Institutul Naţional de Statistică

GHID METODOLOGIC
DE EDITARE A DATELOR

2016

CUPRINS

I. OBIECTIVELE PRINCIPALE

În prezent, nevoia de date statistice de înaltă calitate, cu un grad ridicat de detaliere și produse într-un interval scurt de timp este în creştere, iar cererea de date cuprinde un spectru tot mai divers de domenii ale activităţii sociale şi economice. Conform legii 226/2009 privind organizarea şi funcţionarea statisticii oficiale în România, Institutul Naţional de Statistică, împreună cu direcţiile sale teritoriale, este principala instituţie a sistemului statistic naţional care asigură date statistice actuale, relevante şi de calitate, prin programul statistic naţional. Principalele surse pe baza cărora se realizează colectarea datelor sunt cercetări statistice selective (prin sondaj) şi exhaustive (de tipul recensămintelor). Un factor major care perturbă calitatea datelor colectate este faptul că sursele de date conţin, în mod inevitabil, erori şi valori lipsă, impunând derularea unui proces continuu de editare a datelor.

Viitorul cercetărilor statistice, care prevede utilizarea extensivă a surselor administrative şi a unor surse combinate de date pentru diverse cercetări ridică şi mai multe probleme în ceea ce priveşte utilizarea unor date corecte, coerente şi cât mai aproape de realitatea de zi cu zi. În aceste condiţii, editarea datelor capătă valenţe multiple.

Unul dintre obiectivele principale ale ghidului metodologic este de a promova o mai bună înţelegere a diferitelor metode de editare a datelor statistice rezultate din cercetările statistice, în ceea ce priveşte controlul, validarea şi imputarea datelor produse în cadrul Institutului Naţional de Statistică din România.

Ghidul îşi propune să reprezinte şi un util instrument de perfecţionare a statisticienilor, prin prezentarea unor aspecte esenţiale ale statisticii aplicate, a unor exemple din cercetările statistice implementate de Institutul Naţional de Statistică, alături de aspecte de ordin teoretic.

Prezentul ghid poate servi şi pentru o viitoare standardizare a proceselor statistice în diferite domenii de interes (cercetări în întreprinderi, în gospodării etc.).

De asemenea, ghidul prezintă o scurtă trecere în revistă a metodologiilor utilizate în procesul de editare a datelor - cel mai frecvent utilizate - în alte oficii de statistică naţionale din statele membre ale Uniunii Europene şi din alte state ale lumii (de ex. SUA, Canada).

❖ CE ESTE EDITAREA DATELOR ŞI DE CE ESTE NECESARĂ?

Ce este editarea datelor şi de ce este necesară?

Editarea datelor reprezintă ansamblul operaţiunilor aplicate datelor brute (colectate în cadrul unei cercetări statistice) pentru a îmbunătăţi calitatea rezultatelor (eliminarea erorilor), prin analiza acestora din prisma unui set de criterii sau judecăţi de valoare.

Scopul validării/ editării este asigurarea unui nivel acceptabil de calitate a datelor finale (diseminate). Validarea/ editarea nu pot asigura, în sine, calitatea datelor ci doar îndeplinirea unui set de criterii de calitate în funcţie de care datele pot fi cionsiderate ca fiind „acceptabile”. Se poate spune deci că **validarea nu va avea ca rezultat date „perfecte” sau „adevărate” ci date plauzibile**.

Dimensiunile calităţii avute în vedere de procesul de validare se referă la:

- Acurateţea – se referă la diferenţa dintre valoarea adevărată şi cea estimată a unui parametru ce trebuie măsurat. Această diferenţă este dată atât de erorile de sondaj şi de alte erori (de măsurare, de prelucrare etc.). Erorile de sondaj sunt măsurate prin metode statistice specifice şi nu pot fi corectate de procedurile de validare / editare. Alte tipuri de erori însă, ca de exemplu erorile de măsurare sau cele generate chiar de procesul de producţie statistică (codificare, prelucrare etc.) pot fi identificate şi corectate priin proceduri de validare / editare.
- Coerenţa şi comparabilitatea – se referă la gradul în care datele sunt consistente ca structură, ca evoluţie în timp şi comparabile între zone geografice.

II. DESCRIEREA ETAPELOR PROCESULUI DE EDITARE A DATELOR

Editarea datelor este definită ca procesul care implică controlul şi validarea datelor colectate pe baza cercetărilor statistice selective (anchete statistice) sau exhaustive cu scopul de a îmbunătăţi calitatea rezultatelor. Acest proces se desfăşoară în patru etape:

- Managementul/Organizarea cercetării statistice;
- Introducerea datelor în mediul electronic;
- Validarea datelor;
- Ajustarea datelor.

II.1. Managementul/Organizarea cercetării statistice

Managementul cercetării statistice este o activitate care se desfăşoară, în diferite forme şi cu diferite intensităţi, în fiecare etapă a acesteia.

Unele dintre activităţile de management premergătoare colectării propriu-zise a datelor constau în:

- asigurarea cadrului legal al cercetării statistice;
- verificarea oportunităţii şi necesităţii desfăşurării cercetării respective, precum şi a eficienţei (a concordanţei dintre costuri şi rezultate);
- asigurarea resurselor (umane, financiare şi de timp) necesare desfăşurării ei;
- asigurarea sustenabilităţii în timp a cercetării respective;
- verificarea faptului că instrumentarul proiectat permite colectarea informaţiilor necesare producerii statisticilor necesare şi astfel, că cercetarea îşi va atinge obiectivele generale şi specifice etc.

Ulterior colectării datelor, managementul cercetării statistice constă în următoarele activităţi:

a) **verificarea integrităţii**[1] şi

b) **controlul calităţii**.

[1] *Un exemplu de corelaţii pentru controlul de integritate este prezentat în Anexa 1*

Pentru cercetările statistice selective, **verificarea integrităţii** presupune condiţii de control care verifică gradul de colectare a datelor cercetării, mai exact, dacă s-a colectat ceea ce (şi cât) trebuia colectat. Verificarea integrităţii are o dimensiune cantitativă prin care se verifică că numărul de chestionare completate, respectiv cele aferente unităţilor pentru care colectarea nu s-a realizat din diverse motive, este egal cu numărul unităţilor selectate în eşantionul cercetării.

Dimensiunea calitativă urmăreşte dacă colectarea datelor s-a realizat doar de la unităţile incluse în eşantion şi nu de la alte unităţi (în cazul în care cercetarea selectivă nu permite înlocuirea unităţilor nerespondente).

Un alt aspect al verificării integrităţii priveşte integritatea între chestionare. De exemplu, într-o cercetare statistică în gospodării, pentru fiecare locuinţă ocupată trebuie să existe cel puţin un chestionar de locuinţă (care are completat cel puţin rubrica privind Rezultatul interviului) completat, iar dacă există un chestionar de locuinţă pentru care s-au colectat datele (gospodăria a acceptat interviul), numărul de chestionare individuale completate (care au completată cel puţin rubrica privind Rezultatul interviului) trebuie să fie egal cu numărul membrilor de gospodărie prezenţi sau temporar absenţi din respectiva gospodărie.

Un exemplu privind verificarea integrităţii chestionarelor statistice este prezentat în Anexa 2.

Chestionarele trebuie să conţină o rubrică referitoare la Rezultatul interviului (sau a colectărrii datelor, dacă se utilizează altă metodă de colectare decât cea a interviului) prin care să se poată identifica dacă colectarea datelor a avut loc, dacă respondentul a fost inaccesibil sau a refuzat răspunsul; aceste informaţii potând fi utilizate în cadrul procedurilor de verificare.

Organizarea cercetării statistice include controlul calității procesului de colectare a datelor și măsuri ale impactului activităţii de ajustare asupra rezultatelor. Este un pas în procesul de control al calității totale, care asigură faptul că ipotezele statistice de bază ale unei cercetări statistice selective nu sunt încălcate.

❖ ETAPELE UNEI CERCETĂRI STATISTICE

Etapele şi activităţile principalele care trebuie desfăşurate in fiecare etapă pentru realizarea unei cercetări statistice prin sondaj sunt următoarele:

I) PROIECTAREA CERCETĂRII STATISTICE are în vedere următoarele aspecte:

a. ***Definirea obiectivelor cercetării*** statistice se realizează pornind de la rezultatele ce se doresc a se obţine. In majoritatea cazurilor, o cercetare statistică este iniţiată datorită apariţiei unei nevoi de informaţie statistică, a unui indicator sau pentru a monitoriza un anumit fenomen economic sau social;

b. ***Identificarea bazei de sondaj*** celei mai adaptate la situaţie ţinând seama de aspectele de calitate si cost. Se pot utiliza eventual mai multe baze de sondaj de provenienţe diferite pentru a se realiza baza considerată a fi cea mai adecvată;

c. ***Selecţia eşantionului*** *presupune extragerea unui număr de unităţi dintr-o bază de sondaj, pe baza unor criterii pre-stabilite, de la care urmează a fi colectate datele. Volumul eşantionului se calculează în funcţie de mai multe criterii, cel mai important fiind acela a gradului d ereprezentativitate pe care dorim să-l aibă rezultatele finale ale cercetării statistice;*;

d. ***Proiectarea chestionarului***: pornind de la obiectivele cercetării statistice, de la programul său de observare, se stabilesc variabilele ce urmează a fi colectate şi, pe baza acestora,

se stabileşte conţinutul chestionarului, prin formularea întrebărilor, stabilirea fluxului acestora, a pertinenţei lor, a duratei medii de completare etc;

e. ***Proiectarea mijloacelor de prezentare a rezultatelor cercetării statistice:*** tabele de prezentare a rezultatelor, fişiere de micro-date sau baze de date on-line.

II) DESFĂŞURAREA CERCETĂRII STATISTICE IN TEREN are ca etape principale:

a. ***Pregătirea activităţilor în teren*** - sunt necesare activităţi premergătoare cum ar fi:

- **selectarea operatorilor de interviu** şi a altor categorii de personal se face pe baza unor criterii precum nivelul de instruire, experienţa anterioară în cercetări prin sondaj, abilitatea de a desfăşura un interviu, aspectul fizic etc.;
- **instruirea personalului** în vederea cunoaşterii obiectivelor cercetării, a metodelor celor mai adecvate de contactare a unităţilor din eşantion, a modului de desfăşurare a interviurilor şi de înregistrare a informaţiilor în chestionare;
- **stabilirea celor mai bune practici în colectarea datelor** poate presupune de exemplu, găsirea metodelor de evitare, iar dacă acest lucru nu este posibil, de minimizare a numărului unităţilor care nu au participat la cercetare, pentru limitarea la maxim **a non-răspunsurilor**; de cele mai multe ori operatorii de interviu trebuie să facă vizite repetate până reuşesc să contacteze unităţile pe care nu le-au găsit în prealabil sau care au refuzat iniţial să participe la cercetare; dacă aceştia nu reuşesc să convingă unitatea, o nouă încercare trebuie făcută de către supervizorul acestuia;
- **atribuirea eşantionului fiecărui operator de interviu** presupune identificarea unităţilor care trebuie intervievate de către aceştia;
- **identificarea prealabilă a unităţilor din eşantion** se realizează pentru a face publică în teren desfăşurarea cercetării, în vederea economisirii timpului în perioada de colectare a datelor şi pentru minimizarea non-răspunsurilor;
- **distribuirea materialelor cercetării către operatorii de interviu;**

b. ***Colectarea informaţiilor*** – este etapa de care depinde în mod covârşitor calitatea rezultatelor unei cercetări statistice.

Colectarea se poate realiza prin interviuri faţă-în-faţă (metodă des utilizată în cercetările selective în gospodării din România), în care înregistrarea se realizează mai întâi prin completarea răspunsurilor în chestionarele tipărite pe hârtie (ulterior acestea fiind introduse într-o bază de date) sau prin intermediul mediului online, prin chestionare electronice (care se pot descărca automat în bazele de date). Sursa datelor: persoane fizice și/sau juridice sau surse administrative disponibile.

Pentru cercetarea statistică Intrastat, colectarea datelor de la operatorii economici se face numai în format electronic, în sistem online, pe site-ul dedicat Intrastat, sau în sistem offline, prin aplicaţia Intrastat pusă la dispoziţie cu titlu gratuit de INS operatorilor economici.

Alte metode de colectare a datelor pot fi:

- auto-înregistrarea, în care respondentul completează el însuşi răspunsurile într-un chestionar;
- prin poştă, respondentul primind şi, ulterior completării, returnând chestionarele către organizaţia care a organizat cercetarea prin poştă;
- prin telefon, interviul desfăşurându-se sub forma une convorbiri telefonice între un operator de interviu şi respondent, iar răspunsurile fiind completate fie pe un chetionar pe hârtie, fie direct, într-o bază de date;

- metode mixte, în care se combină mai multe dintre metodele clasice de colectare.

III) PRELUCRAREA DATELOR

a. Principala activitate o reprezintă ***introducerea datelor din chestionare pe suport electronic şi/sau preluarea automată a datelor.***

b. ***Controlul şi validarea datelor*** presupune realizarea unui control de integritate (dacă s-au introdus toate chestionarele şi toate informaţiile conţinute de acestea) precum şi a unui control logic al datelor (controlul coerenţei datelor) şi corectarea erorilor constatate;

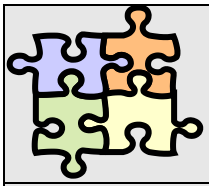
c. În această etapă de prelucrare a datelor are loc ***tratarea non-răspunsurilor***, folosind diverse metode de imputare a datelor lipsă.

IV) CALCULUL COEFICIENŢILOR DE EXTINDERE ŞI AL ESTIMATORILOR

Coeficienţii de extindere sunt necesari pentru a extrapola datele obţinute pe baza unui eşantion la nivelul întregii populaţii. Procesul de ponderare începe cu calculul ponderii de bază care, pentru o anumită unitate din eşantion, este egală cu inversul probabilităţii de selecţie. Ulterior, aceste ponderi de bază sunt ajustate pentru compensarea non-răspunsurilor şi a non-acoperirii şi pentru a face estimaţiile din eşantion conforme cu totalurile din populaţia de referinţă.

Non-răspunsurile totale se pot compensa prin 3 metode, şi anume:

- prin ajustarea ponderilor de bază;
- prin selectarea iniţială a unui eşantion mai mare care, în urma existenţei non-răspunsurilor va determina atingerea unui eşantion realizat de mărimea dorită;
- prin utilizarea substituirii, adică prin înlocuirea unor unităţi care nu răspund cu altele care nu au fost incluse în eşantion şi care sunt similare cu cele non-respondente din punctul de vedere al caracteristicilor de interes.



Caseta 1: Exemplificare din ***STATISTICA SOCIALĂ***

În domeniul statisticii sociale extinderea rezultatelor obţinute din anchetă se realizează pe baza coeficienţilor atribuiţi persoanelor din gospodăriile din eşantion, care au răspuns la interviu. Pentru determinarea acestor coeficienţi este necesară parcurgerea următoarelor etape:

(i) Calculul ponderilor de bază

Probabilităţile de includere ale UP, corespunzătoare primei trepte de eşantionare, au fost calculate conform unei scheme de selecţie stratificată. Criteriile de stratificare utilizate au fost judeţ şi mediu de rezidenţă, prin intersectarea acestora rezultând un număr de 88 straturi (în Mun. Bucureşti selecţia a fost realizată în mod separat pentru fiecare din cele 6 sectoare administrative). În fiecare din cele 88 de straturi au fost calculate probabilităţile de includere în prima treaptă, proporţional cu mărimea unei UP, mărime exprimată în număr de locuinţe permanente, conform următoarei formule de calcul:

$$P_{1hj} = m_h \times \frac{N_{hj}}{\sum_{j=1}^{N_h} N_{hj}}$$

unde:

h = indicele stratului, h=1,..,88

j = indicele UP, j=1,..,4622

P_{1hj} = probabilitatea de includere în prima treaptă a UP j aparţinând stratului h

m_h = volumul eşantionului de UP în stratul h

N_{hj} = Număr de locuinţe permanente în stratul h, UP j

N_h = Număr de locuinţe permanente în stratul h

În a doua treaptă, în interiorul fiecărei UP din totalul celor 792 UP incluse în prima treaptă în EMZOT'2002, au fost extrase câte 36 de locuinţe pe baza unui algoritm de selecţie sistematică cu start aleator. Astfel, toate locuinţele compunând o anumită UP au aceeaşi probabilitate de includere în eşantionul trimestrial al anchetei. Probabilitatea de includere din treapta a doua a fost calculată după cum urmează:

$$P_{2hjk} = \frac{36}{N_j}$$

unde:

h = indicele stratului, h=1,..,88

j = indicele UP , j=1,…,792

k = indicele unei locuinţe aparţinând UP j

P_{2hjk} = probabilitatea de includere în treapta a doua a unei locuinţe permanente k aparţinând UP j din stratul h

N_j = numărul total de locuinţe permanente în UP j

Probabilitatea generală de includere a unei locuinţe k în eşantionul anchetei (PGSk), după cele 2 trepte de eşantionare, este calculată astfel:

PGSk=P1hj*P2hjk

Ponderea de bază a unei locuinţe k, selectată în a doua treaptă de eşantionare din UP j (BWk) este, aşadar, inversa probabilităţii generale de includere a unei locuinţe k:

BWk=1/PGSk

Ponderea de bază a unei locuinţe este ‘împrumutată’ tuturor gospodăriilor din acea locuinţă.

(ii) Ajustarea non-răspunsurilor totale

Pentru a acoperi procentul gospodăriilor care refuză să participe la anchetă, se procedează la re-ponderarea unităţilor respondente, prin ajustarea cu inversul ratei de răspuns.

Experienţa anterioară ne-a dovedit că două variabile pot influenţa decizia unei gospodării de a participa sau nu la anchetă:

- Judeţ;

- Mediul de rezidenţă (urban sau rural).

Ca urmare, tratarea non-răspunsurilor totale nu se face în mod global, pe ansamblul eşantionului, ci în mod diferenţiat, pe grupe de gospodării, generate de intersecţia variabilelor considerate ca variabile explicative ale non-răspunsului: judeţ*mediu de rezidenţă.

