UNIUNEA EUROPEANĂ

Programul Operațional Capacitate Administrativă
Competența face diferența!

Instrumente Structurale 2014-2020

ROMÂNIA

Acord de servicii de asistență tehnică rambursabile privind Consolidarea Sistemului Statistic Național (P167217)

REZULTAT Nr. 10b

Raport privind serviciile de asistență tehnică oferite Beneficiarului privind asistența tehnică și recomandările de bune practici pentru SDC, confidențialitatea datelor și modalități de securizare a microdatelor și a datelor agregate

Octombrie 2022

Revizuit noiembrie 2022

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

Proiect cofinanțat din Fondul Social European prin Programul Operațional Capacitate Administrativă 2014-2020

Declinarea responsabilității

Acest raport este un produs al personalului Băncii Mondiale. Constatările, interpretările și concluziile exprimate în acest document nu reflectă în mod obligatoriu părerile Directorilor executivi ai Băncii Mondiale sau ale guvernelor pe care aceștia le reprezintă. Banca Mondială nu garantează acuratețea datelor incluse în prezentul document și nu își asumă responsabilitatea pentru orice erori, omisiuni sau inconsecvențe în ceea ce privește informațiile sau răspunderea cu privire la utilizarea sau neutilizarea informațiilor, metodelor, proceselor sau concluziilor prezentate. Granițele, culorile, denumirile și alte informații afișate pe orice hartă din acest document nu implică nicio apreciere din partea Băncii Mondiale cu privire la statutul juridic al oricărui teritoriu sau la aprobarea sau acceptarea acestor granițe. Prezentul raport nu reprezintă neapărat poziția Uniunii Europene sau a Guvernului României.

Acest raport a fost transmis în luna octombrie 2022 și revizuit în luna noiembrie 2022, în temeiul Acordului de servicii de asistență tehnică rambursabile privind Consolidarea Sistemului Statistic Național (P167217), semnat între Institutul Național de Statistică din România și Banca Internațională pentru Reconstrucție și Dezvoltare la data de 17 septembrie 2019. Acesta face parte din Rezultatul 10 al acordului menționat mai sus.

Cuprins

Listă de acronime

BNS	Birou(ri) Național(e) de Statistică
BM	Banca Mondială
CKM	Modulul cheii de celulă
INS	Institutul Național de Statistică
NSO	Birouri Naționale de Statistică
RPL	Recensământul Populației și Locuințelor
R	Limbajul de programare R
RAS	Servicii de asistență tehnică rambursabile
SDC	Controlul divulgării datelor statistice
UE	Uniunea Europeană

Introducere

Scopul acestui raport este de a prezenta **asistența tehnică și recomandările de bune practici pentru SDC (Controlul divulgării datelor statistice), confidențialitatea datelor și modalități de securizare a microdatelor și a datelor agregate**. Acesta face parte din livrabilele din cadrul Acordului de servicii de asistență tehnică rambursabile (RAS) privind Consolidarea Sistemului Statistic Național (P167217). Proiectul este implementat de Institutul Național de Statistică (INS) cu sprijinul Băncii Mondiale.

Conform rezultatelor descrise în acest raport, echipa Băncii a oferit asistență tehnică privind modul de realizare a controlului divulgării datelor statistice (SDC), prin intermediul instruirii practice la locul de muncă, pentru a învăța metodele de asigurare a confidențialității datelor și de eliminare a riscului divulgării datelor confidențiale, metodele aplicate la nivelul microdatelor (date individuale), dar și la nivelul datelor agregate. Activitățile s-au desfășurat în perioada 28 iunie - 1 iulie 2022 și 21-22 noiembrie 2022.

Asistența tehnică practică oferită în cadrul acestui rezultat a avut scopul de a înzestra participanții cu instrumente și cunoștințe care să le permită să aplice metodele SDC seturilor de date statistice oficiale și, în special, datelor de recensământ, precum și să le permită să decidă care metode sunt cele mai potrivite în cazuri specifice de utilizare.

Raportul prezintă:

- premisele importante pentru aplicarea SDC; acestea au fost analizate și convenite cu INS și participanții la activități;
- sunt enumerate toate metodele luate în considerare pentru anonimizarea datelor RPL 2021; lista este extinsă la câteva metode de care au fost interesați participanții;
- lista scripturilor R furnizate pentru utilizarea de către personalul INS;
- concluzii și recomandări pentru anonimizarea bazei de date a recensământului populației și locuințelor;
- lista resurselor disponibile furnizată INS ca anexă.

