Direcția Instrumente Inovatoare în Statistică
Institutul Național de Statistică
2017

EXTINDEREA UTILIZĂRII SURSELOR ADMINISTRATIVE ÎN STATISTICA OFICIALĂ DIN ROMÂNIA

DECLARAȚIA DE TVA

Direcția Instrumente Inovatoare în Statistică
Institutul Național de Statistică
2017

Cuprins

Direcția Instrumente Inovatoare în Statistică
Institutul Național de Statistică
2017

I. Introducere

Utilizarea surselor administrative în producţia de date statistice a devenit o axă prioritară în sistemul statistic din România, derivată din necesitatea exploatării informaţiilor deja existente în mediul public şi având ca scop reducerea presiunii şi a sarcinii de răspuns a unitătilor incluse în cercetările statistice.

Declaraţia de TVA este una dintre cele mai importante surse administrative utilizate în producerea de statistici oficiale. Raportul de faţă descrie metode de dezvoltare, detectare şi corectare a erorilor, de validare şi editare a variabilei cifra de afaceri aşa cum o regăsim în sursa administrativă decont de TVA. Metodele si tehnicile statistice utilizate provin din literatura de specialitate conform cercetărilor şi celor mai bune practici internaţionale în acest domeniu.

II. Indicatori de calitate

Indicatorii de calitate sunt calculați înainte de a utiliza sursa administrativă ca un înlocuitor sau ca o sursa de date suplimentară pentru datele obținute din sondaj. Indicatorii de calitate sunt utilizați pentru a evalua fezabilitatea de a trece la folosirea datelor administrative in producția statistică și impactul acestora asupra calității producției statistice, și au fost dezvoltați astfel încât un indicator cu o valoare mică indică o calitate ridicată a datelor, iar un indicator cu o valoare mare indică o calitate slabă a datelor.

Lista indicatorilor de calitate ai surselor administrative

Indicator	Descriere
Procentul unităților cu valori lipsă	Acest indicator trebuie calculat pentru fiecare variabilă cheie. $$\frac{Nr\ unități\ din\ sursa\ administrativă\ cu\ valori\ lipsă\ pentru\ variabila\ X}{Nr\ unităti\ totale\ din\ sursa\ administrativă}\ x\ 100\%$$
Rata de clasificare eronată	Acest indicator ofera informații despre procentul unităților din sursa administrativă care sunt incorect codificate. Activitatea de codificare înregistrată în Registrul Întreprinderii(BR) este considerată corectă. În cazul în care codificarea activității din sursa administrativă nu este folosită de INS, acest indicator nu este relevant. $$\frac{Nr\ unități\ din\ sursa\ administrativă\ cu\ CAEN\ diferit\ fata\ de\ BR}{Nr\ unități\ totale\ din\ sursa\ administrativă}\ x\ 100\%$$
Sub-acoperirea	Acest indicator oferă informații despre unitățile din populația de referință

	(Registrul Întreprinderii) ce ar trebui incluse în sursa administrativă, dar nu sunt incluse(pentru orice motiv). Acest indicator se calculează bazându-se pe numărul unităților relevante din **Registrul Întreprinderii(BR).** $$\frac{Nr\ unități\ care\ sunt\ prezente\ in\ BR\ dar\ nu\ sunt\ prezente\ în\ sursa\ administrativă}{Nr\ unităti\ prezente\ în\ BR}\ x\ 100\%$$
Supra-acoperirea	Acest indicator oferă informații despre unitățile care sunt incluse în sursa administrativă, dar care nu ar trebui incluse. Acest indicator poate fi calculat banându-ne pe numărul unităților relevante din **Registrul Întreprinderilor(BR)**. $$\frac{Nr\ unități\ care\ sunt\ prezente\ in\ sursa\ administrativă\ dar\ nu\ sunt\ prezente\ în\ BR}{Nr\ unităti\ prezente\ în\ BR}\ x\ 100\%$$
Mărimea reviziilor din diferite versiuni ale sursei administrative RAR – Revizii absolute relative	Acest indicator oferă informații despre impactul diferitelor versiuni de surse administrative asupra rezultatelor pentru o anumită perioadă de referință. Datele revizuite pe baza altor informații(survey data) nu sunt incluse in calculul indicatorului. Dacă este primită o singură versiune a datelor administrative, acest indicator nu este relevant. $$\frac{\sum_{t=1}^{T}\lvert X_{Lt}-X_{Pt}\rvert}{\sum_{t=1}^{T}\lvert X_{Pt}\rvert}\ x\ 100,\ X_{Pt} = primele\ date\ primite,$$ $$X_{Lt} = ultimile\ date\ primite$$

III. Metode pentru detectarea erorilor din declaratia decont de TVA

Schema procesului de detectare a erorilor din declaratia decont de TVA

Detectarea și corectarea **erorilor de unitate de măsură**;

Detectarea și marcarea **patternurilor de valori aberante**, astfel încat sa nu fie folosite în aplicarea metodelor de detectare a erorilor aleatoare;

Aplicarea metodelor de detectare a **erorilor aleatoare**.

Au fost identificate trei categorii de erori întâlnite cu regularitate în datele administrative:

a) Erori de unitate de măsură;
b) Patternuri de valori aberante;
c) Erori aleatoare.

1. Metode pentru detectarea erorilor de unitate de măsură

Unele întreprinderi ar putea raporta, din greşeală, cifrele de afaceri în RON în loc de mii RON sau invers. Presupunând că raportarea eronată a întreprinderii nu a fost facută pe două perioade consecutive, astfel de erori pot de multe ori fi detectate prin compararea cu declaraţiile anterioare de TVA pentru aceeaşi intreprindere, utilizând o regulă după formula de mai jos:

$$A < \frac{Cifra\ de\ afaceri\ curentă}{Cifra\ de\ afaceri\ anterioară} < B$$

Valorile parametrilor A si B se pot determina pe baza seriei de date logitudinale.

2. Explorarea datelor pentru detectarea unor patternuri de valorilor aberante

Direcția Instrumente Inovatoare în Statistică
Institutul Național de Statistică
2017

Cu ajutorul datelor longitudinale este posibilă identificarea valorilor aberante declarate în decontul de TVA . Modele pentru identificarea valorilor aberante în datele trimestriale din decontul de TVA:

a) Cifra de afaceri zero în primele trei trimestre si cifra de afaceri pozitivă în trimestrul patru;
b) Cifra de afaceri zero în primul trimestru şi valori pozitive ale cifrei de afaceri în celelalte trei trimestre;
c) Aceeaşi valoare a cifrei de afaceri în toate cele patru trimestre;
d) Aceeaşi valoare a cifrei de afaceri pe trei trimestre si o altă valoare (pozitivă) în al patrulea trimestru;
e) Valoare negativă a cifrei de afaceri în oricare dintre trimestre.

3. Metode pentru detectarea erorilor aleatoare

Metodele de detectare a erorilor aleatoare ale cifrei de afaceri declarate prin decontul de TVA adopta oricare dintre cele două abordări de mai jos:

a) Compararea valorii cifrei de afaceri declarate în perioada curentă faţă de valoarea cifrei de afaceri declarată în perioada anterioară pentru aceeaşi unitate.
b) Compararea valorii cifrei de afaceri pentru aceeaşi perioadă între două unităţi cu aspecte similare. De exemplu, se poate calcula media sau mediana cifrei de afaceri raportate pentru întreprinderi similare. Astfel, obtinem o singură valoare reprezentativă si comparativă pentru un grup de întreprinderi grupate dupa clase de mărime şi clase CAEN.

A. Metoda *Intervalul interquartilic*

Această metodă identifica valori ale cifrei de afaceri neobişnuit de mari sau neobişnuit de mici prin localizarea în distribuţia datelor, a valorilor extreme conform grupării după clasa de mărime, clasa CAEN și periodicitatea de depunere a cifrei de afaceri. Metoda identifică acele intreprinderi pentru care cifrele de afaceri au valori mai mari decât un multiplu stabilit de lungimea intervalului de la a treia quartila sau întreprinderi cu valori mai mici decât acelaşi multiplu stabilit de lungimea intervalului de la prima quartilă. Acestea sunt identificate dupa algoritmul:

dacă *valoarea din* *TVA > Q3 + [C × (Q3 - Mediana)]*

sau

valoarea din *TVA < Q1 - [C × (Mediana - Q1)]*

atunci

valorile sunt considerate extreme

Quartilele şi mediana sunt derivate din datele de TVA în funcție de clasa de mărime, clasa CAEN și periodicitate. Parametrul *C* este derivat din analiza datelor anterioare pentru a stabili pragul care identifică cu succes valorile extreme.

B. Metoda *Raport pe perioadă*

Metoda constă în calcularea periodica a raportului pentru fiecare întreprindere în baza contribuţiei cifrei de afaceri a acesteia în clasa în care este cuprinsă. Clasa va fi definită ca o joncţiune între clasa de marime, clasa CAEN şi periodicitate. Rapoartele pe perioada mai sunt numite si rapoarte de testare. Cifrele de afaceri pentru întreprinderile care au o valoare a TestRatio mai mare decât pragul predefinit sunt considerate aberante.

$$Score = \frac{Cifra\ de\ afaceri\ din\ decontul\ de\ TVA}{Mediana\ pentru\ cifra\ de\ afaceri\ pentru\ clasa\ respectivă}$$

$$TestRatio = \begin{cases} Score_t / Score_{t-1}, & dacă\ Score_t > Score_{t-1} \\ Score_{t-1} / Score_t, & altfel \end{cases}$$

$Score_t$ *este valoarea scorului în perioada t*

$Score_{t-1}$ *este valoarea scorului în perioada t-1.*

C. Metoda *Comparaţie cu raportări anterioare pentru cifra de afaceri*

Metoda compară cifra de afaceri din decontul de TVA cu valori declarate anterior cu scopul determinării valorilor posibil aberante.

Dacă
Cifra de afaceri > 10000 RON
şi
CA > 10 × media cifrei de afaceri pentru intreprindere în ultimele 12 de luni
atunci se va trata ca aberantă.