Această modalitate de tratare a non-răspunsurilor corespunde aşa numitei metode a grupelor de răspuns omogen, care presupune că într-o anumită grupă din eşantion probabilităţile de răspuns sunt egale. În concluzie, pentru fiecare grupă de gospodării, obţinută prin intersectarea variabilelor sus-menţionate şi considerată ca grupă de răspuns omogen, se calculează rata răspunsului, după cum urmează:

$$R_g = \frac{NHH_{2g}}{NHH_{1g}}$$

unde:

g = indicele grupei de răspuns omogen, g=1,…, număr de grupe generate de intersecţia variabilelor judeţ*mediu de rezidenţă

NHH_{1g} = suma ponderilor de bază ale gospodăriilor eşantionate în grupa g, gospodării aparţinând locuinţelor eligibile pentru interviu.

NHH_{2g} = suma ponderilor de bază ale gospodăriilor respondente în grupa g.

În final, ponderea de bază a unei gospodării respondente k este ajustată cu inversul ratei de răspuns, separat pe fiecare grupă de răspuns omogen:

$$W_{adj_k} = BW_k \cdot (1/R_g)$$

Fiecare individ aparţinând unei gospodării respondente primeşte ponderea de bază ajustată a gospodăriei.

(iii) Redresarea eşantionului şi calculul ponderilor finale

Redresarea este realizată cu scopul de a îmbunătăţi calitatea estimaţiilor printr-o ajustare finală a ponderilor rezultate în urma pasului anterior.

Ponderile obţinute în final sunt modificate astfel încât totalurile estimate din eşantion să fie egale cu totalurile în populaţie pentru anumite variabile. În plus, ponderile finale sunt obţinute astfel încât să se îndepărteze cât mai puţin posibil de ponderile iniţiale, prin minimizarea unei funcţii de distanţă dintre cele două ponderi, ceea ce are efect asupra îmbunătăţirii preciziei estimaţiilor.

Această metodă de redresare este cunoscută sub numele de calibrare, în timp ce variabilele utilizate sunt denumite variabile de calaj. Calibrarea se realizează cu ajutorul macro-ului SAS CALMAR (CALaj pe MARje), creat de către INSEE Franţa.

CALMAR calculează ponderea finală utilizând o variabilă auxiliară x (pentru simplificare, presupunem aici o singură variabilă auxiliară disponibilă), ale cărei totaluri sunt cunoscute pentru întreaga populaţie, şi variabila iniţială de ponderare , astfel încât:

$$\sum_{k \in s} G(W_{final_k} / W_{adj_k}) = \min$$

sub restricţia:

$$\sum_{k \in s} W_{final_k} \cdot x_k = X$$

unde:

k = indicele unei gospodării din eşantionul disponibil s

G = este o funcţie de distanţă de argument

X = este totalul variabilei auxiliare în populaţie.

Spre deosebire de alte metode de redresare (de exemplu post-stratificare sau estimatorul prin raport), metodele de calibrare implementate în CALMAR urmăresc, pe lângă un calaj al ponderilor finale pe totalurile populaţiei, şi o minimizare a distanţei dintre ponderile iniţiale şi ponderile finale.

Variabile demografice (populaţia pe sexe şi grupe de vârstă) şi variabile de localizare (populaţia pe regiuni şi medii de rezidenţă) sunt utilizate în ajustarea finală. Structura populaţiei pe variabilele menţionate este cunoscută din surse externe (populaţia rezidentă, disponibilă de două ori pe an: 1 ianuarie şi 1 iulie al anului respectiv)

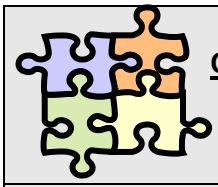
Pentru o regiune se folosesc umătoarele variabile:

- Număr de persoane pe medii de rezidenţă (urban, rural)
- Număr de persoane pe sexe (masculin, feminin)
- Număr de persoane pe categorii de vârstă

După calibrare, totalurile estimate sunt egale cu totalurile în populaţie pentru fiecare din celulele obţinute prin intersecţia variabilelor menţionate mai sus.

Uneori, destul de rar, totuşi, atunci cand anumite celule conţin prea puţine observaţii în eşantion, calajul este dificil sau chiar imposibil de realizat. În aceste cazuri, se recurge la o regrupare a celulelor iniţiale.

La sfârşitul acestei etape, ponderile finale, care sunt diferite de la o gospodărie la alta, sunt obţinute, iar toate persoanele aparţinând unei anumite gospodării primesc ponderea finală a gospodăriei.



Caseta 2: Exemplificare din ***STATISTICA ÎNTREPRINDERILOR***

În cazul anchetei structurale în întreprinderi extinderea rezultatelor obţinute din anchetă se realizează pe baza coeficienţilor atribuiţi fiecărei întreprinderi din eşantion, care a răspuns la chestionar. Comparativ cu ancheta structurală în întreprinderi, în cazul altor anchete în intreprinderi coeficientii de extindere se calculeaza din ponderea de selectie si rata de non-raspuns. Pentru determinarea acestor coeficienţi este necesară parcurgerea următoarelor etape:

- Calculul ponderii de selecţie (π_{ih}) pentru fiecare unitate. Ponderea este de tipul Horvitz-Thompson calculată ca inversul probabilităţi de selecţie.

$$\pi_{ih} = \frac{1}{p_{ih}} = \frac{N_h}{n_h}$$

unde:

p_{ih} = probabilitatea de selecţie a unităţii i din stratul h

N_h = numărul de unităţi din baza de sondaj, în stratul h

n_h = numărul de unităţi în eşantion în stratul h

- Calculul ratei de non-răspuns. Rata de non-răspuns la nivel de strat se calculează pornind de la ipoteza conform căreia unităţile care nu au răspuns sunt similare din punct de vedere statistic cu cele care au răspuns. Se obţin coeficientii finali de extindere la nivel de strat prin raportul Nh/mh unde Nh este numărul total de întreprinderi în stratul h al populaţiei şi mh este numărul de întreprinderi cu date (care au răspuns) din stratul h al eşantionului.

$$c_h = \frac{n_h}{m_h}$$

unde:

n_h = numărul de unităţi din eşantion din stratul h

m_h = numărul unităţi respondente

- Urmează etapa de post-stratificare şi estimare.
- În etapa de estimare se folosesc informaţii auxiliare din surse administrative (situaţii financiare etc). Indicatorii preluaţi din baza de calare sunt cifra de afaceri şi numărul de salariaţi la nivel de unitate. Prin însumarea indicatorilor de interes la nivel de clasa CAEN Rev2 şi clasă de mărime a întreprinderii (stabilită după numărul de salariaţi), din baza de calare, se obţin fişierele de calaj. Cu ajutorul pachetului software CLAN (SAS macro), se calculează coeficienţii de calaj (cc_{ih}). Calcularea acestor coeficienţi la nivel de unitate se face prin ajustarea coeficienţilor finali de extindere, ţinand cont de limitele (marjele) din fişierul de calaj. Aplicând coeficienţii de calaj în locul coeficienţilor de extindere finali, diferenţa dintre suma estimatelor pentru cifra de afaceri/numărul mediu de salariaţi, la nivel de clasa CAEN şi clasă de mărime şi valoarea „marginii" corespunzatoare din fişierul de calaj va tinde catre 0.

Calculul coeficienţilor finali ($COEF_{ih}^{ext}$)

$$COEF_{ih}^{ext} = \pi_{ih} \cdot c_h * cc_{ih}$$

- Calculul coeficienţilor de redresare (ch). Coeficienţii de redresare s-au calculat la nivelul fiecărei celule de selecţie a unităţilor primare ca inversul probabilităţii de răspuns. Coeficientul de redresare are rolul de compensare a unităţilor non-respondente în ipoteza în care aceste unităţi se manifestă similar cu unităţile respondente aferente stratului din care fac parte

$$c_h = \frac{1}{p_{rh}} = \frac{1}{\frac{m_h}{n_h}} = \frac{n_h}{m_h}$$

unde:

p_{rh} = probabilitatea de răspuns din celula de selecţie *h*

n_h = numărul de unităţi primare selectate în eşantion din celula de selecţie *h*

m_h = numărul de unităţi primare selectate în eşantion din celula de selecţie *h* care au răspuns la anchetă.

- Calculul coeficientului de extindere a unităţilor primare($COEF_{ih}^{ext}$) :

$$COEF_{ih}^{ext} = \pi_{ih} \cdot c_h ==$$

Calculul indicatorilor de calitate

Coeficientul de variaţie (CV) este definit ca eroarea standard ($V(\theta)$) împărţită la valoarea calculată a estimatorului ((θ). CV-ul reprezintă eroarea standard în termeni relativi (procente) și cuantifică eroarea de eșantionare.

Coeficienţii de variaţie sunt calculaţi la nivel de celulă utilizând procedura SAS - PROC SURVEYMEANS

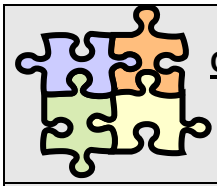
$$SE = \sqrt{\sum (\theta - \hat{\theta})^2} = \sqrt{VAR(\hat{\theta})} \quad ; \quad CV^{EST}(\theta) = \frac{\sqrt{V(\theta)}}{\theta}$$

Estimatorul trebuie să ia în considerare planul de eșantionare și trebui să integreze efectul preciziei ajustărilor cu non-răspuns-ul, corecțiile clasificărilor eronate şi informații auxiliare prin metode de calibrare etc.

V) ANALIZA DATELOR

Analiza datelor se realizează prin calculul unor indicatori medii (de exemplu, consumul mediu pe persoană dintr-un anumit produs, venitul mediu al gospodăriilor, câştigul salarial mediu pe economie etc.) totaluri de variabile repartizate în timp sau spaţiu sau prin prezentarea distribuţiei unor variabile.

Un alt aspect care trebuie detaliat şi documentat în această etapă îl reprezintă ***estimarea calităţii rezultatelor*** obţinute prin calculul intervalelor de încredere şi a coeficienţilor de variaţie pentru cei mai importanţi indicatori calculaţi pe baza datelor cercetării.



Caseta 3: Exemplificare din ***STATISTICA AGRICOLĂ***

În domeniul statisticii agricole, pentru Ancheta Structurală în Agricultură, estimarea calităţii datelor s-a realizat prin calcularea Erorii Relative Standard (ESR).

Eroarea relativă standard (RSE) s-a calculat astfel:

$$\frac{\sqrt{\hat{V}}}{\hat{Y}} \times 100 ,$$

Unde:

$\hat{Y}$ =- este estimarea variabilei şi

$\hat{V}$ = este varianţa estimatorului.

Întrucât, în România, Ancheta Structurală în Agricultură 2013 (ASA 2013) a fost o anchetă selectivă, cu eşantionare stratificată, metoda de estimare a varianţei a ţinut cont de acest lucru. Estimatorul total al lui Y a fost calculat după formula:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \times Y_{hi} = \sum_{h=1}^{H} N_h \times \overline{Y_h},$$

Unde:

H = număr de straturi

N_h = număr (extins) al exploataţiilor agricole din stratul *h*

n_h = numărul exploataţiilor agricole din stratul *h*

Y_{hi} = valoarea variabilei *Y* pentru exploataţia agricolă *i* din stratul *h*

$W_{hi} = \frac{N_h}{n_h}$ = factorul de extindere pentru exploataţia agricolă *i* din stratul *h*

$\overline{Y_h}$ = media valorilor Y_{hi}, i = 1, ..., n_h

Estimatorul varianţei lui $\hat{Y}$ s-a calculat după formula:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \frac{N_h \times (N_h - n_h)}{n_h \times (n_h - 1)} \times \left(\sum_{i=1}^{n_h} Y_{n_i}^2 - \frac{Y_h^2}{n_h} \right)$$

VI) PUBLICAREA REZULTATELOR ŞI EVALUAREA FINALĂ

Scopul final al unei cercetări statistice este *oferirea de informaţii corecte şi în timp util care să permită cunoaşterea unui fenomen, a unui efect al unei cauze sau a unei categorii de populaţie*. Evaluarea finală a unei cercetări prin sondaj trebuie să evidenţieze dacă aceasta şi-a atins obiectivele. Se poate măsura prin gradul de satisfacţie al utilizatorilor rezultatelor unei cercetări, prin numărul de publicaţii vândute, numărul de accesări on-line al bazelor de date care conţin estimaţii obţinute prin sondaj etc.

II.2. Introducerea datelor în mediul electronic

Introducerea datelor într-un mediu electronic reprezintă, în cele mai multe cazuri, conversia datelor din varianta “pe hârtie” în variantă electronică. Această activitate este însoţită de mesaje de eroare care apar în cazul introducerii unor valori eronate (fie erori de înregistrare, fie erori de introducere).

Cheile de control în ceea ce priveşte introducerea datelor pot fi stabilite în două tipuri de abordări: **top down** sau **bottom up**.

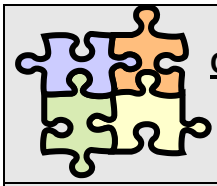
a. Modalitatea **Top down** se referă la introducerea datelor fără detectarea erorilor care pot apare în momentul introducerii. Se utilizează personal care introduce datele cu viteză mare, în mod "heads down". Datele introduse în mod „heads down” sunt adesea re-verificate prin

reintroducerea chestionarului şi compararea celor două variante (copii) introduse ale aceluiaşi chestionar.

b. Modalitatea **Bottom up** se referă la introducerea datelor concomitent cu verificarea lor, în momentul introducerii. Modalitatea „Heads up" de introducere a datelor necesită ca personalul care introduce astfel datele să aibă cunoştinţe în domeniul statistic respectiv. Introducerea datelor este mai lentă, dar revizuirea/ajustarea datelor este mai redusă ca volum deoarece inconsistenţele simple între răspunsuri sunt identificate mai devreme/în faza de început a anchetei. Acest mod este eficient, în special atunci când intervievatorul sau respondentul introduce datele în timpul interviului. Acest lucru este cunoscut sub numele de **Computer Assisted Interviewing** (Intervievare asistată de calculator), care este explicată în detaliu, mai jos.

   Datele pot fi capturate (introduse) prin mai multe metode automate fără a folosi introducerea tradiţională a datelor. Pe măsură ce tehnologia avansează, mai multe instrumente vor deveni disponibile pentru capturarea (introducerea) datelor. Un instrument popular este interviul prin telefon cu sintetizarea vocii asistată de calculator („the touch-tone telephone key-pad with synthesized voice computer-administered interview"). Cititoarele optice de caractere (OCR) pot fi folosite pentru a scana chestionarele în format electronic.

Alegerea modului de introducere a datelor precum și a metodei de ajustare a datelor au cel mai mare impact asupra tipului de personal care va fi necesar și asupra instruirii acestuia.



Caseta 4: Exemplificare din ***STATISTICA ÎNTREPRINDERILOR***

Pentru cercetarea statistică Intrastat, colectarea datelor de la operatorii economici se face numai în format electronic, în sistem online, pe site-ul dedicat Intrastat, sau în sistem offline, prin aplicaţia Intrastat pusă la dispoziţie cu titlu gratuit de INS operatorilor economici.

Există două tipuri de control al datelor colectate prin sistemul Intrastat: validarea primară şi validarea secundară.

Validarea primară reprezintă validarea datelor la nivelul operatorului economici precum şi validarea la nivelul INS din punctul de vedere al corectitudinii codurilor utilizate. Validarea secundară se referă atât la corectitudinea corelaţiilor dintre variabile cât şi a indicatorilor tip medie (ex. preţ mediu).

Validarea primară:

I. La nivelul operatorului economic, în momentul introducerii datelor in declaraţia statistică Intrastat, sunt verificate următoarele:

i. codurile declarate sa fie valide, conform nomenclatoarelor încorporate în aplicaţiile Intrastat offline şi online: codurile de produs din nomenclatorul combinat, codurile de ţări (de destinaţie / expediţie sau origine), termenii de livrare, modul de transport şi natura tranzactiei;

ii. masa netă, unităţile de măsură suplimentare, valoarea facurată şi valoarea statistică să aibă valori pozitive.

II. La nivelul INS – Direcţia Statisticii Comerţului Exterior, sunt monitorizate următoarele:

i. Firmele prioritare (primele 5000 de firme din punct de vedere a valorii cumulate realizate pe fiecare flux în ultimele 12 luni) – se verifică dacă firma a transmis declaraţia Intrastat sau este non-respondent

ii. Erorile de încărcare generate de aplicaţie. Aplicaţia Intrastat online generează pentru fiecare grup de declaraţii Intrastat offline încărcate în baza de date, câte un BATCH conţinând erorile generate automat. Evidenţa acestor BATCH-uri se ţine într-un registru special de BATCH-uri. Tratarea erorilor din aceste BATCH-uri se face progresiv, în Registru fiind înscris numele persoanei care a preluat spre rezolvare respectivele erori. Pentru rezolvarea fiecărei erori de încărcare (validare primară), persoanele responsabile contacteaza firmele prin email sau telefon pentru corectarea şi retransmiterea declaraţiilor Intrastat.

