

METODE DE ANALIZĂ A BAZELOR DE DATE ISTORICE DE TIP RECENSĂMÂNT. APLICAȚIE ASUPRA CATAGRAFIEI ORAȘULUI PLOIEȘTI DE LA 1838

IRINA GAVRILĂ
(Institutul de Istorie „Nicolae Iorga”)

Cercetarea istorică asistată de calculator, ca orice domeniu de avangardă, stârnește critici atât din partea istoricilor, cât și a unor cercetători din alte domenii ale științelor sociale. De multe ori îndreptățite, aceste critici se explică prin faptul că efortul de însușire și aplicare al unor metode sofisticate, greu accesibile umaniștilor, nu conduce întotdeauna la rezultate pe măsura așteptărilor, greșelile și piste false de explorare marcând nu de puține ori încercările de aplicare a noii metodologii.

Totuși, metoda cantitativă, dublată de performanțele calculatorului, departe de a fi abandonată, continuă să ofere celor care o stăpânesc rezultate demne de studii valoroase sau instrumente de lucru utile unei arii largi de investigație istorică.

Pentru lucrările de istorie, tabelele, graficele și chiar formalizările statistice mai complicate sunt acceptate, deși nu întotdeauna privite cu bunăvoință. Utilizarea metodelor cantitative și a datelor cuantificabile este, cu siguranță, un pas înainte pentru cunoașterea istorică. Contribuții majore pentru clarificarea unor probleme de istorie aduce însăși critica noii metodologii, iar folosirea datelor cantitative și a instrumentelor de analiză a lor conduce la producerea unei literaturi istorice de o altă factură, nu întotdeauna apreciată, dar cu siguranță inovativă.

În prezentul articol ne propunem să exemplificăm crearea, modificarea și analiza statistică a unei baze de date întemeiată pe o sursă istorică și anume *Catagrafia orașului Ploiești de la 1838*¹.

Analiza prin interogare a surselor de tip recensământ

Puternic structurate, recensămintele sunt izvoare importante nu numai din punct de vedere istoric, ele fiind adecvate și unei cercetări asistate de calculator.

¹ *Obșteasca catagrafie a Țării Românești. Catagrafia orașului Ploiești de la 1838*, Societatea Culturală Ploiești-Mileniul III, Editura Ploiești-Mileniul III, 2006.

Încă din 1940, înregistrările populației au servit, în țările dezvoltate, drept date de intrare pentru studii care foloseau, la acea vreme, calculatoarele UNIVAC.

Dintre toate catagrafiile perioadei regulamentare, *Catagrafia generală din 1838*, ale cărei acte pregătitoare și rezultate se păstrează la Arhivele Naționale Istorice Centrale din București, este cea mai cuprinzătoare, ceea ce o face extrem de interesantă pentru diverse prelucrări statistice. Volumul de date este impresionant, dosarele cuprinzând înregistrarea nominală a locuitorilor din cele 15 județe ale Țării Românești, din arhive lipsind dosarele pentru județele Mehedinți, Dolj, Vâlcea și Ialomița. După cum se știe, transcrierea acestei importante surse din grafia chirilică în cea latină, ca și costurile ridicate ale tipăririi fac dificil orice demers de publicare a sa².

Scopul acestei catagrafii nu a fost numai unul fiscal, ci s-a dorit ca rezultatele prelucrării sale statistice să servească și în vederea creionării unei strategii de dezvoltare viitoare a țării. Ținta autorităților, care dovedeau o concepție foarte apropiată de cea modernă asupra statisticii, a fost aceea de a cunoaște numărul și structura populației pe unități administrative, județe, plăși, orașe și sate, stadiul dezvoltării agriculturii, meșteșugurilor, construcțiilor, comerțului, situația căilor de comunicație și a mijloacelor de transport sau a resurselor naturale.

Cei care au studiat primii *Catagrafia de la 1838*, cercetătorii Ion Donat și Gheorghe Retegan, au demonstrat în studiile lor că această sursă are toate caracteristicile fundamentale ale recensămintelor moderne și anume: recensarea este inițiată de către autoritatea de stat, pe un teritoriu bine definit, ea fiind exhaustivă, în sensul că pe liste sunt înscrși nu numai plătitorii de impozite, ci și toate categoriile exceptate de la această obligație. Listele trebuie să fie nominale și datele se culeg prin metoda directă, adică prin interogarea fiecărui cap de familie. Datele sunt înregistrate pentru fiecare persoană în parte, condiție introdusă în Franța în 1836, iar înregistrarea este simultană, adică trebuie să se stabilească o zi a recensământului. Concluzia autorilor menționați este aceea că înregistrarea din 1838 are caracteristicile esențiale pentru a fi considerată ca prim recensământ modern al populației³.

Pentru exemplul nostru am creat o mini bază de date cuprinzând 100 de înregistrări extrase din *Catagrafia orașului Ploiești de la 1838*. Ploieștii erau orașul de reședință al județului Prahova, cu 16 mahalale, inegale ca suprafață și număr de case, purtând denumiri de biserici și fiind grupate pe culori sau vopsele⁴. Profesorul Ion Dedu, cel care a transcris informațiile din grafia chirilică în cea latină, menționează în nota sa de început caracteristicile volumului 71, referitor la orașul Ploiești, din *Obșteasca catagrafie a Țării Românești*⁵. Cele 212 file, cuprinzând

² Pentru informații suplimentare asupra catagrafiilor perioadei regulamentare și a rezultatelor parțiale obținute până acum din prelucrarea lor vezi studiul introductiv al lucrării citate anterior.

³ Vezi *Obșteasca catagrafie a orașului Ploiești* și urm., p. XIV.

⁴ *Ibidem*, p. XVI-XVII.

⁵ *Ibidem*, p. VI.

fișele de recensare, însumează 424 de pagini. Pentru fiecare gospodărie, fișele originale conțin 24 de rubrici, dar în procesul de transcriere s-au păstrat doar 8 rubrici, ceea ce nu reduce însă valoarea statistică a sursei originare. Unele rubrici au fost condensate în rubrica *Observații* a transcrierii, în special cele referitoare la animalele deținute în gospodărie, la numărul pogoanelor cultivate și al unor soiuri de pomi fructiferi.

O caracteristică a perioadei este aceea ca indivizii să fie identificați prin prenume și prin numele tatălui, numele de familie după sistemul folosit astăzi nefiind formate, mai ales în Țara Românească (pentru Moldova procesul de formare a numelor de familie era mai avansat). Identificări prin nume și prenume nu găsim nici în cazul boierilor sau al mazelilor, ci doar în cazul unor funcționari, slujbași sau meșteri străini. Pe cale administrativă, numele de familie au fost fixate mai ales după 1860, prin intermediul oficiilor de stare civilă⁶.

O altă rubrică importantă este apartenența etnică, mai ales că orașul Ploiești era caracterizat prin pluritate etnică, diferitele meșteșuguri fiind practicate deopotrivă de români și de străini, ceea ce demonstrează că etniile conviețuiau efectiv. Desigur, prelucrarea statistică poate scoate la iveală o grupare a etniilor pe mahalale, precum și ierarhia lor numerică.

Catagrafia consemnează și vârsta, stabilită probabil de comun acord cu persoana interogată, dat fiind faptul că majoritatea recenzaților erau analfabeți, acte de stare civilă nu existau, iar consultarea registrelor parohiale era o operație migăloasă, greu de presupus a fi fost efectuată.

Rubrica *Birul* consemnează, de fapt, capitația, dare introdusă încă din 1830, care înlocuise toate celelalte obligații fiscale.

În cazul rubricii *Claca*, pentru orașul Ploiești sensul termenului este diferit de cel obișnuit. Această rubrică menționează condiția de proprietar sau chiriaș a celui înregistrat în recensământ.

Ocupația locuitorilor este înscrisă în rubrica *Meșteșugul*, ea incluzând și ocupațiile de funcționar, comerciant sau diferitele ocupații intelectuale.

Catagrafia orașului Ploiești de la 1838 cuprinde, în studiul introductiv al valorosului cercetător Sorin Ifțimi, comentarii extrem de interesante asupra tuturor rubricilor transcrierii în grafia latină și a posibilităților de analiză statistică a acestei importante surse istorice, analiză care poate răspunde unui număr de peste 100 de interogări, oferind astfel o frescă socio-economică, demografică și culturală a orașului de reședință a județului Prahova.

Crearea unei baze de date al cărei input să fie *Catagrafia generală de la 1838* (cel mai complet izvor statistic al perioadei regulamentare), prin posibilitățile de prelucrare statistică automată, ar constitui un inestimabil instrument de lucru pentru istorici. Desigur, realizarea acestui proiect presupune o intensă muncă de echipă, cu caracter interdisciplinar, pe o perioadă îndelungată de timp și care implică un

⁶ *Ibidem*, p. XVII.

considerabil efort financiar. Avantajele creării unei asemenea baze de date sunt însă incontestabile.

Considerații asupra datelor de intrare

Rezultatele prelucrării recensămintelor sunt în general disponibile în volume tipărite, volume în care informația este agregată pentru a da o imagine de ansamblu a societății. Un exemplu de informație agregată este cel din Tabelul 1.

Tabelul 1

Structura etnică a negustorimii din Ploiești

Meșteșug	Români	Ungureni	Bulgari	Greci	Evrei	Albanezi	Ruși	Austrieci	Țigani
Băcani	12	-	12	1		2	-	-	-
Birtași	-	1	-	-		-	-	-	-
Bogasieri	7	1	14	2	5	2	-	-	-
„Brașoveni”	1	-	3	-		-	-	-	-
Bragagii	-	-	3	-		-	-	-	-
Cafegii	1	-	5	1		-	-	1	-
Cramagii	1	-	-	-		-	-	-	-
Cârciumari	51	12	13	7	1	1	-	-	1
Geambași	1	-	1	-		-	-	-	-
Hangii	3	2	-	-		-	1	-	-
Lemnari	-	3	-	-		-	-	2	-
Marchitani	1	-	-	-		-	4	-	-
Mămulari	1	1	4	-	13	-	-	-	-
„Neguțători”	11	2	4	3	1	-	-	-	-
Precupeți	14	1	1	-		-	-	-	-
Rachieri	7	-	3	-		-	-	-	-
Spițeri	-	-	1	-	-	-	-	2	-
Sureccii	6	1	4	-		1	-	-	-
Telali	-	-	-	1		-	-	-	-
Tutungii	5	-	3	2	2	-	-	-	-
Zarafi	-	-	-	-	4	-	-	-	-

Sursa: *Catagrafia orașului Ploiești de la 1838*, p. XXV.

