Metode si tehnici de analiza a calitatii datelor

Lect.dr. Liviu CIORA, prep. Ion BULIGIU
Catedra de Informatica Economica, Universitatea din Craiova

Data Quality Assessment is the scientific and statistical evaluation of data to determine if data obtained from environmental data operations are of the right type, quality, and quantity to support their intended use. This guidance demonstrates how to use data quality assessment in evaluating environmental data sets and illustrates how to apply some graphical and statistical tools for performing this assessment. The guidance focuses primarily on using data quality assessment in environmental decision making.
***Keywords**: data, statistics, methods, techniques, entities.*

Analiza calitatii datelor reprezinta o evaluare stiintifica si statistica realizata pentru a determina daca datele obtinute dintr-un mediu ce utilizeaza si proceseaza date, sunt de tipul potrivit, în cantitatea si calitatea necesara utilizarii lor pentru scopul pentru care au fost obtinute. Aceasta analiza are la baza urmatoarea premisa: conceptul de calitate a datelor este strâns legat de scopul utilizarii acestora.

Analiza calitatii datelor presupune parcurgerea a cinci etape:

1. *Stabilirea obiectivelor analizei calitatii datelor si structura datelor culese* – se face înaintea evaluarii datelor în vederea stabiliri structurii datelor culese din esantionul supus analizei;
2. *Revizuirea datelor preliminare* - studierea formei rapoartelor de evaluare a calitatii, efectuarea calculelor statistice de baza si generarea graficelor pentru aceste date. Informatiile sunt utilizate pentru studiul structurii datelor si pentru identificarea tipurilor acestora, a relatiilor dintre ele sau a potentialelor anomalii;
3. *Selectarea testului statistic* – se va selecta procedura cea mai potrivita pentru efectuarea calculelor si analiza datelor, decizie care are la baza fazele desfasurate anterior. Se vor identifica parametrii cheie care trebuiesc retinuti în vederea validarii procedurilor statistice;
4. *Verificarea previziunilor din testul statistic* – evaluarea parametrilor cheie sau a datelor preliminare, daca sunt caracterizate de valori acceptabile, în functie de datele obtinute sau alte informatii legate de studiul statistic;
5. *Determinarea concluziilor* – se realizeaza calculele necesare testului statistic si întocmirea documentatiei pentru inferentele rezultate din aceste calcule.

Evaluarea calitatii datelor este un proces de natura iterativa si nicidecum lineara. De exemplu, daca în recapitularea datelor preliminare se sesizeaza redundante, anomalii sau inconsistente ale setului de date care va fi supus analizei, atunci se va proceda la revizuirea aspectelor din pasul 1. De asemenea, daca previziunile statistice nu se încadreaza în limitele de evolutie a datelor, atunci se vor revizui pasii anteriori ai procesului de evaluare.

Punctul forte al evaluarii calitatii datelor este acela ca studiul este axat pe determinarea a cât de bine datele analizate se încadreaza în intentia de utilizare a acestora.

Cu toate acestea, trebuie subliniat faptul ca evaluarea calitatii datelor nu poate demonstra la modul absolut ca s-au atins sau nu obiectivele stabilite în timpul etapei de planificare a unui studiu. Aceasta situatie apare deoarece persoana care ia deciziile nu poate sti niciodata valoarea reala a unui element din cadrul intervalului de date supuse analizei. Colectarea datelor furnizeaza investigatorilor numai o estimare a acesteia si nu valoarea exacta. Datorita faptului ca metodele analitice nu sunt perfecte, acestea vor furniza numai o estimare a valorii adevarate dintr-un esantion stabilit. Deoarece investigatorii iau decizii bazate pe estimari si nu pe valori reale, ei îsi

vor asuma un risc de a lua o decizie eronata în privinta elementului de interes.