Erorile generate automat sunt următoarele:

- Cod eroare UPL-1 Fisierul incarcat este corupt sau invalid;
- Cod eroare UPL-2 Declaratia Intrastat este pentru o firma necunoscuta in Registru Intrastat;
- Cod eroare UPL-3 Codul de judet al PDT este invalid;
- Cod eroare UPL-4 Codul de oras al PDT este invalid;
- Cod eroare UPL-5 Corelatie invalida oras / judet;
- Cod eroare UPL-6 Parte declaranta referita in fisierul incarcat nu este utilizatorul curent;
- Cod eroare UPL-7 Nu poate fi incarcata o declaratie revizuita (tip REVIZUIT), daca nu a fost incarcata in prealabil o declaratie de tip NOU;
- Cod eroare UPL-8 Nu poate fi incarcata o a doua declaratie pentru aceiasi perioada de referinta decat marcata ca REVIZUITA;
- Cod eroare UPL-9 Firma referita nu a fost asociata cu PDT-ul referit in Declaratie;
- Cod eroare UPL-10 PDT referit in declaratie nu este cunoscut in Registrul Intrastat;
- Cod eroare UPL-101 Continut invalid al declaratiei…\:Cod NC invalid;
- Cod eroare UPL-102 Continut invalid al declaratiei…\:Cod UMS invalid;
- Cod eroare UPL-103 Continut invalid al declaratiei…\:Masa net nu poate fi 0;
- Cod eroare UPL-104 Continut invalid al declaratiei…\:Cod termeni de livrare invalid;
- Cod eroare UPL-105 Continut invalid al declaratiei…\:Cod natura tranzatiei A invalid;
- Cod eroare UPL-106 Continut invalid al declaratiei…\:Cod natura tranzatiei B lipsa;
- Cod eroare UPL-107 Continut invalid al declaratiei…\:Cod natura tranzatiei B invalid;
- Cod eroare UPL-108 Continut invalid al declaratiei…\:Cod mod de transport invalid;
- Cod eroare UPL-109 Continut invalid al declaratiei…\:Cod tara de origine invalid;
- Cod eroare UPL-110 Continut invalid al declaratiei…\:Cod tara de expeditie invalid;
- Cod eroare UPL-111 Continut invalid al declaratiei…\:Cod tara de destinatie invalid;
- Cod eroare UPL-112 Continut invalid al declaratiei…\:masa neta trebuie sa fie 0)

iii. Firmele care transmit declaraţia Intrastat pe flux invers faţă de obligaţie – Rapoartele din Aplicaţia Intrastat online atenţioneaza asupra firmelor ce au in sistem declaraţii Intrastat pentru alt flux decat pentru cel pentru care au obligaţii legale de declarare (declarare pe flux invers). Firmele sunt contactate şi monitorizate astfel încât să revizuiască şi să retransmită declaraţiile corecte.

iv. Firmele care au greşit perioada de raportare - Rapoartele din Aplicaţia Intrastat online atenţioneaza asupra firmelor ce au în sistem declaraţii Intrastat pentru altă perioadă de raportare decât perioada de referinţă la momentul respectiv. Firmele sunt contactate şi monitorizate astfel încât să revizuiască şi să retransmită declaraţiile corecte.

v. Firmele care au transmis declaraţii duble - Raportul generat odată cu firmele prioritare pentru perioada de referinţă respectivă atenţionează asupra firmelor ce au în sistem declaraţii Intrastat cu valori identice în luna de referinţă curentă faţă de luna de referinţă anterioară sau au valori identice pe cele două fluxuri (achiziţii=expedieri). Firmele sunt contactate şi monitorizate astfel încât să revizuiască şi să retransmită declaraţiile corecte.

vi. Firmele care au transmis declaraţia Intrastat pentru prima dată: declaraţiile acestor firme nu pot fi încărcate în Aplicaţia Intrastat online până nu sunt completate în Registrul Intrastat datele demografice ale acestora. În consecinţă, înainte de încărcarea datelor pentru firmele noi, codul de identificare fiscală al acesteia este transmis responsabilului de la Registrul Intrastat pentru adăugarea datelor demografice şi încărcarea sa în Registrul Intrastat, utilizând funcţia dedicată din Aplicaţia Intrastat online. Dupa această procedură, declaraţia Intrastat a firmei noi este încărcată în baza de date.

Validarea secundară este prezentată în capitolul II.3 *Validarea datelor*.

II.3. Validarea datelor

Validarea datelor constă în trei tipuri de activităţi: detectarea erorilor, analiza erorilor şi a datelor şi corectarea erorilor.

❖ CE SUNT ERORILE, CLASIFICAREA ŞI DESCRIEREA LOR

Eroarea reprezintă o valoare a unei variabile care nu respectă (încalcă) o regulă dinainte stabilită (condiţie logică).

În cercetările prin sondaj se disting 3 categorii principale de erori:

- erori de eşantionare (sampling errors);
- erori nelegate de eşantionare (non-sampling errors) Erorile de eşantionare nu sunt generate de erori în datele colectate şi fac obiectul altei analize (ESSNET pag 7);
- erori de estimare care apar în procesul de extrapolare a rezultatelor obţinute la nivelul eşantionului la nivelul întregii populaţii şi privesc în principal procesul de proiectare şi implementare a eşantionului;

Erorile nelegate de eşantionare se clasifică în:[2]

- erori de acoperire
- erori de măsurare
- erori de procesare (inclusiv introducere, codificare, agregare)
- non-răspunsuri

După sursa care introduce eroarea putem distinge:

[2] *Di Zio M., Fursova N., Gelsema T., Gießing S., Guarnera U., Petrauskiene J., Quensel von Kelben L., Scanu M., Bosch K.O. van der Loo M., Wlsdorfe K. - Methodology for data validation, Essnet Validat Foundation, pag.7, 2015*

- erori ale unităţii observate
- erori ale operatorului de interviu
- erori ale mijloacelor de înregistrare (erori de introducere generate de operator)
- erori generate de metodă (erori de codificare sau erori introduse de programul informatic)
- erori generate de factori externi

O altă clasificare a erorilor este cea în *atenţionări sau erori fatale.*

Eroarea fatală reprezintă constatarea unei situaţii care nu poate exista sub nici o formă în realitate (situaţia reală întâlnită în teren nu poate fi adevărată decât în forma definită de condiţia logică).

Exemple de erori fatale:

- anul de naştere al unei persoane este 1490;
- nivel de instruire superior absolvit de o persoană în vârstă de 12 ani;

Atenţionarea reprezintă constatarea unei situaţii care în general nu poate fi adevărată, dar care ÎN MOD EXCEPŢIONAL, poate exista în realitate (situaţia reală reprezintă încălcarea unei condiţii logice care în general este adevărată, dar care constituie o excepţie a regulii definită de condiţia logică).

Exemple de atenţionări:

- cazul unei persoane care are vârsta de 115 ani;
- cazul unei femei care declară că a născut 15 copii;
- un bărbat care a declarat că statutul său ocupaţional este de persoană casnică;

Altă clasificare a erorilor:

- **sistematice** – sunt erori de acelaşi tip (identice) întâlnite pentru mai mulţi respondenţi şi au de obicei drept cauză probleme în proiectarea cercetării (ambiguităţi în chestionar sau precizările metodologice) sau în implementarea acesteia (instruirea insuficientă a operatorilor de interviu). Deoarece mai mulţi respondenţi greşesc în acelaşi sens, erorile sistematice, necorectate, conduc la obţinerea de estimatori deplasaţi.
- **aleatorii** – sunt erori care apar accidental şi se datorează în principal neatenţiei respondentului, operatorului de interviu sau a persoanei care introduce datele.

❖ ÎN CE ETAPE APAR ERORILE

Dată fiind tipologia vastă a erorilor acestea pot apărea în toate etapele unei cercetări statistice. În continuare sunt descrise cele mai probabile erori care pot apărea, în funcţie de momentul apariţiei lor.

În **faza de proiectare a chestionarului**:

- prin definirea unui salt greşit între întrebările din cuprinsul acestuia, se pot induce non-răspunsuri parţiale, prin omiterea completării unor variabile din cauza saltului greşit definit;
- prin formularea unei întrebări foarte lungi sau, dimpotrivă, a unei întrebări laconice care nu defineşte corect subiectul întrebării, se pot induce erori determinate de către respondent, prin neînţelegerea corectă a întrebării. Acelaşi tip de erori poate fi generat de folosirea unui limbaj prea tehnic sau complicat, fără a defini respondentului noţiunile utilizate;
- prin proiectarea unui chestionar foarte lung, se pot induce non-răspunsuri parţiale la acele întrebări aflate la sfârşitul chestionarului, datorate plictiselii sau oboselii respondentului;

- prin solicitarea unor răspunsuri referitoare la evenimente care au avut loc într-un trecut îndepârtat se pot induce erori de amintire sau non-răspunsuri parţiale etc.

În **faza de colectare a datelor**:

În această fază erorile se pot datora atât respondenţilor, cât şi operatorilor de interviu.

Respondenţii pot induce erori prin declararea unor răspunsuri eronate la întrebările unei cercetări statistice, generate în principal de:

- incapacitatea de a înţelege intrebările;
- definirea incorectă sau ambiguă a întrebărilor;
- folosirea unui limbaj prea tehnic sau complex în cadrul chestionarului;
- imposibilitatea de a-şi aminti corect anumite evenimente din trecutul prea îndepărtat;
- solicitarea de a efectua operaţiuni foarte complexe pentru obţinerea răspunsurilor;
- solicitarea de a răspunde la întrebări cu un caracter intim;
- din neînţelegerea caracterului confidenţial acordat de sistemul statistic informaţiilor colectate de la unităţile statistice şi a principiului confom căruia datele statistice colectate sunt utilizate exclusiv în scopuri statistice etc.

Erorile generate de operatorii se interviu pot fi cauzate de:

- înregistrarea altor variante de răspuns din cauza modului de redactare al chestionarului (economie de spaţiu);
- neînţelegerea noţiunilor utilizate, a definiţiilor sau nefurnizarea unor explicaţii suplientare în cazul în care respondentul le solicită;
- operatorul de interviu nu adresează întocmai respondentului întrebările din chestionar;
- introducerea unui salt „fals” pentru a urma un drum mai scurt în interiorul chestionarului;
- codificarea unor răspunsuri din propria experienţă, fără a mai întreba respondentul etc.

În **faza de validare a datelor**

Chiar şi în faza de validare a datelor se pot introduce erori. Acestea se datorează, în principal, factorului uman.

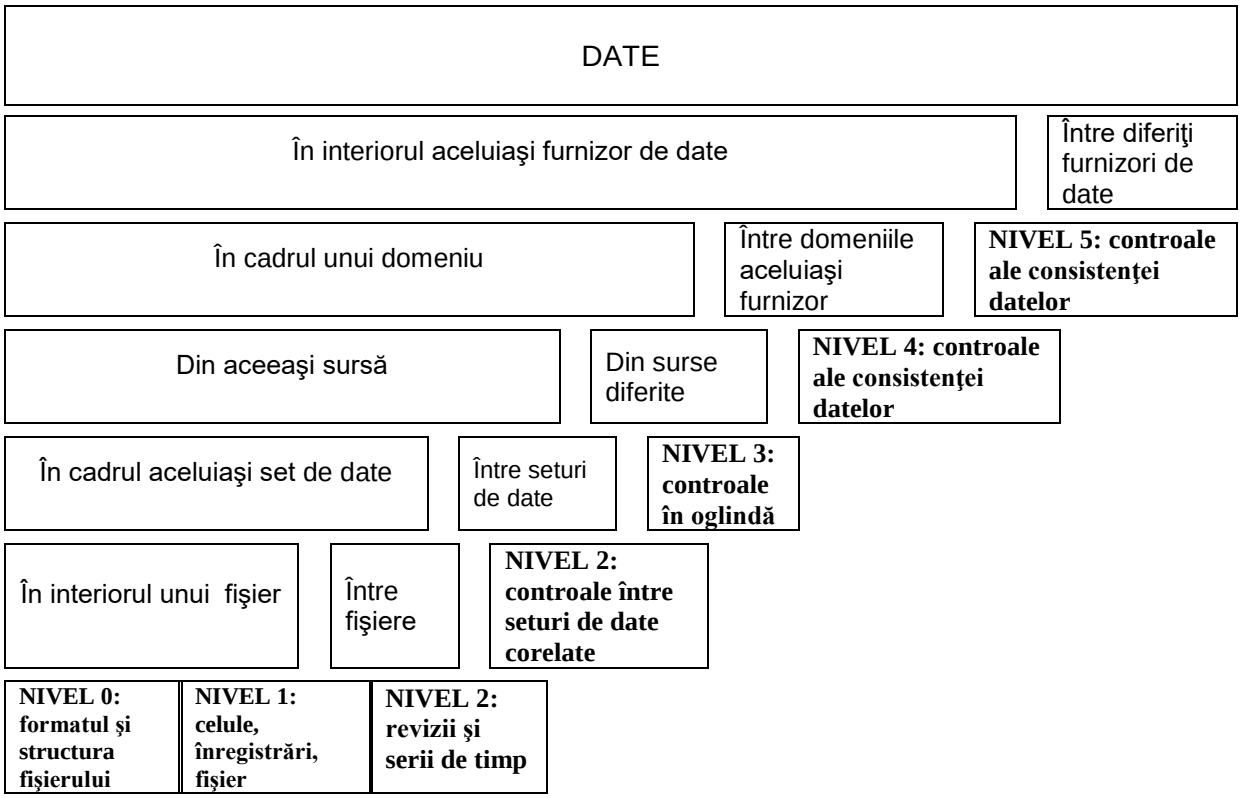
Erori generate de operatorul PC:

- erorile de tastare (introducere) sau de înţelegere greşită a răspunsurilor de pe chestionarele completate;
- erori sistematice de introducere a altor răspunsuri (decât cele completate pe chestioar) care să genereze un flux mai scurt în interiorul chestionarelor (generarea unui salt „fals”);
- alegerea primului răspuns dintr-un nomenclator pre-definit, în cazul codificării automate etc.;

Definirea greşită a unei corecţii automate poate determina introducerea unor erori noi, fie prin faptul că valoarea aplicată prin corecţia automată este una greşită, care intră în contradicţie cu valorile altor variabile, fie din cauza faptului că se pot şterge valorile înregistrate ale unei variabile (pierderea unei informaţii deja existentă în baza de date).

❖ NIVELURI ALE VALIDĂRII

Deoarece validarea datelor este un amalgam de etape şi proceduri care se realizează în diferite etape ale procesului de producţie statistică, ea poate fi ierarhizată pe mai multe niveluri. Acestea sunt descrise într-o formă sistematică în documentul ***Metodologia de validare a datelo***r conform schemei următoare[3]:



Validarea de nivel 0 – presupune verificarea corespondenţei dintre structura fişierelor şi a atributelor acestora şi cerinţele formulate. Regulile de validare pot fi formulate ca:

- numărul de câmpuri (coloane) ale unui fişier este cel specificat
- formatul de date dintr-un câmp (colană) este cel specificat (numeric, caracter, dată etc.)

Validarea de nivel 1 – presupune verificarea consistenţei informaţiilor conţinute într-un singur set de date, la un moment dat. Regulile de validare pot verifica în acest caz:

- la nivel de microdate:
    - dacă valorile intr-un câmp (coloană) sunt valide, de exemplu:
        - numărul din coloana 4 este ne-negativ;
        - anul din coloana 2 este 2011;
        - valoarea din coloana 3 este un cod de activitate valid din Nomenclatorul Activităţilor Economiei Naţionale

[3] Di Zio M., Fursova N., Gelsema T., Gießing S., Guarnera U., Petrauskiene J., Quensel von Kelben L., Scanu M., Bosch K.O. van der Loo M., Wlsdorfe K. - Methodology for data validation, Essnet Validat Foundation, pag.11, 2015

- dacă valorile dintr-un câmp se încadrează într-un interval de valori plauzibile (de exemplu numărul de membrii ai gospodăriei)
- dacă combinaţia de valori înregistrate în 2 sau mai multe câmpuri este permisă sau logică

- la nivel de macrorodate, de exemplu:
  - Total populaţie = Total bărbaţi + Total femei
  - Numărul de femei = (Total populaţie/2) $\pm$ 10%

Validarea de nivel 2 – presupune verificarea consistenţei informaţiilor dintr-un fişier cu:

- alte versiuni ale aceluiaşi fişier, referitoare la aceeaşi perioadă de timp (pentru detectarea reviziilor)
- instante ale aceluiaşi fişier referitoare la alte perioade de timp (pentru verificarea plauzabilităţii evoluţiilor în timp)
- alte fişiere corelate

Validarea de nivel 3 – presupune verificarea consistenţei între date referitoare la acelaşi domeniu dar având surse diferite. În această categorie intra de exemplu statisticile în oglindă. De exemplu, exportul raportat de ţara A în ţara B trebuie sa fie egal cu importul declarat de ţara B din ţara A.

Validarea de nivel 4 – presupune verificarea plauzabilităţii datelor referitoare la acelaşi fenomen dar provenind din domenii diferite, produse de aceeaşi instituţie, având în vedere şi diferenţele de metodologie. De exemplu plauzabilitatea evoluţiei numărului de salariaţi din Ancheta forţei de muncă şi cel provenit din Cercetarea statistica privind costul fortei de munca in unitatile economico-sociale.

Validarea de nivel 5 – presupune verificarea plauzabilităţii statisticilor referitoare la un fenomen dat între diverse instituţii producătoare de statistici.

Există validări de verificare:

1. **a integrităţii** (completitudine) – se verifică dacă s-au colectat toate datele care trebuiau colectate.
2. **a unicităţii a unei înregistrări** (să existe una şi doar o înregistrare pentru fiecare unitate de observare sau element de detaliere);

Verificarea integrităţii se face: în interiorul fiecărui fişier şi între fişiere.

Exemple:

Dacă există completat cel puţin un chestionar de gospodărie (indiferent de rezultatul interviului) pentru toate adresele selectate în eşantionul anchetei din fiecare centru de cercetare.

Se verifică dacă toţi membrii unei gospodării enumeraţi în Componenţa gospodăriei (sau doar cei eligibili conform unor anumite criterii) au chestionare individuale în fişierul individual.

3. **de respectare a „drumurilor" şi a filtrelor în cadrul unui chestionar**
4. **la nivel de variabilă**
   - de identitate (ANUL INTERVIULUI = 2016)
   - de încadrare a variabilei numerice într-un interval plauzibil (0<=VARSTA<=120)
5. **între variabilele aceluiaşi set de date** (dacă STATUT = salariat atunci VARSTA>=15)

❖ CARACTERISTICILE UNEI PROCEDURI DE VALIDARE A DATELOR

O procedură de validare a datelor presupune definirea unui set de reguli de validare. Prin aplicarea acestui set de reguli de validare pe toate înregistrările individuale ale unităţilor statistice incluse într-o cercetare statistică se verifică gradul în care răspunsurile înregistrate (valorile variabilelor colectate) respectă regulile de validare. Cazurile de încălcare ale regulilor de validare constituie erori care, în funcţie de tipul lor (erori fatale sau atenţionări), trebuie corectate sau acceptate.

