UTILIZAREA ANALIZEI CLUSTER ÎN COMPARAȚII TERITORIALE

Babucea Ana-Gabriela, prof. univ. dr.

Universitatea "Constantin Brâncuşi" Târgu Jiu

Abstract: Cluster analysis classifies a set of observations into two or more mutually exclusive unknown groups based on combinations of interval variables and it has proven to be very useful. The classification aim is grouping the objects between their similarities and so providing a synthetic description or a cut of data.

The aim of this paper is using cluster analysis to classify the Romania's regions by monthly average net nominal wages by the activities of the national economy**.**

1. Introducere

Analiza comparativă a regiunilor ţării devine deci cu atât mai importantă cu cât se pune tot mai insistent problema globalizării, ori globalizarea poate fi privită ca manifestare şi la nivel naţional dacă avem în vedere dezvoltarea armonioasă a regiunilor şi reducerea diferenţelor dintre nivelurile de dezvoltare a acestora.

Politica de dezvoltare regională este un concept relativ nou pentru România. Începând cu 1998, ţara a fost structurată în 8 regiuni de dezvoltare, grupând cele 42 judeţe existente. Datele statistice arată că România a intrat în procesul de tranziţie având un nivel relativ scăzut al disparităţilor regionale, comparativ cu alte state. Aceste disparităţi însă au crescut rapid şi în mod deosebit între Bucureşti şi restul ţării.

2. Analiza cluster – descrierea metodei

Termenul de *analiză cluster* (cluster în limba engleză înseamnă mănunchi, ciorchine, grup) a fost utilizat pentru prima dată de către Tryon în 1939 şi se referă la o serie de algoritmi de clasificare care dau posibilitatea grupării unor obiecte (indivizi) în grupe omogene. Totuşi, utilizarea de algoritmi diferiţi are ca efect clasificări diferite.

Analiza de cluster, cunoscută şi ca analiza de segmentare sau de taxonomie are ca scop identificarea unui set de grupe omogene prin gruparea elementelor astfel încât să minimizeze variaţia în cadrul grupe şi să maximizeze variaţia dintre grupe. Analiza cluster este deci, o tehnică de analiză multivariată care cuprinde un număr de algoritmi de clasificare a unor obiecte (elemente sau indivizi) în grupe omogene. Variabilele sau cazurile sunt sortate în grupe (clusteri) astfel încât între membrii aceluiaşi cluster să existe asemănări, similitudini cât mai mari, iar între membrii unor clustere diferite să existe asemanări cât mai slabe. Pentru aceasta se are în vedere în primul rând alegerea distanţei dintre elemente, apoi alegerea algoritmului de grupare şi în final se decide cu privire la nivelul. În analiza de cluster există câteva noţiuni de bază ce se impun a fi amintite pe scurt: clusterul (grupa), distanţa dintre obiecte, distanţa dintre grupe şi algoritmul de cluster.

Distanţa este o funcţie definită pe mulţimea perechilor de obiecte cu ajutorul căreia se apreciază asemănarea sau diferenţele dintre elemente. Formarea grupelor se bazează pe calculul distanţei dintre oricare două obiecte. Distanţa dintre elemente se măsoară apelând la una dintre următoarele distanţe:

- Distanța euclidiană determinată ca rădăcină pătrată din suma pătratului distanțelor dintre x_i şi y_i:

$$d(x,y)=\sqrt{\sum_i (x_i-y_i)^2} \quad (1)$$

- Pătratul distanței euclidiene:

$$d(x,y)=\sum_i (x_i-y_i)^2 \quad (2)$$

- Distanța Chebychev (abaterea maximă):

$$d(x,y)=\max|x_i-y_i| \quad (3)$$

- Distanța City Block sau Manhattan (suma abaterilor) calculată ca diferență medie între dimensiuni:

$$d(x,y)=\sum_i |x_i-y_i| \quad (4)$$

- Distanța Minkovski (distanța euclidiană generalizată):

$$d(x,y)=\left(\sum_i |x_i-y_i|^p\right)^{\frac{1}{p}} \quad (5)$$

- Distanța Power:

$$d(x,y)=\left(\sum_i |x_i-y_i|^p\right)^{\frac{1}{r}} \quad (6)$$

unde p, r parametrii;

Determinarea acestor distanțe presupune un volum de calcul destul de mare şi prin urmare se apelează la prelucrarea automată a datelor cu pachete de programe specializate de tipul SYSTAT, STATISTICA sau SPSS.

Pe baza distanțelor calculate cu una dintre funcțiile distanță de mai sus se construieşte matricea distanțelor care este un tabel în care liniile sunt elementele, iar coloanele sunt variabilele analizate. Această matrice mai poartă numele şi de *matricea similitudinilor* dintre obiecte sau *matricea indicilor de proximitate*. Pasul următor priveşte alegerea algoritmului de grupare şi presupune alegerea unor reguli de determinare a distanțelor dintre clusteri. Există două categorii de metode generale de clustering ierarhice: separatorii şi aglomerative.

