbi X, Y dhe Z.

Matësit e lidhjes ndërmjet Y dhe Z me kusht X nuk mund të vlerësohen dhe zakonisht supozohen të jenë 0. Kjo është e ashtuquajtura supozimi i pavarësisë me kusht (CIA)- një pikë reference për të përcaktuar cilësinë e vlerësuesve të bazuar në kombinim.

Kur ky kusht qëndron, algoritmet e kombinimit do të prodhojnë vlerësime të sakta që reflektojnë shpërndarjen e bashkuar të vërtetë të variablave që janë mbledhur në burime të ndryshme. Për fat të keq, ky supozim i pavarësisë së kushtëzuar rrallë qëndron në praktikë dhe nuk mund të testohet nga bashkësitë e të dhënave. Nëse supozimi i pavarësisë së kushtëzuar, CIA, nuk qëndron dhe nuk ka informacione të tjera shtesë, modeli do të ketë probleme identifikimi dhe bashkësitë e të dhënave artificiale (false) të prodhuara mund të çojnë në konkluzione jo të sakta.

Një metodë për të mbajtur supozimin e pavarësisë me kusht, CIA, është përdorimi i informacionit ndihmës. Informacioni ndihmës zakonisht vjen në një nga llojet e mundshme të mëposhtme:

a) Informacion parametrik ndihmës, i marrë nga variablat "hook" [4];

b) Një bashkësi e tretë të dhënash (C) ose një mbivendosje e dy kampionëve (A, B) që ofrojnë informacion të plotë mbi (X, Y, Z).

Në një metodë parametrike kombinimi makro, informacioni ndihmës, zakonisht i mbledhur nga variablat "hook" ose nëpërmjet kampionit të mëparshëm, arkiva ose të dhëna të mbledhura, mund të jetë veçanërisht i dobishëm. Variablat "hook" mund të kontribuojnë për të rritur fuqinë shpjeguese të variablave të përbashkët dhe në këtë mënyrë për të ulur masën e pasigurisë, gjithashtu mund ta eliminojnë atë plotësisht në disa raste.

Bazat e të dhënave ndihmëse mund të jenë gjithashtu të përdorshme në një metodë kombinimi makro. Funksioni i mundshëm mund të ndahet në dy faktorë dhe skedarët e të dhënave A, B dhe C mund të bashkohen në një skedar. Raporti i fundit i ESSnet[5] mbi Integrimin e të Dhënave identifikon tre metodologji kryesore që fokusohen në përdorimin e bazave të të dhënave ndihmëse me informacion të plotë:

•Singh et al (1993) propozon një procedurë me dy hapa për përdorimin e bazave të të dhënave ndihmëse në kontekstin e metodave "hot deck".

3 Donald Rubin

4 Një hook është një variabël në të cilin mund të ruhen një funksion ose disa funksione të cilat mund të thirren në një rast të veçantë nga një program ekzistues.
5 European Statistical System

Së pari, një vlerë live[6] e variablit Z nga baza e të dhënave C imputohet në çdo njësi në bazën e të dhënave A, duke përdorur një nga procedurat "hot deck". Së dyti, për çdo rekord në A, një vlerë live finale nga B do të imputohet - ajo që i korrespondon fqinjit më të afërt në B me një distancë të llogaritur në vlerën e ndërmjetme të mëparshme.

• Një metodë tjetër për përdorimin e informacionit ndihmës, që merr në konsideratë dizajnime kampionesh komplekse, është ofruar nga Renssen (1998). Renssen identifikon dy mënyra për të ofruar vlerësues nga baza e të dhënave të bashkuara, kryesisht e fokusuar në përshtatjen e peshave:
a) Metoda e kalibrimit që është marrë nën stratifikimin me dy hapa. Kjo metodë konsiston në kalibrimin e peshave në skedarin (C) duke i kufizuar ato për të riprodhuar në C shpërndarjet margjinale të Y dhe Z të vlerësuara nga skedarët që do të kombinohen.
b) Një metodë kombinimi ku një vlerësim më kompleks i P (Y, Z) mund të përftohet nën stratifikimin sintetik me dy hapa. Ajo konsiston në përshtatjen e vlerësuesve të llogaritur nën supozimin e pavarësisë me kusht, CIA, duke përdorur mbetjet e llogaritura në skedarin C ndërmjet vlerave të parashikuara dhe të vrojtuara për Y dhe Z respektivisht.