Direcția Instrumente Inovatoare în Statistică
Institutul Național de Statistică
2017

D. Metoda *Hidiroglou-Berthelot*

Metoda se bazează pe raportul între cifra de afaceri din decontul de TVA pentru o întreprindere în perioada curentă si aceeaşi variabilă din perioada anterioară. Metoda adaugă unele îmbunătăţiri pentru a permite ca dimensiunea întreprinderii după clasa de marime să fie luată în considerare. Metoda este definită după cum urmează:

$$r = \frac{valoarea\ cifrei\ de\ afaceri\ declarată\ în\ perioada\ curentă}{valoarea\ cifrei\ de\ afaceri\ declarată\ în\ perioada\ anterioară}$$

Se calculează mediana raporturilor **r**:

Daca *r < mediana*

atunci se calculeaza

$$t = \frac{r - mediana}{r}$$

altfel se calculeaza

$$t = \frac{r - mediana}{mediana}$$

*E = t * max (cifra de afaceri curenta din TVA, cifra de afaceri anterioara din TVA) * V*

Parametrul *V* poate lua orice valoare intre 0 şi 1, unde 1 conferă o importanţă mai mare întreprinderilor cu valori mai mari ale cifrei de afaceri. Se calculează prima quartilă (Q1), a treia quartilă (Q3) și mediana (Q2) valorilor E. Apoi se calculează:

$$d_{Q1} = max[(Q2 - Q1), |A\ x\ Q2|]$$

$$d_{Q3} = max[(Q3 - Q2), |A\ x\ Q2|]$$

În unele condiţii parametrice de lipsa a distribuţiei normale este posibil ca datele din grupul medianei să identifice eronat valori ale cifrei de afaceri ca fiind aberante. În acest caz este indicată utilizarea unui multiplu A al valorilor medianei E. Parametrul A, în mod frecvent pentru acest scop, primeşte valoarea 0,05. Cifrele de afaceri aberante sunt apoi identificate după cum urmează:

Dacă

$$E < Q2 - C \times d_{Q1}$$

Sau

$$E > Q2 + C \times d_{Q3}$$

cifra de afaceri se va trata ca fiind aberantă.

IV. Metode pentru corectarea valorilor aberante existente în datele de TVA

Următoarele metode sunt aplicate pentru fiecare tip de eroare identificat și sunt testate pentru a stabili care are rezultatele cele mai performante pentru datele din decontul de TVA. Există mai multe opțiuni prin care putem trata erorile sau valorile aberante ce au fost identificate în datele de TVA.

a) **Ignoră potențialele erori**. Aceasta este probabil cea mai puțin folositoare opțiune. Având în vedere utilizarea datelor, se poate considera ca anumite valori nu vor avea un impact semnificativ asupra rezultatelor statistice, indiferent dacă sunt eronate sau nu.

b) **Elimină orice valoare aberantă din setul de date**. Ideea este de a utiliza doar valorile care nu sunt aberante pentru rezultatele statistice. Aceasta poate fi o soluție bună atunci când nu există o necesitate de a utiliza datele din decontul de TVA pentru fiecare unitate în studiul populației. De exemplu, când sunt disponibile alte surse de informație despre cifra de afaceri.

c) **Marchează valoarea ca fiind aberantă, dar nu o modifica**. Aceasta permite oricărui utilizator al datelor să-și aleagă singur modalitatea de a trata valorile aberante. Aceasta oferă o mare flexibilitate pentru diferite utilizări statistice ale datelor din decontul de TVA. Pentru utilizatori cu o insuficientă înțelegere a consecințelor ce vor apărea în urma alegerilor făcute, aceasta opțiune poate conduce la rezultate incorecte.

d) **Modifică valorile aberante manual.** Orice valoare considerată aberantă este inspectată de statistician și modificată pentru a fi în linie cu asteptările. Aceasta poate conduce la o acuratețe mare a datelor, dar metodele folosite pentru modificarea valorilor pot fi inconsistente și subiective, având în vedere percepțiile statisticianului. Aceasta este o opțiune ce consumă multe resurse când există un număr semnificativ de valori aberante ce trebuie modificate, în special pentru seturile mari de date. O variație a acestei opțiuni ar fi doar revizuirea manuală a unui subset de valori aberante, considerat important (datorită marimii setului de date sau impactului așteptat al rezultatelor statistice).

e) **Modifică valorile aberante în mod automat.** Metodele de imputare sunt utilizate să modifice în mod automat valorile aberante, astfel încât să fie în aceeași linie cu așteptările. Din moment ce valorile sunt modificate automat, este important să ai o metodă bine specificată pentru a evita imputarea unui număr mare de erori in setul de date.

f) **Contactează întreprinderile care au cifre de afaceri aberante pentru a confirma valorile depuse**. Contactarea întreprinderilor poate consuma resurse intensiv și poate plasa o sarcină suplimentară asupra întreprinderilor. Aceasta poate fi redusă doar prin contactarea celor mai importante întreprinderi (sau cele cu cele mai multe valori aberante depuse). Aceasta poate să fie opțiunea cu cea mai mare acuratețe pentru întreprinderile contactate, cu presupunerea că aceste întreprinderi sunt dispuse sa ofere cifrele corecte în timpul contactării.

Când alegem o opțiune pentru a trata erorile din datele de TVA este important să luăm în considerare utilizarea statistică a datelor, resursele disponibile pentru curățarea datelor și impactul unor posibile sarcini asupra întreprinderilor. Dacă modificăm valori din setul de date, este recomandat să păstrăm o copie a setului de date original. Aceasta copie ne ajută pentru o analiză viitoare a metodei folosite în modificarea valorilor aberante și ne asigură că setul de date original poate fi re-creat oricând este necesar. În multe cazuri, cea mai sensibilă și mai rentabilă opțiune va fi automatizarea de a schimba valorile aberante utilizând metode de imputare. Metodele de imputare vor fi discutate și testate în cele ce urmează.

1. Metode pentru corectarea erorilor sistematice din datele de TVA.

Corectarea erorilor de unitate de măsură se realizează înmulțind valorile aberante cu 1000 (sau împărțind valorile aberante prin 1000, dacă eroarea este în sensul opus). Sunt necesare surse independente pentru a verifica dacă am identificat erorile de unitate de măsură în mod corect. Dacă valoarea cifrei de afaceri din Declarația de TVA este presupusă aberantă, se poate compara cu valoarea cifrei de afaceri din Registrul Întreprinderii.

2. Metode pentru corectarea patternurilor de date aberante

Patternuri de date aberante observate în datele din decontul de TVA:

- Unități care au depus cifra de afaceri zero în primele trei trimestre si cifra de afaceri pozitivă în trimestrul patru;
- Unități care au depus cifra de afaceri zero în primul trimestru si valori pozitive ale cifrei de afaceri în celelalte trei trimestre;
- Unități care au depus aceeași valoare a cifrei de afaceri in toate cele patru trimestre;
- Unități care au depus aceeași valoare a cifrei de afaceri pe trei trimestre si o alta valoare(pozitivă) în trimestrul patru;
- Valoarea negativă a cifrei de afaceri în oricare dintre trimestre.

Pentru cele cinci patternuri de date aberante se vor folosi metode de imputare diferite, astfel vor fi grupate în următoarele 3 categorii.

Categoria 1

- Unități care au depus cifra de afaceri zero în primele trei trimestre si cifra de afaceri pozitivă în trimestrul patru.
- Unități care au depus aceeași valoare a cifrei de afaceri in toate cele patru trimestre.
- Unități care au depus aceeași valoare a cifrei de afaceri pe trei trimestre si o alta valoare(pozitivă) în al patrulea trimestru.

Categoria 2

- Patternuri de date aberante care sugerează că întreprinderile nu au reușit să raporteze cifra de afaceri intr-un trimestru.

Categoria 3

- Valoarea negativă a cifrei de afaceri în oricare dintre trimestre.

<u>Metode de imputare pentru Categoria 1</u>

Pentru primul pattern, este posibil ca singura valoare pozitivă a cifrei de afaceri să reprezinte o depunere anuală. Pentru cel de-al doilea pattern, pare că cifra de afaceri anuală a fost împărțită în cele patru trimestre. Pentru cel de-al treilea pattern, este posibil ca pentru primele trei trimestre sa fie depusă o cifră de afaceri estimată, iar trimestrul patru să fie folosit ca un element de balansare pentru a da cifra de afaceri corectă pentru tot anul.

În fiecare caz, dacă asumțiile noastre sunt corecte, putem deduce o cifra de afaceri anuala sigură (adunând valorile din cele patru trimestre). Oricum, nu avem nici o informație despre cum ar trebui împărțită cifra de afaceri anuală în cifre de afaceri trimestriale. Pentru a crea date ce vor putea fi folosite in scopuri statistice, este necesar să imputăm cifrele de afaceri depuse trimestrial. Deoarece avem cifrele de afaceri anuale, metoda evidentă este imputarea(determinarea) proporțiilor de cifre de afaceri pentru fiecare trimestru și multiplicarea acestora cu cifra de afaceri anuală.

Următoarele trei metode prezintă imputarea proporțiilor de cifre de afaceri depuse trimestrial, având la baza fie proporțiile cifrelor de afaceri ale unor întreprinderi din aceeași clasă, fie proporțiile din anul anterior pentru aceeași întreprindere. Metodele sunt următoarele:

- Media proporțiilor din clasa omogenă(în același an).
- Mediana proporțiilor din clasa omogenă(în același an).
- Proporțiile pentru aceeași întreprindere din anul anterior.

Media și mediana proporțiilor sunt calculate folosind toate întreprinderile dintr-o clasă omogenă cu valorile cifrelor de afaceri corecte în fiecare trimestru. Se observă că sumând media si mediana proporțiilor nu vom obține întotdeauna 1. Astfel, va fi necesar să rescalăm datele imputate astfel încât suma lor să fie egală cu valoarea anuală a cifrei de afaceri. Rescalarea se va face prin înmulțirea valorilor imputate cu valorile anuale și impărțite la suma valorilor trimestriale imputate.

<u>Metode de imputare pentru Categoria 2</u>

Aceste patternuri sugerează că întreprinderea nu a depus cifra de afaceri intr-un trimestru. Această categorie include numai un singur pattern suspicios: unități care au depus cifra de afaceri zero în primul trimestru si valori pozitive ale cifrei de afaceri în celelalte trei trimestre.

În cele mai multe cazuri, se poate presupune că acest pattern rezultă din faptul că întreprinderea nu depune cifra de afaceri corectă in trimestrul în care a depus cifra de afaceri zero. Menționăm că pot exista excepții unde acest pattern este corect.

Din moment ce acest pattern implică o valoare ce nu poate fi folosită, metodele de imputare descrise mai sus se vor putea folosi și în acest caz.

Metode de imputare pentru Categoria 3

Această categorie are în vedere valoarea negativă a cifrei de afaceri în oricare dintre trimestre. Modalitatea în care tratăm valorile negative ale cifrelor de afaceri poate depinde de modul în care acestea au apărut. Nu este recomandat să înlocuim semnul negativ cu cel pozitiv. Soluția recomandată este să înlocuim cifrele de afaceri negative cu zero sau să tratăm valoarea ca o singură valoare aberantă și să utilizăm una dintre metodele de imputare descrise mai sus pentru a o înlocui.

3. Metode pentru corectarea erorilor aleatoare

Corectarea erorilor aleatoare este o problemă foarte cunoscută în cercetările statistice ale întreprinderilor, astfel a fost dezvoltată o gamă de metode de imputare, metode ce pot fi folosite și pentru corectarea valorilor aberante din datele de TVA. Totuși nu există nici o îndrumare despre ce metodă de imputare funcționează mai bine cu sursele administrative. În continuare vor fi descrise 10 metode de imputare pentru datele de TVA aberante. Acestea țin cont de metodele utilizate în mod curent pentru a edita datele de TVA precum si acoperirea principalelor metode de imputare folosite în cercetările statistice ale întreprinderilor.

