



## ROMÂNIA

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

## REZULTAT Nr. 6b

**Raport privind serviciile de asistență tehnică oferite Beneficiarului privind Metodologia referitoare la metodele de eșantionare la trecerea la un alt tip de cadru de eșantion (de exemplu, de la eșantion de bază (eșantion master) la cadrul bazat pe registre) pentru anchete și statistici în perioadele intercenzitare**

Iulie 2022

Revizuit septembrie 2022



*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 aprobarea sau acceptarea acestor granițe.

Prezentul raport nu reprezintă neapărat poziția Uniunii Europene sau a Guvernului României.

## **Declarația privind drepturile de autor**

Materialul din această publicație este protejat de drepturi de autor. Copierea și/sau transmiterea unor porțiuni din această lucrare fără permisiune poate constitui o încălcare a legilor aplicabile.

Pentru permisiunea de a fotocopia sau imprima orice parte din această lucrare, vă rugăm să trimiteți o solicitare cu informațiile complete către: (i) Institutul Național de Statistică din România (Bd. Libertății nr. 16, Sector 5, București, România) sau (ii) Grupul Băncii Mondiale România (str. Vasile Lascăr nr. 31, etaj 6, București, România).

Acest raport a fost livrat în luna iulie 2022 și revizuit în luna septembrie 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 6 din acordul menționat mai sus.

## **Mulțumiri**

Acest raport a fost întocmit de o echipă a Băncii Mondiale formată din Tiziana Tuoto, Carlo Vaccari și Antonia Ciprian Alexandru, sub coordonarea lui Michael Wild. Echipa dorește să își exprime recunoștința față de funcționarii guvernamentali ai Institutului Național de Statistică (INS), în special față de Silvia Pistică (Manager de Proiect), Lavinia Bălțeanu (Responsabil de Proces), Ruxandra Moldoveanu (Expert), Laura Ichim (Expert), Manuela Vlaicu (Expert), Ana Maria Ciuhu (Expert), Mihaela Ram-Chand (Expert), Florica Frîncu (Expert) pentru colaborarea constructivă.

## Cuprins

|                                                                                                                                                                                                                                                           |           |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>Cuprins.....</b>                                                                                                                                                                                                                                       | <b>4</b>  |
| <b>Listă de figuri .....</b>                                                                                                                                                                                                                              | <b>5</b>  |
| <b>Listă de acronime.....</b>                                                                                                                                                                                                                             | <b>5</b>  |
| <b>Introducere .....</b>                                                                                                                                                                                                                                  | <b>6</b>  |
| <b>0. Premise privind unitatea de referință a unui registru statistic de populație pentru a fi utilizat pentru a produce rezultate de tip recensământ în perioada intercenzitară și drept cadru de eșantionare pentru anchete în domeniul social.....</b> | <b>7</b>  |
| <b>1. Cerințe de calitate ale surselor utilizate pentru construirea bazei de anchetă și metode de evaluare a acestora .....</b>                                                                                                                           | <b>9</b>  |
| <b>2. Metode pentru construirea, menținerea și actualizarea cadrului .....</b>                                                                                                                                                                            | <b>11</b> |
| 2.1. Corelarea înregistrărilor .....                                                                                                                                                                                                                      | 12        |
| 2.2. Metodologia semnelor de viață .....                                                                                                                                                                                                                  | 12        |
| Studiu de caz: Spania .....                                                                                                                                                                                                                               | 13        |
| Studiu de caz: Marea Britanie.....                                                                                                                                                                                                                        | 13        |
| Lecții învățate privind metoda SOL.....                                                                                                                                                                                                                   | 14        |
| Combinație între metoda SOL și anchetele specifice .....                                                                                                                                                                                                  | 14        |
| 2.3. O abordare predictivă .....                                                                                                                                                                                                                          | 15        |
| 2.4. Soluționarea conflictelor/alegerea între surse .....                                                                                                                                                                                                 | 16        |
| 2.5. Metode de evaluare a supraacoperirii și subacoperirii cadrului .....                                                                                                                                                                                 | 17        |
| 2.6. Coerență între stoc și fluxuri .....                                                                                                                                                                                                                 | 19        |
| 2.7. Recomandări.....                                                                                                                                                                                                                                     | 20        |
| <b>3. Conexiunea dintre noul cadru și anchetele prin sondaj: metode de eșantionare și ponderare și posibilități de utilizare a informațiilor din ancheta prin sondaj pentru a îmbunătăți cadrul.....</b>                                                  | <b>21</b> |
| 3.1. Metode de eșantionare: câteva propuneri concrete pentru utilizare la INS.....                                                                                                                                                                        | 21        |
| 3.2. Metode de estimare: câteva propuneri concrete pentru utilizare la INS .....                                                                                                                                                                          | 22        |
| 3.3. Posibile îmbunătățiri pe termen mediu.....                                                                                                                                                                                                           | 23        |
| 3.4. Utilizarea rezultatelor anchetei prin sondaj pentru a îmbunătăți noul cadru: câteva propuneri concrete .....                                                                                                                                         | 23        |
| 3.5. Soluționarea discontinuităților în seriile de date din cauza modificărilor anchetei prin sondaj.....                                                                                                                                                 | 24        |
| <b>Referințe.....</b>                                                                                                                                                                                                                                     | <b>25</b> |

## Listă de figuri

Figura 1 - Exemplu de abordare predictivă folosind rezultatele recensământului ca reper..... 16

## Listă de acronime

|     |                                            |
|-----|--------------------------------------------|
| DBE | Ecuția de echilibrare demografică          |
| BM  | Banca Mondială                             |
| CNP | Cod numeric personal                       |
| DSE | Estimare în sistem dual                    |
| INS | Institutul Național de Statistică          |
| RAS | Servicii de asistență tehnică rambursabile |
| SOL | Metoda „semne de viață”                    |

## Introducere

Scopul acestui raport este să prezinte **metodologia referitoare la metodele de eşantionare la trecerea la alt tip de cadru de eşantion (de exemplu, de la eşantionul principal – eşantion master - la un cadru bazat pe registre) pentru anchete şi statistici în perioadele intercenzitare**. Acesta face parte din livrabilele din cadrul Acordului privind serviciile de asistenţă tehnică rambursabile privind Consolidarea Sistemului Statistic Naţional (P167217). Proiectul este implementat de Institutul Naţional de Statistică (INS) cu sprijinul Băncii Mondiale.

Acest raport îşi propune să ofere indicaţii şi recomandări cu privire la metodele de construire, menţinere şi actualizare a unui cadru bazat pe recensământ şi registre administrative care să servească drept cadru populaţional pentru calcularea numărului populaţiei şi pentru lista de eşantionare pentru anchetele sociale în perioada intercenzitară. În acest scop, raportul proiectează şi ilustrează principalii paşi metodologici, şi anume corelarea înregistrărilor, metodologia semnelor de viaţă, abordări predictive, soluţionarea conflictelor/decizia între surse şi metode de evaluare a supraacoperirii şi subacoperirii.

Raportul conţine exemple de la alte institute naţionale de statistică europene şi recomandări bazate pe cele mai bune practici internaţionale. În plus, raportul oferă indicaţii şi recomandări privind metodele de eşantionare şi ponderare pentru trecerea de la situaţia actuală la inovaţiile legate de lista populaţiei disponibile în perioadele intercenzitare. În acest scop, raportul revizuieste principalele elemente ale procedurilor actuale de eşantionare şi ponderare şi propune câteva îmbunătăţiri concrete, care să fie implementate în mod modular, de la cele mai simple şi cu modificări minime faţă de situaţia actuală, până la cele care necesită expertiză suplimentară şi schimbări mai ample în procedurile actuale.

În cele din urmă, raportul ilustrează şi recomandă câteva propuneri concrete de utilizare a rezultatelor anchetelor curente prin sondaj ca feedback pentru îmbunătăţirea şi actualizarea noului cadru şi unele propuneri de tratare a discontinuităţilor în seriile de date din cauza modificărilor anchetei prin sondaj. Este de remarcat faptul că cele mai multe dintre propunerile concrete ilustrate şi recomandate în acest raport au fost selectate din literatura extinsă privind integrarea datelor, proiectarea şi estimarea eşantionării, validarea datelor, fără prezumţia de exhaustivitate, mai degrabă cu scopul de a satisface nevoia Institutului Naţional de Statistică din România şi de a personaliza propunerile privind procedurile actuale care funcţionează în Institutul Naţional de Statistică din România pentru a facilita implementarea lor practică.