• Metoda e tretë u propozua nga Rubin (1986) dhe konsiston në bashkimin e dy burimeve të të dhënave A dhe B. Në rastin e kampioneve të mbivendosura, vështirësitë në llogaritjen e peshave të lidhura mund të kufizojnë përdorimin e kësaj metode.

Analizimi së bashku i të dhënave të shitjeve neto të mbledhura nga burime të ndryshme është i domosdoshëm. Dy skedarët që i kanë këto të dhëna janë Skedari i Taksës mbi Vlerën e Shtuar (TVSH) dhe Anketa Tremujore pranë Ndërmarrjeve (ATN). Në një anë, skedari i TVSH-së është një burim i rëndësishëm administrativ për mbledhjen e shitjeve neto. Në anën tjetër, ATN-ja, si një burim statistikor, mbledh një sërë variablash për analizat ekonomike.

Artikulli synon të testojë përdorimin e teknikave alternative të bazuara në modele për të integruar informacionin mbi shitjet neto nga TVSH-ja -dhe ATN-ja.

Objektivat janë:

Objektivi 1: Analiza e koherencës ndërmjet statistikave të shitjeve neto të mbledhura nga ATN-ja dhe të dhënat e TVSH-së. Ky krahasim i koherencës së ATN-së me TVSH-në do të ofrojë analiza të rëndësishme mbi cilësinë e informacionit të mbledhur në ATN.

Objektivi 2: Vlerësimi i cilësisë së statistikave të shitjeve neto të marra nëpërmjet kombinimit statistikor bashkërenduar me variablat e mbledhur në ATN.

Në seksionin 2 të kërkimit paraqiten hapat kryesorë të implementimit të kombinimit, duke theksuar rezultatet kryesore në lidhje me dy objektivat e mësipërme. Ndërsa seksioni 3 përmbledh konkluzionet kryesore dhe rekomandimet për zbatimin e teknikave të kombinimit statistikor.

Pyetja e kërkimit është: A ka koherencë ndërmjet statistikave të shitjeve neto të mbledhura nga ATN dhe TVSH?

Për të realizuar objektivat e kërkimit duhen kontrolluar hipotezat mbi mungesën ose jo të koherencës ndërmjet statistikave të shitjeve neto të mbledhura nga ATN dhe TVSH.

2. KOMBINIMI STATISTIKOR: METODOLOGJIA DHE REZULTATET

Dy burimet e të dhënave – TVSH-ja si dhurues dhe ATN-ja si marrës - kanë të përbashkët disa variabla në terma të përkufizimeve, klasifikimeve, shpërndarjeve margjinale dhe periudhës së referencës. Të dy burimet e të dhënave kanë të njëjtën popullatë target - ndërmarrjet.

2.1 METODOLOGJIA

2.1.1 Krahasimi i shpërndarjeve për variablat e përbashkët

Në bazë të popullatës target (ndërmarrjet), është analizuar konsistenca e shpërndarjeve margjinale. Është bërë distanca metrike Hellinger (HD)[7] si një matës i ngjashmërisë së shpërndarjes për variablin e përbashkët të përdorur në procesin e integrimit. Më poshtë paraqiten vlerat e koeficienteve që krahasojnë dy shpërndarje të variablit të përbashkët, NVE (Klasifikimi i Aktiviteteve sipas Nomenklaturës së Veprimtarive Ekonomike, NVE Rev. 2)

6 Vlera live i korrespondon fqinjit më të afërt

7 Në probabilitet dhe statistikë, distanca Hellinger përdoret për të matur ngjashmërinë ndërmjet dy shpërndarjeve probabilitare.

```
$meas
tvd    overlap   Bhatt    Hell
 0          1       1       0
```

Indeksi i pangjashmërisë: Indeksi i pangjashmërisë është përcaktuar si distanca e variancës totale (tvd) ndërmjet shpërndarjeve margjinale dhe varion nga 0 (komplet të ngjashme) deri në 1 (komplet të pangjashme). Ky indeks përfaqëson pjesën e regjistrimeve që po shkaktojnë diferenca ndërmjet shpërndarjeve të krahasuara. Sa më i vogël të jetë indeksi i pangjashmërisë, aq më koherente janë shpërndarjet margjinale të variablit në bashkësinë e të dhënave dhuruese dhe të integruara.
Agresti sugjeron se, për sa kohë që norma e pangjashmërisë është më pak ose e barabartë me 6%, shpërndarjet margjinale të krahasuara mund të konsiderohen të qëndrueshme.