Metodologia de validare a datelor[4] elaborată de Essnet Validat Foundation – Comisia Europeană defineşte următoarele caracteristici pe care trebuie să le îndeplinească o procedura de validare a datelor:

1. ***completitudinea*** – teoretic, o procedură trebuie să cuprindă toate condiţiile posibile de validare. În practică însă, gradul de completitudine a unei proceduri de validare depinde de gradul de general de cunoaştere a fenomenului, de experienţa şi gradul de inteligenţă al persoanei care defineşte regulile şi de gradul de repetativitate al cercetării respective (dacă este o cercetare realizată pentru prima dată sau dacă a mai fost realizată de mai multe ori în trecut. Din perspectiva acestei caracteristici un set de reguli de validare poate fi foarte restrictiv, datorită faptului că include un număr foarte mare de condiţii sau pentru că acestea sunt foarte stricte, sau dimpotrivă, mai „relaxat' cuprinzând un număr mic de reguli, de regulă cele mai elementare;
2. ***redundanţa*** – un set de proceduri de validare nu trebuie să cuprindă reguli redundante. Acestea pot interveni de regulă, prin includerea în diferite categorii de condiţii de control (de salt, de verificare a valorilor aberante etc.), a două sau mai multe reguli de validare care verifică, practic, acelaşi lucru;
3. ***fezabilitatea*** - este caracteristica prin care un set de reguli de validare este capabil să detecteze toate erorile posibile.
4. ***complexitatea*** este definită de varietatea şi cantitatea de informaţie care este necesară pentru definirea unei reguli de validare şi pentru evaluarea, respectiv corecţia erorilor generate de aceasta, precum şi pentru găsirea soluţiei corecte de rezolvare a respectivelor erori.

❖ MODALITĂŢI DE VALIDARE A DATELOR

Există diverse perspective prin prisma cărora se poate realiza validarea datelor, între care cele mai importante fiind[5] :

1. **automatizat sau manual**;

Validarea manuală a datelor poate avea loc înainte de introducerea datelor. Datele pot fi validate și pregătite / corectate înainte de introducere. Această procedură este utilizată în special în cazul introducerii datelor în modul „heads-down".

Validarea automată a datelor poate avea loc fie în mod batch (validare pe pachete de date) fie în mod interactiv (validare individuală a datelor). Este important faptul că datele introduse în mod heads-down pot fi validate ulterior, fie în mod batch, fie în mod interactiv.

- **Validarea datelor în mod batch (validare pe pachete de date)** are loc dupa introducerea datelor si consta in validarea mai multor chestionare intr-un singur batch (pachet). In

[4] Di Zio M., Fursova N., Gelsema T., Gießing S., Guarnera U., Petrauskiene J., Quensel von Kelben L., Scanu M., Bosch K.O. van der Loo M., Wlsdorfe K. - Methodology for data validation, Essnet Validat Foundation, pag.39-52,, 2015

[5] Di Zio M., Fursova N., Gelsema T., Gießing S., Guarnera U., Petrauskiene J., Quensel von Kelben L., Scanu M., Bosch K.O. van der Loo M., Wlsdorfe K. - Methodology for data validation, Essnet Validat Foundation, pag.9, 2015

general, rezulta un fisier cu mesaje de eroare. Acest fisier poate fi tiparit si utilizat pentru corectarea erorilor. Inregistrarile de date pot fi impartite in doua fisiere: un fisier care conține înregistrări "bune" și un fisier conținând înregistrări de date cu erori. Fisierul continand date cu erori poate fi corectat printr-un proces interactiv.

- **Validarea interactivă a datelor** implică validarea chestionarului imediat ce corecţiile/ajustările au fost făcute. Rezultatele validării sunt vizualizate pe un terminal de afişare video iar cel care editează datele poate, fie să revizuiască/ajusteze datele fie, după caz, să ignore marcajul de eroare. Acest proces continuă până când chestionarul este considerat acceptabil de către procesul de validare automată. Apoi, un alt chestionar (următorul) va intra în procesul de validare automată. O caracteristică de dorit a software-ului de editare interactivă a datelor este aceea de a supune atenţiei editorului (vizualiza) doar acele chestionare care necesită corecţii/ajustări.

Validarea realizată în timpul intervievării asistate de calculator (Computer-Assisted Interviewing - CAI) combină validarea interactivă a datelor cu editarea interactivă a datelor în timp ce respondentul este o sursă disponibilă pentru ajustarea/corectarea datelor. Un avantaj suplimentar este acela că introducerea/capturarea datelor (keyentry) are loc în momentul interviului. Această metodă poate fi utilizată atât în timpul intervievări telefonice cât și în cazul interviurilor/colectărilor realizate faţă în faţă cu respondentul, utilizând dispozitive portabile de introducere a datelor.

CAI asistă intervievatorul în formularea întrebărilor şi îndrumă spre întrebările următoare pe baza răspunsurilor anterioare. Este un instrument util pentru reducerea timpului de realizare a unui interviu, ajutând intervievatorii mai puțin experimentați. CAI a fost folosită în principal în interviurile telefonice asistate de calculator (CATI) însă, dat fiind avansul tehnologiilor moderne ce au permis miniaturizarea calculatoarelor personale, interviurile asistate de computerele personale (CAPI) vor fi utilizate din ce în ce mai mult.

2. **obiectiv versus subiectiv** (bazată pe opinia şi experienţa experţilor);
3. **validare structurală** versus **validare de conţinut**;
4. **nivelul la care se realizează validarea**: în cadrul unei înregistrări, între înregistrări, între seturi de înregistrări etc.;

Detectarea erorilor poate avea loc la mai multe nivele.

- **La nivel de linie/articol/la nivel elementar** - Validările la acest nivel sunt în general denumite "range checking" (verificare interval), din moment ce elementele sunt validate pe baza unor intervale. Exemplu: vârsta trebuie să fie > 0 și <120. Întervalul de verificare poate varia în funcție de straturi sau de alt identificator. Exemplu: În cazul în care stratul = "productie agricolă mare" atunci dimensiunea terenului trebuie să fie mai mare decât 500 hectare.
- **La nivel de chestionar** – Acest nivel implică verificarea corelaţiilor dintre variabilele din întregul chestionar. Exemplul 1: Daca căsătorit = 'da' atunci vârsta trebuie să fie mai mare de 16. Examplul 2: Suma tuturor terenuri individuale trebuie sa fie egală cu dimensiunea totală a exploataţiei agricole.
- **Ierarhică** – Acest nivel implică verificarea elementelor din chestionarele aferente unităţilor care compun o altă unitate (de exemplu, chestionarele individuale ale membrilor aceleaşi gospodării). Relaţiile dintre datele de acest tip sunt cunoscute ca „date ierarhice”. În exemplul dat, informaţiile comune la nivel de gospodărie se află pe un chestionar şi informaţiile despre fiecare individ din gospodărie se află pe chestionare diferite/separate. Se fac verificări pentru a asigura faptul că suma datelor individuale pentru un item/individ nu este mai mare decât totalul raportat pentru întreaga gospodărie.

Validare aplicată la nivelul tuturor chestionarelor anchetei (Across Questionnaire level) implică calculul intervalelor de validare pentru fiecare item/element din anchetă sau utilizarea

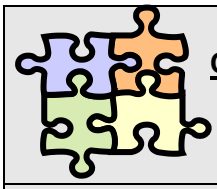
datelor istorice pentru detectarea valorilor aberante (outliers). Rutinele de analiză a datelor, care de obicei sunt rulate în momentul sintezei rezultatelor (totalizării datelor), pot fi incorporate mai uşor în validarea datelor la acest nivel. În acest fel, erorile sunt detectate suficient de devreme pentru a fi corectate în timpul procedurilor uzuale/obişnuite de corectare a datelor. Verificările la nivel de chestionare ar trebui să identifice chestionarul specific care conţine date discutabile. Modificările la nivel de chestionar sunt, în general, de două tipuri: modificări statistice (**statistical edits)** şi modificări macro (**macro edits)**.

- **Modificările statistice (Statistical Edits)** utilizează distribuţia datelor pentru detectarea posibilelor erori. Aceste proceduri utilizează datele curente din mai multe/ din toate chestionarele sau datele istorice ale unităţilor statistice pentru a genera limite acceptabile pentru datele anchetei curente/actuale. Valorile aberante (**Outliers)** pot fi identificate în funcţie de **limitele de acceptabilitate**. **Inliers** sunt date ce se află în interiorul limitelor de acceptabilitate, dar sunt considerate ca fiind suspecte din cauza lipsei variatiilor in timp.

Variabilele aleatoare presupun un grad măsurabil de variaţie. Daca valoarea este prea „consistentă", atunci ea ar putea fi pur si simplu copiată dintr-un chestionar anterior în loc să fie raportată în chestionarul curent. De aceea, testul constă în compararea modificărilor în timp a unei unităţi din eşantion (comparison to the double root residual of a sample unit over time). Dacă testul eşuează, atunci variaţia nu este suficient de aleatoare şi chestionarul ar trebui verificat. În cazul USDA-NASS acest test se aplică datelor privind greutatea animalelor sacrificate. Ipoteza este că numărul de capete de porci sacrificați nu poate varia foarte mult de la o săptămână la alta. Însă, greutatea totală a tuturor porcinelor sacrificate este o variabilă aleatoare și ar trebui să arate un grad de variatie măsurabilă în fiecare săptămână.

- **Modificarile macro (Macro Edits)** asunt validări ale datelor la nivele agregate. Inconsistenţele sunt urmarite la nivelul înregistrărilor individuale implicate. O mare parte din activitatea curentă în acest domeniu este realizată de Leopold Granquist (1991) de la Statistica Suediei. Munca sa se bazează pe convingerea că este preferabil să determini erorile cu impact mare la nivel total şi de a evita ajustările care nu au impact la nivel de total.
- Procesul de revizuire a datelor ar trebui să permită detectarea erorilor de diferite niveluri de severitate. De asemenea, ar trebui să permită decizia privind corectarea sau nu a erorii.

5. **momentul când se face validarea**: în timpul colectării datelor, în timpul introducerii în baza de date, în timpul validării, în etapa de agregare a datelor etc.;
6. **locul în procesul de producţie**: input, thoughput, output;
7. **tipul de regulă de validare**: egalitate, încadrarea într-un interval, condiţie logică etc.



Caseta 5: Exemplificare din ***STATISTICA ÎNTREPRINDERILOR***

În sistemul statistic Intrastat, după colectarea datelor şi validarea primară, este realizată validarea secundară a acestora:

I. La nivel de linie din declaraţie:

Sunt analizate tranzacţiile (liniile cu date) în următoarea ordine:

a. tranzacţiile cu valori foarte mari:

- valoare facturată sau valoare statistică > 5.000.000 LEI
- masa netă > 5.000.000 Kg
- UMS > 1.000.000

b. Codul din Nomenclatorul Combinat (NC) 27090090 – Ţiţei - majoritatea importului este din spaţiul Extra-UE; importul/exportul din/în spaţiul Intra-UE are o probabilitate foarte mică de realizare

c. Codurile NC – 27160000 - Energie electrica şi 27112100 - Gaz natural (prin comparaţie cu informaţiile obţinute din surse administrative)

d. tranzacţiile cu erori importante:

- cod eroare 1006.1 Valoare unitară aberantă (valoare / masa netă)
- cod eroare 1006.2 Valoare unitară aberantă (valoare / cantitate)
- cod eroare 1006.3 Valoare unitară aberantă (masa netă/ cantitate)
- cod eroare 1004 Raportul dintre valoare statistică şi valoare facturata <0.5 sau >1.5
- cod eroare 1009 Masa netă = 0
- cod eroare 1006.1 c>=10 ; eroare 1006.2 c>=10; eroare 1004, raport >4 sau raport<0.25, unde c reprezintă deviaţia faţă de medie
- cod eroare 1009; valori mai mari de 10.000 lei
- cod eroare 1006.1 c<10 ; eroare 1006.2 c<10; eroare 1004 ramase pentru valoare statistica sau facturata>100.000, unde c reprezintă deviaţia faţă de medie
- verificarea tranzacţiilor pentru care cantitatea exprimată în unitate de măsură suplimentară (UMS) = codul din NC pentru respectiva UMS

e. Gramele – rotunjire incorectă gram/Kg

f. Capitolul NC 87 pentru: masa netă/cantitate < 100 şi pentru codurile pentru care cantitatea expimată în UMS = masa netă exprimată în kg

g. Codurile NC 8901, 8902, 8904, 8905, 8906, 8801, 8802, 8805 (analiza strictă a tuturor tranzacţiilor pe aceste coduri)

h. Posibile erori legate de natura tranzacţiei

- natura tranzacţiei 4 la export
- natura tranzacţiei 5 la import
- natura tranzacţiei 6 sau 7

i. Alte erori

- valoarea statistică sau valoarea facturată = masa netă
- valoarea statistică sau valoarea facturată = UMS
- masa netă = UMS
- valoarea statistică = valoarea facturată =1

j. Alte atenţionări semnalate de aplicaţie

- cod atenţionare 1001.A Combinaţie invalidă: mod de transport (transport maritim) şi ţara de expediţie fără ieşire la mare
- cod atenţionare 1001.D Combinaţie invalidă: mod de transport (transport maritim) şi ţara de destinaţie fără ieşire la mare
- cod atenţionare 1002.1 Combinaţie invalidă: mod de transport (propulsie proprie)

şi codul NC al bunului

- cod atenţionare 1002.2 Combinaţie invalidă: mod de transport (postal) şi masa netă > 500 kg
- cod atenţionare 1002.3 Combinaţie invalidă: mod de transport (exclusiv transport prin instalaţii fixe) şi bun (codurile NC 27112100 si 27160000)
- cod atenţionare 1003 Combinatie invalida: tara de origine si bun
- cod atenţionare 1004 Raportul intre valoare statistica si valoare facturata <0.5 sau >1.5 – ramase
- cod atenţionare 1006.1 Valoare unitară aberantă (valoare / masa neta)
- cod atenţionare 1006.2 Valoare unitară aberantă (valoare / cantitate)
- cod atenţionare 1006.3 Valoare unitară aberantă (masa / cantitate)
- cod atenţionare 1007.A Corelaţie invalidă între valoarea facturată şi valoarea statistică în funcţie de termeni de livrare (introduceri)
- cod atenţionare 1007.D Corelaţie invalidă între valoarea facturată şi valoarea statistică în funcţie de termeni de livrare (expedieri)
- cod atenţionare 1008.A Valoareă statistică este obligatorie (introduceri)
- cod atenţionare 1008.D Valoarea statistică este obligatorie (expedieri)
- cod atenţionare 1009 Masa netă = 0

II. Pentru toate declaraţiile dintr-o lună de referinţă şi pentru un flux (export/import): controale de credibilitate (valori aberante, încadrarea datelor transmise între anumite limite a rapoartelor preţ / cantitate, preţ / kg, kg / cantitate, diverse cazuri de incompatibilitate între diverse coduri – cod ţară / produs, cod produs / mijloc de transport etc.);

II.4. Imputarea şi ajustarea datelor

II.4.1. Imputarea non-răspunsurilor

Corectarea/Ajustarea manuală a datelor are loc atunci cand selectarea unei valori mai rezonabile/bune se face de catre o persoana. Aceasta poate implica scrierea corecţiilor, pentru introducere, printr-o procedură de tip batch. Corecţiile "manuale" de date pot avea loc deasemenea în mod interactiv, la fel ca în procesul "heads-up" de introducere a datelor sau validare interactivă a datelor.

Corectarea/ajustarea automata a datelor are loc ca rezultat al acţiunii computerului. O optiune necesară în orice sistem de validare automată a datelor este posibilitatea marcării acelor niveluri pentru care aceste acţiuni să nu fie realizate (skip pentru anumite nivele). Corectarea datelor în batch produce un fişier cu înregistrari corectate (editate/imputate) şi mesaje însoţitoare pentru acțiunile întreprinse de calculator în scopul ajustărilor.

Datele pot fi imputate pe baza unei game variate de metodologii, unele dintre acestea fiind mult mai ușor de programat decât altele. Cea mai simplă presupune realizarea calculelor într-un chestionar (de exemplu, obținerea unei sume lipsă în partea de jos a unei coloane).

Imputarile automate se încadrează, în general, într-unul din următoarele 5 tipuri.

a. **Deterministic** – in cazul in care exista o singură valoare corectă, ca şi în suma lipsă în partea de jos a unei coloane de numere. O valoare este astfel determinată din alte valori aflate pe acelaşi chestionar.

b. **Model based** – utilizarea mediilor, medianelor, ecuaţiilor de regresie, etc, pentru a imputa o valoare.

c. **Deck** – Este folosit un chestionar donator pentru a imputa valoarea lipsă.

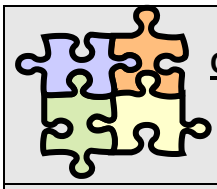
Hot deck - un chestionar donator este găsit în aceeaşi cercetare ca și chestionarul cu elementul lipsă. Tehnica de căutare **"cel mai apropiat vecin" (**"**nearest neighbour**") este adesea folosită pentru a accelera căutarea unui donator de înregistrare. În această tehnică de căutare, pachetul de chestionare donatoare provine din aceeasi cercetare și prezintă similitudini cu înregistrarea primitoare, în cazul în care similitudinea se bazează pe alte date din chestionar, care se corelează cu datele donate. De exemplu: dimensiunea și localizarea similară a unei ferme ar putea fi utilizată pentru donarea prețurilor la combustibili.