De asemenea, raportul are în vedere asistenţa tehnică acordată prin Rezultatul nr. 6d – „*Metodologia pentru o evaluare continuă şi în timp real a calităţii datelor colectate din surse administrative şi geo-spaţiale pentru perioadele intercenzitare*” prevăzute în acelaşi Acord RAS.

Această perspectivă a fost posibilă doar datorită cooperării constructive şi deschise a Institutului Naţional de Statistică.

## 0. Premise privind unitatea de referință a unui registru statistic de populație pentru a fi utilizat pentru a produce rezultate de tip recensământ în perioada intercenzitară și drept cadru de eșantionare pentru anchete în domeniul social

Factorii determinanți în evaluarea ce unități și variabile să includă în registrul statistic al populației sunt, în principiu, utilizările prevăzute. Principalele utilizări pe care le putem avea în vedere pentru registrul statistic al populației sunt producerea de ieșiri de tip recensământ în perioada intercenzitară, adică derivarea numărărilor populației defalcate pe unele variabile demografice de bază; și să servească drept cadru de eșantionare pentru anchetele sociale.

În ceea ce privește unitățile de referință de inclus în registru, să comparăm unitatea elementară „persoană” și unitatea compusă „gospodărie”, care de fapt este definită ca un grup de persoane (inclusiv grupuri compuse dintr-un singur element). În mod evident, numărările de tip recensământ sunt produse în primul rând pentru persoane, iar în unele cazuri și pentru gospodării. Când considerăm registrul ca cadru de eșantionare pentru anchetele sociale, știm că unele anchete se referă în principal la persoană ca unitate de referință, de ex. Ancheta Forței de Muncă are ca scop estimarea condiției de muncă a persoanelor care locuiesc în țară, în timp ce alte anchete se referă în principal la gospodărie, de ex. Sondajul în Venituri și Condiții de viață își propune să evalueze concepte precum venitul și sărăcia care sunt definite la nivel de gospodărie. Prin urmare, este bine să aveți ambele unități, persoane și gospodării, enumerate în cadru. O modalitate de a obține această situație este de a avea o listă de bază de persoane, identificate printr-un PIN și furnizată cu numărul de identificare corespunzător la nivel de gospodărie. În acest fel se păstrează corespondența dintre Persoană și Gospodăriile din care aparțin și se poate exploata cu ușurință registrul atât pentru utilizările la nivel de persoane, cât și de gospodării, într-un mod coerent.

Odată ce cele două tipuri de unități au fost atribuite în registru, este important să le mențineți și să le actualizați pe ambele. Actualizarea unităților de persoane este prima etapă a actualizării și gospodăriilor, din cauza caracterului compus al acestora din urmă. De obicei, sunt necesare surse externe suplimentare pentru a actualiza starea unității la nivel de unitate, cum ar fi Registrul civil al deceselor și nașterilor pentru persoane fizice și alte surse și metode discutate în următoarele secțiuni 1 și 2 (de exemplu, alte surse administrative și metode SOL). Actualizarea compunerilor de gospodărie poate fi destul de greoaie: Registrul civil poate fi folosit într-o oarecare măsură dacă raportează informații despre căsătorii și divorțuri. Uneori, chiar și sursele administrative fiscale oferă unele informații cu privire la componența gospodăriei, totuși este bine cunoscut faptul că gospodăriile reale pot fi destul de departe de gospodăriile înregistrate în Registrul Civil sau Registrul Fiscal. Prin urmare, uneori, pot fi furnizate ajustări/actualizări suplimentare pe baza rezultatelor sondajului, chiar dacă în acest caz se poate oferi îmbunătățiri la nivel de unitate doar pentru unitățile eșantionate, iar modelele statistice sunt utilizate la nivel agregat. De exemplu, după colectarea informațiilor individuale și ale gospodăriei din sondajele sociale, se poate observa că un anumit procent din eșantionul absolut al bărbatului singur de 25-30 de ani se căsătorește și creează o nouă gospodărie chiar dacă această informație nu este înregistrată prompt în sursele administrative utilizate pentru actualizarea registrului. Pentru unitățile eșantionate putem actualiza informațiile din registru și putem adăuga o nouă unitate de gospodărie la aceasta (cu un nou număr de identificare), dar putem oferi doar o estimare a noilor gospodării în ansamblu nereportate în sursele administrative utilizate pentru actualizarea Inregistreaza-te.

Când trecem la utilizarea prin eșantionare a registrului statistic al populației, în mod clar ar trebui să o alegem ca unitate de eșantionare pe cea care are legătură în principal cu fenomenul pe care ne propunem să îl investigăm. Totuși, din motive de funcționare și pentru a reduce costul interviurilor, deseori Institutul Național de Statistică selectează gospodării chiar și pentru anchetele centrate pe persoane. Această situație este luată în considerare în eroarea de eșantionare și, de obicei, un anumit efect de cluster este introdus în evaluarea varianței de eșantionare pentru a lua în considerare faptul că unitățile aparținând aceleiași gospodării nu sunt selectate aleatoriu, ci reprezintă de fapt un cluster.

Când unitatea de eșantionare este gospodăria, în teorie toate unitățile din gospodăria selectată au aceeași pondere de eșantionare, care este la nivel de gospodărie, iar în teorie toate unitățile din gospodăria selectată ar trebui intervievate. În practică, destul de des sunt admise interviuri proxy, adică un singur membru al gospodăriei este interviuat și el/ea oferă și răspunsuri pentru toți ceilalți membri ai gospodăriei, chiar dacă acest lucru poate afecta acuratețea informațiilor raportate.

Utilizările preconizate ale registrului populației ghidează și selecția variabilelor care trebuie incluse în acesta. Ar trebui să cuprindă toate variabilele demografice de bază pe care intenționăm să le folosim pentru producerea numărului populației de tip recensământ și variabilele pentru a le defalca la nivelul necesar. În plus, ar trebui să conțină toate informațiile referitoare la adresa fizică pentru a ajunge la unități sau informații de contact virtuale, cum ar fi numărul de telefon, adresa de e-mail, etc.. Variabile suplimentare disponibile eventual din sursele administrative, pot fi colectate în alt registru statistic și conectate la registrul statistic principal al populației prin intermediul PIN-ului. Chiar și informațiile istorice privind adresele sau atribuirea gospodăriei pot fi eventual colectate și stocate în registre istorice separate.

## 1. Cerințe de calitate ale surselor utilizate pentru construirea bazei de anchetă și metode de evaluare a acestora

Integrarea datelor din diferite surse devine din ce în ce mai comună în statisticile oficiale, datorită avantajelor costurilor reduse și furnizării de informații în timp util în comparație cu anchetele tradiționale. Activitatea de integrare a datelor pe care ne concentrăm în acest raport este construirea, actualizarea și menținerea unui registru statistic al populației, pe baza recensământului populației și a surselor de date administrative. Prin corelarea informațiilor din sursele disponibile la nivel de înregistrare, este posibilă determinarea persoanelor sau gospodăriilor rezidente în țară și a caracteristicilor acestora. Datele integrate din aceste surse devin un registru statistic, și anume o bază de date care poate fi utilizată pentru prelucrare și analiză ulterioară pentru a produce rezultate de tip recensământ în perioada intercenzitară (de exemplu, numărarea populației în funcție de caracteristicile demografice, cadru de eșantionare etc.).

Unele dintre procesele cheie implicate în construirea unui registru statistic al populației sunt:

- (a) identificarea surselor de date care vor fi utilizate,
- (b) corelarea surselor,
- (c) elaborarea și aplicarea unui set de reguli pentru a lua decizii cu privire la ce înregistrări ar trebui incluse în estimările finale,
- (d) soluționarea informațiilor conflictuale dintre sursele corelate.

Procedurile de integrare a datelor joacă un rol important în acest proces. Două dimensiuni cheie ale calității legate de integrarea datelor din diverse surse sunt acoperirea și coerența. Integrarea datelor se face pentru a evalua și, eventual, a reduce eroarea de acoperire. De asemenea, permite și necesită evaluarea coerenței informațiilor între surse și în timp. Aspectele de calitate și metodologiile legate de procesul de integrare a datelor, adică punctele (b)-(d), vor fi ilustrate și discutate în Secțiunea 2. În această Secțiune, ne concentrăm pe cerințele de calitate ale surselor de date în etapa de intrare, adică litera (a), pentru a evalua capacitatea acestora de a fi utilizate în scopul statistic al construirii, actualizării și menținerii unui registru statistic al populației.