Overlap: Overlap është e kundërta e indeksit të pangjashmërisë (shuma e overlap dhe tvd është 1). Vlera e tij varion nga 0 (komplet të pangjashme) deri ne 1 (komplet të ngjashme). Sa më i lartë të jetë overlap, aq më koherente janë shpërndarjet margjinale të krahasuara. Në mënyrë analoge me sugjerimin e konsistencës së shpërndarjeve Agresti (tvd<=0.06), mund të arrihet në përfundimin që një overlap >=0.94 tregon që shpërndarjet e krahasuara mund të konsiderohen si të qëndrueshme.

Distanca Hellinger: Distanca Hellinger është një indeks pangjashmërie që përfaqëson distancën ndërmjet dy shpërndarjeve margjinale, i cili është jo negative, simetrike dhe shtrihet ndërmjet 0 dhe √2. Distanca Hellinger (Hd) matematikisht lidhet me tvd nëpërmjet ekuacionit të mëposhtëm:

$$Hd2 \leq tvd \leq Hd\sqrt{2}$$

Duke pasur parasysh ekuacionin dhe duke marrë të dhënë që tvd<=0.06, mund të derivohet që Hd<=0.042. Në literature, kur distanca Hellinger është më e vogël ose e barabartë me 0.05, dy shpërndarjet konsiderohen të qëndrueshme.

Koeficienti Bhattacharyya: Koeficienti Bhattacharyya (Bhatt) është një matës i ngjashmërisë ndërmjet dy shpërndarjeve dhe varion nga 0 deri në 1. Ky koeficient mund të përdoret për të vlerësuar ngjashmërinë relative ndërmjet dy shpërndarjeve. Sa më e lartë të jetë vlera e koeficientit Bhatt, aq më të ngjashme janë shpërndarjet. Koeficienti Bhatt mund të lidhet matematikisht me distancë Hellinger nëpërmjet ekuacionit të mëposhtëm:

$$Hd = \sqrt{1 - bhatt}$$

Duke marrë në konsideratë kufijtë e një distance Hellinger të pranueshme (<=0.05), koeficienti Bhatt do të ishte i pranueshëm nëse Bhatt>=0.9975.

Për të cilësuar ngjashmërinë ndërmjet shpërndarjeve probabilitare të të dhënave të dhuruesit dhe marrësit, është përdorur distanca metrike Hellinger, e cila merr vlerat ndërmjet 0 dhe 1. Vlera 0 tregon një ngjashmëri absolute ndërmjet dy shpërndarjeve probabilitare, ndërsa vlera 1 tregon një mospërputhje totale. Teknikat e kalibrimit të aplikuara shpjeguan një ngjashmëri absolute për variablin e përbashkët NVE Rev. 2, duke qenë se distanca metrike Hellinger është e barabartë me 0.

2.1.2 Analiza e fuqisë shpjeguese për variablat e përbashkët

Zgjedhja e variablave të kombinimit është një pikë thelbësore në kombinimin statistikor.
Pika referencë është supozimi i pavarësisë me kusht, CIA. Përmbushja e këtij kushti garanton që shpërndarjet e bashkuara të variablave të kombinimit Y dhe Z të jenë të njëjtat si në një procedurë lidhjeje absolute.
Bashkësia e variablave të përbashkët përbëhet nga kodi NVE (Klasifikimi i Aktiviteteve sipas Nomenklaturës së Veprimtarive Ekonomike, NVE Rev. 2).

2.1.3 Metodat e kombinimit

Shpesh qëllimi është të përftojmë një skedar sintetik të dhënash mikro të plotë nëpërmjet imputimit efektiv të vlerave në variablat e pa vrojtuara.
Në analizë është marrë tremujori i katërt i vitit 2016, për të cilin janë testuar disa metoda imputimi si metoda mikse për të kryer kombinimin statistikor dhe "hot deck "[8].