Asistența tehnică oferită completează informațiile și practicile expuse timpul instruirii referitoare la SDC, desfășurată în decembrie 2021, și reflectate în raportul corespunzător privind discuția și explicațiile detaliate cu privire la metodele de anonimizare, furnizat în ianuarie 2022, ca parte a Rezultatului 12 din acordul RAS.

1. Scopul și obiectivele instruirii la locul de muncă privind SDC

Tema principală a instruirii practice la locul de muncă a fost aplicarea metodelor și instrumentelor care pot fi utilizate pentru a proteja seturile de date de recensământ, cu accent pe metode și referire la nevoile de instruire practică ale participanților. Scopul a fost de a înzestra participanții cu instrumente și cunoștințe, care să le permită să aplice metodele SDC seturilor de date statistice oficiale și, în special, datelor de recensământ, precum și să le permită să decidă care metode sunt cele mai potrivite în cazuri specifice de utilizare.

Scopul asistenței oferite a fost de a aprofunda înțelegerea metodelor prin aplicații și instruire practică la locul de muncă. Metodele erau destinate, în special, a fi (și au fost) aplicate bazelor de date RPL și au fost făcute recomandări (a se vedea ultimul capitol).

Obiectivele asistenței tehnice au avut în vedere ca, până la finalul instruiriii la locul de muncă, participanții:

- să cunoască bine principalele metode SDC (metode pre- și post-tabulare), atât din perspectivă teoretică, cât și practică;
- să poată aplica de la început până la sfârșit pașii principali ai unei anonimizări a datelor și, în plus, să fie capabili să explice de ce au optat pentru o anumită metodă SDC, descriind în același timp avantajele și dezavantajele acelei metode; și
- să fie utilizatori competenți ai cel puțin unui instrument software pentru ambele metode, pre- și post-tabulare.

2. Acțiuni preliminare

Înainte de furnizarea de asistență tehnică și în timpul instruirii la locul de muncă, au fost clarificate următoarele condiții preliminare:

a) În general, pentru o ghidare adecvată privind cele mai bune practici pentru INS în vederea aplicării SDC, este important să se cunoască toate statisticile produse și ierarhiile din cadrul datelor, atunci când statisticile calculate pentru zone geografice, de exemplu, mai dificile, se adună la cifre pentru zone geografice cu suprafețe mai mari. S-a clarificat faptul că statisticile detaliate la un sat tip grid de 1 km^2 (200 de locuitori) trebuie furnizate Eurostat cu referire la Regulamentul de punere în aplicare (UE) 2018/1799 al Comisiei privind stabilirea unei acțiuni statistice directe temporare pentru diseminarea tematicilor selectate ale recensământului populației și locuințelor din 2021, geocodate la nivel de griduri de 1 km^2. În total, este vorba de 13 tabele (în funcție de populația totală, sex, grupe de vârstă, persoane angajate, loc de naștere, loc de reședință) care ar trebui distribuite la nivel de griduri (prezentate în Anexa II a Regulamentului nr. 2018/1799). În plus, hipercuburi (date tabelare) care se referă la Regulamentul (UE) 2017/712 al Comisiei din 20 aprilie 2017 de stabilire a anului de referință și a programului de date statistice și de metadate privind recensământul populației și locuințelor prevăzut de Regulamentul (CE) nr. 763/2008 al Parlamentului European și al Consiliului trebuie transmise către Eurostat.

b) Pentru a aplica SDC într-un mod care anonimizează rezultatele potențiale ale unor noi analize și a opțiunilor de partajare a datelor, este important să fim conștienți de ce rezultate potențiale ar putea fi produse în plus față de nevoile actuale. Astfel, o altă cerință a fost să se clarifice dacă obligațiile de livrare (de la/pentru Eurostat) urmează

să fie publicate exclusiv din setul de date sau dacă există și evaluări/tabele suplimentare care ar trebui publicate (la nivel național) și din care ar putea fi generate noi rezultate în viitor. Rezultatul discuției a fost că datele ar trebui să fie pregătite pentru a fi gata de a fi publicate la nivel de sat, etnic și cultural (etnie, limba maternă și religie), la nivelul NUTS3 (municipii, orașe, comune). Numai timpul ne va spune dacă vor fi necesare livrări suplimentare de date.

c) În ceea ce privește microdatele, dar și specificațiile de protecție a datelor tabelare: unele variabile includ o mulțime de categorii, de exemplu, codul ocupațional. Cum și unde pot fi generalizate acele categorisiri pentru a avea la final mai puține? Această întrebare nu a fost clarificată de INS. Deoarece SDC este aplicat pentru prima dată de către INS, acesta trebuie să stabilească cum să se configureze acest cod ocupațional – 6 cifre, dar doar cu 4 cifre publicate (este necesară discuția cu experții). Pentru alte variabile, cum ar fi grupele de vârstă, alegerea grupelor de vârstă a fost clarificată.