De fapt, odată ce datele administrative sunt primite de către Institutul Național de Statistică (INS), calitatea acestora trebuie să fie evaluată de către INS, înainte ca datele să fie prelucrate pentru a fi utilizate în scop statistic. Rezultatul 6d<sup>1</sup> oferă îndrumări pentru identificarea surselor de date care vor fi utilizate într-un registru statistic. Listele de verificare a calității din Rezultatul 6d ajută la identificarea punctelor forte și a limitărilor diferitelor surse de date, precum și a corelării acestora și a potențialei nealinieri conceptuale cu definițiile și conceptele statistice, acoperirea și erorile de măsurare.

Prin urmare, prelucrarea datelor administrative trebuie să fie influențată de constatările obținute din listele de verificare a calității din Rezultatul 6d: de exemplu, evaluarea posibilității de corelare a unei surse administrative influențează modul în care datele sunt corelate, de exemplu, folosind un identificator unic comun al persoanei (CNP) mai degrabă decât o corelare probabilistică bazată pe un set de variabile comune potențial afectate de erori și valori lipsă. O înțelegere a erorii de acoperire

---

<sup>1</sup> Rezultatul nr. 6d - Metodologia pentru o evaluare continuă și în timp real a calității datelor colectate din surse administrative și geo-spațiale pentru perioadele interrecensământ.

în etapa de intrare influențează procesele de integrare a datelor pentru a obține acoperirea necesară registrului populației. Evaluarea acurateții datelor administrative va influența etapele de validare a datelor, inclusiv etapele de editare și intrare, și va oferi informații pentru a sprijini deciziile cu privire la modul în care sursele ar trebui utilizate împreună pentru a construi variabilele necesare. Cu toate acestea, este de menționat faptul că prelucrarea poate introduce o eroare suplimentară (sistematică sau aleatorie), generând astfel subiectivism sau variație în estimările finale. Din acest motiv, este important ca procesele să fie testate și evaluate în mod corespunzător, iar erorile potențiale trebuie gestionate de-a lungul întregului lanț de producție statistică.

În timp ce pentru etapa de intrare ne putem referi la listele de verificare a calității din Rezultatul 6d, următoarea secțiune se va concentra pe Etapa de prelucrare și va oferi o privire de ansamblu asupra metodelor și proceselor cheie utilizate pentru integrarea datelor administrative.

## 2. Metode pentru construirea, menținerea și actualizarea cadrului

Acest capitol se concentrează pe unele dintre cele mai comune metode și proceduri necesare pentru construirea, menținerea și actualizarea unui registru statistic al populației bazat pe integrarea datelor administrative și de recensământ în perioada intercenzitară. În special, vor fi discutate următoarele metode și proceduri:

1. Corelarea înregistrărilor, având ca scop combinarea mai multor surse de date;
2. Metodologia semnelor de viață, adică dezvoltarea și aplicarea unui set de reguli pentru a lua decizii cu privire la ce înregistrări trebuie incluse în estimările finale;
3. Abordarea predictivă, care urmărește evaluarea înregistrărilor care ar trebui incluse în estimările finale prin intermediul modelelor statistice;
4. Soluționarea informațiilor conflictuale dintre sursele corelate;
5. Metode de evaluare a supraacoperirii și subacoperirii cadrului.

În următoarele sesiuni, sunt ilustrate și discutate aceste metode și proceduri, abordând de asemenea și provocările aferente fiecăreia, modalitățile de evaluare a calității datelor prelucrate, pe baza literaturii disponibile și experiențele diferitelor țări, cu unele recomandări privind aplicarea metodelor în viața reală. Merită să reamintim că toți acești pași urmăresc îmbunătățirea acoperirii și coerenței registrului statistic al populației și, prin intermediul acestuia, a estimărilor finale. Deoarece acoperirea joacă un rol crucial în toți pașii de construire, menținere și actualizare a cadrului, merită să descriem și să discutăm pe scurt erorile de acoperire.

Erorile de acoperire sunt relevante atunci când cadrul populației este utilizat pentru a obține rezultate de tip recensământ, cum ar fi numărul populației, precum și atunci când cadrul este utilizat drept cadru de eșantionare. Cadrul de eșantionare este de obicei construit la o dată anterioară și cartografiază populația cât mai precis posibil pentru o anumită perioadă de referință sau zi scadentă. Prin urmare, disponibilitatea în timp util a surselor de intrare este o problemă semnificativă. Decalajul dintre cele două momente ar trebui să fie cât mai scurt posibil, deoarece riscul erorilor de acoperire crește odată cu decalajul.

În mod obișnuit, se disting trei tipuri de erori de acoperire: subacoperire, supraacoperire și înregistrări duplicate. În ancheta pe gospodării, subacoperirea apare din cauza creării de noi familii care nu sunt prezente în cadru, în timp ce supraacoperirea apare din cauza elementelor din cadru care nu (mai) aparțin populației țintă. În special, atunci când folosim registre administrative în scopuri statistice, este posibil ca:

1. indiferent dacă numărul populației dintr-o sursă administrativă este aproape de „adevăr” (cunoscut din altă sursă), acest lucru să nu garanteze că „unitățile țintă” potrivite sunt incluse;
2. sub-raportarea sau întârzierile în raportarea emigrării (statistici privind populația) și „decese” să ducă la „inflația de listă” și la supraacoperirea națională a populației – unități care nu sunt eliminate din sursa de date administrative în timp util, dacă este cazul;
3. unitățile să fie numărate „în locul greșit”, acest lucru reprezentând supraacoperiri și subacoperiri locale (mai degrabă decât naționale) sau să fie înregistrate în mai multe locuri – rezultând în supraacoperire;
4. grupurile de populație care au contact minim cu sistemele administrative să fie subacoperite;
5. din cauza decalajelor de timp, să apară subacoperirea.

Ca o consecință a subacoperirii, unele elemente au probabilitate zero de a fi selectate pentru orice eșantion extras din populația cadrului, în acest fel existând riscul de eșantion subiectiv și estimări subiective. Supracoperirea este uneori considerată o eroare mai puțin gravă în comparație cu subacoperirea, deoarece prima poate fi de obicei evitată dacă este posibil să se identifice elementele eșantionului care nu aparțin populației. De exemplu, cu o examinare atentă a lipsei de contacte în ancheta prin sondaj, este posibil să se facă distincția între supraacoperire și lipsa răspunsurilor. Este de remarcat faptul că erorile de acoperire nu sunt singurele care pot compromite utilizarea unui registru al populației drept cadru de eșantionare, de exemplu unele dintre variabilele din cadru pot fi utilizate ca instrument de stratificare (de exemplu, zonele geografice). Erorile înregistrate în aceste valori variabile înregistrate, apărute din cauza timpilor de referință diferiți sau a altor cauze, pot diminua eficacitatea stratificării și pot reduce dimensiunea eșantionului din cauza imposibilității de a contacta unitățile de eșantionare.

### 2.1. Corelarea înregistrărilor

Aproape fiecare sursă de date administrative necesită o formă de corelare a înregistrărilor cu alte surse de date, pentru validarea datelor sau pentru asigurarea unei acoperiri adecvate a unităților și variabilelor populației. Acest lucru face ca tocmai corelarea înregistrărilor să fie unul dintre cele mai importante procese de utilizare a datelor administrative în producția statistică. Prin urmare, este important și necesar să se evalueze calitatea procesului de corelare, printr-o evaluare a variabilelor sau cheilor de corelare (astfel cum se descrie în listele de verificare a calității din Rezultatul 6d) și printr-o evaluare a procesului în sine. Poate apărea o eroare de corelare prin înregistrările necorelate care ar fi trebuit să fie corelate (cunoscute și sub denumirea de „fals negativ”) și înregistrările corelate care nu ar fi trebuit să fie corelate (cunoscute și sub denumirea de „fals pozitive”).