Metoda 1 – Determinarea mediei

Prin această metodă trebuie să înlocuim valorile aberante prezente în datele trimestriale din decontul de TVA cu media valorilor corecte prezente în datele trimestriale din decontul de TVA, calculată pentru toate întreprinderile. Pentru a îmbunătăți acuratețea metodei, este recomandat să împarțim datele din decontul de TVA în clase omogene și să calculăm media pentru fiecare clasă. Valorile aberante ale întreprinderilor vor fi înlocuite cu media valorilor corecte a clasei omogene din care fac parte. Variabilele CAEN si clasa de mărime sunt folosite pentru a crea clase omogene.

Metoda 2 – Determinarea mediei setului de date ramas dupa eliminarea extremelor care contin outliers

Deoarece mediile sunt influențate de outliers este recomandat să eliminăm valorile extreme înainte de a calcula media. Cele mai mari zece procente din valorile datelor și cele mai mici zece procente din valorile datelor sunt eliminate din fiecare clasă omogenă înainte de a calcula media.

Direcția Instrumente Inovatoare în Statistică
Institutul Național de Statistică
2017

Metoda 3 – Determinarea medianei

Prin această metodă trebuie să înlocuim valorile aberante prezente în datele trimestriale din decontul de TVA cu mediana valorilor corecte prezente în datele trimestriale din decontul de TVA, calculată pentru toate întreprinderile. La fel ca la metoda 1 trebuie să calculăm mediana pentru fiecare clasă omogenă. Mediana este mai robustă decât media. Nu este nevoie să eliminăm valorile extreme.

Metoda 4 – Determinarea raportului: raport al mediilor folosind perioada anterioară

Metoda presupune folosirea unei valori anterioare a cifrei de afaceri pentru întreprindere și inmulțirea ei cu un raport bazat pe creșterea dintre perioada anterioară și perioada curentă pentru aceeași întreprindere.

$$\text{Valoarea determinată} = \text{Valoarea anterioară} \ * \ \text{Raport}$$

$$\text{Raport} = \frac{\sum_{i \in \text{clasă omogenă}} \text{valoarea din perioada curentă pentru întreprinderea } i}{\sum_{i \in \text{clasă omogenă}} \text{valoarea din perioada anterioară pentru întreprinderea } i}$$

Întreprinderile ale căror valori de date curente si anterioare sunt incluse în calculul mediilor sunt toate acele întreprinderi care au valori corecte pentru ambele perioade și care aparțin aceleiaşi clase omogene ca și întreprinderea pentru care se aplică metoda. Observăm că exact aceleași întreprinderi contribuie la cele două medii. De aceea raportul mediilor se reduce la un raport al sumelor.

Metoda 5 – Determinarea raportului: raport al mediilor folosind aceeași perioadă în anul anterior

Metoda presupune determinarea raportului mediilor cu creşterea calculată din aceeași perioadă în anul anterior.

Metoda 6 – Determinarea raportului: media rapoartelor folosind perioada anterioară

Modalitatea de calcul este următoarea:

$$\text{Valoarea determinată} = R \times \text{valoarea anterioară pentru întreprinderea cu date aberante}$$

$$\text{Unde } R = \frac{1}{\text{numărul întreprinderilor din clasa omogenă}} \times \sum_{i \in \text{clasa}} \frac{\text{valoarea din perioada curentă pentru întreprinderea } i}{\text{valoarea din perioada anterioară pentru întreprinderea } i}$$

Observăm că întreprinderile ale caror rapoarte sunt însumate sunt toate acele întreprinderi care au valori corecte pentru ambele perioade și care aparţin aceleiaşi clase omogene ca și întreprinderea pentru care se aplică metoda.

Metoda 7 – Determinarea raportului: media rapoartelor dupa eliminarea extremelor folosind perioada anetrioară

Media raportului poate fi excesiv influențată de întreprinderile cu o valoare extrem de mare a raportului sau cu o valoare extrem de mică a raportului. Datorită acestui fapt, rapoartele extreme sunt deseori eliminate înainte de a calcula media. Astfel vor fi eliminate cele mai mari și cele mai mici zece procente din rapoarte în fiecare clasă omogenă, înainte de a calcula mediile.

Metoda 8 – Determinarea raportului: nedia rapoartelor folosind aceeași perioadă în anul anterior

Pentru completitudine, considerăm media rapoartelor determinate folosind aceeași perioadă în anul anterior, în schimbul perioadei anterioare. Această metodă presupune determinarea raportului folosind media rapoartelor având la bază anul anterior.

Metoda 9 – Determinarea raportului: media rapoartelor dupa eliminarea extremelor folosind aceeași perioadă în anul anterior

Cele mai mari și cele mai mici zece procente din rapoarte în fiecare clasă omogenă sunt eliminate înainte de a calcula mediile.

Metoda 10 – Determinarea donatorului

Această metodă presupune folosirea datelor din decontul de TVA ale unei întreprinderi donatoare, în schimbul valorilor aberante. Donatorul este întreprinderea din aceeași clasa omogena, care are date corecte și care este cea mai asemănătoare cu întreprinderea ce prezintă valori aberante. Întreprinderile asemănătoare sunt identificate prin calcularea distanțelor dintre intreprinderile suspicioase și toti potențialii donatori. Distanțele sunt calculate separat pentru variabile auxiliare specifice. Cifra de afaceri si numarul de salariați din **Registrul Întreprinderii** sunt utilizate ca variabile auxiliare specifice. Distanțele pentru fiecare variabilă care se potrivește sunt agregate împreună și donatorul potențial cu cea mai mică distanță combinată este ales ca donator.

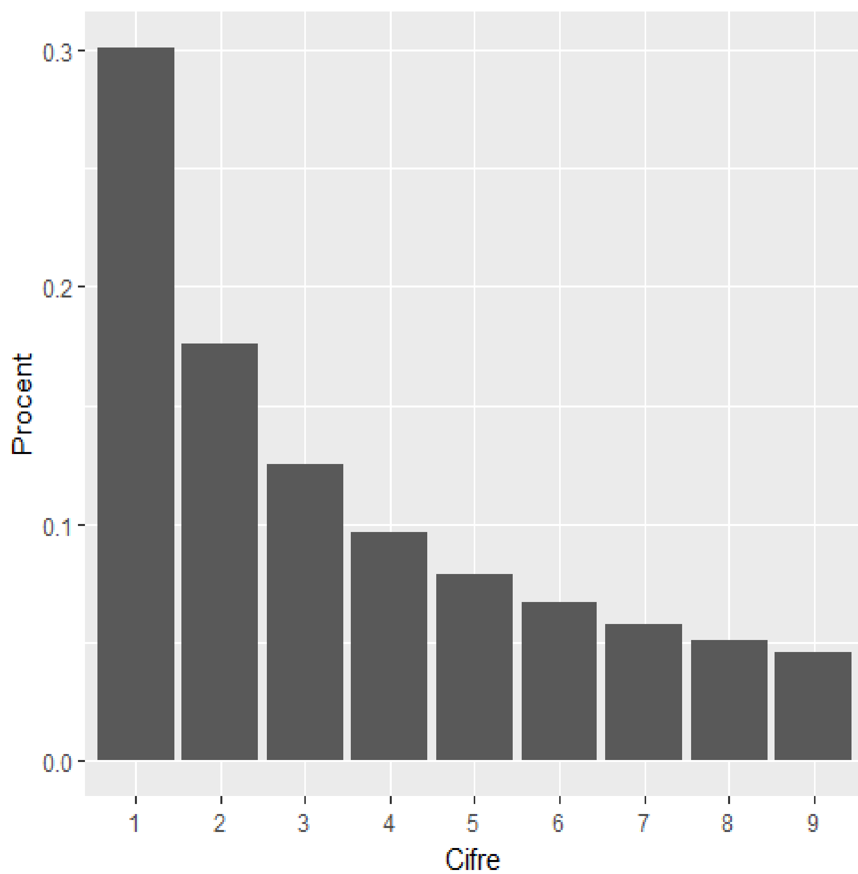
V. Data Quality Mining

Tehnicile de data mining ajută la descoperirea patternurilor de valorilor aberante în seturile mari de date. Legea lui Benford este o tehnică simplă și un exemplu bun pentru descoperirea unui pattern contraintuitiv. Legea se referă la primele cifre ale unei colecții de numere, ca de exemplu, prețurile pieței bursiere sau numărul de locuitori ai orașelor. Desi nu este intuitiv, 30% din date au prima cifra 1, iar 5% din date au prima cifra 9. Distribuția primelor cifre poate fi calculată folosind următoarea formulă:

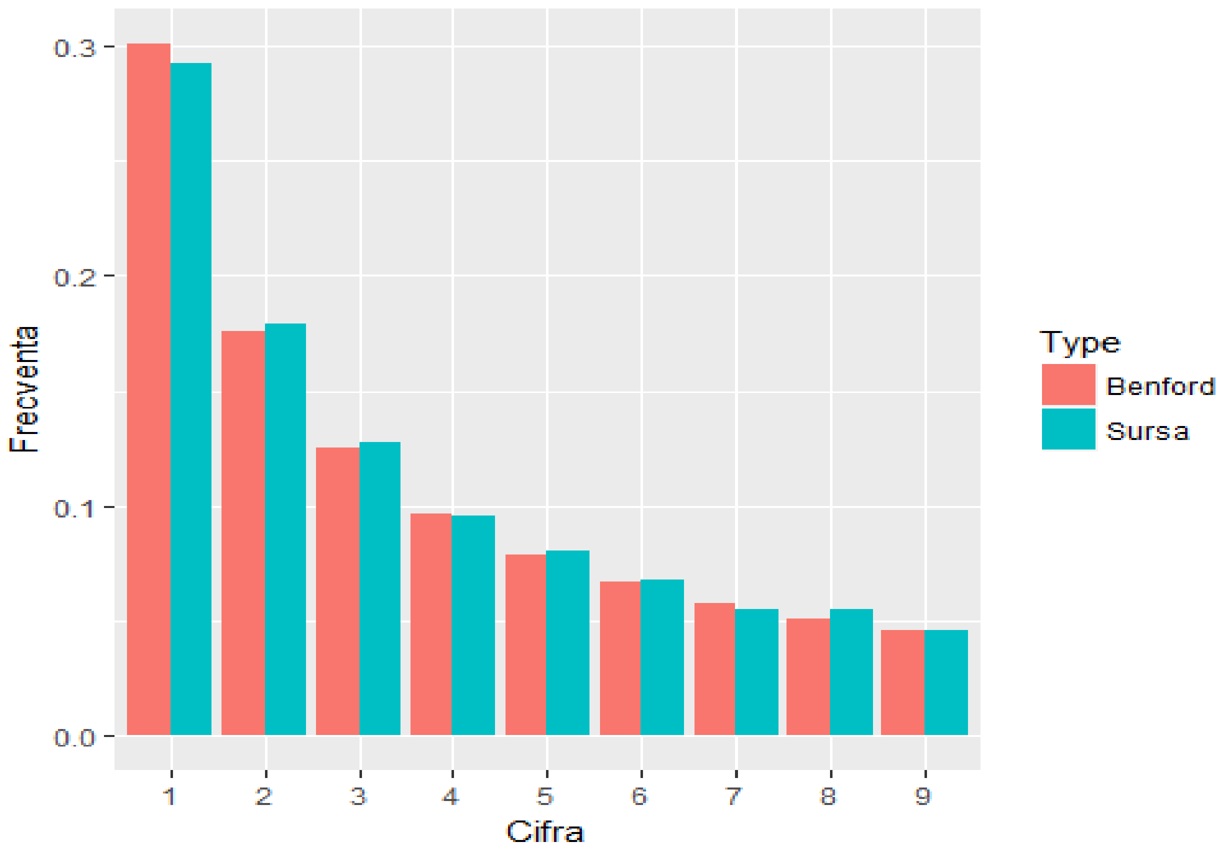
$$P(d) = \log 10 \left[(1 + d) / d \right], d = 1:9$$

Reprezentarea distribuției primelor cifre sub formă de tablel si sub forma grafică:

Cifre	Legea lui Benford
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%



Următorul grafic arată că cifrele de afaceri din decontul de TVA pentru întreprinderile din clasa de mărime cu 250 salariați și peste, respectă legea lui Benford.