d) Nu au existat tendințe (după formarea efectuată în decembrie 2021) în ceea ce privește alegerea metodelor pentru datele tabelare și/sau microdate. În plus, s-a menționat modul în care alte țări aplică metodele SDC pentru datele recensământului populației și locuințelor. Austria, de exemplu, aplică schimbul de date vizate – (eng: target swapping).

e) S-a clarificat necesitatea instalării R și RStudio cu pachetele asociate pentru anonimizarea datelor (sdcTable, sdcMicro, simPop), prin faptul că pachetele R, RStudio și SDC sunt disponibile la INS.

f) Participanții la formarea la locul de muncă sunt aceleași persoane care au participat la formarea desfășurată în anul 2021, pentru a asigura o consolidare a informațiilor și competențelor acestora. Suplimentar au fost incluși trei participanți. Toți au cel puțin cunoștințe de bază în R.

g) S-a clarificat faptul că se acordă sprijin în teorie și practică, iar problemele sunt discutate. De asemenea, s-a clarificat faptul că întreaga activitate de anonimizare a datelor privind populația și locuințele este responsabilitatea INS și va rămâne în sarcina INS.

h) S-a decis să se discute metodele SDC în lumina RPL 2011. În special, SDC ar trebui aplicat pe seturi (eșantioane) de date din 2011 (instruire la locul de muncă), apoi INS ar trebui să repete procesul pentru aceleași eșantioane, iar pentru celelalte seturi, să învețe urmând calea și procesul. S-a luat în considerare aplicarea SDC și asupra datelor colectate efectiv (de exemplu, din localitățile în care recensământul s-a filnalizat în etapa de autorecenzare). INS va replica procesul și va învăța cum să-l pună în aplicare și, odată ce datele sunt colectate la sfârșitul recensământului și validate, va putea aplica metoda la datele reale ale RPL 2021 colectate. Trebuie remarcat faptul că subsetul RPL 2011 a fost livrat experților BM în prima zi a instruirii la locul de muncă.

i) Metodele SDC aplicate pe seturile de date alese de INS, rezultatele și aspectele tehnice specifice observate în timpul testării vor fi discutate și explicate în cadrul a două sesiuni online ce se vor susține în perioada 21-22 noiembrie 2022.

3. Principalele aspecte abordate în timpul instruirii

În profunzime, formarea la locul de muncă a acoperit următoarele subiecte cu aplicații în R asupra bazelor de date ale RPL:

- Schimbul de înregistrări - Record swapping (vizată)
- Metoda cheii de celulă - Cell Key Method (CKM)
- Anonimizarea microdatelor, inclusiv evaluarea riscurilor, metodele de anonimizare și evaluarea utilității datelor
- SDC spațial

Aceste metode au fost explicate în detaliu în raportul transmis în urma instruirii desfășurate în decembrie 2021 (a se vedea Rezultatul 12). În timpul instruirii la locul de muncă, aceste metode au fost aplicate în mod practic.

Datorită interesului mare al participanților, au fost abordate și următoarele subiecte:

- Generarea de date sintetice
- Anonimizarea datelor spațiale

Niciunul dintre subiecte nu a făcut parte din formarea din 2021. Au fost furnizate prezentări și coduri pentru aceste subiecte, a se vedea Anexa.

De asemenea, au avut loc discuții teoretice și practice privind:

- repetarea metodelor de anonimizare pe scurt
- introducerea în conținutul raportului furnizat în urma sesiunii de training din decembrie 2021; menționăm că raportul întocmit în decembrie 2021 nu a fost primit și vizualizat de către participanți până la momentul instruirii practice, prin urmare, s-a discutat și despre conținutul părților principale ale acestui raport.
- avantajele și dezavantajele metodelor pentru anumite seturi de date și selectarea metodelor pentru anumite seturi de date.

4. Materialele în cod R furnizate

În continuare, în această secțiune, sunt prezentate materialele care au fost pregătite, prezentate și furnizate INS în cod R de către echipa Băncii.

A. Pregătirea datelor (fișier prep-data.R)

Script care importă subseturile din RPL 2011 și RPL 2021 în R, inclusiv unele pregătiri de date.

B. Aplicarea metodei cell key - cheii de celule (fișier practical_cellkey_a.R)

Pe eșantionul RPL 2011, a fost prezentată funcționalitatea pachetului cellKey. De reținut că instruirea privind cellkey (cheia de celule) a fost efectuată în prima zi a activității de formare la locul de muncă, când încă nu exista acces la noul set de date RPL 2021. Etapele de formare rămase au fost apoi aplicate pe datele RPL 2021.