Cold deck - Similar ca la hot deck, cu excepția faptului că datele se găsesc într-o cercetare similara efectuata anterior.

d. **Mixed** – In cele mai multe sisteme există, de obicei, o combinaţie de categorii utilizate. GEIS utilizat de Statistica din Canada (Generalized Edit și System imputare), de exemplu, folosește mai întâi o abordare deterministă. În cazul în care nu este de succes, atunci este incercata o abordare hot deckp. Aceasta este urmată de o abordare model based. În cazul în care toate aceste abordări eșuează, atunci are loc o imputare manuală prin intervenţie umana. Explicații mai detaliate ale metodelor a) - d) pot fi găsite în lucrarea lui Giles și Patrick, (1986).

e. **Expert Systems** – Sistemele expert sunt doar recent aplicate pentru editarea datelor și multe cercetări sunt la inceput în acest domeniu. "Un sistem expert" este un program inteligent de calculator ce utilizează proceduri de cunoaștere și de inferență pentru a rezolva probleme destul de dificile pentru o expertiză umană. Fiecare sistem expert este alcătuit din două părți principale: Baza de cunoaștere și motorul de inferență. Baza de cunoaștere conține atât cunoștințe factuale cât și euristice. Cunoaşterea factuală este data de elemente convenite de comun acord experţii într-un anumit domeniu. Cunoașterea euristică este mai puțin riguroasă, mai experimentală și bazată pe reguli de "judecată bună" sau arta de a "ghici"/"presupune", într-un domeniu. O reprezentare utilizat pe scară largă pentru baza de cunoaștere este regula sau IF / THEN. Partea IF afișează un set de condiții într-o anumită combinație logică. O dată ce partea IF a regulii este îndeplinită, partea THEN poate fi încheiata sau problema rezolvata. Sistemele expert cu cunoștințe reprezentate în formă de regulă se numesc **"sisteme bazate pe reguli" (rule based systems)**, (Magnas, 1989). Motorul de inferență face inferente prin determinarea regulilor ce sunt satisfacute prin fapte, prin ordonarea regulilor satisfăcute si executarea regulii cu cea mai mare prioritate.

O altă metodă de imputare a datelor în cazul non-răspunsurilor totale și/sau parțiale este preluarea din surse administrative disponibile.



Caseta 6: Exemplificare din ***STATISTICA SOCIALĂ***

Un exemplu de cercetare statistică în care s-a realizat imputarea totală a unor înregistrări îl reprezintă Recensământul Populaţiei şi al Locuinţelor din 20 octombrie 2011.

În urma procesului de prelucrare a formularelor individuale, sub-înregistrarea constatată în etapa prelucrării rezultatelor provizorii ale RPL 2011 s-a confirmat. Prin urmare, s-au aplicat

metode pentru asigurarea completitudinii datelor de recensământ, folosindu-se metoda colectării indirecte din surse administrative şi metode statistice de imputare a înregistrărilor.

Numărul persoanelor nerecenzate în teren la Recensământul Populaţiei şi al Locuinţelor din 20 octombrie 2011 (RPL 2011) şi identificate în sursele administrative de date a fost 1.183 mii. Pentru aceste persoane s-au imputat total înregistrări individuale şi respectiv, de gospodărie, locuinţă şi clădire (dacă acestea nu existau deja în baza de date).

Sursele de date administrative identificate ca având informaţii utile pentru definitivarea rezultatelor RPL 2011 sunt cele cuprinse în:

- Registrul Naţional de Evidenţa Persoanei - RNEP – gestionat de Direcţia pentru Evidenţa Persoanelor şi Administrarea Bazelor de Date
- Declaraţia privind obligaţiile de plată a contribuţiilor sociale, impozitul pe venit şi evidenţa nominală a persoanelor asigurate - D112 – gestionată de Agenţia Naţională de Administrare Fiscală
- Registrul de Evidenţă a Salariaţilor - IM – gestionat de Inspecţia Muncii
- Baza de date CNPP – gestionată de Casa Naţională de Pensii Publice
- Baza de date CNAS – gestionată de Casa Naţională de Asigurări de Sănătate
- Declaraţia de înregistrare fiscală/Declaraţie de menţiuni pentru persoanele fizice care desfăşoară activităţi economice în mod independent sau exercită profesii libere - D70 – gestionată de Agenţia Naţională de Administrare Fiscală
- Registrul de evidenţă a beneficiarilor de alocaţie de stat pentru copii, alocaţie de susţinere a familiei şi de ajutorul minim garantat – gestionat de Agenţia de Plăţi şi Inspecţie Socială
- Baza de date a elevilor înscrişi în anul şcolar 2011-2012 – gestionată de Ministerul Educaţiei Naţionale.

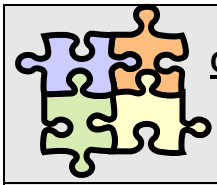
Procedura de colectare indirectă din sursele administrative a avut ca punct de plecare compararea înregistrărilor din baza de date a RPL 2011 (informaţiile înregistrate în formulare individuale completate de recenzori în perioada de colectare în teren) cu înregistrările existente în baza de date de la Evidenţa Populaţiei. Setul de înregistrări găsite la Evidenţa Populaţiei care lipseau din baza de date a RPL 2011 (din toate formularele[6] - P, PPI sau TP) au fost căutate în alte surse administrative aferente lunii octombrie 2011 şi lunilor precedente şi ulterioare din acelaşi an, în conformitate cu conceptul reşedinţei pe teritoriul României pentru o perioadă de cel puţin 12 luni, concept aplicat la RPL 2011 pentru măsurarea populaţiei rezidente (stabile). Doar pentru persoanele identificate în sursele administrative utilizate, pentru care existau dovezi clare că au fost pe teritoriul României în perioada recensământului şi în cea mai mare parte a anului 2011, s-a adăugat (s-a imputat) o înregistrare în baza de date a RPL 2011, pentru care s-au completat ulterior identificatori şi valori pentru variabilele de recensământ. În acest fel, toate înregistrările obţinute prin colectare indirectă au avut aceeaşi structură cu restul înregistrărilor obţinute prin intervievarea persoanelor în perioada de colectare a datelor în teren şi s-au referit la aceeaşi perioadă de referinţă, permiţând agregarea informaţiei pentru întreaga populaţie rezidentă (stabilă) a ţării, indiferent dacă aceasta a fost intervievată de recenzori sau nu.

Pentru adulţii identificaţi (mamă şi tată) pentru care s-au preluat informaţiile din sursele administrative de mai sus, precum şi pentru adulţii (mamă şi tată) care au fost recenzaţi la RPL 2011, au fost căutaţi şi identitificaţi minorii în Registrul Naţional de Evidenţa Persoanei.

[6] P – Persoană (prezentă sau temporar absentă); PPI – Persoană plecată pentru o perioadă îndelungată din gospodărie (în ţară sau în străinătate); TP – persoană temporar prezentă.

Ponderea persoanelor adulte pentru care s-au colectat informaţiile indirect din sursele administrative a fost de: 64,9% din declaraţia D112, 0,5% din registrul de evidenţă a salariaţilor, 2,1% din declaraţia D070, 16,1% din Registrul de evidenţă a beneficiarilor de alocaţie de stat pentru copii, alocaţie de susţinere a familiei şi de ajutor minim garantat, 4,6% din baza de date a persoanelor asigurate la sistemul public de asigurări de sănătate şi 11,8% din combinaţii de mai multe surse.

În Anexa 3 este prezentat un exemplu privind condiţiile de control pentru imputarea non-răspunsurilor variabilei VNET – 2015 din Ancheta asupra forţei de muncă în gospodării.



Caseta 7: Exemplificare din ***STATISTICA AGRICOLĂ***

Tratarea non-răspunsurilor în Ancheta Structurală în Agricultură (ASA) 2013 se realizează diferit, în funcţie de informaţiile înregistrate în cadrul capitolului privind codul de completitudine.

Se presupune că avem următoarele variante de răspuns în cadrul codului de completitudine:

a) Interviu complet
b) Exploataţie agricolă desfiinţată (sau comasată cu altă exploataţie)
c) Exploataţie agricolă temporar fără activitate
d) Interviu refuzat
e) Exploataţie agricolă neidentificată
f) Exploataţie agricolă necontactată
g) Alte situaţii

Baza de eşantionare a fost mai întâi stratificată după criteriile stabilite iniţial (de ex.: judeţ, regiune, clase de mărime, statut juridic etc.). Din această bază a fost extras eşantionul prin metoda alocării proporţionale în cadrul fiecărui strat.

În cele ce urmează este descrisă modalitatea efectivă de tratare a non-răspunsurilor.

I) Asupra exploataţiilor agricole desfiinţate sau comasate cu altă exploataţie (de ex.: având codul de completitudine = b), sau pentru exploataţiile agricole temporar fără activitate (de ex: cod completitudine = c), se acţionează în două moduri:

1. *Nu se efectuează nici o operaţiune (imputare sau ajustare).*

   Se recurge la această metodă, presupunând că valorile indicatorilor statistici aferenţi acestor unităţi vor dispărea după desfiinţarea acestora, sau se vor însuma în cadrul altor unităţi, iar orice altă acţiune de imputare sau ajustare a acestora ar duce la creşterea în mod artificial şi nejustificat a valorii totale a indicatorilor (în cazul variabilelor numerice, însumabile în special).

   2a. Se realizează procedura de imputare a acestor unităţi, prin înlocuirea valorii fiecărui indicator statistic ale acestora cu valorile indicatorilor similari, de la o unitate având aceleaşi caracteristici (de ex: din acelaşi strat) cu unitatea imputată.

   2b. Se realizează procedura de imputare a acestor unităţi, prin înlocuirea valorii fiecărui indicator statistic cu media valorilor indicatorilor similari ale tuturor unităţilor din acelaşi strat cu unitatea imputată.

II) În cazul exploataţiilor agricole cu interviu refuzat (cod completitudine = d), exploataţiilor agricole mutate la o adresă necunoscută (cod completitudine = e), pentru exploataţiile agricole cu interviu nerealizat (cod completitudine = f), precum şi pentru celelalte situaţii se aplică metoda ajustării coeficienţilor de extindere pentru fiecare strat în parte.

Rezultatul ajustării va fi următorul:

$$K_{if} = \frac{N_{i2}}{n_{i2}} \ ,$$

Unde:

N_{i2} = $N_{i1} - X_1$

n_{i2} = $n_{i1} - X_2$

K_{if} = coeficientul final de extindere al stratului „i";

N_{i2} = numărul final de unităţi din baza de eşantionare, aparţinând stratului „i", după eliminarea exploataţiilor agricole desfiinţate sau comasate cu altă exploataţie;

n_{i2} = numărul final de unităţi din eşantion, aparţinând stratului „i", după eliminarea exploataţiilor agricole desfiinţate sau comasate cu altă exploataţie

X_1 = exploataţii agricole desfiinţate (sau comasate cu altă exploataţie) şi cele temporar fără activitate

X_2 = reprezintă toate exploataţiile agricole considerate ca non răspunsuri

Coeficientul iniţal de extindere:

$$K_{i1} = \frac{N_{i1}}{n_{i1}}$$

Coeficientul final de extindere:

$K_{if} = K_{i1} * C_{ir}$

$$C_{ir} = \frac{n_{i1}}{n_{i1} - n_{ix}} \ ,$$

$$K_{if} = \frac{N_{i1}}{n_{i1}} * \frac{n_{i1}}{n_{i1} - n_{ix}}, \quad \Longrightarrow \quad K_{if} = \frac{N_{i1}}{n_{i1} - n_{ix}}$$

unde:

K_{i1} = coeficientul iniţial de extindere al stratului „i";

N_{i1} = numărul iniţial de unităţi din cadrul de eşantionare, aparţinând stratului „i";

n_{i1} = numărul iniţial de unităţi din eşantion, aparţinând stratului „i";

K_{if} = coeficientul final de extindere, aparţinând stratului „i";

C_{ir} = coeficientul final de ajustare, aparţinând stratului „i";

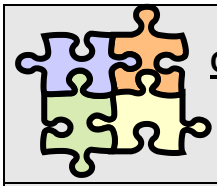
n_{ix} = numărul tuturor non-răspunsurilor (fără exploataţiile agricole desfiinţate sau comasate cu altă exploataţie şi cele temporar fără activitate), aparţinând stratului „i";

II.4.2. Ajustarea valorilor imputate

Această activitate constă în verificarea consistenţei valorilor imputate.

În cele mai multe cazuri, regulile de editare nu sunt luate în considerare de metodele de imputare. Ca o consecință, înregistrările imputate sunt, în general, incompatibile cu regulile de editare.

Această problemă este în prezent rezolvată prin introducerea unei etape de ajustare în care sunt aduse modificări la valorile imputate astfel încât înregistrările să satisfacă toate regulile de editare și ajustările sunt cât mai mici posibil. Această problemă este formulată ca o problemă de programare liniară, minimizând suma valorilor absolute ale ajustărilor sub constrângerea că imputările ajustate rezultate satisfac regulile de editare. Un algoritm pentru a rezolva această problemă este pusă în aplicare în SLICE.



Caseta 8: Exemplificare din ***STATISTICA ÎNTREPRINDERILOR***

Pentru cercetarea statistică Intrastat, se fac următoarele estimari şi imputări de date:

I. Estimarea datelor pentru firmele non-raspondente şi pentru firmele aflate sub pragurile de raportare, utilizând datele fiscale (TVA/VIES):

- utilizarea informaţiilor fiscale (TVA/VIES) pentru estimările Intrastat implică faptul ca datele fiscale şi datele Intrastat sa fie aduse la aceeaşi "sfera de cuprindere" - conţinutul ambelor seturi de date - este similar din punct de vedere metodologic, dar nu egal. Prin urmare, mai întâi se calculează totalurile pentru datele de TVA/VIES şi pentru datele Intrastat pentru acele societăţi care au declarat în sistemul statistic Intrastat, în momentul estimării. Se realizează diverse ajustări pentru a aduce cele două seturi de date la un nivel comparabil. Astfel:
  - Din datele Intrastat se elimină operaţiunile de comerţ intracomunitar cu natura tranzacţiei 3 (tranzacţii care implică transferul de proprietate dar fără beneficii), 4 (operaţiuni în scopul prelucrării pe bază de contract), 5 (operaţiuni ce succed prelucrării pe bază de contract) şi 8 (furnizarea materialelor şi echipamentelor de construcţie).
  - Din datele de TVA se elimină: comertul triunghiular, serviciile, plătile în avans, declaratile TVA si/sau VIES gresite, declaratiile TVA si/sau VIES facute pe alt cui, declaratiile TVA si/sau VIES pentru plinuri combustibil pentru camioane, declaratiile TVA si/sau VIES pentru softul la comanda, declaratiile TVA si/sau VIES pentru valoarea certificatelor de emisii poluanti, declaratiile TVA si/sau VIES pentru licentele achiziţionate pe Internet etc.
- pe baza acestor seturi de date "purificate", se va calcula un raport care va fi folosit pentru a transforma valorile de TVA/VIES ale societăţilor care nu au declarat Intrastat (firme non-raspondente sau firme aflate sub pragurile de raportare) în valorile Intrastat (valori care ar fi trebuit declarate).
- Raportul se calculează pe baza totalurilor declaraţiilor de TVA/VIES "purificate" şi declaraţiile Intrastat "purificate", numai pentru o categorie de societăţi, şi anume pentru acelea care au declarat atât în deconturile de TVA/declaraţiile recapitulative VIES-ul cât şi în declaraţiile statistice Intrastat.
- Aceast raport este necesar deoarece, chiar şi după "purificare", cele două seturi de date produc încă valori diferite pentru categoriile menţionate de societăţi.

- Totalurile de TVA/VIES pentru societăţile care nu au declarat Intrastat sunt multiplicate cu acest raport şi rezulta estimările de TVA/VIES. Nu sunt folosite valorile din TVA/VIES pentru care $\left|\frac{TVA-VIES}{VIES}\right| * 100 \geq 50\%$ (se presupune ca datele de TVA sau cele de VIES sunt eronat completate).

II. Estimare valoare statistică

În România numai operatorii economici peste un anumit prag, sunt obligaţi să furnizeze valoarea statistică a tranzacţiilor, împreună cu valoarea facturată. Din aceast motiv trebuie estimată valoarea statistică pentru firmele sub acest prag. Pentru construirea matricii de estimarea a valorii statistice vom lua în considerare următoarele variabile: condiţiile de livrare şi ţara de destinaţie / expediţie.

Pe baza datelor istorice declarate, referitoare la valoarea statistică şi valoarea facturată, se competeză această matrice cu media rapoartelor pe flux, codiţii de livrare şi ţară de destinaţie / expediţie. Aceste rapoarte vor fi utilizate pentru estimarea valorii statistice pe baza valorii facturate:

Valoare statistică = Valoarea facturata * Medie raport$_{\text{conditie de livrare/tara de destinaţie sau expediţie .}}$

❖ INDICATORI DE MĂSURARE A PERFORMANŢELOR PROCEDURII DE EDITARE

Calitatea sau eficienţa unei proceduri de editare poate fi măsurată cu ajutorul unor indicatori de performanţă care evidenţiază, în principal, cât de eronate au fost datele brute şi cât de multe operaţiuni s-au efectuat asupra acestora pentru a se obţine datele finale corecte.

În *Metodologia de validare a datelor* realizată de către Essnet Validat Foundation sunt prezentaţi următorii indicatori de performanţă ai unei proceduri de editare[7]:

1. Numărul de înregistrări pentru care se înregistrează erori;
2. Numărul minim de variabile care trebuie schimbate pentru ca înregistrările să îndeplinească un set de reguli de validare;
3. Numărul înregistrărilor care au respectat o anumită regulă de validare;
4. Numărul înregistrărilor care au încălcat o anumită regulă de validare;
5. Distribuţia înregistrărilor care au încălcat una, două sau k reguli de validare;
6. Numărul de înregistrări care îndeplinesc/ încalcă/ au non-răspuns parţial pentru fiecare variabilă;
7. Raportul dintre numărul înregistrărilor lipsă şi numărul înregistrărilor greşite;
8. Diferenţa dintre valorile indicatorilor 1-6 din etapa curentă şi cea anterioară de validare.

[7] *Di Zio M., Fursova N., Gelsema T., Gießing S., Guarnera U., Petrauskiene J., Quensel von Kelben L., Scanu M., Bosch K.O. van der Loo M., Wlsdorfe K. - Methodology for data validation, Essnet Validat Foundation, pag.53, 2015*

III. BIBLIOGRAFIE

Bethlehem, J. G. (2009), *Applied Survey Methods. Wiley Series in Survey Methodology*, John Wiley

Fellegi, I. P., Holt, D., (1976), *A Systematic Approach to Automatic Edit and Imputation*. Journal of the American Statistical Association, 71, 17—35.