Două metode comune pentru evaluarea calității corelării sunt:

- (a) estimarea ratelor de fals pozitiv și fals negativ, folosind o analiză administrativă a eșantioanelor de înregistrări corelate și necorelate. Revizuirea documentară poate fi efectuată numai atunci când datele nu sunt indexate prin hashing. Dacă datele sunt indexate prin hashing, INS ar trebui să obțină acces la mostre ale înregistrărilor corelate și necorelate în starea lor inițială, înainte de hashing, pentru a evalua corelarea;
- (b) compararea distribuțiilor caracteristicilor înregistrărilor corelate și necorelate (de exemplu, vârsta, sexul și cetățenia). Diferențele în distribuțiile caracteristice pot sugera că subiectivismul a fost generat prin eroarea de corelare. Aceasta înseamnă că anumite tipuri de subpopulații pot să nu fie corelate, deoarece sunt mai dificil de corelat.

Procedura de corelare aplicată la INS pare a fi facilitată de disponibilitatea unui identificator personal unic (CNP) care este prezent în toate sursele administrative și, în principiu, reduce riscul erorilor de corelare. Cu toate acestea, merită să facem câteva analize pentru a confirma că erorile de corelare pot fi considerate neglijabile și că nu există subpopulații care rămân necorelate.

### 2.2. Metodologia semnelor de viață

Odată ce integrarea datelor a reunit înregistrări din diferite surse, trebuie să dezvoltăm și să aplicăm un set de reguli pentru a lua decizii cu privire la ce înregistrări ar trebui incluse în estimările finale.

Regulile de decizie sau regulile de „activitate” sunt criterii de includere care sunt adesea aplicate la construirea registrelor statistice ale populației pentru a se asigura că numai persoanele care îndeplinesc anumite criterii de rezidență obișnuită predefinite sunt incluse în estimările finale. Acest proces este uneori cunoscut ca metoda „semnelor de viață” (SOL) și este un instrument utilizat pe scară largă pentru a reduce supraacoperirea în registrele statistice (de exemplu, includerea înregistrărilor care nu fac parte din populația rezidentă în mod obișnuit).

Câteva exemple de implementare a metodei SOL în țările europene pot ajuta la identificarea avantajelor și a semnalelor de alarmă.

### Studiu de caz: Spania

Un studiu de caz util este reprezentat de Spania, unde „semnele de prezență” se aplică la patru tipuri de surse administrative:

1. agenția fiscală și dosarele fiscale locale,
2. baza de date privind asigurările sociale compusă din indivizi cu asigurare și beneficiari (angajați și pensionari),
3. surse legate de piața muncii, inclusiv:
  - (i) baza de date a Serviciului Național pentru Șomaj care furnizează un dosar al persoanelor aflate în căutarea unui loc de muncă pentru șomeri,
  - (ii) registrele de afiliere la asigurările sociale pentru informațiile de afiliere ale populației angajate,
  - (iii) baza de date pentru asistență publică care conține informații despre beneficiarii beneficiilor,
4. Baza de date a Registrului central pentru cetățenii străini oferă informații suplimentare despre cetățenii străini care locuiesc în Spania, inclusiv data depunerii cererii pentru permisul de ședere, acordarea sau respingerea permisului de ședere, datele de expirare și verificările rezidenței.

Persoanele care au atins nivelul limită al semnelor de prezență (două din patru surse) sunt considerate „active” și incluse în numărul populației, în timp ce toate celelalte, denumite „inactive”, nu sunt incluse în număr.

Pentru minori, un semn de prezență este dacă un adult, în aceeași gospodărie, asigură el însuși semne de prezență. Minorii care nu îndeplinesc această cerință sunt excluși din numărul populației. Se analizează în prezent posibilitatea utilizării informațiilor despre studenții înscriși pentru studii oficiale.

### Studiu de caz: Marea Britanie

În Marea Britanie, se folosește o abordare similară pentru a decide care înregistrări din surse administrative selectate ar trebui incluse în estimările populației pe o bază administrativă. Într-o versiune anterioară, a fost inclusă o înregistrare în estimările populației dacă era prezentă în două dintre sursele administrative selectate.

Într-o versiune ulterioară, criteriile stricte de includere au fost aplicate fiecărei surse separat (unde înregistrările au fost incluse doar dacă a existat un semn de activitate în ultimele 12 luni) și a fost eliminată regula de includere a înregistrărilor numai dacă sunt prezente pe două surse (cu corelarea datelor folosită pentru a anula duplicarea înregistrărilor care au apărut pe mai multe surse).

Versiunea ulterioară a urmărit reducerea supraacoperirii constatate în versiunea anterioară, în detrimentul creșterii subacoperirii (înregistrări care sunt omise din estimările populației).

Așteptarea era că subacoperirea urma să fie abordată folosind o anchetă de acoperire combinată cu o metodologie de estimare în sistem dual (a se vedea capitolul 2.5).

### Lecții învățate privind metoda SOL

- Strategia SOL, regulile de decizie, criteriile de includere pot fi actualizate, odată ce constatăm că sunt prea stricte sau prea incluzive.
- Succesul unei metode SOL se bazează pe disponibilitatea unor indicatori corecți ai semnelor de activitate în sursele de date administrative individuale sau combinate.
- Aplicarea metodei presupune, de obicei, formularea unor ipoteze, care determină cine este considerat activ și cine nu. INS ar trebui să fie clar cu privire la aceste ipoteze și, acolo unde este posibil, să furnizeze dovezi relevante. În special, alegerea semnelor indicatorilor de activitate (sau a regulilor de decizie) ar trebui să fie influențată de listele de verificare a calității din Rezultatul 6d, precum și de consultarea cu furnizorii de date, validarea încrucișată între surse și în timp și cunoștințele de specialitate.
- O mare parte din evaluarea calității a implicat determinarea stabilirii limitelor pentru scorurile limită ale modelului luând în considerare relevanța, acuratețea, coerența, interpretabilitatea metodologiei și datele produse. Limita pentru includerea înregistrărilor administrative în registrele statistice este stabilită ca un echilibru între maximizarea utilizării datelor administrative pentru a îmbunătăți numărătoarea demografică națională și reducerea la minimum a erorilor (și anume, numărul de persoane recenzate în zona greșită).

### Combinație între metoda SOL și anchetele specifice

Aplicarea metodelor SOL poate fi combinată cu alte metode pentru a evalua și a contabiliza eroarea de acoperire în registrele statistice, ca în cazul Regatului Unit.

În acest caz, se efectuează o anchetă, independentă de registrul statistic, iar informațiile combinate din anchetă și registru sunt utilizate pentru a estima numărul de înregistrări care sunt omise în registru (sau în anchetă) pentru a îmbunătăți estimările din rezultatul final.

Aceasta este o abordare similară cu realizarea unei anchete postrecenzare după recensământul tradițional și aplicarea metodelor de estimare în sistem dual (cunoscute și sub denumirea de captare-recaptare, a se vedea Capitolul 2.5) pentru a evalua nivelul de subacoperire în recensământ. Această strategie este de fapt aplicată în Italia. De fapt, în Italia, SOL sunt utilizate în cadrul fiecărei municipalități pentru a exclude supraacoperirea în registru. Persoanele care nu au un astfel de SOL în municipiu sunt confirmate ca supraacoperire în registru. SOL luate în considerare sunt: calitatea de funcționar public, angajat privat sau independent, beneficierea de pensie, frecventarea școlii (inclusiv preșcolarul) sau a universității, beneficierea de ajutor de șomaj sau venit de bază sau calitatea de a fi membru de familie dependent fiscal de o persoană cu dovezi solide de SOL. În special, unitățile din registrele de populație municipale fără SOL sunt considerate supraacoperire și nu sunt incluse în estimările finale (nu sunt complet eliminate din cadru, mai degrabă etichetate ca supraacoperire pentru perioada de referință). Pentru această evaluare a supraacoperirii în SOL, acestea includ și rezultatele unei cercetări prin sondaj, bazată pe reinterviervarea unităților selectate din registrul însuși și concepută special pentru evaluarea supraacoperirii. Apoi, se efectuează o cercetare prin sondaj distinctă (în care sunt selectate unități teritoriale) – concepută pentru a estima potențiala subacoperire a cadrului corectat și, pe baza rezultatelor sale, se atașează o pondere

fiecărui individ din cadrul. Prin urmare, ponderile finale iau în considerare atât subacoperirea, cât și supraacoperirea. Încă se dezbate dacă estimările finale de numărare derivate de registru ar trebui să utilizeze valorile nominale ale ponderilor de subacoperire și supraacoperire, mai degrabă decât dihotomizarea acestora în  $\{0,1\}$ , determinând alocarea estimată a unității specifice populației de interes.