Formarea cu privire la metoda CK a implicat o abordare pas cu pas a modului de protejare a unui tabel

- a se citi setul de date eșantion + a se adăuga cheile de înregistrare
- a se recodifica setul de date pentru a putea defini un tabel „sex x grupe de vârstă x ocupație"
- a se crea ierarhiile necesare
- a se defini parametrii de perturbare
- a se perturba tabelul
- a se extrage și salva rezultatele perturbate

C. Aplicarea metodei cell key - cheii de celule II (fișier practical_cellkey_b.R)

Unul dintre tabelele obligatorii care trebuie transmise pentru recensământ a fost perturbat folosind metoda cheii de celule, în special punctul 12 din Regulamentul Comisiei (UE) 2017/712, și anume populația în funcție de sex și starea civilă, pe grupe de vârstă - categorii de localități. Din motive didactice, acest script a fost elaborat pe parcursul instruirii cu implicarea activă a participanților și sub îndrumarea experților BM.

D. Pregătirea datelor necesară înainte de aplicarea record swapping - schimbului de înregistrări (fișier records.R)

Se aplică unele înregistrări ale variabilelor.

E. Schimbul de înregistrări vizate – targeted record swapping (fișier recordSwapping_a.R)

Schimbul de înregistrări a fost aplicat pe un subset al bazei de date RPL 2021.

F. SDC spațial (fișier sdcSpatial.R)

Scriptul prezintă cum să se afle care locații sunt considerate sensibile/nesigure pentru publicare și aplicarea unor metode de protecție care reduc sensibilitatea și îmbunătățesc modelele spațiale prin filtrarea, aglomerarea și eliminarea locațiilor sensibile.

G. Aplicarea metodelor tradiționale de anonimizare și simulare a datelor sintetice pe datele de sondaj și datele RPL 2021 (fișier training_2022_class_sdc.R)

Acest script mai lung conține aplicarea metodelor SDC, dar arată și simularea datelor sintetice. Include patru părți:

- anonimizarea tradițională (SDC) pe datele unui singur sondaj
- SDC tradițional pe datele de RECENSĂMÂNT
- date sintetice pe datele unui singur sondaj
- date sintetice pentru datele de RECENSĂMÂNT

H. Funcția de utilitate care se ocupă de gospodării (fișier utils.R)

Această funcție de utilitate este folosită pentru a extrage gospodăriile din baza de date RPL.

5. Concluzii și recomandări

Asistența tehnică și instruirea la locul de muncă au fost necesare pentru a demonstra modul în care metodele SDC sunt aplicate în practică datelor privind recensământul populației și locuințelor și, în general, altor seturi de date.

Echipa Băncii a observat că această instruire practică a fost chiar mai utilă pentru participanți decât formarea desfășurată în decembrie 2021 și a reprezentat de fapt modalitatea de consolidare a cunoștințelor și de aplicare a instrumentelor SDC specifice. Din cauza subiectului vast și a metodelor complexe, a fost totuși necesară și pregătirea cu caracter teoretic oferită inițial, pentru a putea cunoaște elementele de bază și pentru a permite participanților să concentreze pe deplin pe exemplele practice din această instruire. Aceasta ar putea fi o bună practică privind toate instrumentele de producție specifice similare pentru învățare și aplicare continuă în activitatea INS.

Feedbackul primit direct de la participanți a fost în general în foarte mare măsură pozitiv și confirmă abordarea experților conform căreia materialul și exercițiile care au fost pregătite și discutate pe parcursul cursului vor fi, de asemenea, utile în activitatea de zi cu zi a participanților. Participanții ar trebui să fie capabili să poată utiliza codul R dezvoltat de echipa Băncii în problemele lor practice și să anonimizeze baza de date RPL.

Discuțiile care au avut loc și părți din scripturile care au fost furnizate au fost dezvoltate și menținute într-o manieră coordonată împreună cu participanții. Acest lucru a fost realizat din motive didactice, pentru a implica activ participanții într-un mod practic. Acest lucru a permis „feedbackul instantaneu" și a sprijinit discuțiile.

Participanții au fost, de asemenea, interesați de cum să protejeze alte seturi de date. Pe parcursul instuirii practice, au fost oferite și recomandări pentru anumite alte seturi de date ale INS. Acest lucru a sprijinit, de asemenea, discuțiile și a oferit participanților o viziune mai amplă asupra acestui subiect.