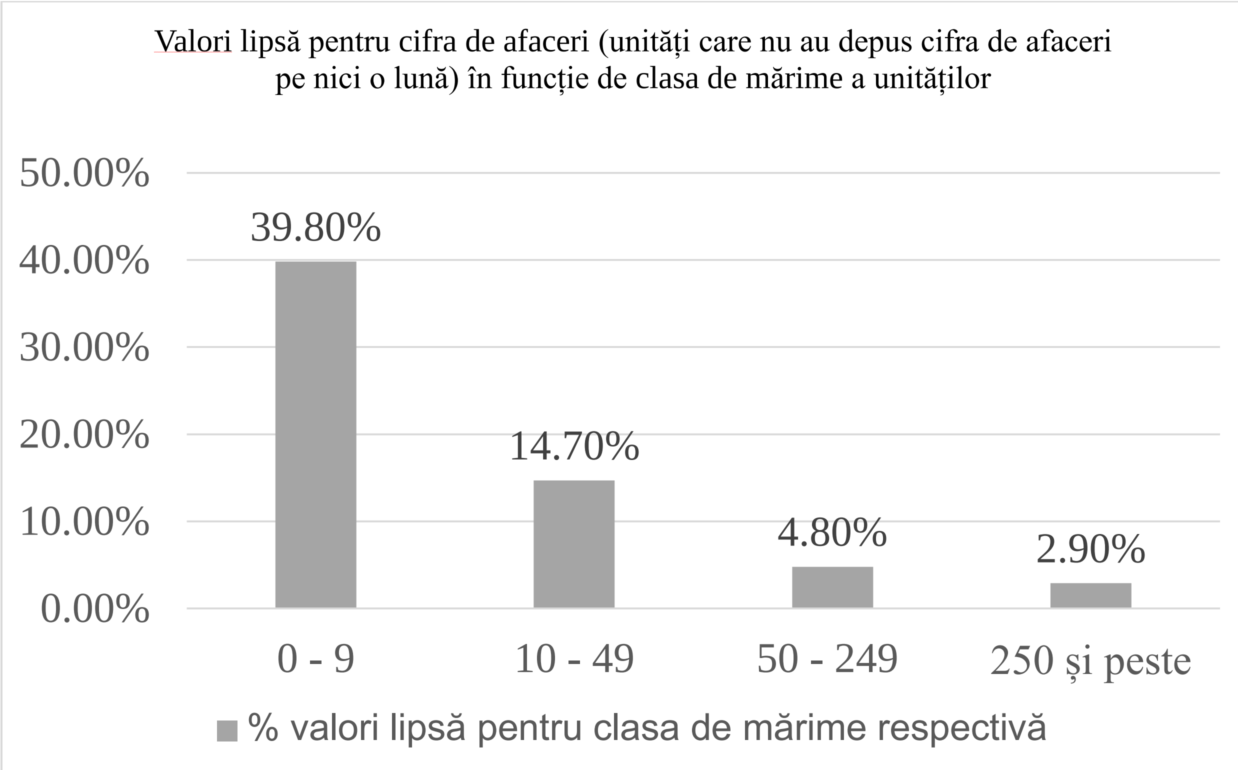


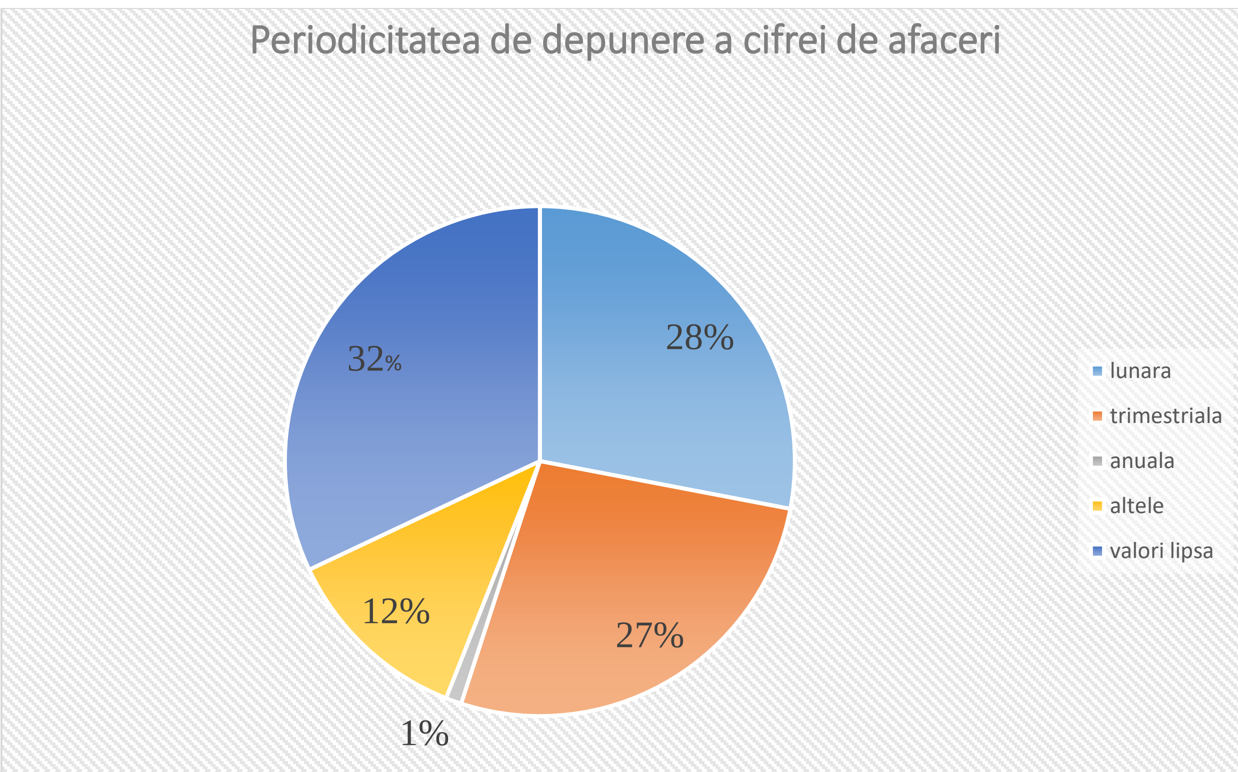
VI. Anexe

1. Indicatorii de calitate obținuți în urma aplicării metodelor prezentate mai sus pentru o versiune a sursei administrative din anul 2016.

Indicator	Procent
Rata de clasificare eronată	24%
Sub-acoperirea	13%
Supra-acoperirea	12%
Mărimea reviziilor din diferite versiuni ale sursei administrative RAR – Revizii absolute relative	49%

Direcția Instrumente Inovatoare în Statistică
Institutul Național de Statistică
2017





2.Codul R folosit pentru a calcula indicatorii de calitate si metodele de depistare a valorilor aberante

Incarcarea librariilor

```
library(dplyr)
```

Indicator - Valori lipsa pe clasa de marime a unitatilor

```
indicator_missing = function(sursa){
  sursa = read.csv(sursa)
  indicator = list()
  for (i in 1:4){
    val_lipsa = subset(sursa, t_r1_r == 0 & t_r2_r  == 0 & t_r3_r  == 0 &
                              t_r4_r == 0 & t_r5_r  == 0 & t_r6_r  == 0 &
                              t_r7_r == 0 & t_r8_r  == 0 & t_r9_r  == 0 &
                              t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 0 &
                              cls_marime == i )
    val_total = subset(sursa, cls_marime == i)
    indicator[i] = nrow(val_lipsa)/nrow(val_total)*100
  }
  indicator_missing = data.frame(cls1 = c(indicator[[1]]),
                                 cls2 = c(indicator[[2]]),
                                 cls3 = c(indicator[[3]]),
                                 cls4 = c(indicator[[4]]))
  return(indicator_missing)
}
```

Exemplu: Aplicarea functiei indicator_missing pentru sursa din anul 2016.

```
indicator_missing("sursa2016rev.csv")
```

```
##      cls1    cls2     cls3     cls4
## 1 39.8503 14.7601 4.867971 2.979666
```

Indicator - Rata de clasificare eronata pentru CAEN

```
indicator_caen = function(sursa, REGIS){
  sursa = read.csv(sursa)
  REGIS = read.csv(REGIS)
  sursa_merged = inner_join(sursa, REGIS, by = c("cod" = "codi"))
  indicator_caen = nrow(subset(sursa_merged, caen != 0 & caen != caen_rev2))/
                   nrow(subset(sursa_merged, caen != 0)) * 100
  return(indicator_caen)
}
```

Exemplu: Aplicarea functiei indicator_caen pentru sursa din anul 2016.

```
indicator_caen("sursa2016rev.csv", "iactiv15.csv")
```

```
## [1] 4.856126
```

Indicator - Sub-acoperirea

```
indicator_undercoverage = function(sursa, REGIS){
  sursa = read.csv(sursa)
  REGIS = read.csv(REGIS)
  indicator_undercoverage = nrow(subset(REGIS,
                            !(REGIS$codi %in% as.list(sursa$cod))))/
                            nrow(REGIS) * 100
  return(indicator_undercoverage)
}
```

Exemplu: Aplicarea functiei indicator_undercoverage pentru sursa din anul 2016.

```
indicator_undercoverage("sursa2016rev.csv", "iactiv15.csv")
```

```
## [1] 13.13882
```

Indicator - Supra-acoperirea

```
indicator_overcoverage = function(sursa, REGIS){
  sursa = read.csv(sursa)
  REGIS = read.csv(REGIS)
  indicator_overcoverage = nrow(subset(sursa,
                           !(sursa$cod %in% as.list(REGIS$codi))))/
                           nrow(REGIS) * 100
  return(indicator_overcoverage)
}
```