Supracoperirea în registrele statistice poate fi evaluată exclusiv prin corelarea registrului cu o anchetă printr-o abordare denumită „anchetare dependentă”, care are drept scop verificarea înregistrărilor administrative în domeniu.

Această abordare a fost utilizată în Italia și în alte țări (de exemplu, Israel) care au trecut cu succes la registrele statistice. INS poate lua în considerare posibilitatea de a susține cu o anchetă dedicată măsura supraacoperirii în registrele statistice ale populației.

### 2.3. O abordare predictivă

Pentru a depăși unele limitări determinate de subiectivismul în judecăți și ipoteze în metoda SOL, se poate adopta o abordare predictivă bazată pe modele statistice.

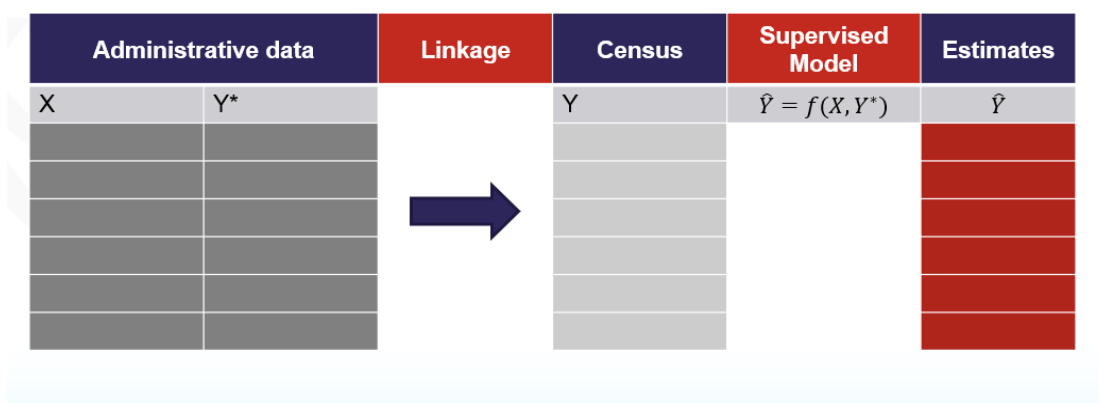
În acest caz, modelele statistice, supravegheate sau nesupravegheate, sunt aplicate pentru a prezice probabilitatea ca registrul administrativ să reflecte statutul real de rezidență obișnuită a unei persoane; persoanele cu scoruri mai mari decât limita pot fi incluse în registrul populației.

Dacă se utilizează un model supravegheat, această abordare presupune corelarea surselor de date administrative integrate cu un set de date „standard de aur” (în multe cazuri, recensământul tradițional), întocmai pentru a construi modele de evaluare a calității datelor administrative și pentru a determina în ce condiții pot fi utilizate datele administrative pentru crearea unui registru statistic al populației. De exemplu, referindu-ne la momentul recensământului, se poate folosi un model logistic pentru a prezice statutul obișnuit de rezidență observat în cadrul recensământului prin SOL derivate de la sursele administrative.

De exemplu, să luăm în considerare variabila țintă privind statutul de rezident disponibil în recensământ, respectiv  $Y$ . În datele administrative utilizate pentru actualizarea și menținerea cadrului, nu măsurăm exact variabila țintă deoarece, de exemplu, putem măsura diferențe de definiții sau decalaj, deci din datele administrative  $Y^*$ , de exemplu modelele de 0 și 1 reprezentând SOL și alte câteva variabile  $X$ , de exemplu, sex, vârstă, câteva date geografice precum rural/urban, județ. Ne propunem să prezicem variabila țintă  $Y$  folosind parametrii modelului estimați folosind datele recensământului, astfel cum este prezentat în Figura 1.

Este de remarcat faptul că predicția se poate face folosind un model parametric, de exemplu, cel logistic, sau unele neparametrice, de exemplu, un arbore de regresie sau o pădure de decizie aleatorie. În principiu, se recomandă testarea diferitelor modele pentru a le analiza și compara rezultatele și performanța.

Figura 1 - Exemplu de abordare predictivă folosind rezultatele recensământului ca reper



Legendă:

Administrative data = Date administrative

Linkage = Corelare

Census = Recensământ

Supervised Model = Model supervizat

Estimates = Estimări

În perioada intercenzitară, în general, nu este disponibil un set de date „standard de aur”; de aceea, ar putea fi necesară realizarea unui sondaj specific dedicat acestui scop. Alternativ, se pot adopta unele modele nesupravegheate, cum ar fi abordările de clasă latentă, în care toate sursele de date sunt considerate ca fiind afectate de erori în identificarea statutului obișnuit de rezidență, adică fiecare sursă este afectată de subacoperire și supraacoperire, dar prezența/absența unei evidențe în sursele administrative este corelată la statutul latent „necunoscut” al unei reședințe obișnuite. Acestea din urmă pot fi estimate prin modele statistice, iar erorile din surse pot fi evaluate.

Atât în abordările supravegheate, cât și în cele nesupravegheate, performanța modelelor statistice este utilizată pentru a determina limita de protecție împotriva subacoperirii, urmărind în același timp reducerea la minimum a supraacoperirii și a sarcinilor de lucru suplimentare, de exemplu, anchete specifice. O atenție deosebită trebuie acordată performanței modelului pe diferite zone geografice, cu diferite concentrații de grupuri de populație.

Limitările acestor abordări apar din cauza subiectivismului privind stabilirea unui scor de limită adecvat al modelului și a robusteții ipotezelor modelului de bază.

## 2.4. Soluționarea conflictelor/alegerea între surse

Când se combină datele administrative pentru a crea registre statistice, pot exista inconsecvențe în valorile variabilelor cheie din diferite surse. De exemplu, odată ce a fost luată o decizie cu privire la ce înregistrări administrative (persoane) trebuie incluse în populația rezidentă în mod obișnuit, dacă adresa unei persoane din două sau mai multe surse este diferită (de exemplu, din cauza întârzierilor în comunicarea unei schimbări de adresă, a întârzierii procesării administrative sau din motive de locuințe secundare/multiple), atunci INS poate fi nevoit să decidă la ce adresă trebuie inclusă persoana. Informațiile conflictuale (sau multiple) referitoare la adresă și orice decizie aferentă pot genera o subacoperire în unele zone și supraacoperire în altele. La nivel agregat (de exemplu, național), aceasta poate să nu fie o problemă, deoarece persoana poate fi numărată o singură dată. Cu toate acestea, la nivel de zonă restrânsă, acest lucru poate avea importanță, dacă

cele două adrese sunt în două zone diferite, deoarece ar putea genera supraacoperire într-o zonă și subacoperire în cealaltă. În plus, adresele incorecte determină scăderea eficienței eșantioanelor extrase de registru atunci când acesta este utilizat drept cadru de eșantion pentru ancheta prin sondaj, introducând non-răspunsuri suplimentare.

În contextul conflictului de adrese, pot fi alese trei abordări pentru a alege între surse:

- (a) înlăturarea înregistrărilor de la populație,
- (b) divizarea înregistrărilor între diferite locații în funcție de ponderi (de exemplu, jumătate dacă există două locații),
- (c) alegerea sursei cu cea mai mare probabilitate de actualizare în funcție de caracteristicile individului sau de variabilele administrative. Această abordare ar putea folosi surse de date suplimentare atunci când apare aceeași persoană.

Prima abordare crește subacoperirea în estimările populației. Celelalte două abordări pot produce estimări acceptabile ale populației la nivel agregat, dar pot genera distorsiuni semnificative din cauza acoperirii și erorii de corelare în estimările la niveluri inferioare de dezagregare, cum ar fi vârsta și sexul.

Pentru a rezuma, atunci când sunt disponibile aceleași atribute în surse diferite, metodele de alegere între surse se bazează de obicei pe reguli de decizie, ca în metodele SOL. Ar trebui luate în considerare și testate de către INS diferite abordări, în funcție de nevoile acestora specifice și pe baza informațiilor de calitate obținute prin listele de verificare din Rezultatul 6d.