Funksioni Mixed.mtc[9] zbaton disa metoda mikse për të kryer kombinimin statistikor ndërmjet dy burimeve të të dhënave, të cilat janë paraqitur në vazhdim:

8 Hot deck është një metodë imputimi që merret me të dhëna që mungojnë në të cilën çdo vlerë e munguar është e zëvendësuar me një vlerë të vrojtuar në një anketë nga një njësi e ngjashme.
9 Ky funksion zbaton disa metoda mikes për të kryer kombinimin statistikor ndërmjet dy burimeve të të dhënave.

1. Në rastin e metodës së vlerësimit nën CIA (rho_YZ|X=0) dhe ku kemi vetëm vlerësim parametrash (micro=FALSE), matrica e korrelacioneve të vlerësuara është si më poshtë:

Tabela 1: Matrica e koeficientëve të korrelacioneve të vlerësuara sipas CIA

	NVE	ATN120	Turnover
NVE	1.00000000		
ATN120	0.04865750	1.000000000	
Turnover	0.04885995	0.002377403	1.00000000

Në Tabelën e mësipërme ATN120 i referohet shitjeve neto në Anketën Tremujore të Ndërmarrjeve ndërsa Turnover i referohet shitjeve neto në skedarin e TVSH-së.

2. Në rastin e metodës së vlerësimit me koefic korrelacioni të pjesshëm (rho_YZ|X=0.5) dhe k kemi vetëm vlerësim parametrash (micro=FAL matrica e korrelacioneve është si më poshtë:

Tabela 2: Matrica e koeficientëve të korrelacioi të pjesshëm ku kemi vetëm vlerësim parametr

	NVE	ATN120	Turnover
NVE	1.00000000		
ATN120	0.04865750	1.000000000	
Turnover	0.04885995	0.501188700	1.00000000

3. Në rastin e metodës së vlerësimit me koeficient korrelacioni të pjesshëm (rho_YZ|X=0.5) dhe ku kemi hap imputimi (micro=TRUE) matrica e korrelacioneve është si më poshtë:

Tabela 3: Matrica e koeficientëve të korrelacioneve të pjesshëm ku kemi hap imputimi

	NVE	ATN120	Turnover
NVE	1.00000000		
ATN120	0.04865750	1.000000000	
Turnover	0.04885995	0.501188690	1.00000000

4. Në rastin e metodës së vlerësimit Moriarity dhe Scheuren nën supozimin e pavarësisë së kushtëzuar, CIA, ku ka vetëm vlerësim parametrash (micro=FALSE), matrica e korrelacioneve është si më poshtë:

Tabela 4: Matrica e koeficientëve të korrelacioneve të pjesshëm ku ka vetëm vlerësim parametrash

	NVE	ATN120	Turnover
NVE	1.00000000		
ATN120	0.04869503	1.000000000	
Turnover	0.04889764	0.002377403	1.00000000

5. Në rastin e metodës së vlerësimit Moriarity dhe Scheuren me koeficient korrelacioni të barabartë me -0.15 (rho_YZ=-0.15), matrica e korrelacioneve është si më poshtë:

Tabela 5: Matrica e koeficientëve të korrelacioneve të pjesshëm në rastin e metodës së vlerësimit Moriarity dhe Scheuren

	NVE	ATN120	Turnover
NVE	1.00000000		
ATN120	0.04869503	1.000000000	
Turnover	0.04889764	-0.150000000	1.00000000

distancës hot deck për të kombinuar rekordet e dy burimeve të të dhënave që kanë të përbashkët disa variabla. Ky funksion gjen dhuruesit më të afërt duke llogaritur distancën Euklidiane mbi variablin NVE. Kjo krijon bashkësinë e të dhënave sintetike duke plotësuar ATN-në me shitjet neto të TVSH-së.

Për arsye se metodat e imputimit kanë zakonisht aftësi të kufizuara për të krijuar vlera në nivel individual, rezultatet vlerësohen në terma të ruajtjes së shpërndarjeve të të dhënave dhe lidhjeve multivariable (Rubin, 1996).

Për të vlerësuar fortësinë e metodave të ndryshme të aplikuara, krahasohet masa në të cilën shpërndarjet e vrojtuara në skedarin dhurues (TVSH) janë ruajtur në skedarin marrës (ATN). Distancat Hellinger përdoren sërish për të matur nivelin e ngjashmërisë së shpërndarjeve të bashkuara të shitjeve neto me variablat çelës.