Exemplu: Aplicarea functiei indicator_overcoverage pentru sursa din anul 2016.

```
indicator_overcoverage("sursa2016rev.csv", "iactiv15.csv")
```

```
## [1] 13.23506
```

Indicator - Revizii relative absolute

```
indicator_rar = function(sursa, sursa_anterioara ){
  sursa = read.csv(sursa)
  sursa_anterioara = read.csv(sursa_anterioara)
  m = inner_join(sursa, sursa_anterioara[,c("cod", "ca_01", "ca_02", "ca_03",
      "ca_04", "ca_05", "ca_06", "ca_07", "ca_08", "ca_09", "ca_10", "ca_11", "ca_12")],
      by = c("cod"))
  indicator_rar = (abs(sum(m$ca_01_r) - sum(m$ca_01)) + abs(sum(m$ca_02_r) - sum(m$ca_02))
                 + abs(sum(m$ca_03_r) - sum(m$ca_03)) + abs(sum(m$ca_04_r) - sum(m$ca_04))
                 + abs(sum(m$ca_05_r) - sum(m$ca_05)) + abs(sum(m$ca_06_r) - sum(m$ca_06))
                 + abs(sum(m$ca_07_r) - sum(m$ca_07)) + abs(sum(m$ca_08_r) - sum(m$ca_08))
                 + abs(sum(m$ca_09_r) - sum(m$ca_09)) + abs(sum(m$ca_10_r) - sum(m$ca_10))
                 + abs(sum(m$ca_11_r) - sum(m$ca_11)))/
```

```
                    (sum(m$ca_01) + sum(m$ca_02) + sum(m$ca_03) + sum(m$ca_04)
                +  sum(m$ca_05) + sum(m$ca_06) + sum(m$ca_07) + sum(m$ca_08)
                +  sum(m$ca_09) + sum(m$ca_10) + sum(m$ca_11) ) * 100
  return(indicator_rar)
}
```

Exemplu: Aplicarea functiei indicator_rar pentru sursa din anul 2016.

```
indicator_rar("sursa2016rev.csv", "sursa2016.csv")
```

```
## [1] 49.32423
```

Periodicitatea de depunere a cifrelor de afaceri de catre intreprinderi

Incarcarea librariilor

```
library(dplyr)
```

Metoda

```
periodicitate = function(sursa){
  sursa = read.csv(sursa)
  periodicitate <- sursa %>% group_by(t_r1_r, t_r2_r, t_r3_r, t_r4_r, t_r5_r,
                              t_r6_r, t_r7_r, t_r8_r, t_r9_r, t_r10_r,
                              t_r11_r, t_r12_r) %>% dplyr :: summarise(n=n())
  periodicitate <- as.data.frame(arrange(periodicitate, desc(n)))
  return(periodicitate[1:6,])
}
# Scrierea fisierului periodicitate
## write.csv(periodicitate, "path/periodicitate.csv" )
```

Exemplu: Aplicarea functiei periodicitate pentru sursa din anul 2016.

```
periodicitate("sursa2016rev.csv")
```

```
##   t_r1_r t_r2_r t_r3_r t_r4_r t_r5_r t_r6_r t_r7_r t_r8_r t_r9_r t_r10_r
## 1      0      0      0      0      0      0      0      0      0       0
## 2      1      1      1      1      1      1      1      1      1       1
## 3      0      0      1      0      0      1      0      0      1       0
## 4      0      0      0      0      0      0      0      0      0       0
## 5      0      0      1      0      0      1      0      0      1       0
## 6      0      0      1      0      0      0      0      0      0       0
##   t_r11_r t_r12_r      n
## 1       0       0 165157
## 2       1       1 147983
## 3       0       1 140788
## 4       0       1   5752
## 5       0       0   5752
## 6       0       0   4556
```

Metoda Erori de unitate de masura

Incarcarea librariilor

```
library(dplyr)
```

```
calculeazaEroriUnitateDeMasura = function(sursa_an_curent, sursa_an_precedent){
  #Fixarea parametrilor
  n = 12 #numar de luni
  A = 0.00065
  B = 0.00135

  sursa_an_curent = read.csv(sursa_an_curent)
  sursa_an_precedent = read.csv(sursa_an_precedent)

  colnames(sursa_an_precedent)[which(names(sursa_an_precedent) == "ca_12_r")] = "ca_12_rr"

  # Join intre sursa_an_curent si sursa_an_precedent:
  # Avem nevoie de luna 12 din sursa_an_precedent
  sursa_an_curent = left_join(sursa_an_curent, sursa_an_precedent[,c("cod","ca_12_rr")],
                              by = c("cod"))

  for(i in 2:n){
    col1 = paste0("Metoda0_",i)
    col2 = paste0("t_r",i, "_r")
    col3 = paste0(ifelse(i <= 9, "ca_0", "ca_"), i, "_r")
    col4 = paste0(ifelse(i-1 <= 9, "ca_0", "ca_"), i-1, "_r")

    sursa_an_curent[,"Metoda0_1"] = ifelse(sursa_an_curent[,"t_r1_r"] == 1 &
                                     sursa_an_curent[,"ca_12_rr"]!= 0 &
                                     sursa_an_curent[,"ca_01_r"]/
                                       sursa_an_curent[,"ca_12_rr"] > A &
                                     sursa_an_curent[,"ca_01_r"]/
                                       sursa_an_curent[,"ca_12_rr"] < B,
                                     1, 0)

    sursa_an_curent$Metoda0_1[is.na(sursa_an_curent$Metoda0_1)] = 0

    sursa_an_curent[,col1] = ifelse(sursa_an_curent[,col2] == 1 &
                              sursa_an_curent[,col4] != 0 &
                              sursa_an_curent[,col3]/sursa_an_curent[,col4] > A &
                              sursa_an_curent[,col3]/sursa_an_curent[,col4] < B , 1, 0)
  }

  # Numarul de outliers pe fiecare luna
  p = colSums(sursa_an_curent %>% select(Metoda0_1:Metoda0_12), na.rm = TRUE)

  # Procent de outliers pe fiecare luna
  procent = vector("numeric", 12L)
  for (i in 1:n){
    col = paste0("t_r", i, "_r")
    procent[i] = paste0( round(p[[i]]*100/sum(sursa_an_curent[,col]),2), "% ")
```

```r
  }

  luna = c("ianuarie", "februarie", "martie", "aprilie", "mai", "iunie", "iulie",
           "august", "septembrie", "octombrie", "noiembrie", "decembrie")
  procent = data.frame(luna, procent)
  return(procent)

  # Scrierea fisierului cu marcarea erorilor in coloanele Metoda0_1:Metoda0_12
  # 1 - cifra de afaceri este aberanta
  # 0 - cifra de afaceri nu este aberanta
  ## write.csv(sursa_an_curent, "path/sursa_metoda0.csv")
}
```

Exemplu: Aplicarea functiei calculeazaEroriUnitateDeMasura() pentru sursa din anul 2016.

```r
calculeazaEroriUnitateDeMasura("sursa2016rev.csv", "sursa2015rev.csv")
```

```
##          luna procent
## 1    ianuarie   0.15%
## 2   februarie   0.01%
## 3      martie   0.01%
## 4     aprilie   0.02%
## 5         mai   0.02%
## 6       iunie   0.01%
## 7       iulie   0.02%
## 8      august   0.03%
## 9  septembrie   0.01%
## 10  octombrie   0.03%
## 11  noiembrie   0.01%
## 12  decembrie   0.01%
```

Patternuri de valori aberante

Incarcarea librariilor

```
library(dplyr)
```

```
calculeazaPatternValoriAberante = function(sursa){
  sursa = read.csv(sursa)
  p1 = subset(sursa, ca_03_r == 0 & ca_06_r == 0 & ca_09_r == 0 & ca_12_r > 0 &
              t_r1_r == 0 &  t_r2_r == 0 &  t_r3_r == 1 &
              t_r4_r == 0 &  t_r5_r == 0 &  t_r6_r == 1 & t_r7_r == 0 & t_r8_r == 0 &
              t_r9_r == 1 & t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 1)
  procent1 = paste0(round(nrow(p1)/nrow(subset(sursa,  t_r1_r == 0 &  t_r2_r == 0 &
    t_r3_r == 1 & t_r4_r == 0 &  t_r5_r == 0 &  t_r6_r == 1 & t_r7_r == 0 & t_r8_r == 0 &
    t_r9_r == 1 & t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 1)) * 100, 2), "% ")

  p2 = subset(sursa, ca_03_r == 0 & ca_06_r > 0 & ca_09_r > 0 & ca_12_r > 0 &
       t_r1_r == 0 &  t_r2_r == 0 &  t_r3_r == 1 &
       t_r4_r == 0 &  t_r5_r == 0 &  t_r6_r == 1 & t_r7_r == 0 &
       t_r8_r == 0 &  t_r9_r == 1 & t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 1)
  procent2 = paste0(round(nrow(p2)/nrow(subset(sursa,  t_r1_r == 0 &  t_r2_r == 0 &
    t_r3_r == 1 & t_r4_r == 0 &  t_r5_r == 0 &  t_r6_r == 1 & t_r7_r == 0 & t_r8_r == 0 &
    t_r9_r == 1 & t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 1)) * 100, 2), "% ")

  p3 = subset(sursa, ca_03_r == ca_06_r & ca_06_r == ca_09_r & ca_09_r == ca_12_r &
                t_r1_r == 0 &  t_r2_r == 0 &  t_r3_r == 1 &
                t_r4_r == 0 &  t_r5_r == 0 &  t_r6_r == 1 & t_r7_r == 0 & t_r8_r == 0 &
                t_r9_r == 1 & t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 1)
  procent3 = paste0(round(nrow(p3)/nrow(subset(sursa,  t_r1_r == 0 &  t_r2_r == 0 &
    t_r3_r == 1 & t_r4_r == 0 &  t_r5_r == 0 &  t_r6_r == 1 & t_r7_r == 0 & t_r8_r == 0 &
    t_r9_r == 1 & t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 1)) * 100, 2), "% ")

  p4 = subset(sursa, ca_03_r == ca_06_r & ca_06_r == ca_09_r & ca_12_r > 0 &
                ca_12_r != ca_09_r &
                t_r1_r == 0 &  t_r2_r == 0 &  t_r3_r == 1 &
                t_r4_r == 0 &  t_r5_r == 0 &  t_r6_r == 1 & t_r7_r == 0 & t_r8_r == 0 &
                t_r9_r == 1 & t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 1)
  procent4 = paste0(round(nrow(p4)/nrow(subset(sursa,  t_r1_r == 0 &  t_r2_r == 0 &
    t_r3_r == 1 & t_r4_r == 0 &  t_r5_r == 0 &  t_r6_r == 1 & t_r7_r == 0 & t_r8_r == 0 &
    t_r9_r == 1 & t_r10_r == 0 & t_r11_r == 0 & t_r12_r == 1)) * 100, 2), "% ")


  pattern = cat(
  "Procentul unitatilor care au depus cifra de afaceri 0 in primele trei trimestre
   si cifra de afaceri pozitiva in trimestrul patru este:",procent1, "\n",
  "Procentul unitatilor care au depus cifra de afaceri 0 in primul trimestru si valori
   pozitive ale cifrei de afaceri in celelalte trei trimestre este:", procent2, "\n",
  "Procentul unitatilor care au depus aceeasi cifra de afaceri in toate cele patru
   trimestre este:",procent3,"\n",
  "Procentul unitatilor au depus aceeasi cifra de afaceri pe trei trimestre si o alta
   valoare pozitiva in trimestrul patru este:", procent4)
}
```