Datele nominale, adică informațiile referitoare la fiecare individ (de exemplu faptul că Dinu Neajlu era dogar) sunt însă confidențiale, ele nefiind puse la dispoziția cercetătorilor, în cazul Marii Britanii de pildă, decât după 100 de ani de la efectuarea respectivului recensământ.

Deși datele agregate sunt de mare însemnătate, constituind cea mai importantă sursă de studiu, din diferite motive cercetătorul dorește să cunoască și datele asupra fiecărui individ. Istoricii din ziua de azi au capacitatea de a chestiona un recensământ într-un mod mult mai complex și mai precis decât au făcut-o recenzorii epocilor trecute. De multe ori se dorește corelarea informațiilor din recensământ cu cele din afara sa, ceea ce nu se poate realiza doar cu datele agregate. Pentru a obține datele individuale, cercetătorul este nevoit să se întoarcă

la sursa originală, adică cea produsă de recenzor. Datele înscrise pe formulare în timpul recensământului devin astfel foarte importante. De obicei, aceste formulare erau colectate de recenzori și transcrise de aceștia, alcătuind adevărate volume care se păstrează în arhivele lumii. Formularele de recensământ denotă clar că structurarea informației nu este apanajul secolului al XX-lea sau al apariției calculatoarelor. O bună parte a informației pe care istoricul o găsește în arhive este deja structurată, desigur maniera în care a fost realizat acest proces variind de la o perioadă la alta. Am putea defini structurarea ca fiind obținerea de informații asupra unui număr de obiecte, în cazul recensămintelor indivizi, despre care se cunosc întotdeauna anumite lucruri (nume, prenume, vârstă, ocupație etc.).

Informațiile referitoare la fiecare individ recenzat reprezintă un rând, iar coloanele reprezintă categoriile uniforme care pot fi aplicate fiecărui individ. Această formă de prezentare, care apare și în sursa la care ne referim, a mijlocului de secol XIX, este utilă pentru a înțelege principiile fundamentale ale unei baze de date structurate. Rândurile din formularele de recenzare pot fi transformate în *înregistrări*, înregistrarea fiind nucleul unei baze de date. În limbajul bazelor de date, de multe ori înregistrarea mai poartă și numele de *rând*, *observație*, *caz*.

Coloanele includ numele și prenumele capului gospodăriei, adresa, vârsta, sexul, ocupația, relația individului cu capul gospodăriei, starea civilă, locul nașterii etc⁷. Coloanele reprezintă aceleași categorii de informații pentru fiecare înregistrare. Ele constituie *câmpurile* bazei de date, cunoscute uneori și sub numele de *variabile* sau *attribute*, caracterizate prin valori. Nu este neapărat necesar ca valorile să fie numerice, ele pot fi și calitative. De exemplu, dacă *starea civilă* este unul din câmpuri, valorile sale sunt *căsătorit*, *necăsătorit*, *văduv*, *divorțat*.

Structura informației dintr-un recensământ nu se rezumă doar la existența rândurilor și a coloanelor. Ea depinde și de măsura în care recenzorul a respectat instrucțiunile de culegere a informației, de exemplu cele de caracterizare a diferitelor ocupații. Pentru crearea unei baze de date istorice este extrem de importantă existența unei standardizări consistente.

Pasul următor este transformarea sursei, în cazul nostru a fragmentului extras din *Catagrafia orașului Ploiești*, într-o bază de date. Procesul de introducere a datelor variază în funcție de software-ul ales, dar principiile de bază sunt aceleași. În cazul nostru vom recurge la crearea unei baze de date de tip flat-file, iar fișierului care conține datele i se va atribui un nume din cel mult 8 litere și vom conveni ca cel creat de noi să se numească CATAGRAF (Anexa I). El poate fi salvat pe o dischetă sau pe un stick de memorie, în orice fază a creării sale, întocmai ca și un fișier word. Odată ce structura bazei de date a fost creată și a fost introdus un număr de înregistrări, procesul poate fi oricând întrerupt și apoi reluat, continuându-se, după caz, introducerea datelor sau analiza lor statistică.

⁷ Lucrarea lui Edward Higgs, *Making Sense of the Census*, London, 1989, ne oferă informații complete asupra acestui subiect.

Asignării unui nume îi urmează o sarcină mai complicată, aceea a selectării informației din sursa originală și a modului în care ea va fi înregistrată, fiind utilă chiar crearea manuală a unei versiuni pilot, cu 10-20 de înregistrări, în temeiul căreia să hotărâm care va fi structura definitivă a bazei de date. Este adevărat că această structură poate fi schimbată și după începerea introducerii datelor, dar este cu mult mai bine ca ea să fie stabilită de la bun început. În acest fel se realizează de fapt un tipar (template) potrivit căruia datele vor fi introduse, înregistrare cu înregistrare.

Categoriile de informație constituie câmpurile sau variabilele și fiecareia i se atribuie un nume, de obicei scurt, deoarece un nume lung este incomod de utilizat atunci când se fac interogări. Pentru câmpuri cum ar fi *Relația cu capul familiei* este bine să se folosească o prescurtare, de exemplu RELCF. În general, numele câmpurilor sunt compuse din 4-8 caractere. Mai trebuie să stabilim și tipul câmpurilor, de exemplu șir de caractere, adică alfanumeric, numeric, dată etc. Un câmp va fi numeric dacă asupra valorilor înscrise în el se vor face operații matematice. Dacă asignăm unul din câmpuri pentru anul nașterii, considerat ca valoare numerică, în acest câmp trebuie să indicăm un singur an, de exemplu „1834”, formulări de genul „1834 sau 1835” nefiind admise pentru că apare cuvântul „sau”. Cele mai multe sisteme de baze de date au un tip special de informație pentru dată, de genul *ziua/luna/anul* sau *luna/ziua/anul*.

Oricare ar fi tipul de câmp, trebuie să precizăm și numărul maxim de caractere pe care acesta îl poate conține. Această decizie este importantă din două puncte de vedere. În primul rând, spațiul de stocare a informației este limitat, iar multe dintre sistemele de gestionare a bazelor de date numără și blankurile ca fiind caractere. Dacă se alocă unui câmp 60 de caractere și valorile introduse în el nu depășesc niciodată 15 caractere, înseamnă că 45 de caractere pentru fiecare înregistrare sunt irosite. Un caracter ia un byte din spațiul fișierului. Dacă avem 1000 de înregistrări, înseamnă că am irosit 45 000 de byți.

Una dintre principalele probleme care se ridică atunci când creăm o bază de date istorice este aceea dacă datele de intrare trebuie să fie exact așa cum sunt în sursa originală, în cazul nostru în documentele de arhivă. Alterarea datelor înseamnă, într-o mai mare sau mai mică măsură, pierderea unor informații din sursa inițială. Aici intervine codificarea, care are, la nivelul stadiului de introducere a datelor, un număr de avantaje și de dezavantaje.

La o primă examinare a sursei, în cazul nostru *Catagrafia orașului Ploiești de la 1838*, chiar cel care a efectuat transcrierea din grafia chirilică în cea latină, Ion Dedu, și-a dat seama că utilizarea tuturor celor 24 de câmpuri ar îngreuna mult parcurgerea și prelucrarea materialului. El a restrâns informația la 8 câmpuri. Prin urmare, redarea mecanică nu este recomandată, chiar și în cazul unor surse bine structurate și scrise într-o formă ușor de descifrat. Întrebări se ridică și atunci când intervin inconsistențe în scrierea numelor, când nu știm dacă este mai corect să înscriem numele prescurtat, așa cum apare în sursă, sau pe cel întreg. Într-o

gospodărie pot să apară adulți fără valori de intrare în câmpul *Stare civilă*, de pildă. Au fost acești indivizi cu adevărat necăsătoriți sau este o greșeală de înregistrare a recenzorului? Presupunând că putem clarifica anomaliile din sursa originală, este sau nu recomandabil să păstrăm această formă? Dacă dorim să executăm o sortare, spre exemplu, și am alocat un singur câmp pentru nume și prenume, este normal ca ordinea să fie *nume de familie, prenume* (de exemplu *Neajlu, Dinu*), pentru ca indivizii să poată fi sortați alfabetic. Cel mai bine este ca numelui de familie și prenumelui să le fie alocate câmpuri separate pentru a putea găsi rapid persoanele cu același nume de familie, de pildă. În același mod, câmpul *Locul nașterii* poate fi disociat în două câmpuri, *Orașul* și *Județul*. Unele intrări pot fi abbreviate pentru a micșora lungimea câmpului ce le este alocat. Fiecărei intrări i se atribuie un număr unic, care identifică în mod unic înregistrarea, pentru că este foarte posibil ca două persoane diferite să aibă același nume de familie și același prenume. Acest număr unic poate fi considerat drept cheie primară.

Procesul de modificare a valorilor din cadrul câmpurilor este cunoscut sub numele de *codificare*. Ca trăsătură pozitivă, putem spune că acest proces face ca datele de intrare să fie mai compacte, fără a se pierde din informație. Oricum, trebuie să existe o codificare standard pentru toate tipurile de relații cu capul gospodăriei, toate variantele stării civile, cât și pentru sex. Dificultățile cele mai mari apar mai ales în cazul codificării ocupațiilor.

Prin urmare, un prim pas în proiectarea bazei de date este acela de a stabili numărul câmpurilor, numele câmpurilor, tipul și lungimea fiecărui câmp, principiile codificării. Bineînțeles, documentația trebuie să fie însoțită de descrierea precisă a abrevierilor și a codurilor utilizate.

O structură posibilă a bazei de date CATAGRAF este cea din Tabelul 2.

Tabel 2

Structura posibilă a tablei CATAGRAF

Nume câmp	Tip date	Exemplu
ID	Numeric (3)	10
Nume familie	Șir de caractere (15)	Neajlu
Prenume	Șir de caractere (15)	Dinu
Adresă	Șir de caractere (25)	Str. Parfumului, 25
Gospodăria	Număr (2 caractere)	1
RELCF	Șir de caractere (15)	Soție
STC	Șir de caractere (2)	C
Sex	Șir de caractere (1)	M
Vârsta	Numeric (3)	40
Ocupația	Șir de caractere (30)	Croitor
Observații	Șir de caractere (25)	Are 3 cai

Lista abrevierilor și a codurilor: ID = identificator; RELCF = relația cu capul familiei; STC = starea civilă; C = căsătorit; N = necăsătorit; V = văduv (ă); D = divorțat; M = masculin; F = feminin.