Granquist, L., (1995), *Improving the Traditional Editing Process. In: Business Survey Methods* (eds. Cox, Binder, Chinnappa, Christianson and Kott), John Wiley &Sons, New York, pp. 385-401

Hartwig, P. (2009), *How to Use Edit Staff Debriefings in Questionnaire Design*. Paper presented at the

Hoogland, J., van der Loo, M., Pannekoek, J., and Scholtus, S. (2011), *Data Editing: Detection and Correction of Errors*. Methods Series Theme, Statistics Netherlands, The Hague.

Lindgren, K. (2012), *The Use of Evaluation Data Sets when Implementing Selective Editing*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.

Norberg, A. (2011), The Edit. Working Paper, *UN/ECE Work Session on Statistical Data Editing*,

Norberg, A. (2012), *Tree Analysis – A Method for Constructing Edit Groups*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.

Pannekoek, J., (2009), *Research on edit and imputation methodology: the throughput programme,* Statistics Netherlands, ISSN: 1572-0314

Scholtus, S., (2008), *Algorithms for detecting and resolving obvious inconsistencies in business survey data*. DMV-2008-04-25-SSHS, Discussion paper, CBS.

https://ec.europa.eu/eurostat/cros/sites/crosportal/files/Statistical%20Data%20Editing-05-M-Manual%20Editing%20v1.0_5.pdf

http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

http://www.cbs.nl/nr/rdonlyres/693e4b18-9322-4ac2-99fd-db61f03637b2/0/200818x10pub.pdf

https://ec.europa.eu/eurostat/cros/sites/crosportal/files/Statistical%20Data%20Editing-05-M-Manual%20Editing%20v1.0_5.pdf

https://books.google.ro/books?id=WmX-0yU7a5EC&pg=PA400&lpg=PA400&dq=netherland+Guide+for+data+editing&source=bl&ots=CA0kiX1S8H&sig=4xLNMRCJZFdi4WZAzScmMhXV8CM&hl=en&sa=X&ved=0ahUKEwjQw-Ltw8fLAhXm73IKHRx8DowQ6AEIKDAC#v=onepage&q=netherland%20Guide%20for%20data%20editing&f=false

IV. ANEXE

Anexa 1. CORELATII PENTRU CONTROLUL DE INTEGRITATE – Ancheta forţei de muncă în gospodării – 2016

Controlul între fişierele CL şi CI se face la nivel de COD CENTRU, COD LOCUINTA, NR. CL si NR. PERSOANA.
Controlul între fişierele CL, CI şi LG se face la nivel de COD CENTRU şi COD LOCUINTA.

COD REFUZ	DESCRIERE
RI1:	CENTR şi/sau LOC şi/sau CL sunt în CI şi nu sunt în CL
RI2:	SIL din CL diferita de col 1 (SIL) din LG
RI3:	CL din chestionar CL diferit de col 2 (CL) din LG
RI4:	NTG din CL (I4) diferit de col 3 (NTG) din LG
RI5:	PA CL (I1 cap.2) diferit de cel din col 4 din LG
RI6:	Numărul de persoane cu SEX = 1 si PREZ = 1...9 din CL diferit de col 5 (MASCT) din LG în cadrul aceleiasi locuinte
RI7:	Numărul de persoane cu SEX = 2 si PREZ = 1...9 din CL diferit de col 6 (FEMT) din LG în cadrul aceleiasi locuinte
RI8:	Numarul de persoane cu SEX = 1 si vârsta >= 15 ani şi PREZ = 1...9 din CL diferit de col 7 (MASC) din LG în cadrul aceleiasi locuinte
RI9:	Numarul de persoane cu SEX = 2 si vârsta >= 15 ani şi PREZ = 1...9 din CL diferit de col 8 (FEM) din LG în cadrul aceleiasi locuinte
RI10:	CENTR şi/sau LOC din CL 1 nu este în LG
RI11:	CENTR şi/sau LOC din LG nu se află în CL 1
RI12:	NRP cu CI nu se afla în componenta gospodariei (Cap.2) la locuinta si centrul respectiv
RI13:	NRP din CL (Cap.2) cu vârsta >= 15 ani şi PREZ = 1...6 nu are CI
RI14:	Persoana NRP din CI are PREZ = 7...9 în CL 2 (sit. incorectă)
RI15:	Să existe toate judeţele încărcate pentru LG, CL 1, CL 2 şi CI.
RI16:	La nivelul fiecărui judeţ numărul de centre trebuie să fie cel specificat în listă, pentru LG, CL 1, CL 2 şi CI
RI18:	Gospodăriile din CL 2 în componenţa cărora sunt persoane cu vârsta >= 15 ani şi PREZ = 1...6 trebuie să apară obligatoriu în CI.
RI19:	Toate locuinţele din CL 1 cu PA = 1 trebuie să apară obligatoriu în CL 2.

Anexa 2. CORELAŢII DE VERIFICARE A CORECTITUDINII "DRUMURILOR" LOGICE DIN CHESTIONARUL CI – Ancheta forţei de muncă în gospodării – 2016

Dacă	Atunci	si	Altfel	Cod
LUCRM (CI.1) = 1	STAP (CI.5) ≠ 0	ABST(CI.2),SITSP(CI.3),LUCREX(CI.4)=0	R	C7.1
LUCRM (CI.1) = 2	ABST (CI.2) ≠ 0		R	C7.2
ABST (CI.2) = 1,…,9	STAP (CI.5) ≠ 0	SITSP (CI.3), LUCREX (CI.4) = 0	R	C7.3
ABST (CI.2) = 10	SITSP (CI.5) ≠ 0		R	C7.4
SITSP (CI.3) = 1, 2,….,7	LUCREX (CI.4) ≠		R	C7.6
LUCREX (CI.4) = 1	STAP (CI.5) ≠ 0		R	C7.7
LUCREX (CI.4) = 2	LUCRU (CI. 67)	STAP (CI.5), AGRACT (CI.6), AGRPROP (CI.7),	R	C7.8
STAP (CI.5) = 1	STAPSPEC (CI.8)	AGRACT (CI.6), AGRPROP (CI.7), AGRVZ	R	C7.9
STAP (CI.5) = 2	NPERS (CI. 19)	AGRACT (CI.6), AGRPROP (CI.7), AGRVZ	R	C7.10
STAP (CI.5) = 3,4	AGRACT (CI.6)		R	C7.11
STAP(CI.5) = 5,6	ACT (CI. 20) ≠ 0	AGRACT (CI.6), AGRPROP (CI.7), AGRVZ	R	C7.12
AGRACT (CI.6) = 1,2…7	AGRPROP (CI.7)		R	C7.177
AGRACT (CI.6) = 8	PERS (CI.18) ≠ 0	AGRPROP (CI.7), AGRVZ (CI.8), AGRCONST	R	C7.178
AGRPROP (CI.7) = 1,2	AGRVZ (CI.8) ≠		R	C7.179
AGRPROP (CI.7) = 3	PERS (CI.18) ≠ 0	AGRVZ (CI.8), AGRCONST (CI.9), AGRCONSA	R	C7.180
AGRVZ (CI.8) = 1,2	AGRCONST		R	C7.181
AGRCONST (CI.9) = 1	PERS (CI.18) ≠ 0	AGRCONSA (CI.10), STAPSPEC (CI.11),	R	C7.182
AGRCONST (CI.9) = 2	AGRCONSA		R	C7.183
AGRCONSA (CI.10) = 1,2	PERS (CI.18) ≠ 0	STAPSPEC (CI.11), ANGSAL (CI.12),	R	C7.184
STAPSPEC (CI.11) = 1,2	ANGSAL (CI.12)		R	C7.185
ANGSAL (CI.12) = 1,2,3,4	CONTR (CI.16)	MOTEMP (CI.13), STAPLA (CI.14), ANGT	R	C7.186
ANGSAL (CI.12) = 5,6,7,8,9	MOTEMP		R	C7.187
MOTEMP (CI.13) = 1,2	STAPLA (CI.14)		R	C7.188
MOTEMP (CI.13)= 3,4,5	ANGT (CI.15) ≠	STAPLA (CI.14) = 0	R	C7.18
STAPLA (CI.14) = 1,2	ANGT (CI.15) ≠		R	C7.189
ANGT (CI. 15) = 1,….,8	CONTR (CI. 16)		R	C7.19
CONTR (CI. 16) = 1,2	PROP (CI. 17) ≠		R	C7.20
PROP (CI. 17) = 1,…., 5	PERS (CI. 18) ≠		R	C7.25
PERS (CI. 18) = 1	NPERS (CI. 19)		R	C7.21
PERS (CI. 18) = 2,….,7	ACT (CI. 20) ≠ 0	NPERS (CI. 19) = 0	R	C7.22
DURE (CI. 32) ≠ 0	OSUPLIM (CI.		R	C7.23
NPERS (CI. 19) ≠ 0	ACT (CI. 20) ≠ 0		R	C7.24
ACT (CI. 20) ≠ 0	TARA1 (CI. 21)		R	C7.26
TARA1 (CI. 21) = 1	JUD (CI. 21) ≠ 0	TARA2 (CI. 21)= 0∩REG (CI. 21) =0	R	C7.27
TARA2 (CI. 21) ≠ 0		TARA1 (CI. 21)= 0∩JUD (CI. 21) = 0		C7.29
TARA1 (CI.21) ≠ 0 ∪	OCUP (CI.22) ≠		R	C7.28
OCUP≠ 0	COORD (CI. 23)		R	C7.30
COORD (CI. 23) = 1,2,3	LUNIN, ANIN		R	C7.31
LUNIN, ANIN (CI. 24) ≠ 0	MODG (CI. 25)		R	C7.32
MODG (CI. 25) = 1,2	PROG (CI. 26) ≠		R	C7.33
PROG (CI. 26) = 1	EDURO (CI. 29)	MOPARM(CI. 27), CAUZAPP (CI. 28) = 0	R	C7.34
PROG (CI. 26) = 2	MOPARM (CI.		R	C7.35
MOPARM (CI. 27) =	EDURO (CI. 29)		R	C7.36
MOPARM (CI. 27) = 3,4,5	CAUZAPP (CI.		R	C7.37
CAUZAPP (CI. 28) ≠ 0	EDURO (CI. 29)		R	C7.38
EDURO (CI. 29) = 1	DURO (CI. 31) ≠	TOMP (CI. 30) = 0	R	C7.39
EDURO (CI. 29) = 2	TOMP(CI. 30) ≠		R	C7.40
TOMP(CI.30) ≠ 0	DURE (CI. 32) >	DURO (CI. 31) = 0	R	C7.41
DURO (CI. 31) ≠ 0	DURE (CI. 32) >		R	C7.42
OSUPLIM (CI. 33) ≠ 0 si	MOTMU (CI.		R	C7.43
OSUPLIM (CI. 33) ≠ 0 si	LUCDOM (CI.	MOTMU (CI.34), MODUREM (CI. 35),	R	C7.44
OSUPLIM (CI.33) ≠ 0 si	MODUREM (CI.	MOTMU (CI. 34) = 0	R	C7.45
OSUPLIM (CI. 33) ≠ 0 si	MOTNLUC (CI.	MOTMU (CI.34), MODUREM (CI. 35),	R	C7.46

Dacă	Atunci	si	Altfel	Cod
MOTMU (Cl. 34) ≠ 0	LUCDOM (Cl.	MODUREM (Cl. 35), COSMI (Cl. 36),	R	C7.47
MODUREM (Cl. 35) =	LUCDOM (Cl.	COSMI (Cl. 36), MOTNLUC (Cl. 37), COSMII	R	C7.48
MODUREM (Cl. 35) = 4,5,6	COSMI (Cl. 36)		R	C7.49
COSMI (Cl. 36) ≠ 0	LUCDOM (Cl.	MOTNLUC (Cl. 37), COSMII (Cl.38), LUNSF	R	C7.50
MOTNLUC (Cl. 37) = 1,2,3	LUCDOM (Cl.	COSMII (Cl. 38), LUNSF (Cl.39), ANSF (Cl.	R	C7.51
MOTNLUC (Cl. 37) =	COSMII (Cl. 38)		R	C7.52
MOTNLUC (Cl. 37) = 6	LUNSF, ANSF	COSMII (Cl. 38) = 0	R	C7.53
COSMII (Cl. 38) ≠ 0	LUNSF, ANSF		R	C7.54
LUNSF, ANSF (Cl. 39) ≠ 0	REV (Cl. 40) ≠ 0		R	C7.55
REV (Cl. 40) = 1	ABSDURT (Cl.		R	C7.56
REV (Cl. 40) = 2	CAUTNOCM	ABSDURT (Cl. 41), ABSPL (Cl. 42), ABSCONT	R	C7.57
ABSDURT (Cl. 41) = 1	LUCDOM (Cl.	ABSPL (Cl. 42), ABSCONT (Cl. 43) = 0	R	C7.58
ABSDURT (Cl. 41) = 2 si STAP	ABSPL (Cl. 42) ≠		R	C7.59
ABSDURT (Cl. 41) = 2 si STAP	ABSCONT (Cl.	ABSPL (Cl. 42) = 0	R	C7.60
ABSDURT (Cl. 41) = 2 si STAP	CAUTNOCM	ABSPL (Cl. 42), ABSCONT (Cl. 43), LUCDOM	R	C7.61
ABSPL (Cl. 42) = 1	LUCDOM (Cl.	ABSCONT (Cl. 43) = 0	R	C7.62
ABSPL (Cl. 42) = 2,3	CAUTNOCM	ABSCONT (Cl. 43), LUCDOM (Cl. 44), LUCSE	R	C7.63
ABSCONT (Cl. 43) = 1,2,3,4	LUCDOM (Cl.		R	C7.64
ABSCONT (Cl. 43) = 5	CAUTNOCM	LUCDOM (Cl. 44), LUCSE (Cl. 45), LUCNO	R	C7.65
LUCDOM (Cl. 44) ≠ 0	LUCSE (Cl. 45) ≠		R	C7.66
LUCSE (Cl.45) ≠ 0	LUCNO (Cl. 46)		R	C7.67
LUCNO (Cl. 46) ≠ 0	LUCSAM (Cl.		R	C7.68
LUCSAM (Cl.42 47) ≠ 0	LUCDUM (Cl.43		R	C7.69
LUCDUM (Cl.43 48) ≠ 0	SCH (Cl.44 49)		R	C7.70
SCH (Cl. 49) ≠ 0	TWE (Cl.50) ≠ 0		R	C7.71
TWE (Cl.50) ≠ 0	TWI (Cl.51) ≠ 0		R	C7.190
TWI (Cl.51) ≠ 0	TWLOC (Cl.52)		R	C7.191
TWLOC (Cl.52) ≠ 0	TWDUR (Cl.53)		R	C7.192
TWDUR (Cl.53) ≠ 0	TWDE (Cl.54) ≠		R	C7.193
TWDE (Cl.54) = 1,2…5	TWDI (Cl.55) ≠	TWDI (Cl.55), TWS (Cl.56)=0	R	C7.194
TWDE (Cl.54) = 6	ASEC (Cl.57) ≠		R	C7.195
TWDI (Cl.55) ≠ 0	TWS (Cl.56) ≠ 0		R	C7.196
TWS (Cl.56) ≠ 0	ASEC (Cl.57) ≠		R	C7.197
ASEC (Cl.45 57) = 1	STAPS (Cl. 58) ≠		R	C7.72
ASEC (Cl. 57) = 2	DOROREM (Cl.	STAPS (Cl. 58), PROPS (Cl. 59), ACTS (Cl.	R	C7.73
STAPS (Cl. 58) = 1	PROPS (Cl. 59)		R	C7.74
STAPS (Cl. 58) = 2,…,6	ACTS (Cl. 60) ≠	PROPS (Cl. 59) = 0	R	C7.75
PROPS (Cl. 59) ≠ 0	ACTS (Cl. 60) ≠		R	C7.76
ACTS (Cl. 60) ≠ 0	OCUPS (Cl. 61)		R	C7.77
OCUPS ≠ 0	DURES (Cl. 62)		R	C7.78
DURES (Cl. 62) ≠ 0	DOROREM (Cl.		R	C7.79
DOROREM (Cl. 63) ≠ 0	ORT (Cl.64) ≠ 0		R	C7.80
ORT (Cl. 64) ≠ 0	CAUTALT (Cl.		R	C7.81
CAUTALT (Cl. 65) = 1	MOCAUTM (Cl.		R	C7.82
CAUTALT (Cl. 65) = 2	DISP (Cl. 100) ≠	MOCAUTM (Cl.66), LUCRU (Cl. 67),	R	C7.83
MOCAUTM (Cl. 66) ≠ 0	LOCM (Cl. 79) ≠	LUCRU (Cl. 67), MONLUCRUM (Cl. 68),	R	C7.84
LUCRU (Cl. 67) = 1	MONLUCRUM		R	C7.85
LUCRU (Cl. 67) = 2	CAUTNOCM	INLUCRU (Cl. 69), ANII (Cl. 70), LUNU (Cl.	R	C7.86
MONLUCRUM (Cl. 68)≠ 0	INLUCRU (Cl.		R	C7.87
INLUCRU (Cl. 69) = 1	ANII (Cl. 70) ≠ 0		R	C7.88
INLUCRU (Cl. 69) = 2	LUNU, ANU (Cl.	ANII (Cl. 70) = 0	R	C7.89
ANII (Cl. 70) ≠ 0	CAUTNOCM	LUNU (Cl. 71), ANU (Cl. 71), STAPU (Cl. 72),	R	C7.90
LUNU, ANU (Cl. 71) ≠ 0	STAPU (Cl. 72)		R	C7.91
STAPU (Cl. 72) = 1	PROPU (Cl. 73)		R	C7.92
STAPU (Cl. 72) = 2,…,6	ACTU (Cl. 74) ≠	PROPU (Cl. 73)= 0	R	C7.93
PROPU (Cl. 73) ≠ 0	ACTU (Cl. 74) ≠		R	C7.94
ACTU (Cl. 74) ≠ 0	OCUPU (Cl. 75)		R	C7.95
OCUPU (Cl. 75) ≠ 0	CAUTNOCM		R	C7.96