Stabilirea obiectivelor analizei calitatii datelor si structura datelor culese

Analiza calitatii datelor începe prin studierea iesirilor rezultate în urma fazelor din ciclul de viata a datelor: obiectivele calitatii datelor, asigurarea calitatii, planul proiectului si a altor documente asociate acestora. Obiectivele calitatii datelor furnizeaza contextul pentru întelegerea scopului colectarii datelor si stabilirea criteriilor calitative si cantitative pentru asigurarea calitatii setului de date stabilit în scopul utilizarii lui. Prin studierea metodelor prin care datele sunt colectate, masurate si raportate, analistul pregateste etapele preliminare ale întregului proces de analiza a calitatii datelor.

Aceasta prima etapa are ca scop stabilirea datelor rezultate în urma studiului, structura esantionului de date studiate si a documentatiei pentru colectia de datelor supuse analizei. Activitatile parcurse sunt: *studierea obiectivelor studiului, transformarea obiectivelor în ipoteze pentru studiul statistic*, *definirea limitelor de toleranta pentru erorile decizionale* si *stabilirea structurii esantionului de date*.

Obiectivele vor fi studiate pentru a furniza contextul procedurilor de analiza a datelor, obiectivele procesului calitatii datelor fiind urmatoarele:

1. Definirea problemei, identificarea componentilor echipei de analisti, examinarea bugetului necesar, planificarea operatiunilor intermediare;
2. Starea deciziei, identificarea cerintelor studiului si definirea alternativelor de actiune;
3. Identificarea datelor de intrare necesare deciziei (sursele informatiilor, se pun bazele pentru nivelul de actiune urmator, se stabilesc metodele de selectare a esantioanelor reprezentative);
4. Definirea unei reguli de decizie – prin care se stabilesc parametrii statistici (media si mediana), specificarea nivelului decizional, elaborarea schemei logice a actiunilor ce urmeaza a fi efectuate;
5. Specificarea limitelor de toleranta pentru erorile decizionale – stabilirea intervalului acceptabil pentru aceste erori relativ la implicatiile rezultate (efecte, costuri);
6. Optimizarea structurii de analiza pentru obtinerea datelor – selectarea planului de analiza si a resurselor esantionului din care sunt culese datele, astfel încât per total proiectul sa îndeplineasca criteriile de performanta.

Transformarea obiectivelor în ipoteze pentru studiul statistic presupune utilizarea obiectivelor utilizatorilor pentru a crea o structura a ipotezelor primare ce vor testa mediul datelor intrate în studiu. Structura ipotezelor statistice primare cuprinde o ipoteza de referinta, care este o conditie de baza „0”, ce se presupune a fi adevarata în absenta unui element decisiv care sa dovedeasca contrariul si o ipoteza alternativa, care verifica toate celelalte variante. Cu alte cuvinte, ipoteza de referinta se pastreaza numai daca ipoteza alternativa nu este considerata adevarata, în urma existentei unor dovezi covârsitoare.

În general, aceste ipoteze au în componenta urmatoarele elemente:

- o populatie de parametri de interes, care descriu proprietati ale mediului de date investigat;
- o valoare numerica cu care parametrul va fi comparat, cum ar fi valoarea medie, valoarea de risc sau valoarea unui alt parametru din alta locatie sau din alt moment de timp (de obicei se compara cu o valoare a parametrului dintr-un moment anterior);
- relatia (cum ar fi „este egal” sau „este mai mare decât”) care specifica precis comparatia dintre parametru si valoarea numerica.

Scopul activitatii de stabilire a limitelor pentru erorile decizionale este acela de a preciza toleranta utilizatorului datelor la erorile decizionale de tip *falsa respingere (tip I)* sau *falsa acceptare (tip II)* ca rezultat al incertitudinii datelor. Eroarea de *falsa respingere* apare ori de câte ori ipoteza de referinta, numita si ipoteza nula, este respinsa când este adevarata. Eroarea de *falsa acceptare* apare când ipoteza nula nu este respinsa când este falsa.

Pasii care vor fi urmati în continuare (figura 1):

- Specificarea „zonei cenusii” (grey region), în care aparitia unei erori decizionale de tip

„falsa acceptare" este relativ minima. „Zona cenusie" este delimitata de zona de valori a parametrului pentru care riscul pentru aparitia unei erori de tip falsa acceptare devine semnificativ. Latimea acestei zone este importanta pentru cei care iau decizii deoarece reprezinta intervalul de restrictie.