Odată ce structura bazei de date a fost stabilită, poate începe operația de introducere a datelor, o muncă care, deși cere un grad înalt de calificare, nu-i oferă istoricului mari satisfacții. Soluția ideală este existența unui colectiv de cercetare care să-și împartă sarcina ingrată a introducerii datelor, ca în cazul proiectelor de istorie locală. O altă modalitate, dacă datele nu sunt în manuscris ci tipărite, este introducerea lor electronică prin scanare optică. În fine, există și varianta înregistrării doar a unei părți din sursă, cu alte cuvinte a unui eșantion reprezentativ, care să poată conduce la concluzii pertinente pentru întreaga populație.

Interogări ale bazei de date de tip recensământ

În acest paragraf vom examina modul în care poate fi interogată, din punct de vedere istoric, baza de date, de tip flat-file, CATAGRAF. Desigur, informațiile obținute în urma recensământului din 1838 au fost agregate și la vremea respectivă, evident fără a face uz de calculator. Și astăzi, în cazul unui volum mic de informații, datele pot fi înregistrate pe fișe, care la rândul lor pot fi sortate după diferite criterii: în ordinea alfabetică a numelor de familie, după locul nașterii, sex, grupe de vârstă etc. Chiar dacă numărul de înregistrări este mic, munca va fi anevoioasă, devenind profund supusă erorilor și chiar imposibilă dacă volumul de date este mare, caz în care se înscriu catagrafiile și recensămintele. Rezultate corecte vor fi obținute instantaneu dacă informația a fost stocată într-o bază de date.

O primă operație este extragerea unei înregistrări din baza de date. Pentru a fixa terminologia reamintim că înregistrarea este unitatea de bază, ea are câmpuri sau variabile, iar variabilele iau valori. CATAGRAF cuprinde 100 de înregistrări, fiecare înregistrare având 12 câmpuri. Un exemplu de înregistrare este cel din Tabelul 3.

Tabel 3

O înregistrare din baza de date CATAGRAF

ID	Nr. gosp.	Prenume	Fiul lui	RELCG	Sex	Etnie	Vârstă	Birul	Claca	STC	Ocupația
1		Stoica	Stoian	Cap gos.	M	Român	30	Birnic	Chiriaș	C	Florar

După cum am mai spus, un sistem de gestionare a bazelor de date este un software care va furniza utilizatorului modalitatea de a extrage înregistrări din baza de date, acest proces fiind cunoscut și sub numele de *filtrare*. Denumirea reflectă de fapt ceea ce face software-ul, o filtrare doar a ceea ce este necesar din ceea ce nu este de folos. Interogarea bazelor de date este, de fapt, găsirea tuturor înregistrărilor care au anumite caracteristici specificate de către utilizator.

Să presupunem că vrem să interogăm baza de date CATAGRAF cu privire la numărul indivizilor sub 30 de ani. Aceasta înseamnă că toate înregistrările din baza de date vor fi examinate pentru a fi găsite cele care au în câmpul *Vârstă* o valoare mai mică decât numărul 30. Cu alte cuvinte, calculatorul scanează fiecare înregistrare, caută în câmpul *Vârstă* și pune de-o parte înregistrarea dacă valoarea acestui câmp este un număr cuprins între 0 și 29. Dacă valoarea respectivă este 30 sau mai mare de 30, calculatorul ignoră respectiva înregistrare. Este apoi examinată următoarea înregistrare și procesul se repetă până la epuizarea înregistrărilor.

Există o diferență netă, deși uneori greu de sesizat, între extragerea înregistrărilor și vizualizarea unor câmpuri ale înregistrărilor extrase. Calculatorul are capacitatea de a expune toate informațiile, cu alte cuvinte toate câmpurile unei înregistrări. Poate fi vizualizat numai conținutul unor anumite câmpuri și acest proces este independent de procesul anterior, al filtrării. De exemplu, criteriul de extragere poate fi „bărbați sub 30 de ani”, în timp ce utilizatorul dorește numai vizualizarea câmpurilor *Prenume*, *Vârstă*. Criteriul de selectare al înregistrărilor nu are nimic în comun cu câmpurile vizualizate. Sub același criteriu de filtrare ne poate interesa afișarea altor câmpuri, de pildă *Prenume*, *Fiul lui*, *Ocupația*. În fine, calculatorul poate fi instruit să nu expună niciunul dintre câmpuri, ci doar să furnizeze numărul bărbaților care au vârsta sub 30 de ani.

Cu cât software-ul de gestionare a bazelor de date este mai sofisticat, cu atât el are mijloace mai variate de prezentare a rezultatelor sub forma unor rapoarte. Unele programe de gestionare a bazelor de date acționează la comenzi, în sensul că utilizatorul tastează o serie de cuvinte sub forma unei comenzi. Acest gen de software este foarte flexibil, dar presupune din partea utilizatorului cunoașterea comenzilor. Alte pachete de programe sunt acționate alegând comenzi dintr-o listă, un meniu, furnizat de software. În ambele cazuri, calculatorul primește, de fapt, o comandă clară, care în cazul nostru este: „Alege toate înregistrările din baza de date CATAGRAF pentru care valoarea din câmpul *Sex* este M, iar valoarea din câmpul *Vârstă* este mai mică decât 30. Pentru înregistrările extrase vizualizează valorile din câmpurile *Prenume* și *Vârstă*.”

Într-un sistem acționat prin comenzi va trebui să tastăm comanda în mod precis și într-o formă care să fie înțeleasă de calculator. Vom utiliza comenzi ale limbajului SQL⁸, forma comenzii fiind:

```
SELECT Prenume, Sex, Vârstă INTO Result
FROM CATAGRAF
WHERE Sex = „M” AND Vârstă < 30
```

Rezultatul interogării este prezentat în Tabelul 4.

⁸Vezi Irina Gavrilă, *Baze de date istorice. Cercetarea istorică asistată de calculator. Rezultate statistice ale comparării automate a bazelor de date istorice*, București, Editura Oscar Print, 2007, p. 41.

Tabel 4

Bărbați sub 30 de ani din baza de date CATAGRAF

Result		
Prenume	Sex	Vârstă
Gheorghe	M	14
Panaete	M	20
Costandin	M	16
Ianache	M	20
Nicolae	M	15
Costandin	M	25
Preda	M	3
Gheorghe	M	2
Lică	M	25
Ioniță	M	5
Ene	M	5
Ștefan	M	3
Nicolae	M	0
Baros	M	9
Stoica	M	10
Marin	M	6
Ioniță	M	3
Stan	M	16
Voicu	M	12
Tănase	M	8
Dumitru	M	6
Minea	M	3
Șărban	M	13
Ivanciu	M	25
Nicolae	M	5
Dumitru	M	15
Ștefan	M	20
Dumitru	M	8
Iancu	M	25
Tincu	M	28
Neagu	M	8
Andrei	M	4
Dumitru	M	18
Ion	M	13
Dumitru	M	2

Putem extrage toate înregistrările pentru care valoarea din câmpul *Vârstă* este mai mică decât 30, indiferent de sexul indivizilor. Comanda SQL va fi:

```
SELECT Prenume, Vârstă INTO Result1
FROM CATAGRAF
WHERE Vârstă < 30
```

Rezultatul filtrării este prezentat în Tabelul 5.

Tabel 5

Persoane cu vârsta sub 30 de ani din baza de date CATAGRAF

Result1	
Prenume	Vârstă
Uța	19
Safta	20
Maria	7
Gheorghe	14
Anica	10
Ioana	7
Catrina	3
Zița	2
Luxandra	18
Ilinca	3
Dobrița	1
Panaete	20
Costandin	16
Drăgaica	13
Voica	1
Ianache	20
Maria	18
Ioana	2
Nicolae	15
Costandin	25
Stanca	20
Preda	3
Gheorghe	2
Lică	25
Sevasta	20
Ioniță	5
Ene	5

Result1	
Prenume	Vârstă
Ștefan	3
Maria	0
Nicolae	0
Stana	18
Safta	0
Rada	0
Baros	9
Stoica	10
Marin	6
Stana	3
Maria	25
Ioniță	3
Stan	16
Voicu	12
Tănase	8
Dumitru	6
Minea	3
Șarban	13
Maria	7
Zmaranda	4
Ivanciu	25
Maria	20
Nicolae	5
Pătrana	2
Dumitru	15
Ștefan	20
Neacșa	18
Maria	13
Rada	25
Costandina	6
Dumitru	8
Haila	12
Iancu	25
Maria	20
Tincu	28
Catrina	20
Neagu	8
Andrei	4

Result1	
Prenume	Vârstă
Maria	1
Dumitru	18
Luxandra	15
Dica	25
Ion	13
Dumitru	2
Vasilca	4

Procesul de filtrare se poate concentra nu numai asupra valorilor numerice, ci și asupra șirurilor de caractere. De exemplu, criteriul de filtrare poate fi: „Alege din baza de date CATAGRAF acele înregistrări pentru care valoarea din câmpul *Ocupația* este Croitor”. Comanda SQL prin care realizăm această filtrare va fi:

```
SELECT Prenume, Vârstă, Ocupația INTO Result2
FROM CATAGRAF
WHERE Ocupația = „Croitor”
```

Rezultatul se regăsește în Tabelul 6.

Tabel 6

Persoane a căror ocupație este *croitor*, din baza de date CATAGRAF

Result	
Prenume	Ocupația
Ivanciu	croitor
Stoica	croitor

Valoarea câmpului *Ocupația* trebuie scrisă exact așa cum apare ea în baza de date. Prin urmare *CROITOR* sau *croitor* nu este același lucru cu *Croitor*.

Folosind wildcards⁹, majoritatea sistemelor au posibilitatea de a găsi șiruri de caractere care încep cu anumite litere sau conțin, de pildă, o anumită literă pe a patra poziție din șir ș.a.m.d., ceea ce este deosebit de util în cazul codificărilor.

Interogări de ordin istoric din ce în ce mai interesante rezultă atunci când sunt examinate mai multe câmpuri în același timp. De exemplu, putem fi interesați de indivizii sub 30 de ani, de etnie sârbă. Interogarea implică șiruri de caractere și cifre, iar criteriul de filtrare va fi: „Alege din baza de date CATAGRAF toate

⁹ Un character wild-card este un caracter folosit în căutări, care are altă semnificație decât cea obișnuită și, de obicei, înseamnă „orice șir de caractere”.