Në një strukturë parametrike, supozimi i pavarësisë me kusht siguron që të dhënat të jenë të mjaftueshme për të vlerësuar parametrat e modelit.

2.2 REZULTATET

Vlerësimi i cilësisë në kontekstin e nevojave të kombinimit përbën një procedurë. Secili nga hapat si cilësia dhe koherenca e burimeve të të dhënave, teknikat e modelimit, algoritmet e kombinimit/ imputimit ka ndikim të rëndësishëm në cilësinë e rezultateve.

Kur vlerësojmë rezultatet bazuar në burimet e të dhënave të integruara, duhet marrë në konsideratë qëllimi i analizës dhe interpretimi i tyre në përputhje me objektivat e studimit. Në analizë u konsideruan tre kritere kryesore:

1) Konsistenca e shpërndarjeve të bashkuara (shitjet neto me variablat integrues) ndërmjet TVSH-së së vrojtuar, ATN-së së vrojtuar, ATN-së së imputuar. Hapat e ndjekur janë:

a) Krahasimi i TVSH-së së vrojtuar dhe ATN-së së vrojtuar ndihmon për të kontrolluar koherencën e variablave të përbashkët. Në këtë rast, ai ndihmon gjithashtu për të vlerësuar cilësinë e informacionit të shitjeve neto të mbledhur në ATN me TVSH, si përgjigje e objektivit 1.

b) Krahasimi i TVSH-së së vrojtuar dhe ATN-së së imputuar shërben si një kriter cilësie të kombinimit, me referencë objektivin 2.

c) Krahasimi ndërmjet ATN-së së vrojtuar dhe ATN-së së imputuar ndihmon për të kuptuar si realizohet kombinimi, në krahasim me informacionin e mbledhur në ATN.

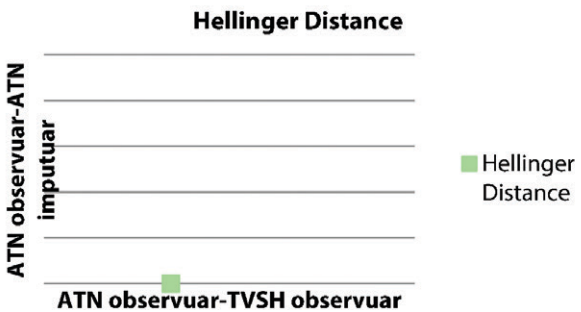
2) Konsistenca e parametrave të ndryshëm si, totalet, mesataret, etj.

3) Testimi i supozimit të pavarësisë së kushtëzuar, CIA-s, për variabla target specifikë: shitje neto.

Objektivat e studimit për këtë artikull janë si më poshtë:

OBJEKTIVI 1 - Vlerësimi i cilësisë së shitjeve neto në ATN me TVSH, si pikë referimi

Figura 1: Ngjashmëria mesatare e shpërndarjeve të bashkuara të shitjeve neto të ATN



OBJEKTIVI 2 - Vlerësimi i cilësisë së shitjeve neto të marra nëpërmjet kombinimit

Për të vlerësuar cilësinë e rezultateve të marra nëpërmjet kombinimit statistikor, i referohemi dy kritereve kryesore:
• Ruajtja e shpërndarjeve dhe parametrave kryesore ndërmjet dhuruesit dhe marrësit.
• Kapja e shpërndarjeve reale të bashkuara dhe korrelacioneve për variablat jo të mbledhura së bashku.

Pas një analize të informacionit të shitjeve neto të imputuar, në përgjithësi u vu re se parametrat e shpërndarjes së variablit të shitjes neto, si dhe shpërndarja e saj e bashkuar me variablat integruese, janë zakonisht në përputhje ndërmjet dhuruesit (TVSH e vrojtuar) dhe marrësit (ATN e imputuar). Për shembull, Figura 2 krahason pikat e këputjes për shitjet neto ndërmjet TVSH-së së vrojtuar dhe ATN-së së vrojtuar për tremujorin e katërt të vitit 2016. Siç shihet, rezultatet e kombinimit dhe mbledhjes së të dhënave janë të njëjta.

Kufizimi kryesor i kombinimit statistikor është mbështetja e tij në supozime të nënkuptuara. Kur shitja neto duhet të analizohet me variabla shtesë të mbledhura vetëm në TVSH, një kusht thelbësor për sukses është ekzistenca e variablave mirëshpjegues që japin lidhjen ndërmjet këtyre variablave.