Exemplu: Aplicarea functiei calculeazaPatternValoriAberante() pentru sursa din anul 2016.

```
calculeazaPatternValoriAberante("sursa2016rev.csv")
```

```
## Procentul unitatilor care au depus cifra de afaceri 0 in primele trei trimestre
##    si cifra de afaceri pozitiva in trimestrul patru este: 2.09%
##  Procentul unitatilor care au depus cifra de afaceri 0 in primul trimestru si valori
##    pozitive ale cifrei de afaceri in celelalte trei trimestre este: 4.37%
##  Procentul unitatilor care au depus aceeasi cifra de afaceri in toate cele patru
##    trimestre este: 11.65%
##  Procentul unitatilor au depus aceeasi cifra de afaceri pe trei trimestre si o alta
##    valoare pozitiva in trimestrul patru este: 2.25%
```

Quartila1, Mediana, Quartila3 in functie de diviziunea CAEN, clasa de marime si periodicitate

Incarcarea librariilor

```
library(dplyr)
library(lazyeval)
```

```
calculeazaQuartileMediana = function(sursa){
  sursa = read.csv(sursa)
  #Fixarea parametrilor
  n = 12 #numar de luni

  # Definim un dataframe gol
  randuri = read.csv(text="div_caen, cls_marime, periodicit, q1, m, q3, luna")

  for(clasa_marime in 1:4) {
    for (i in 1:n) {
      coloana_ca = paste0(ifelse(i <= 9, "ca_0", "ca_"), i, "_r");
      conditie_t_r = interp(~ col_tr == 1, col_tr = as.name(paste0("t_r", i, "_r")));
      un_rand <- sursa %>% filter_(~cls_marime == clasa_marime,
                                   ~periodicit == 1, conditie_t_r,
                                   ~div_caen != 0 ) %>% group_by(div_caen) %>%
      summarise_(cls_marime = clasa_marime, periodicit =1,
           q1 = interp(~quantile(col_ca, probs = 0.25),
           col_ca = as.name(coloana_ca)), m = interp(~median(col_ca),
           col_ca = as.name(coloana_ca)), q3 = interp(~quantile(col_ca, probs = 0.75),
           col_ca = as.name(coloana_ca)), luna = i);
      randuri = bind_rows(randuri, un_rand); #adaugam la sfarsitul dataframe-ului
    }
  }

  for(clasa_marime in 1:4) {
    for (i in c(3,6,9,12)) {
      coloana_ca = paste0(ifelse(i <= 9, "ca_0", "ca_"), i, "_r");
      conditie_t_r = interp(~ col_tr == 1, col_tr = as.name(paste0("t_r", i, "_r")));
      un_rand <- sursa %>% filter_(~cls_marime == clasa_marime, ~periodicit == 2,
                                   conditie_t_r, ~div_caen != 0 ) %>%
      group_by(div_caen) %>% summarise_(cls_marime = clasa_marime, periodicit =2,
      q1 = interp(~quantile(col_ca, probs = 0.25), col_ca = as.name(coloana_ca)),
      m = interp(~median(col_ca), col_ca = as.name(coloana_ca)),
      q3 = interp(~quantile(col_ca, probs = 0.75), col_ca = as.name(coloana_ca)),
           luna = i);
      randuri = bind_rows(randuri, un_rand); #adaugam la sfarsitul dataframe-ului
    }
  }
  return(head(randuri))

  # Scrierea fisierului
  ## write.csv(randuri, "path/q1_m_q3.csv")
}
```

Exemplu: Aplicarea functiei calculeazaQuartileMediana() pentru sursa din anul 2016.

```
calculeazaQuartileMediana("sursa2016rev.csv")
```

```
##   div_caen cls_marime periodicit     q1      m        q3 luna
## 1        1          1          1      0      0   9986.50    1
## 2        2          1          1      0   9582  46029.25    1
## 3        3          1          1      0   2156  25048.00    1
## 4        5          1          1 325020 325020 325020.00    1
## 5        6          1          1      0      0      0.00    1
## 6        7          1          1      0      0   8543.50    1
```

Metoda Intervalul interquartilic

Incarcarea librariilor

```
library(tidyverse)
library(dplyr)
library(reshape2)
library(lazyeval)
library(stringi)
library(stringr)
```

```
calculeazaMetodaIntervalulInterquartilic = function(sursa, quartile_mediana){
  sursa = read.csv(sursa)
  quartile_mediana = read.csv(quartile_mediana)
  #Fixarea parametrilor
  n = 12 #numar de luni
  c = 10

  model1 = function(data) {

    ndata = names(data)
    n = sum(str_detect(ndata, 'LB'))

    for (i in 1:n){
      col1 = paste0(ifelse(i <= 9, "ca_0", "ca_"), i, "_r")
      col2 = paste0('LB_', i)
      col3 = paste0('UB_', i)

      new_col_name = paste0('Metoda1_', i)

      mutate_call = lazyeval::interp(~ ifelse(a < b | a > c, 1, 0), a = as.name(col1),
                                     b = as.name(col2),
                                     c = as.name(col3))
      data = data %>% mutate_(.dots = setNames(list(mutate_call), new_col_name))
    }

    return(data)
  }

  # Inlocuire caen NA cu 0
  sursa$div_caen[is.na(sursa$div_caen)] = 0

  quartile_mediana2 = quartile_mediana %>% mutate(LB = q1 - c*(m-q1),
                                                  UB = q3+c*(q3-m))

  quartile_mediana2 = quartile_mediana2[, !names(quartile_mediana2)%in%c("q1", "m", "q3")]

  quartile_mediana_melt2 = melt(quartile_mediana2[,-1], id.vars = c("div_caen",
                                                   "cls_marime", "periodicit", "luna"))
  quartile_mediana_cast = dcast(quartile_mediana_melt2, div_caen + cls_marime +
                                periodicit ~ variable+luna)
```

```
  sursa_merged_lj = left_join(sursa, quartile_mediana_cast, by = c("div_caen",
                                                                   "cls_marime",
                                                                   "periodicit"))

  sursa_merged_lj_mutate = model1(sursa_merged_lj)

  sursa_merged_lj_mutate = sursa_merged_lj_mutate[, !names(sursa_merged_lj_mutate)%in%
                         c("LB_1", "LB_2", "LB_3", "LB_4", "LB_5", "LB_6", "LB_7",
                           "LB_8", "LB_9", "LB_10", "LB_11", "LB_12", "UB_1",
                           "UB_2", "UB_3", "UB_4", "UB_5", "UB_6", "UB_7",
                           "UB_8", "UB_9", "UB_10", "UB_11", "UB_12")]

  # Inlocuire NA cu 0
  sursa_merged_lj_mutate[c("Metoda1_1", "Metoda1_2", "Metoda1_3", "Metoda1_4",
                         "Metoda1_5", "Metoda1_6", "Metoda1_7", "Metoda1_8",
                         "Metoda1_9", "Metoda1_10", "Metoda1_11",
                         "Metoda1_12")][is.na(sursa_merged_lj_mutate[c("Metoda1_1",
                         "Metoda1_2", "Metoda1_3", "Metoda1_4", "Metoda1_5",
                         "Metoda1_6", "Metoda1_7", "Metoda1_8", "Metoda1_9",
                         "Metoda1_10","Metoda1_11","Metoda1_12")])] <- 0

  # Calcularea numarului de outliers pe fiecare luna
  p = colSums(sursa_merged_lj_mutate %>% select(Metoda1_1:Metoda1_12), na.rm = TRUE)
  p

  # Calcularea procentului de outliers
  procent <- vector("numeric", 12L)
  for (i in 1:12){
    col = paste0("t_r", i, "_r")
    procent[i] <- paste0( round(p[[i]]*100/sum(sursa[,col]),2), "%")
  }
  procent

  luna = c("ianuarie", "februarie", "martie", "aprilie", "mai", "iunie", "iulie",
           "august", "septembrie", "octombrie", "noiembrie", "decembrie")
  procent = data.frame(luna, procent)
  return(procent)

  # Scrierea fisierului cu marcarea erorilor in coloanele Metoda1_1:Metoda1_12
  # 1 - cifra de afaceri este aberanta
  # 0 - cifra de afaceri nu este aberanta
  ## write.csv(sursa_merged_lj_mutate, path/sursa_metoda1.csv")
}
```

Exemplu: Aplicarea functiei calculeazaMetodaIntervalulInterquartilic() pentru sursa din anul 2016

```
calculeazaMetodaIntervalulInterquartilic("sursa2016rev.csv", "q1_m_q3.csv")
```

```
##          luna procent
## 1    ianuarie   2.11%
## 2   februarie   2.05%
## 3      martie   1.22%
## 4     aprilie   1.94%
## 5         mai   1.94%
```

```
## 6        iunie   1.19%
## 7        iulie   1.91%
## 8       august   1.88%
## 9  septembrie   1.26%
## 10  octombrie    1.9%
## 11  noiembrie   1.98%
## 12  decembrie   1.37%
```