În cele din urmă, putem lua în considerare cazul în care sursele externe de informații sunt disponibile la nivel agregat, mai degrabă decât la nivel de unitate, așa cum s-a avut în vedere până acum. De exemplu, putem avea repere derivate din programe comune cu țări străine, de exemplu, totalul emigranților către Italia și Spania, disponibil la nivel agregat, dar nu la nivel de unitate. Putem urmări să „reproducem” aceste totaluri cunoscute (marje), păstrând în același timp relațiile și dependențele dintre variabilele auxiliare observate în cadru. În această situație, se pot aplica metode statistice care sunt constrânse de marje, păstrând în același timp dependențele în tabelul de contingență. De exemplu, în acest scop, este folosită în multe aplicații și este ușor de implementat IPF – Iterative Proportional Fitting (*Ajustare proporțională iterativă*), având în vedere disponibilitatea sa în mai multe software generalizate; un dezavantaj al acestor metode este că trebuie să fie replicate pentru fiecare tabel în care ne propunem să realizăm acest tip de reconciliere a datelor.

## 2.5. Metode de evaluare a supraacoperirii și subacoperirii cadrului

Odată ce înregistrările din diferite surse de date au fost integrate în cadru, s-au decis regulile de evaluare a statutului de rezident (adică eligibilitatea unităților în populație) și au fost rezolvate conflictele dintre variabile, este în continuare nevoie să se evalueze și măsoare erorile de acoperire ale cadrului rezultat.

Pentru a ajusta erorile de acoperire, pot fi utilizați estimatori de calibrare, similari non-răspunsurilor, dacă sunt disponibile un cadru actualizat sau totaluri externe pentru populația țintă. Pentru măsurarea erorilor de acoperire, se pot folosi diferite instrumente, care se aplică la cadrul de eșantionare, registre, liste de recensământ, orice tip de listă care vizează recenzarea completă a unei populații țintă.

Pentru a măsura subacoperirea, putem alege între două alternative, în funcție de disponibilitatea unei surse absolut fiabile, așa-numitul „standard de aur”. În această situație neobișnuită, atât subacoperirea, cât și supraacoperirea pot fi măsurate pur și simplu prin comparație cu această sursă perfectă.

Pe de altă parte, atunci când toate sursele disponibile pot fi supuse subacoperirii, cea mai răspândită metodă de măsurare a subacoperirii este Estimarea în sistem dual (DSE), alias metoda de captare-recaptare. Această metodologie a fost dezvoltată inițial pentru a evalua dimensiunea populației de animale sălbatice (Petersen, 1896 și Lincoln, 1936), iar în prezent are o îndelungată tradiție și în cazul populației umane, fiind aplicată pentru a evalua sub-numărarea în recensăminte (a se vedea Wolter 1986, Fienberg 1991, Hogan 1992).

DSE necesită o anchetă prin sondaj specifică dedicată, care să fie efectuată independent de registrul/lista pentru care dorim să evaluăm subacoperirea. Din acest motiv, ancheta de subacoperire, adesea denumită studiu post-recensământ în contextul recensământului, se bazează destul de des pe un eșantion zonal. În plus, DSE necesită îndeplinirea unor ipoteze, și anume

1. corelarea perfectă între cadru și anchetă (și anume, putem identifica unitățile din ambele liste fără erori);
2. independența cadrului și a anchetei;
3. omogenitatea probabilităților de includere cel puțin într-o listă (de exemplu, ancheta);
4. închiderea populației și lipsa unităților în afara domeniului de aplicare în cadru și în anchetă (și anume, fără supraacoperire sau care a fost deja eliminată).

În acest cadru, DSE ne permite să estimăm dimensiunea necunoscută reală a populației, precum și subacoperirea cadrului.

De obicei, ancheta de subacoperire selectează un eșantion de zone dintr-un cadru de suprafață, de obicei lista zonelor de recensare identificate de recensământ. Eșantionul de suprafață garantează că ancheta este independentă de listă (recensământul sau noul cadru) față de care urmărim estimarea subacoperirii (hp. 2). În zonele selectate, se realizează o recensare completă a persoanelor rezidente, prin vizitele tradiționale din ușă în ușă, de către anchetatori cu experiență, care au fost pregătiți cu atenție pentru a evalua statutul de rezident al persoanelor întâlnite. Estimările finale aplică ponderile directe la raportul numărărilor identificate atât de ancheta de subacoperire, cât și de cadrul care depășește numărul identificat doar de anchetă, în fiecare zonă eșantionată, conform unor grupuri omogene în ceea ce privește probabilitatea de a fi observate în anchetă (hp. 3). Nu este necesară ajustarea ponderilor directe pentru non-răspuns, având în vedere natura estimărilor țintă, adică dimensiunea populației. Deoarece aceste estimări sunt în mod substanțial estimări bazate pe model, ponderile eșantionului joacă de fapt un rol limitat în rezultatele finale, adică estimările neponderate și ponderate diferă ușor.

Pentru a măsura supraacoperirea, ar putea fi util să se selecteze un al doilea eșantion/o a doua anchetă, diferit/ă de eșantionul selectat pentru a măsura subacoperirea. Obiectivul acestui eșantion este de a identifica recenzi eronate care rezultă din numărarea persoanelor neeligibile în cadru. Prin urmare, cadrul de eșantionare este acum cadrul în sine, care se presupune că este format din persoane eligibile și neeligibile. Pentru a reduce costurile și a simplifica colectarea datelor, este convenabil ca eșantionul de supraacoperire să cuprindă aceleași zone de recensare ca eșantionul de subacoperire. Totuși, anchetele privind subacoperirea și supraacoperirea rămân două anchete

distincte, prima bazată pe un eșantion zonal, cea de-a doua pe un eșantion de listă, cu două scopuri foarte diferite, primul fiind de a identifica câte persoane sunt omise de cadru, al doilea fiind de a identifica dacă în cadru există unități care nu aparțin populației țintă.

Zona de anchetă privind supraacoperirea este cunoscută și sub denumirea de anchetă de monitorizare, anchetare dependentă. Abordarea bazată pe model pentru estimarea și ajustarea pentru supra-numărare se bazează pe cinci ipoteze de bază. În plus față de primele trei ipoteze ale modelului clasic DSE, avem două ipoteze suplimentare: (A4) numărul de persoane neeligibile numărate de cadru într-o zonă de recensare este distribuit conform distribuției Poisson și (A5) numărările de persoane eligibile și neeligibile din zona de recensare sunt independente reciproc. Rețineți că ipoteza (A4) afirmă că ratele numărărilor neeligibile sunt omogene la nivelul zonelor de recensare. Acesta este un model multinomial-Poisson, propus și aplicat de exemplu în Israel.

Odată ce erorile de subacoperire și de supraacoperire sunt evaluate, estimările populației din cadrul populației pot fi corectate de acești coeficienți atât pentru subacoperire, cât și pentru supraacoperire.

## 2.6. Coerență între stoc și fluxuri

Când noul cadru integrat va fi pe deplin operațional ca sursă pentru numărarea populației și lista de eșantionare, va fi de dorit ca estimările derivate similare recensământului să fie în concordanță cu informațiile disponibile despre evenimentele demografice. În special, estimarea dimensiunii populației la momente de referință diferite ar trebui să îndeplinească ecuația de echilibrare demografică (DBE), care afirmă că numărul final al populației este egal cu numărul populației inițiale plus suma creșterii naturale și a migrației nete:

$$P^{(t+1)} = P^{(t)} + N + M$$

unde creșterea naturală,  $N = B - D$ , este diferența dintre nașteri și decese, iar migrația netă,  $M = I - E$ , este diferența dintre imigranți și emigranți.

Fiecare componentă a DBE poate fi estimată independent; în special cifrele privind nașterea, decesul și migrația pot fi obținute din datele administrative, în timp ce estimările privind dimensiunea populației la momentele de referință ulterioare pot fi derivate din noul cadru, care integrează datele recensământului, datele administrative și datele anchetei prin sondaj.

Luând în considerare erorile de eșantionare și non-eșantionare care afectează toate datele implicate, este posibil ca DBE să nu fie satisfăcută în mod trivial. Prin urmare, sunt necesare metode adecvate pentru a obține estimări finale consistente. Aceste metode ar trebui să ajusteze simultan atât (i) estimările inițiale ale dimensiunilor populației, cât și (ii) cifrele demografice brute, astfel încât datele rezultate să îndeplinească exact DBE.

Merită să subliniem că se poate folosi DBE pentru a aplica în comun (i) consistența în timp a numărului estimat al populației la care se face referire în momentele ulterioare în timp, precum și (ii) consistența spațială dintre cifrele creșterii naturale, cifrele migrației nete și estimările privind dimensiunea populației la care se face referire în diferite zone geografice.