înregistrările pentru care valoarea din câmpul *Vârstă* este mai mică decât 30 și valoarea din câmpul *Etnie* este *Sârb*". Comanda SQL este în acest caz:

```
SELECT Prenume, Vârstă, Etnie INTO Result3
FROM CATAGRAF
WHERE Vârstă < 30 AND Etnie = „Sârb”
```

Comanda de mai sus presupune că înregistrarea îndeplinește ambele criterii. Rezultatul este ilustrat de Tabelul 7.

Tabel 7

Persoanele sub 30 de ani, de etnie sârbă din baza de date CATAGRAF

Result3		
Prenume	Vârstă	Etnie
Ivanciu	25	sârb
Dumitru	15	sârb
Rada	25	sârb
Tincu	28	sârb

Este de asemenea posibil să alegem înregistrări care îndeplinesc unul sau altul din cele două criterii, de exemplu indivizi care sunt sau sub 30 de ani sau sunt de etnie sârbă: „Alege din baza de date CATAGRAF toate înregistrările pentru care valoarea din câmpul *Vârstă* este mai mică decât cifra 30 sau valoarea din câmpul *Etnie* este *Sârb*". Realizăm această filtrare prin comanda SQL:

```
SELECT Prenume, Vârstă, Etnie INTO Result4
FROM CATAGRAF
WHERE Vârstă < 30 OR Etnie = „Sârb”
```

Tabelul 8 prezintă rezultatul filtrării.

Tabel 8

Persoane sub 30 de ani sau de etnie sârbă din baza de date CATAGRAF

Result4		
Prenume	Vârstă	Etnie
Uța	19	
Safta	20	român
Maria	7	

Result4		
Prenume	Vârstă	Etnie
Gheorghe	14	
Anica	10	
Ioana	7	
Catrina	3	
Zița	2	
Luxandra	18	
Ilinca	3	
Dobrița	1	
Panaete	20	
Costandin	16	
Drăgaica	13	
Voica	1	
Ianache	20	român
Maria	18	
Ioana	2	
Nicolae	15	
Costandin	25	român
Stanca	20	
Preda	3	
Gheorghe	2	
Lică	25	român
Sevasta	20	
Ioniță	5	
Ene	5	
Ștefan	3	
Maria	0	
Nicolae	0	
Stana	18	
Safta	0	
Rada	0	
Baros	9	
Stoica	10	
Marin	6	
Stana	3	
Nenu	35	sârb
Maria	25	
Ioniță	3	
Stan	16	

Result4		
Prenume	Vârstă	Etnie
Voicu	12	
Tănase	8	
Dumitru	6	
Minea	3	
Șarban	13	
Maria	7	
Zmaranda	4	
Ivanciu	25	sârb
Maria	20	
Nicolae	5	
Pătrana	2	
Dumitru	15	sârb
Ștefan	20	român
Neacșa	18	
Maria	13	
Rada	25	sârb
Costandina	6	
Dumitru	8	
Haila	12	evreu
Iancu	25	evreu
Maria	20	
Tincu	28	sârb
Catrina	20	
Neagu	8	
Andrei	4	
Maria	1	
Cârstea	50	sârb
Dumitru	18	
Luxandra	15	
Dica	25	
Ion	13	
Dumitru	2	
Vasilca	4	

Chiar și în interiorul aceluiași câmp, criteriile de selecție pot varia. De exemplu, putem extrage din baza de date toate înregistrările pentru care valoarea din câmpul *Ocupația* este sau *Croitor* sau *Căruțaș*.

Posibilitățile de filtrare sunt variate, în funcție de sistemul de gestionare a bazei de date utilizat. Majoritatea acestor sisteme vor folosi mijloace standard pentru interogări complexe. Un exemplu SQL de acest gen poate fi cel de mai jos:

```
SELECT Prenume, Vârstă, Ocupația INTO Result5
FROM CATAGRAF
WHERE Vârstă > 20 AND (Ocupația = „Cârciumar” OR Ocupația = „Croitor”)
```

Ultima linie se bazează pe convenția bine cunoscută din matematică, prin care operațiile dintre paranteze sunt efectuate primele. Rezultatul este prezentat în Tabelul 9.

Tabel 9

Persoane cu vârsta peste 20 de ani a căror ocupație este cârciumar sau croitor

Result5		
Prenume	Vârstă	Ocupația
Nenu	35	cârciumar
Ivanciu	25	croitor
Stoica	37	croitor

Un alt mod util de manipulare a datelor este sortarea și listarea lor într-o anumită ordine. Majoritatea bazelor de date vor permite listarea întregii informații pe care o conțin sau a rezultatelor interogărilor în ordinea dorită, în funcție de unul sau mai multe câmpuri. Datele pot fi listate după numele de familie, prenume, locul nașterii etc. Indivizii pot fi listați în funcție de vârstă, de la cel mai tânăr la cel mai vârstnic, adică în ordine ascendentă, sau de la cel mai vârstnic la cel mai tânăr, adică în ordine descendentă. Sortarea poate fi efectuată și în funcție de mai multe câmpuri, de pildă după numele de familie (în cazul tablei CATAGRAF nu e posibil), iar apoi, dacă mai mulți indivizi au același nume de familie, în funcție de prenume.

Procesul de filtrare poate fi utilizat pentru a crea noi baze de date. Spre exemplu, poate fi necesară o bază de date care să conțină doar persoanele de sex feminin. Operațiunea se realizează printr-o interogare care cere câmpurile dorite și prin salvarea rezultatelor într-un fișier nou, în timp ce fișierul original va rămâne intact.

Agregarea informației și reprezentările grafice

Interogările la care ne-am referit până acum au produs doar liste. Există situații în care cercetătorul nu este interesat de liste de nume, ci dorește informații

cu caracter general asupra unui grup de indivizi, precum o numărătoare a bărbaților sub 40 de ani, valoare minimă și maximă a vârstei celor incluși în recensământ sau valori centrale ale datelor cum ar fi media și mediana.

Căutarea numărului de înregistrări care au o anumită valoare poartă numele de *distribuție de frecvență*. Astfel, putem fi interesați, în cazul nostru, de distribuția stării civile a adulților, considerând indivizii cu vârsta de peste 18 ani. Rezultatul este prezentat în Tabelul 10.

Tabelul 10

Distribuția de frecvență a indivizilor peste 18 ani din baza de date CATAGRAF

Codul	Starea civilă	Frecvență
C	Căsătorit	42
N	Necăsătorit	50
V	Văduv/ă	8

Distribuția de frecvență de mai sus poate fi prezentată grafic foarte simplu, de exemplu printr-un grafic sub formă de batoane.

Există situații în care suntem interesați de gruparea datelor. În cazul câmpului *Vârstă*, o distribuție de frecvență a tuturor vârstelor posibile, de la 0 la 60 de ani, extremele înregistrate în baza de date CATAGRAF, va reprezenta 60 de categorii diferite, fără vreo semnificație din punct de vedere istoric. Analiza va fi cu mult mai clară dacă stabilim grupe de vârstă, care pot fi egale (de exemplu 0-4; 5-9; 10-14; 15-19 etc.) sau inegale (de exemplu 0-14; 15-29; peste 30). Un software bun va da utilizatorului flexibilitatea de a face astfel de grupări și de a număra înregistrările din fiecare grup. Se va crea un nou câmp, numit de pildă *VârstăGrup*, iar calculatorul va fi instruit să parcurgă baza de date înregistrare cu înregistrare. În cazul grupelor egale, de fiecare dată când va întâlni în câmpul *Vârstă* o valoare situată între limitele „0-4”, va pune valoarea „0-4” a respectivei înregistrări în câmpul nou creat *VârstăGrup*. În mod analog, de câte ori va întâlni în câmpul *Vârstă* o valoare situată în intervalul „5-9”, va pune valoarea „5-9” în câmpul *VârstăGrup*, ș.a.m.d. Câmpul *VârstăGrup* va fi declarat ca șir de caractere și nu ca un câmp numeric, pentru că valorile lui nu sunt numere asupra cărora să se opereze matematic. Distribuția de frecvență a noului câmp este prezentată în Tabelul 11, pentru grupe de vârstă egale, respectiv Tabelul 12, pentru grupe de vârstă inegale.

Tabel 11

Distribuție de frecvență pentru grupe de vârstă egale

Grupe egale de vârstă	Frecvență
0-4	18
5-9	13
10-14	9

15-19	11
20-24	9
25-29	8
30-34	6
35-39	5
40-44	5
45-49	1
50-54	4
55-59	0
60-64	6
65-69	0
70-74	1

Tabel 12

Distribuție de frecvență pentru grupe de vârstă inegale

Grupe inegale de vârstă	Frecvență
0-14	40
15-29	28
Peste 30	28

Calcululele matematice sunt, de asemenea, importante. În tabla CATAGRAF există doar un singur câmp numeric și anume *Vârstă*. Din păcate, asupra variabilei *Vârstă* nu se pot face calcule numerice care să conducă la rezultate pertinente. În cazul unor variabile cum ar fi suprafața cultivabilă a moșiilor, numărul animalelor deținute într-o gospodărie, venitul etc., se pot face calcule de tipul numărul indivizilor care dețin o anumită suprafață de pământ, sau numărul total al animalelor deținute, pe categorii, sau venitul mediu pe gospodărie, calcule care întregesc imaginea din punct de vedere economic a comunității. Aceste operații diferă de o simplă numărătoare a frecvențelor și sunt cu mult mai importante din punct de vedere istoric. Majoritatea sistemelor de gestionare a bazelor de date realizează astfel de calcule instantaneu, pentru un număr impresionant de înregistrări.

În fine, trebuie să amintim și un alt mod de analiză, care ia în considerație două variabile în același timp. De exemplu, putem analiza, printr-un tabel de contingență, două variabile și anume *Stare civilă* și *Sex*. Rezultatele sunt prezentate, pentru tabla CATAGRAF, în Tabelul 13.

Tabel 13

Tabel de contingență al variabilelor *Stare civilă* și *Sex*

SEX	Căsătoriți	Necăsătoriți	Văduvi
Masculin	21	28	1
Feminin	21	22	7
Total	42	50	8

Scopul celor expuse mai sus a fost acela de a introduce interogările standard la care poate fi supusă o tablă utilizând un sistem de gestionare a bazelor de date.

Codificare și structurare

Ne-am referit așadar la modul de introducere a datelor în cazul unei surse cu o structură simplă, care poate fi computerizată, fiind practic foarte puțin modificată. În cele ce urmează vom încerca să prezentăm avantajele pe care cercetătorul le poate avea extrapolând informația transcrisă din sursa originală.