În cele din urmă, sunt făcute câteva recomandări. În special, aceste recomandări sunt incluse și în raportul transmis în decembrie 2021 (a se vedea Rezultatul 12), dar au fost restrânse și se bazează acum pe noile informații primite de la INS.

Principala recomandare ar putea fi folosirea schimbului de înregistrări pentru protejarea bazei de date RPL. Există mai multe motive care fundamentează recomandarea, dar decizia finală rămâne la latitudinea INS, care ar putea lua în considerare următoarele:

a) Este evident că pe lângă obligațiile de livrare, se vor solicita date suplimentare (variabile suplimentare, alte defalcări, ierarhii etc.); acest lucru înseamnă, de asemenea că, în realitate, nu este clar ce se va publica de fapt din microdatele de recensământ. Toate tabelele și alte obligații de livrare (nivel de hipercub + grid) pentru Eurostat vor fi doar o parte a întregii publicații.

b) Aceasta înseamnă că singura soluție de a anonimiza datele în mod consecvent este o (singură) perturbare a microdatelor (cum ar fi schimbul - swapping). Acest set de date perturbat este apoi cel din care sunt produse toate celelalte publicații, tabele, hipercuburi și altele.

c) Această abordare rezolvă și problema „satelor", care fac parte din localități, și alte probleme de ierarhie; în caz contrar, este imposibil să se producă o anonimizare consistentă (și sigură).

d) Ar trebui să se precizeze clar, că nu are sens să se lucreze cu eliminarea primară și secundară în hipercuburi sau griduri, pentru a elimina o mulțime de informații și ulterior să nu se poată garanta că: a) celulele legate au întotdeauna același status (deschis/eliminat); și b) prin urmare, celulele eliminate individuale pot fi descoperite. O astfel de abordare (eliminarea celulelor) consumă extrem de mult timp și duce adesea la întrebări, care sunt dificil de explicat (de exemplu, inconsecvențe).

e) Există un motiv pentru care alte BNS preferă să folosească schimbul de înregistrări (eng.: records swqapping) pentru datele de recensământ. Reglementările, privind hipercuburile de date necesare, duc la unele dintre cuburi de data care sunt în esență microdate (procent mare de celule cu unități/persoane unice). În plus, multe celule din cuburile de date se suprapun.

f) Un beneficiu imens al acestei variante este, de asemenea, ce ar putea oferi mai târziu cercetătorilor, de exemplu, statistici la nivel de grid restrâns, fără a fi nevoie de nicio schimbare.

g) Ar trebui să se acorde o atenție deosebită comunicării modului de anonimizare către manageri, părți interesate, avocați și public.

h) Un dezavantaj al schimbului de înregistrări (eng: records swapping) este că aceasta este cea mai dificilă metodă de aplicat, deoarece software-ul este destul de dificil de utilizat.

i) Mai trebuie făcute clarificări pentru INS cu privire la detaliile categoriilor unor variabile. Acestea includ în special posibila recodificare a nivelurilor de cod de 6 cifre în funcție de ocupație, de exemplu, coduri de 2 sau 4 cifre. Experții BM recomandă insistent ca INS să evalueze, de asemenea, utilitatea datelor și riscul de divulgare după recodificare.

Acțiunile principale care trebuie întreprinse includ adaptarea codului furnizat de experți pentru baza de date completă a RPL și decizia finală privind alegerea metodei.

Anexe

1. Participanții la formarea la locul de muncă au primit codurile R menționate în Secțiunea 4 (A-H) și toate cadrele dintr-un folder Google Drive partajat de INS: https://drive.google.com/drive/folders/1XOS09XB5DOEjnk2VWlhS6ZORfbzw3kJ5?usp=sharing_eip_m&ts=62bc30b (acces doar pe bază de invitație de la INS). Acesta include, de asemenea, cadre detaliate pe subiecte noi, cum ar fi generarea de date sintetice și SDC spațial, deoarece aceste două metode nu au făcut parte din formarea din 2021.

2. Codurile pentru sdcMicro, sdcTable și simPop păstrate și scrise de echipa Băncii, sunt disponibile la adresa: https://cran.r-project.org

3. Codul pentru metoda cheii de celule (enf: cell key method) este disponibil la adresa: https://github.com/sdcTools/cellKey .

4. Raportul privind formarea desfășurată în decembrie 2021, care detaliază metodele, a fost pus la dispoziție și este, de asemenea, inclus în folderul Google Drive de mai sus.

Competența face diferența!
Proiect selectat în cadrul Programului Operațional Capacitate Administrativă, cofinanțat de Uniunea Europeană, din Fondul Social European