În ceea ce privește obiectivul de consistență în timp, datele de referință ale oricăror două lansări ulterioare stabile de producție ale noului cadru par candidații firești pentru a juca rolul (t) și (t + 1) în cadrul DBE. În ceea ce privește dimensiunea spațială, intuitiv, ar fi de dorit să se folosească DBE pentru a obține coerența între creșterile naturale, migrațiile nete și estimările dimensiunii populației

la cel mai bun nivel teritorial posibil. Cu toate acestea, complexitatea de calcul a ajustărilor necesare este, desigur, de așteptat să crească odată cu cardinalitatea clasificării teritoriale adoptate. De exemplu, clasificarea NUTS3 ar putea fi un bun compromis. În acest caz, este important să se studieze fezabilitatea tehnică a aplicării simultane a DBE pentru o ierarhie imbricată a zonelor geografice.

Pentru a rezolva această problemă, metodele cele mai utile sunt cele adoptate în mod obișnuit în Institutele Naționale de Statistică pentru echilibrarea sistemelor de mari dimensiuni de conturi naționale. Într-adevăr, diviziile de Conturi Naționale ale majorității Institutelor Naționale de Statistică utilizează în mod obișnuit estimări inițiale independente care se caracterizează prin grade diferite de fiabilitate și trebuie ajustate pentru a satisface un set mare de identități contabile. O lucrare de referință importantă despre problemele de echilibrare în Conturile Naționale este Stone et al. (1942). În plus, este posibil să se recunoască impactul erorilor de măsurare asupra estimărilor inițiale și să se urmeze ideea că estimările inițiale mai puțin fiabile ar trebui să sufere ajustări mai mari. Din păcate, soluția în formă închisă propusă de Stone și colab. (1942), derivată în esență din metoda generalizată a celor mai mici pătrate, este atât de solicitantă din punct de vedere computațional, încât nu poate fi aplicată niciunei probleme de echilibrare la scară largă de interes practic. Ca o alternativă viabilă, o abordare iterativă de optimizare constrânsă este propusă în Byron (1978), care exploatează algoritmul de gradient conjugat. Această abordare este eficientă din punct de vedere computațional, chiar și pentru matrice foarte mari și este în prezent adoptată ca standard în cadrul multor divizii de Conturi Naționale. Atunci când aceste metode sunt aplicate pentru DBE, ele sunt de obicei chemate să producă Matrici de Contabilitate Socială (SAM) (vezi, Eurostat, 2003).

## 2.7. Recomandări

Câteva recomandări se aplică atunci când se utilizează surse de date administrative pentru construirea, menținerea și actualizarea cadrului populației:

- (a) acuratețea și caracterul complet al variabilelor de corelare ar trebui să fie evaluate înainte de a corela datele din diferite surse;
- (b) pentru procedura de corelare, ar trebui evaluate și raportate ratele generale de corelare și ratele de fals pozitiv/negativ;
- (c) ar trebui să fie evaluată și contabilizată eroarea de acoperire în registrul statistic al populației. Acest lucru poate fi realizat folosind comparații cu alte surse, inclusiv metodologia SOL și/sau folosind anchete (care pot fi concepute special pentru a se ajusta pentru supra sau subacoperire);
- (d) alegerea indicatorilor „SOL” (sau a regulilor de decizie) la construirea registrelor statistice ar trebui să fie bazată pe o evaluare a calității surselor, astfel cum este prezentată în Rezultatul 6d, și ar trebui testate diferite metode (și ipotezele de bază);
- (e) modelele pot fi utilizate atât pentru a evalua calitatea datelor administrative în scopul recensării unităților de populație (față de un set de date care este considerat ca fiind „reale”), cât și pentru a determina când și cum să se utilizeze datele administrative în acest scop;
- (f) atunci când se decide între surse, când aceleași atribute sunt disponibile în acestea, ar trebui luate în considerare și testate abordări diferite, în funcție de nevoile specifice recensământului și pe baza informațiilor de calitate obținute din listele de verificare a calității propuse în Rezultatul 6d.

### 3. Conexiunea dintre noul cadru și anchetele prin sondaj: metode de eșantionare și ponderare și posibilități de utilizare a informațiilor din ancheta prin sondaj pentru a îmbunătăți cadrul

Odată ce noul cadru va fi creat folosind recensământul ca piatră de temelie și va fi menținut și actualizat în mod regulat integrând registrul administrativ, vor fi posibile, în perioada intercenzitară, obținerea din acesta a unor rezultate de tip recensământ, cum ar fi numărul populației, precum și utilizarea ca listă de referință pentru anchetele sociale prin sondaj. Într-adevăr, pentru a efectua anchete prin sondaj în perioada intercenzitară, se va putea trece de la un cadru zonal, actualul cadru principal al ariilor teritoriale, la un cadru bazat pe registrul de recensământ al persoanelor/gospodăriilor și se va putea utiliza în scopul eșantionării, asigurând coerența cu numărul populației derivat din cadrul bazat pe registrul de recensământ, în plus față de eșantioane mai eficiente, întrucât va fi posibilă exploatarea în scopuri de eșantionare a tuturor informațiilor disponibile în cadru.

Crearea, menținerea și actualizarea noului cadru poate necesita acțiuni și metodologii relativ noi, de exemplu, corelări, SOL, rezolvarea conflictelor, evaluarea supraacoperirii și subacoperirii, astfel cum este descris în secțiunea anterioară. În ceea ce privește practica și metodologiile de introducere a noului cadru în scopuri de eșantionare în procesul de producție, merită să se aplice o strategie modulară, bazându-se pe puncte de continuitate cu practica actuală cât mai mult posibil în stadiul incipient și introducerea inovației odată ce adoptarea noului cadru ajunge într-un stadiu matur.

În următoarele secțiuni, sunt ilustrate și discutate câteva propuneri concrete pentru proiectarea și estimarea eșantionării. Discuția a beneficiat considerabil de activitatea comună cu echipa INS. Deoarece pentru eșantionare și estimare există mai multe metodologii disponibile și discutate în mare măsură în literatura de specialitate, următoarele sesiuni oferă doar sfaturi și ilustrații generale, precum și referințe pentru alte analize aprofundate. În plus, am încercat să facem distincție între acțiunile pe termen scurt și mediu, pentru a planifica inovațiile în metodologii și practică într-un mod care să minimizeze impactul acestora asupra proceselor de producție actuale.

#### 3.1. Metode de eșantionare: câteva propuneri concrete pentru utilizare la INS

În prezent, majoritatea anchetelor prin sondaj la INS se bazează pe o eșantionare probabilistică în două etape. Unitățile primare de eșantionare (PSU) sunt zonele, concepute printr-o procedură complexă pe baza informațiilor culese la ultimul recensământ. PSU sunt stratificate pe județe și rural/urban, echilibrate de numărul de persoane angajate în zonele respective. Unitățile de eșantionare secundare (SSU) sunt locuințele, selectate prin eșantionare aleatorie simplă sistematică.

În perioadele intercenzitare, se recomandă menținerea unei eșantionări probabilistice în două etape, modificând în mod corespunzător PSU și SSU. În cazul PSU, se recomandă selectarea unităților administrative care să eficientizeze organizarea muncii pe teren și care sunt, de asemenea, legate de estimarea domeniului. Opțiunea obișnuită în Europa sunt provinciile și/sau municipiile, unde munca de teren poate fi ușor organizată și monitorizată. O opțiune extremă este de a continua utilizarea zonelor actuale, după proiectarea pe baza noului recensământ și cartografierea în noul cadru. De fapt, această opțiune nu prezintă avantaje din perspectiva metodologică și nici operațională, în timp ce definirea și harta zonelor pe noul cadru par destul de costisitoare din punctul de vedere al timpului și al resurselor.

În ceea ce privește stratificarea PSU, aceasta este încă utilă, putând fi introduse noi variabile de stratificare, în funcție de disponibilitatea în cadru și de nevoile de estimare. De asemenea, este util să se adopte în continuare un eșantion echilibrat pentru PSU. Avantajul echilibrării pe anumite date demografice puternice disponibile în cadru, de exemplu, sexul și clasele de vârstă sau unele variabile derivate din alte anchete prin sondaj, de exemplu, numărul de persoane angajate, va îmbunătăți eficiența eșantionului și coerența estimărilor cu numărul populației derivat de cadru.