- Specificarea limitelor de toleranta pentru probabilitatea unei aparitii a erorilor de tip falsa respingere si falsa acceptare, care reflecta de fapt limitele de toleranta pentru luarea deciziilor incorecte de catre factorul decizional uman.

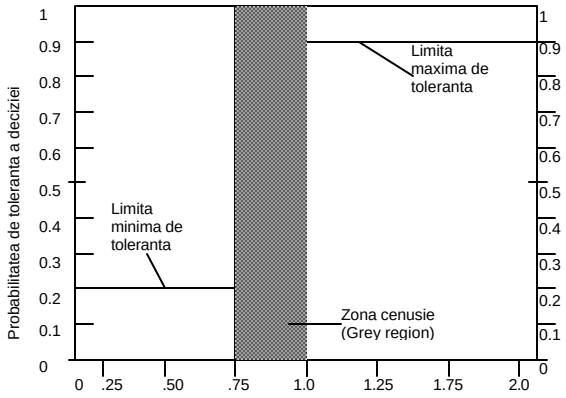

Figura 1

Stabilirea structurii esantionului de date este o etapa care furnizeaza baza analizei esantioanele de date supuse studiului. Aceasta analiza difera din punct de vedere al tehnicilor de analiza si a procedurilor ce urmeaza a fi aplicate pentru diferite seturi de date supuse evaluarii.

Exista doua tipuri de selectare a datelor: selectionare experta si selectionare probabilistica.

Selectionarea experta a datelor consta în asistarea de catre un expert, care cunoaste procesul din care provin datele si care coordoneaza culegerea datelor, precizând de unde si cum se realizeaza aceasta, precum si dimensiunea esantioanelor de date.

Acest tip de selectionare trebuie aleasa numai când obiectivele investigarii nu au o natura statistica sau când studiul este concentrat numai asupra locatiilor în sine ale esantioanelor culese. În general, concluziile rezultate din selectionarea experta se aplica doar unor esantioane individuale si agregarea acestora poate duce la concluzii eronate.

Selectionarea probabilistica este caracterizata de faptul ca fiecare element al populatiei de date de tip destinatie are o probabilitate cunoscuta de a fi inclus în esantion. Aceasta selectionare poate fi de variate forme, dar toate folosesc metoda factorului întâmplator (random), care permite determinarea de elemente probabilistice valide legate de calitatea estimarilor.

Revizuirea datelor preliminare

În aceasta faza a procesului de analiza a calitatii datelor, analistul procedeaza la o evaluare preliminara a setului de date prin efectuarea unor calcule statistice si examinarea datelor utilizând reprezentari grafice.

Revizuirea preliminara a datelor studiate trebuie efectuata când datele sunt utilizate fara a lua în considerare daca este utilizat sau nu un suport de decizie si se realizeaza estimarea unor parametri ai populatiei de date. Prin revizuirea atât a datelor de tip numeric sau prin reprezentarile grafice ale acestora, se va studia structura lor si totodata identificarea abordarilor si limitarilor datelor studiate.

Sunt doua elemente principale ale revizuirii datelor preliminare: (1) marimi statistice de baza si (2) reprezentari grafice ale datelor. Marimile statistice sunt functii ale datelor care descriu numeric setul de date. Ele sunt utilizate pentru a furniza o imagine privind setul de date si sunt utile pentru realizarea de inferente privind populatia de date.

Reprezentarile grafice sunt utilizate în identificarea modelelor privind datele si relatiile dintre acestea, în confirmarea sau infirmarea

ipotezelor statistice si identificarea potentialelor probleme.

Faza de revizuire a datelor preliminare are scopul de a familiariza analistul cu mediul de date pe care îl analizeaza. Acest proces are rolul de a identifica anomaliile ce indica elementele de incertitudine care pot influenta analiza datelor. Rezultatele acestei faze sunt utilizate pentru selectarea unei proceduri pentru testarea ipotezei statistice în vederea sustinerii deciziei.