Dacă istoricul transcrie informații pe fișe, în scopul indexării, cu siguranță că își face adnotări care să îl ajute să sorteze datele pe categorii. Același lucru poate fi făcut și în cazul unei baze de date, prin adăugarea de câmpuri care să conțină prescurtări și coduri analitice. În cazul procesului automat, codificarea are un grad de consistență și acuratețe mult mai mare.

Introducerea câmpurilor de codificare într-o bază de date trebuie să fie făcută cu multă grijă, datorită faptului că spațiul de stocare nu mai este o problemă, fiind vorba de fapt de o post-codificare. Aceasta înseamnă că informația este reținută în câmpuri într-o formă cât mai apropiată de putință de sursa originală. În acest fel nu se pierde din informație și schema de codificare poate fi verificată în orice moment sau modificată ulterior. Acest gen de tehnici sunt utile pentru o mare varietate de surse istorice, dar ne vom opri la exemplul utilizat în paragrafele anterioare. Desigur, ni se poate deja imputa pierderea unor informații de importanță pentru cercetarea istorică¹⁰, prin ignorarea corecțiilor făcute de recenzori, a schimbării scrisului sau a altor însemnări care apăreau pe fișele de recensământ găsite în arhive¹¹. Obiectivul principal al cercetării este însă, în cazul nostru, clarificarea unor probleme de istorie socio-economică a secolului al XIX-lea, ceea ce face ca informațiile care ne interesează să fie cele despre structura gospodăriei sau ocupații, nu amănuntele privind modul de lucru al recenzorilor.

Există patru posibilități de codificare, fie că ea este de tipul pre-codificare sau post-codificare și anume:

- înlocuirea informației dintr-un singur câmp cu un cod standard de prescurtare;
- adăugarea unui nou câmp codificator, care clasifică sau impune o structură asupra informației din alt câmp;
- adăugarea unui nou câmp care conține informații derivate din sursă, dar care nu apar în sursă în mod explicit;
- adăugarea unui câmp codificat, care grupează înregistrări potrivit informației care se găsește în mai multe câmpuri ale uneia sau mai multor surse originare.

¹⁰ Vezi Edward Higgs, *op. cit.*

¹¹ Vezi P.M.Tillot, *Sources of Inaccuracy in the 1851 and 1861 Censuses*, în Edward Anthony Wrigley (ed.), *Nineteenth Century Society*, Cambridge, 1972, p. 82-133.

Scopul codificării este întotdeauna acela de a simplifica și/sau lărgi posibilitățile de căutare¹². Ne vom referi pe rând la fiecare dintre tipurile de codificare menționate mai sus.

Cel mai simplu este codul de prescurtare, ca acela folosit în cazul variabilei *Starea civilă*. Cele mai multe aplicații de baze de date îi vor da utilizatorului posibilitatea să facă substituții, înlocuind una sau mai multe valori dintr-un anumit câmp cu alte valori, mai simple. Dacă, inițial, informația a fost introdusă exact așa cum apare ea în sursă, programul poate produce o listă a tuturor variantelor, în ordinea lor de apariție. O asemenea listă, în care cea de-a doua coloană reprezintă codul adăugat, conduce la standardizarea informației:

Tabel 14

Standardizarea informației prin codificare

Stare civilă	Cod
Căs	C
Căsăt	C
Necunoscut	X
Necăs	N
Necăsăt	N
Văd	V
Vă	V

În acest caz este convenabil să substituim informația din sursa originală cu coduri de o literă, fără a pierde din semnificația ei. În momentul utilizării, o asemenea codificare va asigura consistență și viteză maximă în regăsirea informației privind starea civilă a indivizilor.

Multe sisteme de gestionare a bazelor de date au posibilitatea fie de a modifica informația dintr-o anumită coloană (de exemplu schimbă *Căsătorit* cu *C*), fie de a plasa codul într-o nouă coloană, fără a afecta informația așa cum a fost ea inițial introdusă. Cea din urmă opțiune este utilă mai ales atunci când codificarea este menită să rezume o informație prea diversă pentru o analiză cu caracter general. Un exemplu ar fi variabila *Vârstă*, ale cărei valori între 0 și 60 de ani sunt toate reprezentate în tabla CATAGRAF. În loc să folosească vârsta individuală a populației recensate, istoricul poate prefera adăugarea coloanei *VârstăGrup*, dând apoi calculatorului comanda de a insera în coloana respectivă cifra 0 pentru orice individ a cărui vârstă este sub 10 ani, cifra 1 pentru orice individ a cărui vârstă este mai mare ca 9 și mai mică decât 20 ș.a.m.d. Dacă scopul a fost acela de a explora relația existentă între vârstă și starea civilă, pot fi adoptate grupări mai mici de

¹² Vezi Konrad Hugo Jarausch, Kenneth A. Hardy, *Quantitative Methods for Historians*, Chapel Hill, 1991, p. 37-41.

10 ani. Codificarea de grupare poate fi oricând modificată sau chiar se poate renunța la ea în orice moment al cercetării, dacă nu mai este necesară.

O altă abordare poate fi luată în considerație pentru variabila *Relația cu capul gospodăriei* (RELCG). Această variabilă este vitală pentru orice cercetare asupra structurii gospodăriei și codificarea trebuie să fie flexibilă și clară. Putem fi tentați să adoptăm o codificare prin prescurtare, de exemplu „FI” pentru fiică, „NO” pentru noră etc. Astfel de coduri se dovedesc a nu fi utile în procesul de regăsire a informației. În locul unei codificări prin șiruri de litere este de preferat adoptarea celei prin secvențe de câte două cifre, care să ilustreze relația cu capul gospodăriei: astfel, numerele de la 1 la 20 vor reprezenta relațiile din familia nucleară (un fiu poate fi codificat 11, un fiu vitreg 12, o fiică 14 ș.a.m.d.), cele între 30 și 69 vor reprezenta rudele mai îndepărtate de capul gospodăriei și soția lui, iar cei care nu le sunt rude vor fi codificați prin numere mai mari ca 70. Numărul 99 va fi rezervat pentru înregistrări care nu dau nicio informație sau în care valoarea câmpului este *necunoscut*.

Dacă codificarea relației cu capul gospodăriei este bine aleasă și dacă sursa în sine este consistentă, înlocuirea informației inițiale cu un cod este acceptabilă. Codificarea poate conduce și la corecții ale informațiilor care nu sunt consistent plasate, așa cum se întâmplă adesea în majoritatea surselor istorice. De exemplu, calitatea de *ucenic*, care poate apare uneori ca relație cu capul gospodăriei sau cel mai adesea ca ocupație, va fi trecută doar ca ocupație. Recenzorul poate fi de cele mai multe ori imprecis în ceea ce privește anumite relații, mai ales în cazul cuscilor, nurorilor, ginerilor sau a rudelor îndepărtate. În această situație este mai bine să adăugăm o nouă coloană, a relațiilor rezultate din cercetarea întregii surse și a unor surse adiacente. Astfel de coloane adăugate (de pildă pentru un cuplu căsătorit care trăiește în respectiva gospodărie, ocupându-se de treburile acesteia) trebuie să suplimenteze informația extrasă din sursa originală și nicidecum să o înlocuiască.

Dacă există dubii în privința informațiilor de natură geografică, o schemă similară poate fi adoptată pentru câmpurile bazei de date *Orașul* și *Comuna*, componente ale variabilei *Locul nașterii*. În acest caz este necesară o schemă de codificare și mai atentă. Un cod dintr-un singur caracter va fi insuficient pentru o analiză amănunțită, de finețe. Mai potrivită poate fi o abordare ierarhică, ce permite regăsirea informației la diferite nivele geografice. Un cod numeric din șase cifre poate determina țara prin primele sale două cifre, județul prin următoarele două și comuna prin ultimele două cifre. Cifra 9 poate fi păstrată pentru cazuri nespecificate. Avantajul unui sistem de codificare ierarhic este acela că, folosind wildcards, fiecare nivel geografic poate fi extras separat în timpul procesului de regăsire a informației.

O astfel de schemă de codificare poate fi suficientă pentru a înlocui toate variantele din coloana de origine, *Locul nașterii*. Sistemul poate fi adoptat și ca sistem de post-codificare, informația inițială fiind păstrată. În ambele cazuri, lista

denumirilor geografice și a codurilor alocate va fi utilă și pentru codificarea altor surse. Schema de codificare poate fi lărgită pentru a reflecta schimbările istorice ale limitelor unităților administrative.

Schemele folosite pentru codificarea informației din sursa originală, păstrându-se în același timp și această informație, sunt foarte utile pentru descrierea unui tip de informație complex, precum profesiile sau ocupațiile¹³. Recenzorul se bazează pe declarația capului de gospodărie, se presupune deci că pot exista erori intenționate sau neintenționate, variante sau omisiuni. Ne putem întreba dacă înregistrări ca *servitori*, *menajere*, *băiat în casă* trebuie sau nu să fie tratate ca o singură categorie, cea de *servitori*, diferența fiind dată doar de alte variabile, ca sexul și vârsta? O altă întrebare care se ridică este aceea dacă respectivele categorii înseamnă același lucru atât la țară cât și la oraș. Sau ne putem întreba cum facem distincția între ucenicul dintr-un atelier de fierărie și muncitorul care îl asistă pe fierar. Trebuie oare să facem diferența între sora capului de gospodărie, care se ocupă de menaj, și o menajeră plătită pentru această muncă, dar care nu este rudă cu capul gospodăriei?

Mai există și problema indivizilor cu mai multe ocupații, care ar trebui să fie listate în ordinea importanței. Toate sursele istorice sunt pline de inconsistențe în ceea ce privește denumirea ocupațiilor. Complicații suplimentare apar dacă cercetătorul dorește să compare ocupațiile de-a lungul timpului sau din regiuni geografice cu tradiții diferite¹⁴. Desigur, instrucțiunile oficiale de recenzare, privind nomenclatorul ocupațiilor, s-au modificat în timp, de la un recensământ la altul, și nu toate aceste modificări reprezintă cu acuratețe schimbările petrecute în viața socio-economică a secolului al XIX-lea. Toate aceste ambiguități nu sunt ușor de rezolvat. Pentru Anglia, de exemplu, abia în 1911 au început să fie introduse instrucțiuni clare în acest sens.