3. KONKLUZIONE DHE REKOMANDIME

Në studimin për këtë artikull janë kryer metoda të kombinimit statistikor. Rezultatet janë bazuar në të dhëna reale nga skedari i TVSH-së dhe skedari i ATN-së, dhe për përpunimin e tyre është përdorur gjuha e programimit R. Bazuar në një rast studimi, janë kryer disa analiza të parakushteve dhe të rezultateve të metodave të kombinimit statistikor, dhe janë përmendur përfitimet e përdorimit të informacioneve ndihmëse. Studimi për këtë artikull u bazua në dy objektiva dhe konkluzionet e tyre janë si më poshtë:

OBJEKTIVI 1: Në bazë të analizës, koherenca e shitjeve neto të TVSH-së dhe ATN-së është e mirë.

OBJEKTIVI 2: Një faktor i rëndësishëm i analizës së bashkuar dhe kombinimit të ATN-së me TVSH-në është një koherencë e mirë e variablave. Diferencat e shpërndarjeve për variablat e përbashkët, të përdorur në kombinim, mund të shkaktojnë mospërputhje për vlerësuesit e shitjeve neto.

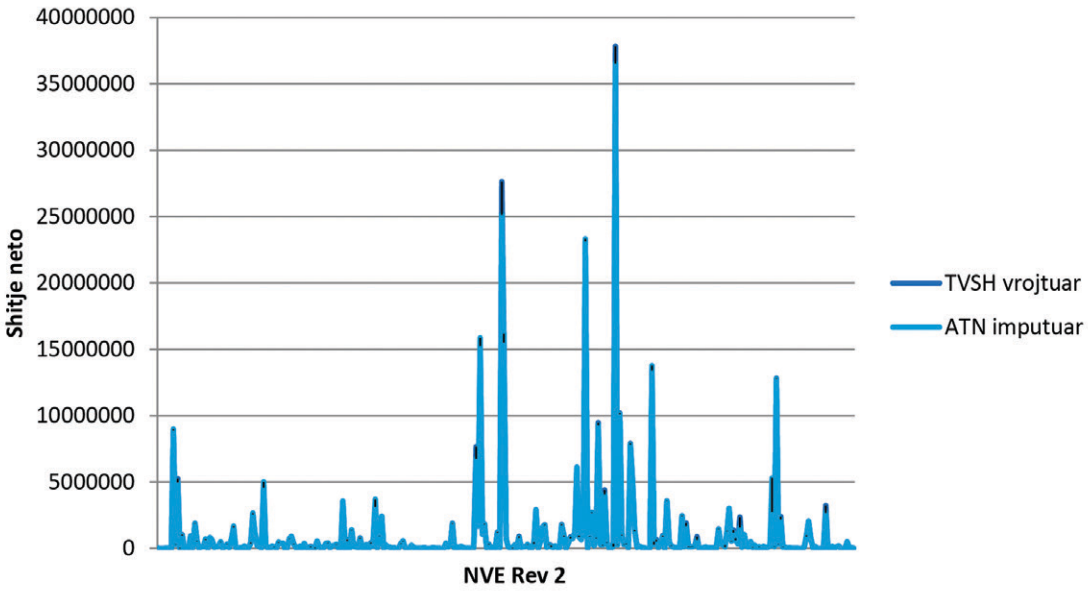
Metodat specifike të kombinimit u vërtetuan se ishin më të mirat. Megjithatë, rezultatet tentojnë të jenë të ngjashme dhe zakonisht vlerësuesit nga kombinimi janë më të ndjeshëm ndaj parakushteve të koherencës dhe variablat e përdorura në model, sesa metoda e përdorur.
Rezultatet tregojnë se, kur parakushtet e koherencës arrihen, kombinimi ofron rezultate të mira për shpërndarjet margjinale dhe shpërndarjet e bashkuara që përfshijnë dimensione të kontrolluara në model. Kur supozimet e modelit qëndrojnë, kombinimi statistikor mund të ofrojë konkluzione të mira për vlerësuesit specifikë.

Në rastin e marrë në studim, parakushtet për kombinimin janë arritur. Dy burimet e të dhënave – TVSH-ja si dhurues dhe ATN-ja si marrës – janë në përputhje, në terma të përkufizimeve, klasifikimeve, shpërndarjeve margjinale dhe periudhës së referencës, dhe të dy burimet e të dhënave kanë të njëjtën popullatë target.