Metoda Raport pe perioada

Incarcarea librariilor

```
library(tidyverse)
library(plyr)
library(reshape2)
library(lazyeval)
library(stringi)
library(stringr)
library(dplyr)
library(data.table)
```

```
calculeazaMetodaRaportPerioada = function(sursa_an_curent, sursa_an_precedent,
                                quartileMediana_curent, quartileMediana_precedent){
  sursa_an_curent = read.csv(sursa_an_curent)
  sursa_an_precedent = read.csv(sursa_an_precedent)
  quartileMediana_curent = read.csv(quartileMediana_curent)
  quartileMediana_precedent = read.csv(quartileMediana_precedent)

  #Fixarea parametrilor
  n = 12 #numar de luni

  sursa_an_curent$div_caen[is.na(sursa_an_curent$div_caen)] = 0 #inlocuire caen NA cu 0

  setnames(sursa_an_precedent, "ca_12_r", "ca_12_rr" ) #modificare nume coloana

  # Join sursa_an_curent cu sursa_an_precedent dupa coloana cod
  sursa_an_curent = left_join(sursa_an_curent, sursa_an_precedent[,c("cod","ca_12_rr")],
                             by = c("cod"))

  # Extragere date din fisierul quartileMediana_precedent numai pentru luna 12
  sursa_m15 <-  quartileMediana_precedent[,c(2,3,4,6,8)] %>% filter(luna == 12)

  setnames(sursa_m15, "m", "m15" ) #modificare nume coloana

  for(i in 1:n){
    col1 = assign(paste0("sursa_m",i),
                 quartileMediana_curent[,c(2,3,4,6,8)] %>% filter(luna == i))
    setnames(col1, "m", paste0("m",i) )
  }

  sursa_merged = join_all(list(sursa_an_curent, sursa_m1, sursa_m2, sursa_m3, sursa_m4,
                             sursa_m5, sursa_m6, sursa_m7,   sursa_m8, sursa_m9,
                             sursa_m10, sursa_m11, sursa_m12, sursa_m15),
                             by = c("div_caen", "cls_marime", "periodicit"),
                             type='left')
  sursa_merged = sursa_merged[, !names(sursa_merged)%in%c("luna")]

  # Crearea coloanelor score
  sursa_merged[,"score15"] <- sursa_merged[,"ca_12_rr"]/sursa_merged[,"m15"]
  for (i in 1:n){
```

```
  col1 = paste0("score",i)
  col2 = paste0("m",i)
  col3 = paste0(ifelse(i <= 9, "ca_0", "ca_"), i, "_r")
  sursa_merged[,col1] <- sursa_merged[,col3]/sursa_merged[,col2]
}

# Eliminare date de tipul NaN si Infinite din coloanele score
sursa_merged[,"score15"][is.nan(sursa_merged[,"score15"])] <- NA
sursa_merged[,"score15"][is.infinite(sursa_merged[,"score15"])] <- NA
for(i in 1:n){
  col = paste0('score',i)
  sursa_merged[,col][is.nan(sursa_merged[,col])] <- NA
  sursa_merged[,col][is.infinite(sursa_merged[,col])] <- NA
}

# Crearea coloanelor test
sursa_merged[,"test1"] <- ifelse(sursa_merged[,"score1"] > sursa_merged[,"score15"],
                                 sursa_merged[,"score1"] / sursa_merged[,"score15"],
                                 sursa_merged[,"score15"] / sursa_merged[,"score1"])
for(i in 1:(n-1)){
  col1 = paste0("test",i+1)
  col2 = paste0("score", i+1)
  col3 = paste0("score", i)
  sursa_merged[,col1]<- ifelse(sursa_merged[,col2] > sursa_merged[,col3],
                               sursa_merged[,col2] / sursa_merged[,col3],
                               sursa_merged[,col3] / sursa_merged[,col2])
 }

# Inlocuirea datelor de tip NaN si Inf din test cu NA
for(i in 1:n){
  col = paste0('test',i)
  sursa_merged[,col][is.nan(sursa_merged[,col])] <- NA
  sursa_merged[,col][is.infinite(sursa_merged[,col])] <- NA
}

# Marcarea erorilor cu 1
for(i in 1:n){
  col1 = paste0("Metoda2_",i)
  col2 = paste0("test",i)
  sursa_merged[,col1] <- ifelse(!is.na(sursa_merged[,col2]) &
                                sursa_merged[,col2] > 30 , 1, 0)
}

# Eliminarea coloane nefolositoare din sursa
sursa_merged = sursa_merged[, !names(sursa_merged)%in%c("m1", "m2",
               "m3", "m4", "m5", "m6", "m7", "m8", "m9", "m10", "m11",
               "m12", "m15", "score1", "score2", "score3", "score4",
               "score5", "score6","score7", "score8", "score9","score10",
               "score11","score12","score15", "test1", "test2", "test3",
               "test4", "test5", "test6", "test7", "test8", "test9",
               "test10","test11","test12" )]

# Calcularea numarului de outliers
```

```r
p = colSums(sursa_merged %>% select(Metoda2_1:Metoda2_12), na.rm = TRUE)

# Calcularea procentului
procent <- vector("numeric", 12L)
for (i in 1:n){
  col = paste0("t_r", i, "_r")
  procent[i] <-  paste0( round(p[[i]]*100/sum(sursa_an_curent[,col]),2), "% ")
}

luna = c("ianuarie", "februarie", "martie", "aprilie", "mai", "iunie", "iulie",
           "august", "septembrie", "octombrie", "noiembrie", "decembrie")
procent = data.frame(luna, procent)
return(procent)


# Scrierea fisierului cu marcarea erorilor in coloanele Metoda2_1:Metoda2_12
# 1 - cifra de afaceri este aberanta
# 0 - cifra de afaceri nu este aberanta
## write.csv(sursa_merged, "path/sursa_metoda2.csv")

}
```

Exemplu: Aplicarea functiei calculeazaMetodaRaportPerioada() pentru sursa din anul 2016

```r
calculeazaMetodaRaportPerioada("sursa2016rev.csv", "sursa2015rev.csv",
                                "q1_m_q3.csv","q1_m_q315.csv")
```

```
##          luna procent
## 1    ianuarie   3.7%
## 2   februarie  1.35%
## 3      martie  0.48%
## 4     aprilie  0.92%
## 5         mai  0.84%
## 6       iunie  0.47%
## 7       iulie  0.99%
## 8      august  0.99%
## 9  septembrie   0.5%
## 10  octombrie  1.01%
## 11  noiembrie   0.9%
## 12  decembrie  0.54%
```

Metoda Comparatie cu raportari anterioare pentru cifra de afaceri

Incarcarea librariilor

```
library(dplyr)
```

```
calculeazaMetodaComparatie = function(sursa_an_curent, sursa_an_precedent){

  sursa_an_curent = read.csv(sursa_an_curent)
  sursa_an_precedent = read.csv(sursa_an_precedent)

  #Fixarea parametrilor
  n = 12 #numar de luni
  c = 10000

  # Modificare denumire coloana ca_12_r din sursa_an_precedent
  # Pentru a nu se face confuzie cu ca_12_r din sursa_2016
  colnames(sursa_an_precedent)[which(names(sursa_an_precedent) == "ca_12_r")] = "ca_12_rr"

  # Media cifrelor de afaceri pe ultimele 12 luni - sursa_an_precedent
  sursa_an_precedent[,"media"] = 10 * rowMeans(sursa_an_precedent[, c("ca_01_r","ca_02_r",
                        "ca_03_r","ca_04_r","ca_05_r", "ca_06_r","ca_07_r",
                        "ca_08_r", "ca_09_r", "ca_10_r","ca_11_r", "ca_12_rr")])

  sursa_an_curent = left_join(sursa_an_curent, sursa_an_precedent[,c("cod","media")],
                            by = c("cod"))

  for(i in 1:n){
    col1 = paste0("Metoda3_",i)
    col2 = paste0("t_r",i,"_r")
    col3 = paste0(ifelse(i <= 9, "ca_0", "ca_"), i, "_r")

    sursa_an_curent[,col1] = ifelse(sursa_an_curent[,col2] == 1 &
                      sursa_an_curent[,col3] > sursa_an_curent[,"media"] &
                      sursa_an_curent[,col3] > c , 1, 0)
  }

  p = colSums(sursa_an_curent %>% select(Metoda3_1:Metoda3_12), na.rm = TRUE)

  # Calcularea procentului
  procent = vector("numeric", 12L)
  for (i in 1:n){
    col = paste0("t_r", i, "_r")
    procent[i] = paste0( round(p[[i]]*100/sum(sursa_an_curent[,col]),2), " % ")
  }

  luna = c("ianuarie", "februarie", "martie", "aprilie", "mai", "iunie", "iulie",
             "august", "septembrie", "octombrie", "noiembrie", "decembrie")
  procent = data.frame(luna, procent)
  return(procent)
```

```
  # Scrierea fisierului cu marcarea erorilor in coloanele Metoda3_1:Metoda3_12
  # 1 - cifra de afaceri este aberanta
  # 0 - cifra de afaceri nu este aberanta
  ##write.csv(sursa_an_curent, "path/sursa_metoda3.csv")
}
```

Exemplu: Aplicarea functiei calculeazaMetodaComapartie() pentru sursa din anul 2016.

```
calculeazaMetodaComparatie("sursa2016rev.csv", "sursa2015rev.csv")
```

```
##          luna procent
## 1    ianuarie 0.49 %
## 2   februarie 0.78 %
## 3      martie  3.1 %
## 4     aprilie 1.21 %
## 5         mai 1.36 %
## 6       iunie 4.54 %
## 7       iulie  1.8 %
## 8      august 1.85 %
## 9  septembrie 5.82 %
## 10  octombrie 2.19 %
## 11  noiembrie 2.36 %
## 12  decembrie 7.24 %
```