Scopul acestei faze este de a genera elemente statistice în urma calculelor si reprezentari grafice pentru descrierea datelor. Informatiile sunt utilizate pentru studiul structurii datelor si identificarea tuturor modelelor de date si a relatiilor formate în cadrul populatiei de date.

Pasii urmati pe parcursul acestei faze sunt: *revizuirea rapoartelor de certificare a calitatii*, *calcularea marimilor de baza statistice*, *reprezentarea grafica a datelor*.

Primul lucru care se urmareste în revizuirea preliminara a datelor este parcurgerea oricaror rapoarte cu relevanta ce privesc analiza de calitate si descriu colectia de date si procesele deja implementate care lucreaza cu aceasta colectie de date. La studiul acestor rapoarte, o atentie particulara trebuie acordata informatiilor ce vor fi utilizate pentru verificarea presupunerilor facute în procesul de studiul al obiectivelor calitati datelor. O atentie deosebita este data anomaliilor din cadrul datelor înregistrate, valorilor lipsa, deviatiilor de la procedurile de operare standard si utilizarii metodologiilor nestandardizate a colectiilor de date.

Calculul marimilor de baza efectueaza o caracterizare a setului de date utilizând metode clasice de calcul statistic, structura esantionului de date determinând tipul calculelor statistice efectuate în obtinerea marimilor. Lista marimilor statistice utile analistului în studiul efectuat include: numarul observatiilor; masurarea tendintei centrale – prin calculul mediei sau medianei; masurarea dispersiei cu ajutorul variatiei, deviatiei standard sau coeficientului de variatie; masurarea simetriei distributiei sau forma acestora etc. Aceste marimi sunt utilizate pentru descrierea si testarea ipotezelor statistice privind populatia din care provin datele.

Reprezentarea grafica a datelor are ca scop identificarea modelelor si tendintelor din cadrul esantionului de date prin utilizarea unor metode pure de calcul numeric. Graficele pot fi utilizate în identificarea modelelor si tendintelor, pentru confirmarea sau infirmarea ipotezelor statistice, pentru descoperirea unor noi procese, identificarea problemelor potentiale si pentru sugerarea masurilor de corectie. În plus, anumite reprezentari grafice pot fi utilizate pentru a înregistra si memora date într-o forma compacta sau pentru a o transmite pentru o alta utilizare.

Selectarea testului statistic

Acest pas furnizeaza informatiile necesare analistului pentru selectarea testului potrivit pentru ipoteza statistica – ce va fi folosit pentru concretizarea concluziilor privind studiul de analiza a calitatii datelor. Etapele parcurse sunt *selectarea testului pentru ipoteza statistica* si *identificarea presupunerilor sublinia-te de testul statistic*.

Precizarea unui anumit test în cadrul procesului de stabilirea a obiectivelor de studiu a calitatii datelor, în planul de analiza a calitatii sau de catre un program sau studiu particular, determina analistul sa utilizeze rezultatele revizuirii datelor preliminare în vederea determinarii daca testul statistic este potrivit pentru colectia de date supuse analizei. Daca acest test nu întruneste aceasta conditie, analistul trebuie sa documenteze aceasta nepotrivire în sensul argumentarii cauzelor sau caracteristicilor ce determina acest lucru si sa selecteze un test potrivit, posibil dupa o consultare prealabila cu factorul decizional uman.

Toate testele statistice induc presupuneri despre date, fiind necesara identificarea presupunerilor rezultate din aceste teste statistice. Testele parametrice presupun ca datele sa aiba o forma uniform distribuita, pe când testele non-parametrice nu fac aceasta presupunere. Oricum, ambele tipuri de teste verifica daca datele sunt independente statistic sau daca urmeaza sau nu o tendinta în repartizarea lor. În timpul examinarii datelor, analistul trebuie întotdeauna sa expuna sub forma unui

raport presupunerile evidentiate în cadrul testului ipotezei, cum ar fi distributia, dispersia s.a. O alta caracteristica importanta a testelor statistice este senzitivitatea (non-robustetea) la abaterile de la valorile rezultate în urma concluziilor. O procedura statistica este considerata robusta daca performanta ei nu este afectata într-o masura considerabila de valorile moderate ale deviatiilor (abaterilor) de la presupunerile rezultate. Analistul trebuie sa retina orice presupunere senzitiva relativa la micile deviatii care ar putea afecta validitatea rezultatelor testelor.