Kombinimi statistikor duhet të përdoret sepse është një metodë e dobishme për të optimizuar burimet e të dhënave. Ai lejon të përdoret një madhësi e vogël kampioni duke përdorur analiza të stratifikuara a priori me madhësi kampioni më të vogla, krahasuar me një kampion të pa kombinuar me analiza të stratifikuara të bëra a posteriori.

Kombinimi shmang një analizë të stratifikuar me shumë strata dhe me të vërtetë, në një rast studimi të pa kombinuar, ndërsa kryejmë regresin logjistik, mund të përfundojmë me një strata boshe.

Figura 2: Paraqitja e vlerave të TVSH-së së vrojtuar dhe ATN-së së imputuar për tremujorin e katërt 2016



BIBLIOGRAFI

Andridge R.R., Little R.J.A. (2009) "The Use of Sample Weights in Hot Deck Imputation". Journal of Official Statistics, 25(1), 21–36.

Andridge R.R., Little R.J.A. (2009) "The Use of Sample Weights in Hot Deck Imputation". Journal of Official Statistics, 25(1), 21–36.

Andridge R.R., Little R.J.A. (2009) "The Use of Sample Weights in Hot Deck Imputation". Journal of Official Statistics, 25(1), 21–36.

Andridge R.R., Little R.J.A. (2010) "A Review of Hot Deck Imputation for Survey Nonresponse". International Statistical Review, 78, 40–64.

Andridge R.R., Little R.J.A. (2010) "A Review of Hot Deck Imputation for Survey Nonresponse". International Statistical Review, 78, 40–64.

Andridge R.R., Little R.J.A. (2010) "A Review of Hot Deck Imputation for Survey Nonresponse". International Statistical Review, 78, 40–64.

D'Orazio, M. (2014), StatMatch: Statistical Matching (aka data fusion). R package version 1.2.2. http://CRAN.R-project.org/package=StatMatch.

D'Orazio M. (2010) "Statistical matching when dealing with data from complex survey sampling", in Report of WP1. State of the art on statistical methodologies for data integration, ESSnet project on Data Integration, 33–37, http://www.essnet-portal.eu/sites/default/files/131/ESSnetDI_WP1_v1.32.pdf

D'Orazio M. (2010) "Statistical matching when dealing with data from complex survey sampling", in Report of WP1. State of the art on statistical methodologies for data integration, ESSnet project on Data Integration, 33–37, http://www.essnet-portal.eu/sites/default/files/131/ESSnetDI_WP1_v1.32.pdf

D'Orazio M. (2010) "Statistical matching when dealing with data from complex survey sampling", in Report of WP1. State of the art on statistical methodologies for data integration, ESSnet project on Data Integration, 33–37, http://www.essnet-portal.eu/sites/default/files/131/ESSnetDI_WP1_v1.32.pdf

D'Orazio M., Di Zio M., Scanu M. (2006a) "Statistical matching for categorical data: Displaying uncertainty and using logical constraints". Journal of Official Statistics 22, 137–157.

D'Orazio M., Di Zio M., Scanu M. (2006a) "Statistical matching for categorical data: Displaying uncertainty and using logical constraints". Journal of Official Statistics 22, 137–157.

D'Orazio M., Di Zio M., Scanu M. (2006a) "Statistical matching for categorical data: Displaying uncertainty and using logical constraints". Journal of Official Statistics 22, 137–157.

D'Orazio M., Di Zio M., Scanu M. (2006b) Statistical matching: Theory and practice. Wiley, Chichester

D'Orazio M., Di Zio M., Scanu M. (2006b) Statistical matching: Theory and practice. Wiley, Chichester

D'Orazio M., Di Zio M., Scanu M. (2006b) Statistical matching: Theory and practice. Wiley, Chichester

D'Orazio M., Di Zio M., Scanu M. (2008) "The statistical matching workflow", in: Report of WP1: State of the art on statistical methodologies for integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 25–26. http://cenex-isad.istat.it/

D'Orazio M., Di Zio M., Scanu M. (2008) "The statistical matching workflow", in: Report of WP1: State of the art on statistical methodologies for integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 25–26. http://cenex-isad.istat.it/

D'Orazio M., Di Zio M., Scanu M. (2008) "The statistical matching workflow", in: Report of WP1: State of the art on statistical methodologies for integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 25–26. http://cenex-isad.istat.it/

D'Orazio M., Di Zio M., Scanu M. (2010) "Old and new approaches in statistical matching when samples are drawn with complex survey designs". Proceedings of the 45th "Riunione Scientifica della Societa' Italiana di Statistica", Padova 16–18 June 2010.