Dacă	Atunci	si	Altfel	Cod
CAUTNOCM (Cl. 76) = 1	LOCM (Cl. 79) ≠	MONCAUTM (Cl. 77), CAUZANE (Cl. 78) = 0	R	C7.97
CAUTNOCM (Cl. 76) = 2	MONCAUTM		R	C7.98
MONCAUTM (Cl.77) = 1, 2,	LOCM (Cl. 79) ≠	CAUZANE (Cl. 78) = 0	R	C7.99
MONCAUTM (Cl.77) = 5, 6	MET1 (Cl.82)	CAUZANE (Cl.78), LOCM (Cl.79), PROGCM	R	C7.100
MONCAUTM (Cl.77) = 7	[LUNC(Cl. 81) ≠	CAUZANE (Cl.78), LOCM (Cl.79), PROGCM	R	C7.101
MONCAUTM (Cl. 77)= 8,9,10	CAUZANE (Cl.		R	C7.102
MONCAUTM (Cl. 77) =		CAUZANE (Cl.78), LOCM (Cl. 79), PROGCM	R	C7.103
CAUZANE (Cl.78) = 1,2,3,4	MET1 (Cl.82)	LOCM (Cl.79), PROGCM (Cl. 80), LUNC (Cl.	R	C7.105
LOCM (Cl.79) = 1	PROGCM		R	C7.106
LOCM (Cl.79) = 2	[LUNC (Cl.81) ≠	PROGCM (Cl.80) = 0	R	C7.107
PROGCM (Cl.80) ≠ 0	[LUNC (Cl.81) ≠		R	C7.108
DURAC (Cl.81) = 1	SITANT (Cl. 97)	MET1 (Cl.82), MET2 (Cl.82), MET3 (Cl.82),	R	C7.109
LUNC (Cl. 81) ≠ 0	ANC(Cl. 81) ≠ 0	DURAC (Cl.81) = 0	R	C7.110
LUNC, ANC, LUNFC, ANFC	MET1 (Cl.82)		R	C7.111
MET 18 (Cl. 96) = 1,2	SITANT (Cl. 97)		R	C7.113
SITANT (Cl.97) ≠ 0	CAUT (Cl. 98) ≠		R	C7.115
CAUT (Cl.98) ≠ 0	DOR (Cl. 99) ≠ 0		R	C7.116
DOR (Cl.99) = 1, 2	DISP (Cl. 100) ≠		R	C7.117
LUNFC (Cl.81) ≠ 0	ANFC (Cl.81) ≠	DURAC (Cl.81) = 0	R	C7.118
DISP (Cl.100) = 1	MOREF (Cl.	MONDISP (Cl. 101) = 0	R	C7.119
DISP (Cl.100) = 2	MONDISP		R	C7.120
MONDISP (Cl.101) = 1,….,7	OFMS (Cl. 103)	MOREF (Cl.102) = 0	R	C7.121
MOREF (Cl.102) = 1,…,11	OFMS (Cl.103)		R	C7.122
OFMS (Cl.103) = 1	ALOCM (Cl.		R	C7.123
OFMS (Cl.103) = 2	STAPP (Cl. 105)	ALOCM (Cl.104) = 0	R	C7.124
ALOCM (Cl. 104) =,1,2	STAPP (Cl.105)		R	C7.125
STAPP (Cl.105) ≠ 0	STAPPAP ≠ 0		R	C7.126
STAPPAP (Cl.118) = 1 si STAP	STAPAP (Cl.		R	C7.142
STAPPAP (Cl.118) = 2,…., 7 si	NETA (Cl.121) ≠	STAPAP (Cl.119), ACTAP (Cl. 120) = 0	R	C7.143
STAPAP (Cl.119) ≠ 0	ACTAP (Cl.120)		R	C7.144
STAPPAP (Cl.118) = 1 si STAP	STAPAP	NETA (Cl.121), INETA (Cl. 122),	R	C7.145
STAPPAP (Cl. 118) = 2,., 7 si	DIFICIL (Cl. 124)	STAPAP (Cl.119), ACTAP (Cl.120), NETA	R	C7.146
RASP (Cl. 125) = 1,….3	ZI, LUN, DURM		R	C7.149
OSUPLIM (Cl. 33) = 1	OSUPLIM (Cl.		R	C7.150
OSUPLIM (Cl. 33) 2 ≠ 0	OSUPLIM (Cl.		R	C7.151
OSUPLIM (Cl. 33) 3 ≠ ‘ ’	MOTMU (Cl.		R	C7.152
INETA (Cl.122) = 35, 36	DIFICIL (Cl.124)	LINTR (Cl.123) = 0	R	C7.154
LINTR (Cl.123) ≠ 0	DIFICIL (Cl.112)		R	C7.155
LINTR (Cl.123) = 2	LINTR (Cl.123)		R	C7.156
NETA (Cl.121) = 1	NETA (Cl.121)		R	C7.157
NETA (Cl.121) 1 ≠ 0	LINTR (Cl.123)	INETA (Cl. 122) = 0	R	C7.158
NETA (Cl.121) = 2	INETA (Cl.122)		R	C7.159
INETA (Cl.122) = 1,…, 34	LINTR (Cl.123)		R	C7.161
DIFICIL (Cl.124) = 1,….,5	RASP (Cl.125) ≠		R	C7.162
MET1 (Cl.82) =1 ∪ MET2	MET5 (Cl. 83) ≠		R	C7.163
MET 5 (Cl. 83) = 1,2	MET 6 (Cl.84) ≠		R	C7.164
MET 6 (Cl.84) = 1,2	MET 7 (Cl. 85) ≠		R	C7.165
MET 7 (Cl.85) = 1,2	MET 8 (Cl. 86) ≠		R	C7.166
MET 8 (Cl.86) = 1,2	MET 9 (Cl. 87) ≠		R	C7.167
MET 9 (Cl.87) = 1,2	MET 10 (Cl. 88)		R	C7.168
MET 10 (Cl.88) = 1,2	MET 11 (Cl. 89)		R	C7.169
MET 11 (Cl.89) = 1,2	MET 12 (Cl. 90)		R	C7.170
MET 12 (Cl.90) = 1,2	MET 13 (Cl. 91)		R	C7.171
MET 13 (Cl.91) = 1,2	MET 14 (Cl. 92)		R	C7.172
MET 14 (Cl.92) = 1,2	MET 15 (Cl. 93)		R	C7.173
MET 15 (Cl. 93) = 1,2	MET 16 (Cl. 94)		R	C7.174
MET 16 (Cl.94) = 1,2	MET 17 (Cl.95)		R	C7.175
MET 17 (Cl.95) = 1,2	MET 18 (Cl. 96)		R	C7.176

Anexa 3. IMPUTAREA NON-RASPUNSURILOR pentru variabila VNET - Ancheta forţei de muncă în gospodării – 2015

Scop: Toţi salariaţii trebuie să aibă o valoare corectă (≥724 RON) la VNET.
Inregistrarile din fisier pot fi clasificate în:
INDIFERENTE: CAT≠A ∪ STAP≠1 ∪ [CAT=A ∩ STAP=1 ∩ MOTNLUC=4 ∩ (VNET ≥ 724)] adica inactivi sau non-salariati sau salariati dar în concediu pentru crestere a copilului care au declarat un salariu corect
DONORI: CAT=A ∩ STAP=1 ∩ VNET ≥ 724 ∩ MOTNLUC≠4 adica salariati pentru care am o valoare corectă la VNET şi care nu sunt în concediu pentru crestere a copilului[8]
PRIMITORI: CAT=A ∩ STAP=1 ∩ [0 ≤ VNET < 724] adică salariati pentru care VNET este incorect
Reuniunea celor 3 categorii = total înregistrări
Pentru înregistrărille care sunt DONORI sau INDIFERENTI nu se face nimic. Pentru înregistrările care sunt PRIMITORI (adică ar trebui să aibă valori pentru VNET dar nu au sau valorile sunt incorecte) se caută între între înregistrările care au o valoare corectă pentru VNET (DONORI) un "geamăn" de la care să se preia datele pentru câştiguri.
"Geamănul" este căutat folosind mai întâi criteriul cel mai strict (acelaşi sex, acceasi grupa de varsta, acelaşi nivel de instruire, aceeaşi activitate, aceeaşi ocupatie, aceeasi regiune) şi care are acelaşi program de lucru (PROG) şi care nu şi-a mai donat anterior valoarea unui PRIMITOR. Dacă se gaseşte un "geamăn", PRIMITORUL preia datele acestuia, prin procedura de imputare 1 şi se trece la următoare înregistrare. Dacă nu gaseşte, se continuă cautarea folosind acealşi criteriu dar făcând abstracţie de programul de lucru (PROG). Dacă se gaseşte un "geamăn", PRIMITORUL preia datele acestuia, prin procedura de imputare 2 şi se trece la următoare înregistrare. Dacă nu gaseşte, se continuă cautarea folosind un criteriu mai relaxat (acelaşi sex, acceasi grupa de varsta, aceeaşi instruire – doar pe cele 3 categorii superior, mediu, scazut, aceeaşi activitate, aceeaşi ocupatie) şi care are acelaşi program de lucru (PROG) şi care nu şi-a mai donat anterior valoarea unui PRIMITOR. Dacă se gaseşte un "geamăn", PRIMITORUL preia datele acestuia, prin procedura de imputare 1. Se procedează în felul acesta, aplicând criterii din ce în ce mai relaxate şi cautând prima dată printre DONORII care îndeplinesc criteriul şi au acelaşi program de lucru şi abia apoi (dacă nu găseşte) făcând abstracţie de programul de lucru, până când printre DONORI se găseşte un "geamăn" care nu şi-a mai donat anterior valoarea unui PRIMITOR şi de la care se pot prelua datele pentru câştiguri
Criteriile de căutare sunt ierarhizate, de la 1 la 17, criteriul 1 fiind cel mai strict. Fiecare dintre criteriile 2, 3, ... este mai relaxat decât cel anterior, în sensul că fie se renunţă la precizie pentru una dintre variabilele după care se face căutarea (ex.la criteriul 3 "nivelul de instruire" devine "instruire" – superior/mediu/scazut, la criteriul 5 "grupa minoră de ocupaţii" devine "grupa majoră de ocupaţii" ş.a.m.d) fie se renunţă la una dintre variabilele după care se face potrivire (ex. la criteriul 3 "regiunea" nu mai apare, la criteriul 9 nu mai apare "grupa de ocupaţii" ş.a.m.d). Criteriul 17 cel mai relaxat (practic, la aplicarea criteriului 17 este imposibil să nu se găsească un DONOR care nu şi-a mai donat anterior valoarea unui PRIMITOR şi de la care să se poată prelua datele).
Criteriile folosite pentru a identifica un "geamăn":
CRITERIUL 1:
SEX
GRUPA DE VARSTA_1
NIVEL DE INSTRUIRE
ACTIVITATE
GRUPA MINORA DE OCUPATII
REGIUNE
CRITERIUL 2:

---

[8] Aceste persoane, fiind în absenta de lunga durata de la locul de muncă au declarat ultimul salariu primit (posibil cu mai mult de 1 an in urmă) → nu sunt folosite da DONORI

SEX
GRUPA DE VARSTA_1
INSTRUIRE
ACTIVITATE
GRUPA MINORA DE OCUPATII
REGIUNE
CRITERIUL 3:
SEX
GRUPA DE VARSTA_1
NIVEL DE INSTRUIRE
ACTIVITATE
GRUPA MINORA DE OCUPATII
CRITERIUL 4:
SEX
GRUPA DE VARSTA_1
INSTRUIRE
ACTIVITATE
GRUPA MINORA DE OCUPATII
CRITERIUL 5:
SEX
GRUPA DE VARSTA_1
NIVEL DE INSTRUIRE
ACTIVITATE
GRUPA MAJORA DE OCUPATII
CRITERIUL 6:
SEX
GRUPA DE VARSTA_1
INSTRUIRE
ACTIVITATE
GRUPA MAJORA DE OCUPATII
CRITERIUL 7:
SEX
GRUPA DE VARSTA_1
NIVEL DE INSTRUIRE
GRUPA DE ACTIVITATE
GRUPA MAJORA DE OCUPATII
CRITERIUL 8:
SEX
GRUPA DE VARSTA_1
INSTRUIRE
GRUPA DE ACTIVITATE
GRUPA MAJORA DE OCUPATII
CRITERIUL 9:
SEX
GRUPA DE VARSTA_1
NIVEL DE INSTRUIRE
ACTIVITATE
CRITERIUL 10:
SEX
GRUPA DE VARSTA_1

INSTRUIRE
ACTIVITATE
CRITERIUL 11:
SEX
GRUPA DE VARSTA_1
NIVEL DE INSTRUIRE
GRUPA DE ACTIVITATE
CRITERIUL 12:
SEX
GRUPA DE VARSTA_1
INSTRUIRE
GRUPA DE ACTIVITATE
CRITERIUL 13:
SEX
GRUPA DE VARSTA_1
NIVEL DE INSTRUIRE
CRITERIUL 14:
SEX
GRUPA DE VARSTA_1
INSTRUIRE
CRITERIUL 15:
SEX
GRUPA DE VARSTA_1
CRITERIUL 16
SEX
GRUPA DE VARSTA_ 2
CRITERIUL 17:
SEX
GRUPA DE VARSTA_ 3
unde:
- GRUPA DE VARSTA_1 → VARSTA
15-19 ani
20-24 ani
25-29 ani
30-34 ani
35-39 ani
40-44 ani
45-49 ani
50-54 ani
55-59 ani
60-64 ani
65 ani si peste
GRUPA DE VARSTA_2 → VARSTA:
15-24 ani
25-34 ani
35-44 ani
45-54 ani
55-64 ani
65 ani si peste
- GRUPA DE VARSTA_3 → VARSTA

15-24 ani
25-54 ani
55 ani si peste
NIVELUL DE INSTRUIRE → NIVS
INSTRUIRE → NIVS
superior → 1, 2, 3, 4
mediu → 5, 6, 7, 8, 9, 10
scazut → 11, 12, 13, 14, 15, 16
ACTIVITATE → ACT
GRUPA DE ACTIVITATE → ACT - grupele din publicaţie (atentie – vezi grupele corespunzatoare CAEN Rev 2)
GRUPA MINORA DE OCUPATII → OCUP - primele 3 caractere
GRUPA MAJORA DE OCUPATII → OCUP - primul caracter (grupele din publicaţie)
Pentru a avea o evidenţă a înregistrărilor asle căror valori au fost imputate, a celor care au fost folosite ca donori si a criteriilor de potrivire folosite, pentru fiecare înregistrare se vor adauga 3 noi variabile:
- IMPUTAT - se macheaza in cazul în care pentru înregistrarea în cauza valorile pentru castiguri au fost imputate – adica inregistrarea este PRIMITOR si s-a realizat efectiv imputarea.

- DONOR - se macheaza in cazul în care înregistrarea în cauza este DONOR şi a donat efectiv valori
-CRITERIU – numai pentru PRIMITORI pentru care s-a realizat efectiv imputarea- se atribuie valori de la 1 la 17 corezpunzator criteriului (1, 2, 3...,17) care s-a folosit pentru realizarea efectiva a imputarii.

Modul de lucru:
Pentru fiecare înregistrare i din fisier (i de la 1 la n, unde n este numărul de înregistrari din fisier)
1. dacă i este DONOR sau INDIFERENT → nimic de făcut şi treci la urmatoarea înregistrare
1. dacă este i este PRIMITOR → există o înregistrare j care este DONOR şi CRITERIUL1j=CRITERIUL1i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea ?
2. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
2. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL1j=CRITERIUL1i şi nu şi-a mai donat anterior valoarea ?
3. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
3. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL2j=CRITERIUL2i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea ?
4. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
4. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL2j=CRITERIUL2i şi nu şi-a mai donat anterior valoarea ?
5. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
5. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL3j=CRITERIUL3i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea ?
6. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
6. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL3j=CRITERIUL3i şi nu şi-a mai donat anterior valoarea ?
7. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
7. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL4j=CRITERIUL4i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
8. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
8. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL4j=CRITERIUL4i şi nu şi-a mai donat anterior valoarea?
9. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare

9. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL5j=CRITERIUL5i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
10. dacă exista → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
10. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL5j=CRITERIUL5i şi nu şi-a mai donat anterior valoarea?
11. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
11. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL6j=CRITERIUL6i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
12. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
12. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL6j=CRITERIUL6i şi nu şi-a mai donat anterior valoarea?
13. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
13. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL7j=CRITERIUL7i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
14. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
14. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL7j=CRITERIUL7i şi nu şi-a mai donat anterior valoarea?
15. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
15. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL8j=CRITERIUL8i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
16. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
16. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL8j=CRITERIUL8i şi nu şi-a mai donat anterior valoarea?
17. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
17. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL9j=CRITERIUL9i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
18. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
18. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL9j=CRITERIUL9i şi nu şi-a mai donat anterior valoarea?
19. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
19. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL10j=CRITERIUL10i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
20. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
20. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL10j=CRITERIUL10i şi nu şi-a mai donat anterior valoarea?
21. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
21 . dacă nu → există o înregistrare j care este DONOR şi CRITERIUL11j=CRITERIUL11i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
22. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
22. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL11j=CRITERIUL11i şi nu şi-a mai donat anterior valoarea?
23. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
23 . dacă nu → există o înregistrare j care este DONOR şi CRITERIUL12j=CRITERIUL12i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
24. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
24. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL12j=CRITERIUL12i şi nu şi-a mai donat anterior valoarea?
25. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare

25 . dacă nu → există o înregistrare j care este DONOR şi CRITERIUL13j=CRITERIUL13i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
26. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
26. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL13j=CRITERIUL13i şi nu şi-a mai donat anterior valoarea?
27. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
27 . dacă nu → există o înregistrare j care este DONOR şi CRITERIUL14j=CRITERIUL14i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
28. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
28. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL14j=CRITERIUL14i şi nu şi-a mai donat anterior valoarea?
29. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
29 . dacă nu → există o înregistrare j care este DONOR şi CRITERIUL15j=CRITERIUL15i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
30. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
30. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL15j=CRITERIUL15i şi nu şi-a mai donat anterior valoarea?
31. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
31 . dacă nu → există o înregistrare j care este DONOR şi CRITERIUL16j=CRITERIUL16i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
32. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
32. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL16j=CRITERIUL16i şi nu şi-a mai donat anterior valoarea?
33. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
33 . dacă nu → există o înregistrare j care este DONOR şi CRITERIUL17j=CRITERIUL17i şi PROGj=PROGi şi nu şi-a mai donat anterior valoarea?
34. dacă da → PROCEDURA DE IMPUTARE_1 şi treci la urmatoarea înregistrare
34. dacă nu → există o înregistrare j care este DONOR şi CRITERIUL17j=CRITERIUL17i şi nu şi-a mai donat anterior valoarea?
35. dacă da → PROCEDURA DE IMPUTARE_2 şi treci la urmatoarea înregistrare
35 . dacă nu → nimic de făcut şi treci la urmatoarea înregistrare

OBS. Dacă la o căutare se găsesc mai multe înregistrări care satisfac criteriile de cautare, datele se preiau de la prima înregistrare găsită.