Verificarea previziunilor din testul statistic

În aceasta faza, analistul trebuie sa evalueze validitatea testului statistic ales în faza anterioara, prin examinarea presupunerilor rezultate în perspectiva noului mediu de date generat. Punctul forte al acestei sectiuni este determinarea când datele sunt verificate de presupunerile rezultate în urma testului statistic sau daca sunt necesare modificari în cadrul esantionului de date înaintea analizei statistice.

Aceasta determinare poate fi realizata cantitativ utilizând analiza statistica a datelor pentru a admite sau respinge presupunerile ce decurg din orice test statistic. Aproape întotdeauna tehnicile cantitative trebuie sustinute de criterii calitative bazate pe teorii stiintifice. Reprezentarile grafice ale datelor vor furniza informatii calitative importante despre presupunerile rezultate. Documentatia la acest pas este important, în special când judecatile subiective joaca un rol important în acceptarea rezultatelor analizei. Daca datele suporta toate criteriile testului statistic, atunci analiza calitatii datelor poate continua cu pasul urmator si anume faza de determinare a concluziilor pentru datele studiate. De câte ori una sau mai multe presupuneri ridica întrebari, acest lucru determina o reevaluare a unuia din pasii anteriori.

Acest tip de iteratie în analiza calitatii datelor este o verificare importanta pentru validitatea si aplicabilitatea practica a rezultatelor.

Ca actiuni ale acestei faze se disting *determinarea abordarii pentru verificarea presupunerilor, efectuarea testelor pentru presupunerile rezultate din testul statistic, determinarea corectiilor ce trebuie efectuate*.

În majoritatea cazurilor, presupunerile privind forma distributiei datelor, independenta acestora si dispersia, pot fi verificate formal utilizând testele statistice descrise în sectiunile de documentare tehnica sau, în unele cazuri, informatii din faza de revizuire preliminara a datelor, care pot furniza o evidenta suficient de fundamentata pentru sustinerea presupunerilor. Ca parte a acestei activitati, analistul trebuie sa identifice metodele ce verifica daca tipul si volumul de date necesar pentru efectuarea testelor necesare sunt disponibile. Datele de iesire ale acestei activitati trebuie sa includa o lista a testelor specifice care vor fi utilizate pentru verificarea presupunerilor statistice. Pentru fiecare test statistic este necesar ca pentru procedurile de investigare sa fie precizat *nivelul de semnificatie*. Pentru ipoteza de referinta zero (ipoteza nula) pentru verificarea testului considerat, nivelul de semnificatie este probabilitatea ca aceasta ipoteza sa fie respinsa, sa fie nula. Alegerea nivelului de semnificatie depinde de experienta investigatorului. În cazul în care sunt selectate mai multe teste statistice, este recomandat sa se aleaga o valoare numerica scazuta pentru nivelul de semnificatie, pentru prevenirea acumularilor potentialelor erori. Nivelul de semnificatie pentru un test statistic este prin definitie acelasi lucru cu eroarea de falsa respingere.

Efectuarea testelor pentru presupunerile (previziunile) testului statistic este faza în care investigatorul va evalua cât de rezonabile sunt aceste presupuneri în relatie cu structurile componentelor, facând în acest sens o raportare.