Este de remarcat faptul că noul cadru ar putea permite o actualizare mai frecventă a PSU, comparativ cu cea actuală. În principiu, PSU pot fi actualizate (și alternate) cu fiecare ocazie a unei anchete. În practică, este important să se ia în considerare organizarea muncii pe teren și modul în care este organizată și monitorizată rețeaua de anchetatori pentru a planifica actualizarea PSU astfel încât să se găsească un compromis între proprietățile statistice dorite și aplicarea lor concretă. De exemplu, în multe țări europene, anchetele sociale prin sondaj aplică o alternare a municipiilor restrânse din eșantion, cu fiecare ocazie, pentru a asigura reprezentativitatea tuturor zonelor, în timp ce orașele principale sunt incluse cu o probabilitate de câte unul în fiecare eșantion, adică sunt considerate auto-reprezentative. Această alegere este motivată de particularitățile specifice ale celor mai mari orașe față de restul țării și este probabil să fie cazul și în România.

Pentru SSU, gospodăriile (sau indivizii, conform opțiunilor operaționale din cadru) pot înlocui locuințele, în acest fel reducând, dar fără a elimina complet, efectul de cluster. Din nou, poate fi adoptat un eșantion aleatoriu simplu pentru SSU.

Pentru a rezuma, poate fi adoptată aceeași metodologie pentru eșantionare la trecerea la noul cadru: eșantion probabilistic în două etape, stratificat și echilibrat în prima etapă și eșantion aleatoriu simplu în a doua etapă. Acest lucru asigură că pot fi utilizate pentru alocare și selecție instrumentele și raționamentul actual. Principalele modificări vor viza unitățile alese atât ca PSU (unități administrative care înlocuiesc suprafețele), cât și ca SSU (gospodării care înlocuiesc locuințele) și variabilele utilizate pentru alocare, stratificare, echilibrare.

### 3.2. Metode de estimare: câteva propuneri concrete pentru utilizare la INS

În prezent, procedura de estimare se bazează pe ajustările ponderilor directe pentru răspunsul conversiei și în final pentru calibrare. Ajustările pentru non-răspuns identifică combinația dintre NUTS3 și rural/urban ca grupuri omogene care explică non-răspunsul; pentru calibrare, procedura actuală utilizează NUTS2, clase de vârstă, sex, totalul populației rurale/urbane.

În special pentru non-răspuns, noul cadru ar putea furniza cu ușurință variabile corelate într-o mai mare măsură cu non-răspunsul la nivel de gospodărie, de exemplu dimensiunea gospodăriei, vârsta sau locul de muncă al capului gospodăriei. Acest lucru va produce grupuri de ajustare mai omogene, corelate într-o mai mare măsură cu factorii care afectează non-răspunsul, asigurând o ajustare generală mai eficientă a non-răspunsului.

Pe lângă ajustarea pentru non-răspuns, etapa de calibrare va garanta coerența între estimările privind dimensiunea populației derivată din cadru și estimările anchetei.

Pentru a rezuma, aceeași metodologie și aceleași instrumente pot fi adoptate pentru etapa de estimare, ajustând ponderile directe ale eșantionului pentru non-răspuns și pentru calibrare. Modificările relevante vor consta în adaptarea strategiei actuale la noile unități de eșantion (adică

atât PSU, cât și SSU) și, eventual, alegerea unor variabile mai eficiente pentru ajustarea non-răspunsurilor.

### 3.3. Posibile îmbunătățiri pe termen mediu

Odată ce noul cadru va fi bine stabilit și adoptat atât pentru numărarea populației, cât și ca listă de eșantionare, pot fi luate în considerare inovații suplimentare pentru a îmbunătăți calitatea anchetelor sociale. De exemplu, dacă în noul cadru sunt disponibile informații de contact diferite, de exemplu, adresă, e-mail, telefon, va fi posibilă proiectarea unei scheme de colectare a datelor multimodale, astfel încât costul și timpul petrecut pentru colectarea datelor să se reducă, iar rata de răspuns să crească.

În plus, pot fi aplicate tehnici de estimare a suprafețelor mici pentru a produce estimări la nivel dezagregat, exploatând informațiile raportate în noul cadru ca variabile auxiliare în modelele cu suprafețe mici, folosind prin urmare noul cadru pentru a câștiga forță pentru unitățile neeșantionate.

Deoarece aceste metodologii nu sunt încă familiare în rândul personalului INS și nici aplicate în procesul de producție actual, se recomandă introducerea lor odată ce personalul INS s-a familiarizat deja cu noul cadru, cu întreținerea acestuia și utilizarea lui în procesele de producție curente.

### 3.4. Utilizarea rezultatelor anchetei prin sondaj pentru a îmbunătăți noul cadru: câteva propuneri concrete

Ca de obicei, anchetele prin sondaj vor produce informații precise despre unitățile eșantionate (în mod rezonabil, gospodăriile) în ceea ce privește variabilele relevante pentru noul cadru în sine: de exemplu, statutul de rezident, locația și adresa corecte și datele demografice de bază, plus informații suplimentare specifice anchetelor. Informațiile colectate pot fi corelate la noul cadru și utilizate pentru:

1. actualizarea noului cadru, înlocuind valorile incorecte pentru unitățile eșantionate;
2. estimarea modelelor statistice pentru statutul de rezidență, îmbunătățirea metodologiei SOL cu abordarea predictivă care utilizează rezultatele anchetei ca reper (a se vedea Secțiunile 2 și 3 din secțiunea anterioară, precum și Figura 1, unde recensământul trebuie înlocuit cu ancheta);
3. estimarea modelelor statistice pentru supraacoperirea (locală) a noului cadru (a se vedea capitolul 2.5);
4. estimarea subacoperirii în cadrul gospodăriilor, necesitând câteva instrucțiuni suplimentare oferite anchetatorilor (a se vedea capitolul 2.5);
5. evaluarea calității noului cadru pentru informațiile demografice de bază, dimensiunea gospodăriei, componența gospodăriei etc.;
6. estimarea modelelor statistice pentru a „prezice” variabilele specifice anchetei, folosindu-le pe cele care pot fi derivate din datele administrative (de exemplu, condiția de angajat etc.) ca variabile auxiliare.

### 3.5. Soluționarea discontinuităților în seriile de date din cauza modificărilor anchetei prin sondaj

Adoptarea noului cadru ca listă de eșantionare pentru anchetele sociale și pentru calibrarea la totaluri cunoscute derivate de cadrul însuși poate genera unele întreruperi și discontinuități în rezultatele anchetelor longitudinale, de exemplu, ancheta privind forța de muncă.

Pentru a gestiona această discontinuitate și pentru a evalua pe deplin efectele modificărilor, varianta de preferat ar fi să avem în paralel cele două serii, cu metodologia/cadrul nouă/nou și veche/vechi. Aceasta este de fapt o soluție costisitoare și uneori nu este posibilă punerea sa în aplicare pentru toate anchetele sau pe termen lung. Această soluție poate fi luată în considerare, de exemplu, pentru ancheta privind forța de muncă și pentru un an de referință. În principiu, este posibil să se aplice metode statistice pentru a gestiona discontinuitățile de primă speță. Aceste analize pot fi efectuate intern, iar alegerea diseminării lor către utilizatorii finali trebuie luată în considerare cu atenție.

Este fundamental să se comunice în mod clar utilizatorilor de date toate schimbările de metodologie, care stau la baza unei calități îmbunătățite pentru estimările rezultate, a costului redus pentru comunitate și alinierii la standardele internaționale și cele mai bune practici.

## Referințe

- Byron, R., The estimation of large social account matrices, *Journal of the Royal Statistical Society A* (1978), vol. 141(3), pp. 359-367.
- Fienberg S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika* 59, 3, 591-603
- Hogan, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *The American Statistician*, 46, 261-269
- Lincoln, F. C. (1936). Returns of Banded Birds: Third Paper (Some Recoveries of Water Birds from Latin America). *Bird-banding*, 7(4), 139-148.
- Petersen, C.G.J. (1896). The yearly immigration of young plaice into the Limfiord from the German Sea. *Report of the Danish Biological Station*, 6, 5–84
- Stone, R., Champernowne, D.G., and Meade, J.E., The Precision of National Income Estimates, *Review of Economic Studies* (1942), vol. 9 (2), pp. 111-125.
- Wolter, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346



### **Competența face diferența!**

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