Anexa 4. Imputarea non-răspunsurilor în EU-SILC

Problemele apar în special în determinarea venitului total al gospodăriei din cauza lipsei informațiilor asupra unor componente de venit, dar pot să apară și când nu există toate informațiile pentru toţi membrii dintr-o gospodărie.

Există două motive majore pentru imputarea datelor:

- din punct de vedere *statistic* se impune imputarea pentru minimizarea abaterii medii pătratice a estimărilor din anchetă, în particular pentru deplasarea (bias) de non-răspuns când lipsa datelor nu este întâmplătoare;
- din punct de vedere *practic* se impune imputarea atunci când nu există coerență între rezultate din analize diferite.

Lipsa datelor în EU-SILC

Erori legate de acoperire şi de selectare a eșantionului

Acestea apar atunci când unitățile din populația țintă nu sunt reprezentative în planul de eșantionare sau când probabiltățile de selectare a unităților sunt distorsionate. Câteva corecții pot fi posibile pe baza informațiilor externe planului de eșantionare. Astfel de ajustări sunt numite post-stratificări, calibrări, analiză comparativă etc.

Unitatea non-răspuns

Se referă la absența informațiilor pentru toate unitățile (gospodării și/ sau persoane) selectate din eșantion. De regulă, impactul unităților non-răspuns este redus prin atașarea ponderilor adecvate la cazurile care au răspuns.

Unitatea non-răspuns parțială

In EU-SILC există două nivele de analiză: gospodării și persoane. Analiza care implică distribuția unităților la oricare dintre cele două nivele, non-răspunsul poate f tratat prin ponderare. O caracteristică aparte a EU-SILC este aceea că un număr de variabile la nivel de gospodărie nu sunt colectate în mod direct, ci sunt construite prin agregarea informațiilor culese la nivel de persoană, de fiecare membru al gospodăriei. Termenul de unitate non-răspuns parțială este introdus pentru a descrie situația în care doar câțiva, nu toți membrii gospodăriei selectați în anchetă au fost anchetați cu succes. Există două posibilități pentru a rezolva această problemă:

- ajustarea ponderii eșantionului persoanelor anchetate din gospodărie în scopul de a compensa lipsa celorlalți membrii din gospodărie;
- construirea variabilelor pentru fiecare persoană neanchetată din gospodărie prin imputare.

Item non-răspuns

Se referă la situația în care unitățile din eșantion au fost anchetate cu succes, însă nu au fost obținute toate informațiile. În anumite situații, când incidența de non-răspuns este infimă este de preferat ca aceste cazuri să fie ignorate și să se facă analiza doar pentru cazurile care au informații complete.

Tratarea datelor lipsă

Date lipsă	Măsuri de compensare	
Erori legate de acoperire şi selectare a eşantionului	Bench-marking (analiză comparativă), post-stratificare, calibrare	
Unitatea non-răspuns	Bench-marking (analiză comparativă), post-stratificare, calibrare	Ponderare
Unitatea non-răspuns parţială	Ponderare	Imputare
Item non-răspuns	Imputare	Ponderare
Variabile ţintă (componentele venit brut)		Modelare

Construirea variabilelor țintă de venit

În EU-SILC componentele de venit trebuie să fie în forma brută. Componentele care sunt disponibile (colectate) în formă netă trebuie să fie convertite în formatul brut cerut.

In conformitate cu Regulamentul (CE) nr. 1177/2003 al Parlamentului European si al Consiliului din 16 iunie 2003 privind Statisticile comunitare referitoare la venit și la condițiile de viață (EU-SILC), Statele Membre transmit Comisiei (Eurostat), sub formă de fișiere de microdate, datele transversale și longitudinale ponderate, verificate integral, editate și imputate în funcție de venit.

Regulamentul (CE) nr. 1982/2003 al Comisiei din 21 octombrie 2003 de punere în aplicare a Regulamentului (CE) nr. 1177/2003 al Parlamentului European și al Consiliului privind statisticile comunitare referitoare la venit și la condițiile de viață (EU-SILC), în ceea ce privește regulile de eșantionare și urmărire prevede:

1. În cazul în care non-răspunsul la variabilele venitului la nivelul componentei are drept rezultat date lipsă, se aplică metodele de imputare statistică corespunzătoare.
2. În cazul în care o variabilă a venitului brut la nivelul componentei nu este culeasă direct, se aplică metode de imputare statistică și/sau modelare corespunzătoare pentru a obține variabilele țintă necesare.
3. În cazul în care apare un non-răspuns la un chestionar individual într-o gospodărie eșantion, se utilizează proceduri statistice de ponderare și/sau imputare corespunzătoare pentru estimarea venitului total al gospodăriei.
4. Factorii de ponderare se calculează după cum este necesar pentru a se lua în considerare probabilitatea de selectare a unităților, non-răspunsurile și, dacă este cazul, pentru a se ajusta eșantionul la datele externe referitoare la distribuția gospodăriilor și a persoanelor în populația țintă, de exemplu după sex, vârstă (grupe de vârstă de cinci ani), dimensiune și compoziție a gospodăriei și regiune (nivel NUTS II) sau referitoare la datele privind venitul din alte surse naționale, în cazul în care Statele Membre în cauză consideră că datele externe respective sunt suficient de fiabile.
5. Statele Membre furnizează Comisiei (Eurostat) toate informațiile necesare privind organizarea și metodologia anchetei și, în special, indică criteriile adoptate în alegerea planului de eșantionare și a mărimii eșantionului.

Anexa 5. DESCRIEREA UNEI CERCETĂRI STATISTICE SELECTIVE - Ancheta forţei de muncă în gospodării

1. Obiectivele anchetei

Obiectivul principal al Anchetei forţei de muncă în gospodării (AMIGO) îl constituie asigurarea informaţiilor necesare pentru evaluarea situaţiei existente pe piaţa forţei de muncă din România, măsurarea dimensiunilor şi evoluţiei fenomenelor de ocupare, şomaj şi inactivitate.

Începând cu anul 1996, ancheta forţei de muncă în gospodării se realizează trimestrial, ca o cercetare continuă, permiţând astfel obţinerea de date conjuncturale asupra mărimii şi structurii ofertei de forţă de muncă şi evidenţierea fenomenelor cu caracter sezonier care se manifestă pe piaţa forţei de muncă. Metodologia anchetei este armonizată cu standardele Uniunii Europene,respectiv cu ancheta europeană Labour Force Survey (LFS) iar rezultatele sunt comparabile cu cele din statele membre.

Cercetarea statistică se realizează în conformitate cu normele euroepene, respectiv cu Regulamentul (CE) nr. 577/98 al Consiliului privind organizarea unei anchete prin sondaj asupra forţei de muncă din Comunitate, cu modificările ulterioare, Regulamentul (UE) nr. 545/2014 al Parlamentului European şi al Consiliului de modificare a Regulamentului (CE) nr. 577/98 al Consiliului privind organizarea unei anchete prin sondaj asupra forţei de muncă din Comunitate, Regulamentul (CE) nr. 377/2008 al Comisiei de punere în aplicare a Regulamentului (CE) nr. 577/98., în ceea ce priveşte codificarea utilizată pentru transmiterea datelor începând cu anul 2009 şi utilizarea unui subeşantion pentru culegerea datelor referitoare la variabilele structurale şi definirea trimestrelor de referinţă, cu modificările ulterioare.

2. Sfera de cuprindere

Ancheta se desfăşoară pe întreg cuprinsul ţării.

Doar gospodăriile individuale din locuinţele permanente sunt intervievate. Unităţile de locuit în comun (cămine de bătrâni, de handicapaţi, cămine muncitoreşti, sanatorii etc.) şi persoanele care locuiesc permanent în astfel de unităţi nu sunt cuprinse în anchetă. De asemenea nu sunt cuprinse în anchetă locuinţele sezoniere.

Fac obiectul anchetei persoanele rezidente - temporar sau permanent - în România, membre ale gospodăriilor din locuinţele selectate. Se consideră membri ai gospodăriei şi persoanele plecate din localitate pentru o perioadă mai mare de 6 luni, care se află în ţară sau străinătate[9], dacă acestea păstrează legături familiale cu gospodăria din care fac parte, precum: militarii în termen, elevii şi studenţii plecaţi la studii, persoanele plecate la lucru, deţinuţii şi arestaţii, persoanele spitalizate sau aflate temporar în sanatorii pentru tratament sau recuperare.

3. Unitatea de observare

Unitatea de observare este persoana.

4. Periodicitatea şi perioada de referinţă

Ancheta se realizează cu periodicitate trimestrială, iar rezultatele se prezintă trimestrial şi anual conform regulamentelor europene.

[9)] Dacă durata absenţei din ţară este mai mică de 12 luni.

5. Metoda şi perioada de înregistrare a datelor

Datele sunt colectate prin metoda interviului faţă-în-faţă. Înregistrarea informaţiilor în chestionarele anchetei se realizează prin intervievarea persoanelor de 15 ani şi peste, la domiciliul gospodăriilor din locuinţele cercetate.

Interviurile sunt repartizate uniform de-a lungul trimestrului, ancheta realizându-se ca o cercetare continuă.

6. Planul de sondaj

Planul de sondaj folosit pentru ancheta AMIGO este un plan de sondaj în două trepte: construirea, **în prima treaptă**, a Eşantionului Multifuncţional de Zone Teritoriale (eşantionul "master" EMZOT'); în **a doua treaptă**, au fost selectate sistematic, din EMZOT, clusteri (grup de trei locuinţe), eşantionul final aferent unui trimestru constând în 28080 locuinţe[10]. Toate gospodăriile dintr-o locuinţă sunt incluse.

Eşantionul este reprezentativ la nivelul ţării şi pe regiuni. Reprezentativitatea se referă atât la structura gospodăriilor, cât şi la distribuţia populaţiei pe medii, sexe şi grupe de vârstă. Diferenţele de structură, care apar datorită situaţiei din teren la momentul realizării anchetei, sunt anulate prin aplicarea unor procedee de ajustare în funcţie de rata de non-răspuns pe medii de rezidenţă şi în funcţie de distribuţia populaţiei pe medii, sexe şi grupe de vârstă, distribuţii obţinute din surse exhaustive de cercetare demografică.

Eşantionul este construit pe baza unui procedeu de înnoire parţială a eşantionului trimestrial ("schema rotaţională 2-2-2"), care are ca principiu de bază următoarea tehnică: o locuinţă este cercetată două trimestre consecutive, este scoasă temporar din cercetare următoarele două trimestre, este reintrodusă în cercetare următoarele două trimestre, apoi este scoasă definitiv din cercetare. Aşadar, o locuinţă este gestionată 6 trimestre, fiind intervievată de 4 ori. Conform schemei aplicate, în fiecare trimestru se păstrează în eşantion 50% din locuinţele trimestrului anterior, 25% din locuinţele cercetate cu două trimestre în urmă, iar restul de 25% sunt locuinţe absolut noi. Astfel se păstrează şi o acoperire de 50% a eşantioanelor de la un trimestru al unui an la acelaşi trimestru al anului precedent.

6.1. Baza de sondaj

În lipsa unor registre adecvate (registru de locuinţe, registru al populaţiei etc), Ancheta forţei de muncă în gospodării se bazează pe utilizarea unui eşantion master, ceea ce impune aplicarea unor planuri de sondaj multistadiale. Eşantionului Multifuncţional de Zone Teritoriale (eşantionul "master" EMZOT) este constituit din 780[10] centre de cercetare (unităţi primare de eşantionare), repartizate în toate judeţele şi sectoarele Municipiului Bucureşti. Începând cu anul 2015 se utilizează eşantionul master EMZOT realizat pe baza datelor de la Recensământului Populaţiei şi Locuinţelor din anul 2011.

6.2. Unitatea de selecţie

Unitatea primară de eşantionare, corespunzătoare primei trepte de eşantionare (selecţia eşantionului master), a fost un grup de secţii de recensământ.

Unitatea secundară de eşantionare, corespunzătoare celei de a doua trepte de eşantionare (selecţia eşantionului anchetei), a fost locuinţa.

6.3. Variabilele de stratificare

---

[10] Esantionul anchetei a cuprins 18036 locuinţe (din 501 centre de cercetare) pe trimestru – până în anul 2003 şi 28080 locuinţe (din 780 centre de cercetare) pe trimestru – în perioada 2004-2014.

EMZOT este un eşantion stratificat. Criteriile de stratificare au fost judeţul şi mediul de rezidenţă, obţinându-se 88 de straturi.

6.4. Mărimea şi alocarea eşantionului

Volumul eşantionului este de 28080 locuinţe/trimestru (112320 locuinţe pe an).

Conform metodologiei adoptate pentru anchetele în gospodării realizate de către INS-România, o locuinţă poate conţine una sau mai multe (cazuri rare, totuşi) gospodării. Toate gospodăriile aparţinând locuinţelor selectate, precum şi toate persoanele de 15 ani şi peste, aparţinând gospodăriilor din locuinţele selectate, sunt anchetate.

În aceste condiţii, volumul eşantionului de gospodării, ca şi volumul eşantionului de persoane, sunt variabile aleatoare, depinzând de eşantionul de locuinţe selectat în treapta a doua.

6.5. Extragerea eşantionului

Pentru extragerea unităţilor primare, în interiorul fiecărui strat s-a utilizat metoda de extragere echilibrată, folosind macro SAS Cube.

6.6. Extinderea rezultatelor

Extinderea rezultatelor obţinute din anchetă, la nivelul întregii ţări, se realizează pe baza ponderilor atribuite persoanelor din gospodăriile care fac parte din eşantion şi au răspuns la interviu. Pentru determinarea acestor coeficienţi se parcurg următoarele etape:

- calculul ponderilor de bază: ponderea de bază atribuită unei locuinţe reprezintă inversul probabilităţii generale de includere a locuinţei în eşantionul anchetei; toate gospodăriile dintr-o locuinţă "împrumută" ponderea de bază a locuinţei respective;
- tratarea non-răspunsurilor totale: se realizează cu ajutorul metodei grupelor de răspuns omogen, utilizând ca variabile explicative judeţul şi mediul de rezidenţă; în această etapă, ponderile de bază ale gospodăriilor repondente sunt ajustate cu inversul ratei de răspuns;
- redresarea eşantionului şi calculul ponderilor finale: redresarea este realizată în scopul de a îmbunătăţi calitatea estimaţiilor printr-o ajustare finală a ponderilor în etapa precedentă: metoda de redresare folosită este cunoscută sub numele de calibrare. Calibrarea se realizează la nivel de regiune de dezvoltare utilizând ca variabile populaţia pe sexe şi grupe de vârstă, mediul de rezidenţă şi numărul total de gospodării. Utilizarea acestei metode conduce la creşterea gradului de precizie al estimaţiilor.

7. Chestionarul anchetei

Informaţiile sunt colectate pe chestionare identice pe întrega perioadă a anului. Colectarea datelor se face utilizând trei chestionare statistice şi anume: ***CL*** *- chestionarul locuinţei;* ***CI*** *- chestionarul individual.*

8. Clasificări utilizate

Ocupaţia: se defineşte şi codifică conform Clasificării Ocupaţiilor din România (COR 2008) armonizată cu Clasificarea internaţională standard a ocupaţiilor ISCO-08[11)].

Activitatea: se defineşte şi codifică conform Clasificării Activităţilor din Economia Naţională[12)] (CAEN Rev.2) armonizată cu clasificarea europeană a activităţilor economice NACE Rev.2.

[11)] *Anterior anului 2011 s-a utilizat clasificarea COR armonizată cu clasificarea internaţională standard a ocupaţiilor ISCO-COM (88).*

Profil teritorial: se codifică conform criteriilor Regulamentului 1059/2003, privind stabilirea unei clasificări comune a unităţilor teritoriale statistice şi a reglementărilor europene emise de EUROSTAT, corespunzătoare *Nomenclatorului Unităţilor Teritoriale pentru Statistică (NUTS).*

Statutul profesional: se defineşte şi codifică în conformitate cu clasificarea internaţională **ICSE-93**.

Nivelul de instruire: gruparea datelor după nivelul de instruire absolvit s-a făcut având în vedere: nivelul de învăţământ absolvit, corespondenţa între nivelurile de învăţământ (stabilite conform legislaţiei naţionale) şi nivelurile de educaţie definite conform Clasificării Internaţionale Standard a Educaţiei în vigoare la data respectivă[13)].

În publicaţii şi alte medii de diseminare, datele privind nivelul de instruire absolvit pot fi grupate şi sub forma: ***scăzut***: fără şcoală absolvită, primar, gimnazial; ***mediu***: liceal (ciclul superior sau inferior), profesional, complementar sau de ucenici, postliceal de specialitate sau tehnic de maiştri; ***superior***: universitar de scurtă durată (colegii universitare, secţii de subingineri/conductori arhitecţi ale instituţiilor de învăţământ superior) şi de lungă durată (învăţământ universitar de lungă durată, licenţă şi masterat), postuniversitar, doctorat, postdoctorat.

---

12) Anterior anului 2011, datele au fost colectate si diseminate astfel: anterior anului 2003 - conform CAEN armonizat cu NACE Rev.1, 2003-2007 conform CAEN Rev.1 - armonizat cu NACE Rev.1.1; în anul 2008 datele au fost colectate si diseminate în dublă clasificare CAEN Rev.1 şi CAEN Rev 2

13) *ISCED97 – până în anul 2013 (inclusiv) şi ISCED 2011 începând cu anul 2014*