D'Orazio M., Di Zio M., Scanu M. (2010) "Old and new approaches in statistical matching when samples are drawn with complex survey designs". Proceedings of the 45th "Riunione Scientifica della Societa' Italiana di Statistica", Padova 16–18 June 2010.

D'Orazio M., Di Zio M., Scanu M. (2010) "Old and new approaches in statistical matching when samples are drawn with complex survey designs". Proceedings of the 45th "Riunione Scientifica della Societa' Italiana di Statistica", Padova 16–18 June 2010.

D'Orazio M., Di Zio M., Scanu M. (2012) "Statistical Matching of Data from Complex Sample Surveys". Proceedings of the European Conference on Quality in Official Statistics - Q2012, 29 May–1 June 2012, Athens, Greece.

D'Orazio M., Di Zio M., Scanu M. (2012) "Statistical Matching of Data from Complex Sample Surveys". Proceedings of the European Conference on Quality in Official Statistics - Q2012, 29 May–1 June 2012, Athens, Greece.

D'Orazio M., Di Zio M., Scanu M. (2012) "Statistical Matching of Data from Complex Sample Surveys". Proceedings of the European Conference on Quality in Official Statistics - Q2012, 29 May–1 June 2012, Athens, Greece.

D'Orazio M., Di Zio M., Scanu, M. (2005) "A comparison among different estimators of regression parameters on statistically matched files trough an extensive simulation study". Contributi Istat, 2005/10

D'Orazio M., Di Zio M., Scanu, M. (2005) "A comparison among different estimators of regression parameters on statistically matched files trough an extensive simulation study". Contributi Istat, 2005/10

D'Orazio M., Di Zio M., Scanu, M. (2005) "A comparison among different estimators of regression parameters on statistically matched files trough an extensive simulation study". Contributi Istat, 2005/10

D'Orazio, M. (2014), StatMatch: Statistical Matching (aka data fusion). R package version 1.2.2. http://CRAN.R-project.org/package=StatMatch.

D'Orazio, M. (2014), StatMatch: Statistical Matching (aka data fusion). R package version 1.2.2. http://CRAN.R-project.org/package=StatMatch.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), Statistical Matching: Theory and Practice. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), Statistical Matching: Theory and Practice. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), Statistical Matching: Theory and Practice. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

Moriarity C., Scheuren F. (2001) "Statistical matching: a paradigm for assessing the uncertainty in the procedure". Journal of Offcial Statistics, 17, 407–422.

Moriarity C., Scheuren F. (2001) "Statistical matching: a paradigm for assessing the uncertainty in the procedure". Journal of Official Statistics, 17, 407–422.

Moriarity C., Scheuren F. (2001) "Statistical matching: a paradigm for assessing the uncertainty in the procedure". Journal of Official Statistics, 17, 407–422.
Moriarity C., Scheuren F. (2003). "A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation", Jour. of Business and Economic Statistics, 21, 65–73.

Moriarity C., Scheuren F. (2003). "A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation", Jour. of Business and Economic Statistics, 21, 65–73.

Moriarity C., Scheuren F. (2003). "A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation", Jour. of Business and Economic Statistics, 21, 65–73.

R¨assler S. (2002) Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. Springer Verlag, New York.

R¨assler S. (2002) Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. Springer Verlag, New York.

R¨assler S. (2002) Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. Springer Verlag, New York.

Renssen R.H.(1998) "Use of statistical matching techniques in calibration estimation". Survey Methodology 24, 171–183.

Renssen R.H.(1998) "Use of statistical matching techniques in calibration estimation". Survey Methodology 24, 171–183.

Renssen R.H.(1998) "Use of statistical matching techniques in calibration estimation". Survey Methodology 24, 171–183.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation". Survey Methodology, No 24, pp. 171-183.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation". Survey Methodology, No 24, pp. 171-183.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation". Survey Methodology, No 24, pp. 171-183.