Cu siguranță, nu există răspunsuri simple la întrebări ca cele de mai sus, deciziile luându-se, pentru fiecare caz, în mare măsură în funcție de scopul căruia îi servesc datele. O cercetare cantitativă, care urmărește agregarea datelor și este asistată de calculator, va trebui să evite ambiguitățile și utilizarea variantelor, ceea ce într-adevăr știrbește din fidelitatea față de sursa originală. Așa s-a procedat în Anglia unde, legat de fiecare recensământ în parte, *General Register Office* a compilat liste detaliate de cuvinte cu reguli de clasificare ale sectoarelor economice. Dezbaterile pe tema adoptării versiunilor modificate ale acestor clasificări pentru o codificare pe scară largă a ocupațiilor continuă.

¹³ Vezi William A. Armstrong, *The Use of Information About Occupation*, în Edward Anthony Wrigley (ed.), *op. cit.*, p. 191-310.

¹⁴ Vezi D. Mills, J. Mills, *Occupation and Social Stratification Revisited: the Census Enumerators' Books of Victorian Britain*, în „Urban History Yearbook”, 1989, p. 63-77.

Codificarea nu este numai un mod de standardizare a informației din câmpuri individuale¹⁵, fiind utilă și ca stadiu intermediar în procesul de analiză a datelor pentru identificarea unor categorii de înregistrări. Un sistem bun de gestionare a bazelor de date îi va permite utilizatorului să introducă noi câmpuri codificate, bazate pe informația din alte câteva câmpuri. De exemplu, dacă valoarea câmpului 1 este x, iar a câmpului 2 este mai mică decât y sau mai mare decât z, programului i se poate cere să insereze codul AA într-un nou câmp, destinat special codificării.

Cercetătorul poate fi interesat de compararea ocupațiilor femeilor din diferite grupe de vârstă, extinzând cercetarea chiar la recensăminte diferite. În acest scop, poate fi utilă separarea femeilor cu vârste sub 15 ani și gruparea celor cu vârste de 15 ani și peste în categorii diferite, în funcție de criterii ca: existența unei responsabilități față de o familie proprie (măritate și/sau cu copii), traiul cu o rudă mai în vârstă, existența servitorilor sau a altui gen de asistență domestică, existența unor soți al căror venit este suficient pentru a le garanta siguranța financiară. Chiar dacă recensământul analizat nu răspunde satisfăcător criteriilor de analiză dorite, înainte de a fi consultate și alte surse se pot face, pe baza lui, o serie de filtrări referitoare la informația din alte câmpuri ale aceleiași înregistrări¹⁶.

Astfel de codificări, chiar dacă se bazează numai pe câteva câmpuri din recensământ, pot produce interogări mult mai flexibile decât căutarea manuală prin sursa respectivă. Viteza de lucru și posibilitățile de analiză statistică a sistemelor moderne de gestionare a bazelor de date permit testarea rapidă și eficientă a unui mare număr de ipoteze cu caracter istoric. În decursul acestui proces, istoricul trebuie, în primul rând, să nu dea credit unor rezultate eronate, care nu pot fi susținute de sursa originală. Rolul codificării este acela de a ușura procesul de regăsire a informației și de a îmbogăți potențialul de căutare într-un set de date, și nu acela de a distorsiona rezultatele. În al doilea rând, toate operațiile de codificare trebuie să fie însoțite de o documentație clară, precum și de materiale de arhivă, astfel încât, în cazul în care anumite rezultate se dovedesc a nu fi pertinente, calea către datele necodificate să poată fi refăcută.

Datele dintr-un recensământ conțin și informații care nu sunt explicit exprimate, dar care pot fi cu ușurință extrase și înregistrate în câmpuri separate. Astfel, în cazul gospodăriilor, istoricul poate să deducă numărul de persoane din fiecare gospodărie și din fiecare familie, număr care se înregistrează în câmpurile *Mărimea gospodăriei* și *Mărimea familiei*. Aceste informații vor permite calculul numărului mediu de persoane pe gospodărie, respectiv familie, pentru diferite tipuri de gospodărie, în același timp fiind utile schemei de codificare privind rolul femeii în gospodărie.

¹⁵ Vezi K. Schürer, *The Historical Researcher and Codes*, în Evan Mawdsley (ed.), *History and Computing III*, Manchester, 1990, p. 74-82.

¹⁶ Vezi Edward Higgs, *Structuring the Past: the Occupational and Household Classification of 19th Century Census Data*, în Evan Mawdsley (ed.), *op. cit.*, p. 67-73.

Mai poate fi adăugat și un cod al structurii gospodăriei, care constă dintr-o reprezentare vizuală a fiecărei gospodării, în termenii elementelor sale constitutive: A pentru fiecare adult prezent, cap de gospodărie sau soție; C pentru fiecare copil; N pentru fiecare nepot; R pentru rude din afara familiei nucleare; CH pentru fiecare chirieș; S pentru fiecare servitor ș.a.m.d. O intrare pentru o gospodărie cu doi adulți, trei copii, o rudă, un chirieș și doi servitori va fi codificată astfel: AACCCRCHSS. Introdus pentru fiecare gospodărie, un astfel de sistem de codificare reprezintă un rezumat grafic al structurii gospodăriilor din întreaga comunitate asupra căreia s-a operat recensământul¹⁷.

Sursele istorice de tipul recensămintelor ne dau informații la diferite niveluri. Cu cea mai mare frecvență apar informațiile asupra fiecărui individ. În același timp, există date privind unități mai mari, de pildă gospodăria sau așa numitele unități co-rezidențiale, care reprezintă un grup de indivizi ce împart permanent masa și casa. Mai apar și informații privind clădirile existente, atunci când este menționată și adresa la care locuiește fiecare familie. În cazul mediului urban, dens populat, acest tip de informație nu este omogen. În mod evident, datele au o structură bine definită: fiecare individ aparține unei unități co-rezidențiale și fiecare unitate de acest tip locuiește într-o clădire.

Pentru ca baza de date să reflecte această structură și pentru a se evita repetarea unor informații ca adresa, mărimea gospodăriei sau mărimea familiei în mod inutil pentru fiecare înregistrare, sistemele de baze de date relaționale distribuie datele ierarhic, între mai multe table. În mod logic, va exista o tablă pentru fiecare nivel. În cazul nostru ar putea fi trei table, una pentru gospodărie, una pentru unitatea co-rezidențială și una pentru indivizi.

Pentru recensămintele timpurii, informația asupra clădirilor nu este concludentă, prin urmare trebuie consultate alte surse din care ar putea fi extrase date referitoare la clădiri. Dacă istoricul hotărăște să-și extindă cercetarea și în această direcție, modelul va distribui informația din recensământ în trei table, legate între ele prin identificatori unici, unul pentru fiecare clădire și unul pentru fiecare unitate co-rezidențială. La fiecare nivel pot fi create legături cu alte tipuri de surse istorice, care conțin date despre gospodărie, familie etc.

Scindarea sursei în trei table diferite înseamnă transformarea ei dintr-o bază de date de tipul flat-file, într-o bază de date relaționale. La prima vedere aceasta ar putea să creeze senzația distorsionării sursei originare. În realitate, această abordare este chiar mai fidelă sursei și scopului în care a fost creată. Recensămintele secolului al XIX-lea nu se doreau a fi doar simple evidențe administrative, explorarea făcându-se la nivel de gospodărie, familie, urmărindu-se structura acestora, a comunității sau chiar potențialul militar, semnificația lor fiind socio-economică.

¹⁷ Vezi R.J. Morris, *Occupational Coding: Principles and Examples*, în „Historical Social Research”, XV, 1990, p. 3-29.

Modelul relațional favorizează, de cele mai multe ori, examinarea critică a unei surse aparent simplă. Practic, avantajele acestui model sunt multiple. Întrucât informația poate fi stocată în table diferite, cu o structură bine precizată, legate între ele printr-o cheie, adică printr-un identificator unic, va fi mai simplu ca regăsirile să se facă la niveluri distincte. Totodată, va fi mai simplă extragerea informațiilor din sursele auxiliare. Sistemul de gestionare a bazei de date poate fi utilizat în așa fel încât să evite contradicțiile interne care pot să apară dacă informația dintr-o tablă a fost editată fără informația sa complementară, care se află în altă tablă. De asemenea, poate fi inclus un software care să verifice automat validitatea datelor, de exemplu să nu existe indivizi cu vârste peste 110 ani sau a căror vârstă la căsătorie să fie sub 12 ani.

Deși majoritatea sistemelor de gestionare a bazelor de date pot fi utilizate cu un grad mare de complexitate, pentru anumite surse nu este necesar să depășim stadiul structurii unui flat-file. Desigur, utilizarea unui sistem de gestionare a bazelor de date capabil să clarifice și să restructureze sursa originală, potrivit scopului cercetării, rămâne o atracție¹⁸.

Există câteva principii generale care se desprind din cele expuse până acum. Toate sursele de date istorice, nu numai recensămintele, sunt structurate într-o mai mică sau mai mare măsură, iar structura surselor merită a fi examinată atât din punctul de vedere a ceea ce indică despre nașterea sursei, cât și din perspectiva utilizării datelor. De asemenea, chiar în timpul procesului de introducere a datelor, sursa poate fi îmbogățită, potențialul său fiind astfel mărit. Această operație conduce și spre identificarea altor surse, relevante pentru fiecare interogare a bazei de date. Dezvoltarea software-ului creat anume pentru analiza surselor de date istorice s-a intensificat în ultimii ani și este evident că modelul relațional, utilizat cu precauție, poate gestiona informații dintr-o mare varietate de surse istorice. Bineînțeles, fiecare proiect de cercetare trebuie să se limiteze la resursele documentare existente, dar procesul de introducere a datelor și analiza asistată de calculator conduc către o examinare critică sistematică a sursei, de multe ori imposibilă prin metode tradiționale.

Considerații asupra analizei statistice a bazelor de date istorice

Prin natura formației lor, dar și datorită controversei încă existente între adepții metodologiei tradiționale și cei ai istoriei cantitative, istoricii evită de cele mai multe ori analiza statistică a surselor. Chiar și atunci când adepții metodologiei tradiționale sunt de acord cu concluziile unui studiu cantitativ, ei obiectează că o laborioasă analiză statistică nu a făcut altceva decât să confirme ceea ce era evident. Tradiționaliștii au de cele mai multe ori dreptate când afirmă că abordarea

¹⁸ Vezi David Spaeth, *A Guide to Software for Historians*, Glasgow, 1991. Lucrarea prezintă modalități alternative de tratare a surselor.

statistică umbrește eleganța unei argumentații narative. În afară de asta, sunt și alte multe motive care îi fac pe istorici să evite statistica, motive care țin de necesitatea de a stăpâni abilități speciale pentru a aplica noua metodologie și a-i interpreta rezultatele, abilități care, în general, nu fac parte din programul de instruire al istoricului.