Presupunerile sau previziunile ce trebuie investigate includ urmatoarele:

1. *Se poate asuma ipoteza ca erorile (deviatiile fata de model) sunt normal distribuite?*
2. *Se poate verifica daca erorile sunt necorelate?*
3. *Se poate presupune ca rezonabila ideea ca erorile sunt în regim aditional si au o variabilitate constanta?*

Câteodata, presupunerile rezultate în urma testului primar statistic pot sa nu satisfaca cer-

rintele si câteva tipuri de corectii sunt necesare înainte de a lansa procedurile de evaluare calitativa. În anumite cazuri, o conversie a datelor va corecta problema presupunerilor distribuite. În alte cazuri datele utilizate pentru verificarea unor presupuneri (previziuni) cheie este posibil sa nu fie disponibile, iar informatiile existente sa nu suporte o justificare teoretica a validitatii acestor previziuni. În aceasta situatie, este necesara colectarea aditionala pentru verificarea presupunerilor rezultate în urma testului statistic. Daca aceste presupuneri rezultate din testarea ipotezei nu sunt satisfacute si conversiile de date sau alte modificari nu se dovedesc fezabile, atunci este necesar sa se ia în consideratie un alt test statistic.

Determinarea concluziilor

În finalizarea procesului de analiza a calitatii datelor, analistul va efectua testul ipotezei statistice si determinarea concluziilor care vizeaza obiectivele utilizatorilor datelor supuse studiului. Acest pas reprezinta punctul culminant al fazelor de planificare, implementare si evaluare a esantionului de date.

Procedurile acestei faze sunt *efectuarea testului pentru ipoteza statistica*, *trasarea concluziilor studiului* si *evaluarea performantelor studiului efectuat.*

Scopul fazei de efectuare a testului pentru ipoteza statistica este coordonarea acestuia. Calculele ce se efectueaza trebuie documentate foarte clar si usor verificabile. Documentatia rezultatelor testului trebuie sa fie usor de înteles astfel încât rezultatele sa poata fie comunicate eficient celor care o utilizeaza în procesul de luare a deciziilor. În cazul în care calculele sunt efectuate prin intermediul unui software dedicat, se asigura ca procedurile sunt documentate adecvat si algoritmii utilizati au codul proiectat si dezvoltat specific pentru proiectul în care sunt utilizate.

Pasul urmator consta în translatarea rezultatelor testului statistic, astfel încât utilizatorul datelor obtinute poate sa deduca o concluzie din aceste date.

Rezultatele testului statistic pot fi unul din urmatoarele:

1. respingerea ipotezei nule – caz în care analistul este preocupat de o posibila eroare decizionala de tip falsa respingere;
2. esuarea în respingerea ipotezei nule – analistul îsi concentreaza atentia asupra unei posibilitati de aparitie a erorii de falsa acceptare.

În primul caz, datele furnizeaza necesitatea de respingere a ipotezei nule, astfel decizia poate fi luata în siguranta si fara proceduri de analiza suplimentara. Acest lucru se datoreaza faptului ca testul statistic controleaza rata erorii de falsa respingere în limite tolerabile, furnizând ideea ca presupunerile testului au fost verificate corect.

În cel de-al doilea caz, datele nu furnizeaza suficiente dovezi pentru respingerea ipotezei nule, iar datele trebuie analizate prin prisma criteriului de a situa limitele de toleranta pentru eroarea de falsa acceptare în valori acceptabile.

Pentru evaluarea performantelor studiului efectuat, analistul va efectua o analiza statistica pentru estimarea capacitatii testului statistic de a efectua operatii cu valori situate dincolo de limitele parametrilor standard. Performanta testului statistic este data de probabilitatea de respingere a ipotezei nule când aceasta este falsa. Prin aceste operatiuni, analistul va determina buna functionare a testului statistic si va compara performantele acestuia cu ale altor tipuri de teste.

Bibliografie

- Guidance for Data Quality Assessment, Practical Methods for Data Analysis, Quality Staff Office of Environmental Information, U.S. Environmental Protection Agency, 2000, NewYork- Berthouex,
- P.M., and L.C. Brown, 1994. Statistics for Environmental Engineers. Lewis, Boca Raton, FL.
- U.S. Environmental Protection Agency, 1994a. Guidance for the Data Quality Objectives Process. EPA/600/R-96/055. Office of Research and Development.
- Cleveland, W.S., 1993. Visualizing Data. Hobart Press, Summit, NJ.