Metoda Hidiroglou - Berthelot

Incarcarea librariilor

```
library(tidyverse)
library(dplyr)
library(reshape2)
library(lazyeval)
library(stringi)
library(stringr)
```

```
calculeazaMetodaHidiroglou = function(sursa_an_curent, sursa_an_precedent){

  sursa_an_curent = read.csv(sursa_an_curent)
  sursa_an_precedent = read.csv(sursa_an_precedent)
  #Fixarea parametrilor
  n = 12 #Nr de luni
  v = 1
  c = 250
  a = 0.05

  # Modificare denumire coloana ca_12_r din sursa_an_precedent
  # Pentru a nu se face confuzie cu ca_12_r din sursa_2016
  colnames(sursa_an_precedent)[which(names(sursa_an_precedent) == "ca_12_r")] = "ca_12_rr"

  # Join sursa_an_curent cu sursa_an_precedent dupa cod pentru ca am nevoie de ca_12
  sursa_merged = left_join(sursa_an_curent, sursa_an_precedent[,c("cod","ca_12_rr")],
                           by = c("cod"))

  # Calcularea rapoartelor r
  for(i in 2:n){
    sursa_merged$r1 = ifelse(sursa_merged$t_r1_r == 1 & sursa_merged$ca_12_rr != 0,
                             sursa_merged$ca_01_r/sursa_merged$ca_12_rr, NA)
    col1 = paste0("r",i)
    col2 = paste0("t_r",i, "_r")
    col3 = paste0(ifelse(i-1 <= 9, "ca_0", "ca_"), i-1, "_r")
    col4 = paste0(ifelse(i <= 9, "ca_0", "ca_"), i, "_r")

    sursa_merged[,col1] = ifelse(sursa_merged[,col2] == 1 &
                                 sursa_merged[,col3] != 0,
                                 sursa_merged[,col4]/sursa_merged[,col3], NA)
  }

  # Calcularea medianei
  mediana = vector("numeric", 12L)
  for (i in 1:n){
    col = paste0("r",i)
    mediana[i] = median(sursa_merged[,col], na.rm = TRUE)
  }

  # Calcularea rapoartelor t
  for(i in 1:n){
```

```
  col1 = paste0("t",i)
  col2 = paste0("r",i)
  sursa_merged[,col1] = ifelse(sursa_merged[,col2] < mediana[i],
                               (sursa_merged[,col2] - mediana[i])/sursa_merged[,col2],
                               (sursa_merged[,col2]-mediana[i])/mediana[i])
}

# Eliminarea datelor de tipl NaN si Inf
for(i in 1:n){
  col = paste0('t',i)
  sursa_merged[,col][is.nan(sursa_merged[,col])] = NA
  sursa_merged[,col][is.infinite(sursa_merged[,col])] = NA
}

# Calcularea variabilei E
for(i in 2:n){
sursa_merged$e1 = ifelse(sursa_merged$t_r1_r == 1, sursa_merged$t1 *
                  pmax(sursa_merged$ca_01_r,sursa_merged$ca_12_rr) * v, NA)
  col1 = paste0("e",i)
  col2 = paste0("t_r",i, "_r")
  col3 = paste0(ifelse(i-1 <= 9, "ca_0", "ca_"), i-1, "_r")
  col4 = paste0(ifelse(i <= 9, "ca_0", "ca_"), i, "_r")
  col5 = paste0("t",i)
sursa_merged[,col1] = ifelse(sursa_merged[,col2] == 1,
                             sursa_merged[,col5] * pmax(sursa_merged[,col4],
                                                        sursa_merged[,col3]), NA)
}

memory.limit(size=5000) #setarea limitei de memorie
# Crearea functiei metoda5
metoda5 = function(data){
  for(i in 1:n){
    col = paste0("e",i)
    new_col_name = paste0('Metoda5_', i)

    q1 = quantile(data[,col], probs = 0.25, na.rm = TRUE)
    q2 = median(data[,col], na.rm = TRUE)
    q3 = quantile(data[,col], probs = 0.75, na.rm = TRUE)

    a  = 0.05
    c  = 250

    d1 = max((q2-q1), abs(a*q2))
    d2 = max((q3-q2), abs(a*q2))

    data[,new_col_name] = ifelse((data[,col] < q2 - c*d1 | data[,col] > q2 + c*d2)
                                 & !is.na(data[,col]), 1, 0)
  }
   return(data)
}

# Aplicarea functiei metoda5
sursa_merged = metoda5(sursa_merged)
```

```
  # Eliminarea coloanelor care nu sunt relevante
  sursa_merged = sursa_merged[, !names(sursa_merged)%in%
                c("r1", "r2", "r3", "r4", "r5", "r6", "r7", "r8", "r9", "r10", "r11",
                  "r12", "t1", "t2", "t3", "t4", "t5", "t6", "t7", "t8",
                  "t9", "t10", "t11", "t12", "e1", "e2", "e3", "e4", "e5",
                  "e6",  "e7", "e8", "e9", "e10", "e11", "e12")]

  # Calcularea numarului de outliers pe fiecare luna
  p = colSums(sursa_merged %>% select(Metoda5_1:Metoda5_12), na.rm = TRUE)


  # Calcularea procentului de outliers pe fiecare luna
  procent = vector("numeric", 12L)
  for (i in 1:n){
    col = paste0("t_r", i, "_r")
    procent[i] = paste0( round(p[[i]]*100/sum(sursa_an_curent[,col]),2), " % ")
  }
  luna = c("ianuarie", "februarie", "martie", "aprilie", "mai", "iunie", "iulie",
                 "august", "septembrie", "octombrie", "noiembrie", "decembrie")
  procent = data.frame(luna, procent)

  return(procent)

  # Scrierea fisierului cu marcarea erorilor in coloanele Metoda5_1:Metoda5_12
  # 1 - cifra de afaceri este aberanta
  # 0 - cifra de afaceri nu este aberanta
  ## write.csv(sursa_merged_lj_mutate, path/sursa_metoda5.csv")
}
```

Exemplu: Aplicarea functiei calculeazaMetodaHidiroglou() pentru sursa din anul 2016.

```
calculeazaMetodaHidiroglou("sursa2016rev.csv", "sursa2015rev.csv")
```

```
##          luna procent
## 1    ianuarie 1.67 %
## 2   februarie 1.44 %
## 3      martie 0.68 %
## 4     aprilie 1.38 %
## 5         mai  1.3 %
## 6       iunie  0.8 %
## 7       iulie  1.6 %
## 8      august 1.58 %
## 9  septembrie 0.77 %
## 10  octombrie 1.62 %
## 11  noiembrie 1.49 %
## 12  decembrie  0.8 %
```

Legea lui Benford

Incarcarea librariilor

```
library(ggplot2)
```

Graficul de reprezentare pentru legea lui Benford

```
compareBenford <- function(coloana){
  digits <- coloana[!is.na(coloana)]
  digits <- substr(stringr::str_extract(as.character(abs(digits)),
                                        pattern = "[^0\\.]"),1,1)
  digits <- factor(digits, levels = 1:9) # ensure all digits represented
  depth <- prop.table(table(digits))
  ben <- log10(1 + (1/(1:9)))
  dat2 <- data.frame(ben, depth)
  names(dat2) <- c("Benford","Digit",deparse(substitute(coloana)))
  dat2L <- reshape2::melt(dat2,id.vars="Digit", variable.name = "Type",
                          value.name = "Frequency")
  ggplot(dat2L, aes(x=Digit, y=Frequency, fill=Type)) +
  geom_bar(stat = "identity", position = "dodge")
}
```

3. Structura fisierului sursa administrativa

Camp	Lungime			Descriere	Sursa
COD	Numeric	13		Cod fiscal	TVA sau D112
CAEN01	Numeric	4		Cod CAEN luna xx	TVA
CAEN02	Numeric	4			
CAEN03	Numeric	4			
CAEN04	Numeric	4			
CAEN05	Numeric	4			
CAEN06	Numeric	4			
CAEN07	Numeric	4			
CAEN08	Numeric	4			
CAEN09	Numeric	4			
CAEN10	Numeric	4			
CAEN11	Numeric	4			
CAEN12	Numeric	4			
CA_01	Numeric	20		Cifra de afaceri luna xx	TVA
CA_02	Numeric	20			
CA_03	Numeric	20			
CA_04	Numeric	20			
CA_05	Numeric	20			
CA_06	Numeric	20			
CA_07	Numeric	20			
CA_08	Numeric	20			
CA_09	Numeric	20			
CA_10	Numeric	20			
CA_11	Numeric	20			
CA_12	Numeric	20			
R7_T01	Numeric	10		Venit brut luna xx	
R8_T01	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T01	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T01	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	
R12_T01	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T01	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T01	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T01	Numeric	10		TIMPUL luna xx	
R7_T02	Numeric	10		Venit brut luna xx	
R8_T02	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T02	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T02	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	D112

R12_T02	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T02	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T02	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T02	Numeric	10		TIMPUL luna xx	
R7_T03	Numeric	10		Venit brut luna xx	
R8_T03	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T03	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T03	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	
R12_T03	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T03	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T03	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T03	Numeric	10		TIMPUL luna xx	
R7_T04	Numeric	10		Venit brut luna xx	
R8_T04	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T04	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T04	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	
R12_T04	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T04	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T04	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T04	Numeric	10		TIMPUL luna xx	
R7_T05	Numeric	10		Venit brut luna xx	
R8_T05	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T05	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T05	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	

R12_T05	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T05	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T05	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T05	Numeric	10		TIMPUL luna xx	
R7_T06	Numeric	10		Venit brut luna xx	
R8_T06	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T06	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T06	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	
R12_T06	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T06	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T06	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T06	Numeric	10		TIMPUL luna xx	
R7_T07	Numeric	10		Venit brut luna xx	
R8_T07	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T07	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T07	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	
R12_T07	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T07	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T07	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T07	Numeric	10		TIMPUL luna xx	
R7_T08	Numeric	10		Venit brut luna xx	
R8_T08	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T08	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T08	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	

R12_T08	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T08	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T08	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T08	Numeric	10		TIMPUL luna xx	
R7_T09	Numeric	10		Venit brut luna xx	
R8_T09	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T09	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T09	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	
R12_T09	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T09	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T09	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T09	Numeric	10		TIMPUL luna xx	
R7_T10	Numeric	10		Venit brut luna xx	
R8_T10	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T10	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T10	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	
R12_T10	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T10	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T10	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T10	Numeric	10		TIMPUL luna xx	
R7_T11	Numeric	10		Venit brut luna xx	
R8_T11	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T11	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T11	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	

R12_T11	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T11	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T11	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T11	Numeric	10		TIMPUL luna xx	
R7_T12	Numeric	10		Venit brut luna xx	
R8_T12	Numeric	10		Contribuţia salariaţilor la bugetul asigurărilor pentru şomaj luna xx	
R9_T12	Numeric	10		Contribuţia individuală de asigurări sociale de stat	
R10_T12	Numeric	10		Contribuţia salariaţilor pentru asigurările sociale de sănătate luna xx	
R12_T12	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII (excl. salariaţii cu contract de muncă/ raport de serviciu suspendat) luna xx	
R13_T12	Numeric	6		EFECTIVUL SALARIAŢILOR LA SFÂRŞITUL LUNII CU CONTRACT DE MUNCĂ /RAPORT DE SERVICIU SUSPENDAT luna xx	
R14_T12	Numeric	10	2	NUMĂRUL MEDIU AL SALARIAŢILOR luna xx	
R16_T12	Numeric	10		TIMPUL luna xx	
DENI	Charact	250		Denumirea unitatii	
JUD	Numeric	2		Judetul	
FOJU	Numeric	2		Forma juridica	
FORMP	Numeric	2		Forma de proprietate	
TIP_UN	Numeric	2		Tipul unitatii	
FEL	Numeric	3		Fel	
STARE	Numeric	1		Stare unitate	
D_STARE	Date	8		Data starii	
END_DATE	Date	8		Data sfarsit activitate	
UP_DATE	Date	8		Data actualizare	
CAEN	Numeric	4		Cod CAEN	
OBCO	Numeric	4		Activitate prinicpala declarata la RECOM	REGIS
TIP_CONTR	Numeric	1		Tip contribuabil (mare, mic, mijlociu)	ANAF
T_R1	Numeric	1			
T_R2	Numeric	1			
T_R3	Numeric	1			
T_R4	Numeric	1			
T_R5	Numeric	1			
T_R6	Numeric	1		Marcaj de depunere TVA in luna xx	
T_R7	Numeric	1			
T_R8	Numeric	1			
T_R9	Numeric	1			
T_R10	Numeric	1			
T_R11	Numeric	1			
T_R12	Numeric	1			

DATA_I01	Date	8		Data depunerii TVA	
DATA_I02	Date	8			
DATA_I03	Date	8			
DATA_I04	Date	8			
DATA_I05	Date	8			
DATA_I06	Date	8			
DATA_I07	Date	8			
DATA_I08	Date	8			
DATA_I09	Date	8			
DATA_I10	Date	8			
DATA_I11	Date	8			
DATA_I12	Date	8			
D_R1	Numeric	1		Marcaj de depunere D112in luna xx	
D_R2	Numeric	1			
D_R3	Numeric	1			
D_R4	Numeric	1			
D_R5	Numeric	1			
D_R6	Numeric	1			
D_R7	Numeric	1			
D_R8	Numeric	1			
D_R9	Numeric	1			
D_R10	Numeric	1			
D_R11	Numeric	1			
D_R12	Numeric	1			
P_COD	Numeric	10		Cod fiscal parinte	
TIPI_A	Character	4		Tip ancheta UNICA	

SURSA
ESSnet ADMIN – Utilizarea surselor administrative de date în statistica întreprinderilor
Livrabil 2.4 – SGA II