Există însă o problemă reală, care privește atât specificul datelor istorice, cât și condițiile riguroase impuse de metoda statistică. Este adevărat că pentru anumite perioade de timp arhivele abundă în statistici care, prelucrate, se concretizează în diagrame sugestive și în alte rezultate de sinteză cărora suntem gata să le dăm crezare, dar se uită că datele primare au fost culese de persoane care au făcut această operație mai mult sau mai puțin corect¹⁹.

Istoria economică, demografia istorică, istoria socio-politică sunt domenii în care analiza statistică și modelarea matematică își spun cuvântul, existând chiar reviste specializate ale căror studii și articole ilustrează aplicarea celor mai sofisticate și complexe metode de acest gen. În aceste domenii cercetătorul nu poate evita analiza statistică, ușurată considerabil acum de răspândirea pe cale largă a calculatoarelor personale și a software-ului specializat, cum ar fi SPSS sau Minitab.

Scopul nostru nu este acela de a prezenta diferitele pachete de programe de prelucrare statistică a datelor, ci acela de a explica tehnicile statistice de bază necesare pentru prelucrarea surselor istorice în general și a bazelor de date istorice în special. Ne vom referi la procedeele de descriere a variabilelor sau a entităților istorice, la măsurarea relației dintre două atribute și la formularea de concluzii cu caracter general, valide din punct de vedere istoric, pe baza unei submulțimi din totalitatea datelor.

Utilizarea corectă a metodelor statistice cere o practică îndelungată și studierea unor lucrări de specialitate, care nu sunt întotdeauna ușor de parcurs pentru istoric. Necesitatea publicării unor lucrări destinate istoricilor, în care metodele statistice să fie adaptate specificului datelor istorice, este din ce în ce mai acută. Există o serie de astfel de lucrări, majoritatea destinate însă altor științe sociale decât istoriografia²⁰.

Vocabularul de bază al statisticii matematice

Vocabularul de bază al statisticii matematice este apropiat de cel al bazelor de date. *Populația* este grupul de obiecte supus analizei și nu trebuie confundat cu populația în sensul numărului de locuitori dintr-un anumit spațiu geografic. În termenii unei baze de date, populație se poate numi chiar baza de date, o mulțime de date sau o tablă. Baza de date CATAGRAF este o populație.

¹⁹ Vezi Richard E. Beringer, *Historical Analysis*, New York, 1978, p. 198.

²⁰ Vezi Konrad Hugo Jarausch, Kenneth Hardy, *op. cit.* și Jarol B. Manheim, R. Rich, *Empirical Political Analysis*, New York, 1991.

Fiecare obiect al populației, fiecare entitate se numește *caz* sau *observație*. Într-o bază de date, fiecare înregistrare reprezintă un caz. O selecție de cazuri care este extrasă în scopul de a reprezenta o populație se numește *eșantion*. Caracteristicile sau atributele entităților din populație se numesc *variabile* (de exemplu vârsta, suprafața moșiei, numărul de animale, cuantumul birului), iar variabilele iau *valori* (35 ani, 50 ha, 5 animale, 10 ducăți). Într-o bază de date, variabilele sunt câmpurile sau coloanele. Schematic, terminologia analizei statistice este reprezentată în Tabelul 15.

Tabel 15

Terminologia de bază a analizei statistice

Baza de date	Statistică	Exemplu din baza de date CATAGRAF
Bază de date (Fișier, Tablă)	Populație	Vopseaua de Verde
Înregistrare	Caz (Observație)	Lazăr sin Ion
Câmp (Coloană)	Variabilă	Etnie
Valoare	Valoare	Român

Natura fiecărei variabile este importantă. O primă diferențiere se face între variabile numerice – exprimate prin cifre – și variabile calitative – exprimate prin cuvinte. De multe ori, după cum am mai spus, poate fi necesară codificarea numerică a variabilelor calitative (Căsătorit = 11, Masculin = 1 etc.), atunci când informația trebuie să fie analizată cu ușurință de un software. Dar, chiar dacă toată informația calitativă a fost codificată numeric, există diferențe fundamentale între variabilele nominale, ordinale și de interval. Aplicarea metodelor statistice diferă de la un tip de variabilă la altul.

Tipul datelor, măsuri ale tendinței centrale în date, testarea existenței unei relații între două variabile

Datele nominale sunt acele date în care cazurile sunt clasificate în categorii neordonate, dându-li-se un nume. Un exemplu de date nominale din baza de date CATAGRAF este *Etnia*. Indivizii pot fi clasificați în grupe, după etnie, fără să fie necesară o ordonare numerică logică. Variabila *Etnie* poate fi codificată numeric, dar nu se pot face operații matematice asupra ei.

În cazul *datelor ordinale*, informația este divizată în grupuri care pot fi ordonate, cărora li se poate, cu alte cuvinte, atribui un rang, dar care nu sunt împărțite în intervale egale.

În cazul *datelor de interval*, intervalul dintre valori este egal. Acestea sunt datele care pot fi exprimate numeric, asupra lor fiind posibilă efectuarea operațiilor matematice. În baza de date CATAGRAF, valorile variabilei *Vârsta* reprezintă date

de interval. Dacă informația primară din recensământ este îmbogățită prin indicarea numărului de persoane din fiecare gospodărie și familie, acest gen de informație suplimentară reprezintă tot date de interval.

Metodele statisticii descriptive, important capitol al statisticii matematice, sunt utilizate în descrierea variabilelor. Una dintre cele mai simple operații statistice este aceea de a găsi una sau mai multe cifre care să rezume un grup. Un exemplu simplu este *distribuția de frecvență* a variabilei nominale *Stare civilă* din Tabelul 10, care prezintă numărul indivizilor căsătoriți, necăsătoriți și divorțați. Dacă exprimăm cifrele și procentual, pentru fiecare categorie, informația devine și mai utilă.

Este de dorit să avem un număr care să reprezinte o informație importantă asupra întregului grup de indivizi. Acest număr este, cel mai adesea, o *valoare medie*. *Media aritmetică* este considerată a fi una dintre măsurile tendinței centrale a datelor. Alte măsuri de acest tip sunt *modul*, adică valoarea care apare în date cu cea mai mare frecvență, respectiv *mediana*, ea fiind valoarea de mijloc din lista datelor ordonate de la cea mai mică la cea mai mare valoare²¹.

Nu este necesar ca media să coincidă cu mediana, după cum demonstrează exemplul ipotetic din Tabelul 16, referitor la numărul de clăcași de pe 9 moșii.

Tabel 16

Exemplu ipotetic privind numărul de clăcași pe 9 moșii

Moșia	Număr clăcași
Moșia 1	10
Moșia 2	25
Moșia 3	15
Moșia 4	10
Moșia 5	20
Moșia 6	10
Moșia 7	20
Moșia 8	10
Moșia 9	100
Total	220

Ordonând valorile crescător: 10, 10, 10, 10, 15, 20, 20, 25, 100, valoarea mediană (de mijloc), va fi 15. Media aritmetică este aproximativ 24 (220:9), deci cu mult mai mare, asta datorită valorii 100, înregistrată pentru Moșia 9. Modul, adică valoarea care apare cu cea mai mare frecvență, este 10.

Măsuri ale dispersiei în date se adaugă celor care exprimă tendința lor centrală. Aceste măsuri nu arată ceea ce este tipic în date, ci cât de mare este

²¹ Vezi Angelo Porro, *Storia e statistica*, Roma, La Nuova Italia Scientifica, 1989, p. 151-158; T.G. Connolly, Wladyslaw Sluckin, *An Introduction to Statistics for the Social Sciences*, Great Britain, MacMillan, 1971, p. 23-60.

variația lor. *Amplitudinea* datelor este exprimată prin diferența dintre cea mai mare și cea mai mică valoare din date (în cazul nostru $100 - 10 = 90$). Ca și în cazul tendinței centrale, este util să exprimăm gradul de dispersie a datelor printr-un număr, acesta fiind *deviația standard*. Această măsură statistică este extrem de sugestivă atunci când comparăm două populații care au aceeași medie și aceeași mediană, dar amplitudini diferite. Software-ul statistic modern este capabil să facă astfel de calcule în câteva secunde.

De cele mai multe ori, un caz comportă mai multe variabile, de exemplu vârsta, locul nașterii, etnia etc. O întrebare care se ridică în mod natural este aceea a existenței sau inexistenței unei legături între două variabile. De pildă, indivizii cu o vârstă mai înaintată au un venit mai ridicat? Sau: există o legătură între variabila *Etnie* și variabila *Ocupație*?

Metodele statisticii multivariate se ocupă de configurarea relațiilor dintre variabile. Existența unor relații poate fi testată, ipoteza testului fiind aceea că între cele două variabile există o relație. Ipoteza nulă, pe de altă parte, afirmă contrarul, și această ipoteză trebuie să fie respinsă înainte ca ipoteza alternativă să fie admisă. Un test statistic examinează evidența și pe această bază ipoteza nulă este sau nu respinsă. Procesul de luare a deciziei pe baza testării ne conduce la noțiunea de *nivel de semnificație*, adică la probabilitatea de a respinge în mod incorect ipoteza nulă. Cu cât nivelul de semnificație este mai mic, cu atât gradul de siguranță este mai mare. Acest gen de abordare a problemei poate fi folosit cu orice tip de variabile, testele diferind de la un tip la altul.

Pentru variabilele nominale, *tabelul de contingență* este una dintre metodele de bază care ilustrează relațiile dintre variabile. Un exemplu pentru baza de date CATAGRAF a fost dat deja în Tabelul 13. O modalitate precisă de testare a existenței unei relații între două variabile dintr-un tabel de contingență este testul χ^2 , generat de orice software statistic care produce un tabel de contingență. Tabelele de contingență sunt utile în studiile de demografie istorică, atunci când se analizează structura familiei²².

Datele de interval permit reprezentări grafice simple și un alt gen de teste. Un exemplu poate fi creat chiar din baza de date CATAGRAF. Pornind de la această informație, putem deriva o altă tablă, numită CUPLURI, în care cazul (înregistrarea) nu este individul ci cuplul căsătorit. Vor exista două variabile de interval, *Vârsta soțului* și *Vârsta soției*, care pot fi comparate și reprezentate grafic ca *diagramă de împrăștiere*. Într-o cercetare asistată de calculator, operația devine foarte simplă. Software-ul este instruit ce variabile să reprezinte în diagramă și ce indici statistici să fie calculați. Fiecare punct din diagramă reprezintă o înregistrare, cu alte cuvinte un cuplu căsătorit. Vârsta soțului va fi măsurată pe axa verticală, iar vârsta soției va fi măsurată pe axa orizontală. În baza de date CATAGRAF există

²² Vezi Michael Anderson, *Family Structure in Nineteenth Century Lancashire*, 1971, p. 79-95.

20 de cupluri căsătorite, prin urmare în diagramă vor fi 20 de puncte, fiecare aflându-se la intersecția perpendicularelor ridicate pe axa orizontală, respectiv verticală, din punctul care reprezintă vârsta soției, respectiv a soțului.

Tabel 17

Variabile din tabla CUPLURI

Variabila	Exemplu	Tipul variabilei
ID soț	1	Interval
Prenume soț	Stoica	Nominal
Fiul lui	Stoian	Nominal
Vârsta soț	30	Interval
ID soție	2	Interval
Prenume soție	Uța	Nominal
Vârsta soție	19	Interval

În afară de a exprima modul de răspândire a datelor, diagrama de împrăștiere ne ajută să înțelegem măsurile statistice utilizate, măsuri care de fapt rezumă diagrama. De fapt, reprezentările grafice, prin varietatea și complexitatea lor, sunt un mod sugestiv, rapid și simplu de primă analiză a surselor istorice.

Concluzii

Recensămintele, surse de importanță majoră pentru cercetările demo-statistice sau de istorie socială, îl forțează pe cercetător să ia în considerație nu numai datele agregate, rezultat al prelucrărilor, ci și datele referitoare la fiecare individ în parte, pentru corelarea informațiilor din recensământ cu cele din surse suplimentare, edite sau inedite. De aici rezultă importanța pe care o au informația culeasă pe formularele de recensare a populației și măsura în care ea a fost sau nu inițial structurată.

Mobilitatea geografică pe termen scurt a influențat procesul de înregistrare a indivizilor și a condus la apariția celor două metode de recensare: *de jure*, respectiv *de facto*. Unificarea sistemelor de recensare nu s-a reușit nici până în ziua de azi, deși există o intensă preocupare de comparare a recensămintelor din țări vecine, dar cu sisteme diferite de recensare a populației.

Bazele de date de tip recensământ, în funcție de scopul cercetării și de erudiția celui care le proiectează și gestionează, permit o multitudine de interogări care se pot realiza prin limbajul SQL.

Gândită întotdeauna pentru a simplifica și a extinde posibilitățile de explorare, codificarea se face acum de obicei după introducerea datelor, astfel încât informația să fie înregistrată în baza de date într-o formă cât mai fidelă sursei primare.

Reprezentările grafice, analiza corelațiilor, metodele de eșantionare și metodele statisticii neparametrice vin să întregesc tabloul complexelor prelucrări statistice ce se pot realiza asupra informației înscrisă într-o bază de date istorice de tip recensământ.

Anexa I

CATAGRAF												
IDentificator	Nr Gosp	Nume	Prenume	Fiul lui	RELCF	Sex	Etnie	Vârstă	Bir	Claca	STC	Ocupația
1	1		Stoica	Stoian	cap fam	M	român	30	birnic	chiriaș	C	florar
2	1		Uța		soție	F		19			C	
3	1		Safta		soră	F	român	20	scutit		V	
4	1		Maria		nepoată	F		7			N	
5	2		Lazăr	Ion	cap fam	M	român	40	pat cl III	prop	C	pescar
6	2		Despina		soție	F		35			C	
7	2		Gheorghe		fîu	M		14			N	
8	2		Anica		fiică	F		10			N	
9	2		Ioana		fiică	F		7			N	
10	2		Catrina		fiică	F		3			N	
11	2		Zița		fiică	F		2			N	
12	3		Zane	Stamate	cap fam	M	albanez	30	pat cl III	prop	C	mămular
13	3		Luxandra		soție	F		18			C	
14	3		Ilinca		fiică	F		3			N	
15	3		Dobrița		fiică	F		1			N	
16	3		Panaete		slugă	M		20			N	
17	3		Slamnea		slugă	F	român	60	scutit		V	
18	4		Năstase	Ghioca	cap fam	M	român	50	birnic	prop	C	pescar
19	4		Maria		soție	F		30			C	
20	4		Costandin		fîu	M		16			N	
21	4		Drăgaica		fiică	F		13			N	
22	4		Voica		fiică	F		1			N	
23	5		Ianache	Dumitrache	cap fam	M	român	20	birnic	prop	C	pescar
24	5		Maria		soție	F		18			C	
25	5		Ioana		fiică	F		2			N	
26	5		Calița			F		35			V	
27	5		Nicolae		slugă	M		15			N	

CATAGRAF												
IDentificator	Nr Gosp	Nume	Prenume	Fiul lui	RELCF	Sex	Etnie	Vârstă	Bir	Claca	STC	Ocupația
28	6		Costandin	Gheorghe	cap fam	M	român	25	pat	prop	C	pescar
29	6		Stanca		soție	F		20			C	
30	6		Preda		fîu	M		3			N	
31	6		Gheorghe		fîu	M		2			N	
32	6		Rafla		mamă	F	român	70	scutit		V	
33	7		Lică	Stan	cap fam	M	român	25	pat cl III	prop	C	pescar
34	7		Sevasta		soție	F		20			C	
35	7		Ioniță		fîu	M		5			N	
36	7		Ene		fîu	M		5			N	
37	7		Ștefan		fîu	M		3			N	
38	8		Stoica	Radu	cap fam	M	român	30	birnic	prop	C	cojocar
39	8		Maria		soție	F		0			C	
40	8		Nicolae		fîu	M		0			N	
41	8		Neda		mamă	F		60	scutit		V	
42	9		Ion	Toma	cap fam	M	român	40	pat cl III	prop	C	pescar
43	9		Stana		soție	F		18			C	
44	9		Safta		fiică	F		0			N	
45	9		Rada		fiică	F		0			N	
46	9		Stan		socru	M		60	N O R		V	
47	10		Stanciu	Crăciun	cap fam	M	țigan	60	birnic	prop	C	herar
48	10		Dumitra		soție	F		30			C	
49	10		Baros		fîu	M		9			N	
50	10		Stoica		fîu	M		10			N	
51	10		Marin		fîu	M		6			N	
52	10		Stana		fiică	F		3			N	
53	11		Nenu	Dinu	cap fam	M	sârb	35	plăt Gh	chiriaș	C	cârciumar
54	11		Maria		soție	F		25			C	
55	11		Ioniță		fîu	M		3			N	
56	12		Pârvu	Radu	cap fam	M	țigan	60	birnic	prop	C	cantaragiu
57	12		Păuna		soție	F		40			C	
58	12		Stan		fîu	M		16			N	

CATAGRAF												
IDentificator	Nr Gosp	Nume	Prenume	Fiul lui	RELCF	Sex	Etnie	Vârstă	Bir	Claca	STC	Ocupația
59	12		Voicu		fiu	M		12			N	
60	12		Tănase		fiu	M		8			N	
61	12		Dumitru		fiu	M		6			N	
62	12		Minea		fiu	M		3			N	
63	12		Șarban		fiu	M		13			N	
64	12		Maria		fiică	F		7			N	
65	12		Zmaranda		fiică	F		4			N	
66	13		Ivanciu	Nicolae	cap fam	M	sârb	25	birnic	prop	C	croitor
67	13		Maria		soție	F		20			C	
68	13		Nicolae		fiu	M		5			N	
69	13		Pătrana		fiică	F		2			N	
70	13		Dumitru		slugă	M	sârb	15			N	
71	13		Ilinca		slugă	F	român	45	scutit		V	
72	14		Ștefan	Neculai	cap fam	M	român	20	birnic	prop	C	căruțaș
73	14		Neacșa		soție	F		18			C	
74	14		Ștefan	Dinu	cap fam	M	român	50	birnic	chiriaș	C	muncitor
75	14		Stoiana		soție	F		35			C	
76	14		Maria		fiică	F		13			N	
77	14		Rada	Dima	cap fam	F	sârb	25	scutit		V	
78	14		Costandina		fiică	F		6			N	
79	14		Dumitru		fiu	M		8			N	
80	15		David	Avram	cap fam	M	evreu	60	scutit	chiriaș	C	
81	15		Rifca		soție	F	evreu	50			C	
82	15		Haila		fiică	F	evreu	12			N	
83	16		Iancu	Maria	cap fam	M	evreu	25	birnic	chiriaș	C	sticlar
84	16		Maria		soție	F		20			C	
85	17		Tincu	Nițu	cap fam	M	sârb	28	pat cl III	prop	C	pescar
86	17		Catrina		soție	F		20			C	
87	17		Neagu		fiu	M		8			N	
88	17		Andrei		fiu	M		4			N	
89	17		Maria		fiică	F		1			N	
90	18		Cârstea	Păun	cap	M	sârb	50	pat cl	prop	C	dulgher

CATAGRAF												
IDentificator	Nr Gosp	Nume	Prenume	Fiul lui	RELCF	Sex	Etnie	Vârstă	Bir	Claca	STC	Ocupația
					fam				V			
91	18		Ioan		soție	F		40			C	
92	18		Dumitru		fîu	M		18			N	
93	18		Luxandra		fiică	F		15			N	
94	19		Bute	Bute	cap fam	M	ungurean	40	scutit	chiriaș	C	morar
95	19		Păuna		soție	F		30			C	
96	20		Stoica	Țighira	cap fam	M	arnăut	37	pat cl III	prop	C	croitor
97	20		Dica		soție	F		25			C	
98	20		Ion		fîu	M		13			N	
99	20		Dumitru		fîu	M		2			N	
100	20		Vasilca		fiică	F		4			N	

METHODS OF CENSUS-TYPE HISTORICAL DATABASES ANALYSIS. APLICATION TO THE 1838 CATAGRAPHY OF THE TOWN OF PLOIEȘTI

Abstract

This article presents the construction, management, and statistical analysis of a historical database, whose data source is the *Catagraphy of the Town of Ploiești* of 1838. This effort has all the main characteristics of a modern census. One of its goals was to count the possible tax-payers. Another goal was to serve as a starting point for conceiving a strategy for the future development of the country.

The article starts with a few considerations about the input data, then it emphasizes the necessity to corroborate the census with other historical sources. This task is not possible using only the aggregated data. The article continues by explaining how to codify and interrogate the source data by using commands specific to the relational databases.

The graphics, the correlation analysis, the sampling methods and non-parametric statistical methods applied to the source data and presented in the last part of the article complete the array of complex statistical procedures that can be applied on census-type data stored in a relational database.