Tehnicile separatorii încep prin preluarea unui singur grup, fragmentând acel grup în subgrupe, acele subgrupe la rîndul lor în subgrupe şi aşa mai departe până când fiecare individ (obiect) formează propriul subgrup. *Tehnicile aglomerative* încep cu fiecare obiect reprezentând un grup şi apoi acestea se combină în grupe asemănătoare până când se ajunge la un singur grup. În orice caz, rezultatul aplicării acestor tehnici sunt cel mai bine

descrise de dendograme sau de arborii binari, unde obiectele sunt reprezentate ca noduri, iar ramurile indică grupurile conținând acel obiect.

Lungimea ramurii indică distanța dintre subgrupurile pe care le unește. Distanța dintre grupe poate fi calculată la rândul ei aplicând mai multe metode. Cele mai uzuale dintre metodele de calcul a distanțelor dintre grupe sunt:

- *metoda legăturii simple* sau a celui mai apropiat vecin (SINGLE sau nearest neighbor în SPSS/WIN) calculează distanța dintre două subgrupuri ca distanța minimă între oricare dintre doi membri ai subgrupurile distincte respective;
- *metoda legăturii totale* sau a celui mai îndepărtat vecin (COMPLETE sau furthest neighbor în SPSS/WIN) implică calculul distanțelor dintre grupe la fiecare pas ca maximul distanței dintre oricare două obiecte din grupe diferite;
- *metoda legăturii centrale* (Centroid method în SPSS/WIN) calculează distanțele dintre subgrupuri la fiecare pas ca medie a distanțelor dintre obiectele a două subgrupuri.
- *metoda BAVERAGE* (Between groups linkage în SPSS) care implică calculul mediei distanțelor dintre elementele celor două grupe;
- *metoda WAVERAGE* (Within-groups linkage în SPSS) care presupune alegerea acelei perichi de clusteri pentru care media distanțelor dintre elementele posibilului cluster reunit pentru fiecare pereche de clusteri existenți la acel moment este cea mai mică;
- *metoda WARD* (Ward's method în SPSS) în care se determină pentru fiecare cluster media fiecărei variabile, distanța dintre clusteri fiind determinată ca medie a distanțelor de la elementul mediu la toate elementele celuilalt cluster;
- *metoda MEDIAN* (Median clustering în SPSS) care presupune determinarea distanței dintre mediile corespunzătoare celor doi clusteri.

Dendograma este una dintre metodele de reprezentare a grupărilor putând furniza o sinteză cu privire la clasificare. O dendogramă care diferențiază clar grupele de obiecte va avea distanțe mici la ramurile mai îndepărtate ale arborelui şi diferențe mari la ramurile apropiate. Când distanțele dintre ramurile îndepărtate sunt mari comparativ cu ramurile apropiate, atunci gruparea nu este chiar eficace, iar dendograma va trebui interpretată cu prudență. Dendograma poate fi utilă şi pentru identificarea acelor obiecte care nu pot fi alăturate nici unui grup fiind excepții ale structurii de grupare şi care nu se alătură nici unui grup până la ultimul pas.

Analiza cluster presupune deci parcurgerea următoarelor etape:

- identificarea şi înregistrarea variabilelor semnificative în gruparea elementelor;
- calculul distanțelor dintre elemente şi determinatrea matricei similarităților;
- alegerea algoritmului de cluster pentru generarea grupelor şi interpretarea dendogramei.

3. Aplicație în analiza comparativă a regiunilor României

În continuare vom aplica algoritmul cluster pentru cele 8 regiuni ale României. Vom avea în vedere câştigul salarial nominal mediu lunar al populației civile ocupate în domeniile de activitate ale economiei naționale (agricultură, silvicultură, industrie, construcții, comerț, hoteluri şi restaurante, transporturi, poştă şi telecomunicații, activități bancare, financiare şi de asigurări, tranzacții imobiliare, administrație publică, învățământ, sănătate şi alte activitățişi ale economiei naționale, conform tabelului nr. 1) şi deci vom considera 14 variabile corespunzătoare.

Datele necesare analizei sunt prezentate în tabelul 1. Aplicând la analiza cluster vom analiza disparitățile între cele 7 regiuni ale țării şi Bucureştiul în funcție de variabilele alese cu privire câştigul salarial nominal mediu lunar din domeniile de activitate ale populației ocupate în ipoteza că fiecare regiune are funcție de aşezare geografică un anumit

specific.

Pentru construirea matricei similitudinilor ca măsură a distanței s-a folosit pătratul distanței euclidiene, iar ca algoritm de grupare metoda Ward apelându-se la pachetul de programe SPSS. Matricea similitudinilor are 8 linii şi 8 coloane.

În tabelul nr. 2 este prezentată această matrice, pe baza căreia se vor forma grupele.

În figura 1 se pot observa posibilitățile de grupare a celor 8 regiuni ale României dacă avem în vedere asemănările privind câştigul salarial nominal mediu lunar al populației civile ocupate în diverse domenii de activitate ale economiei naționale.

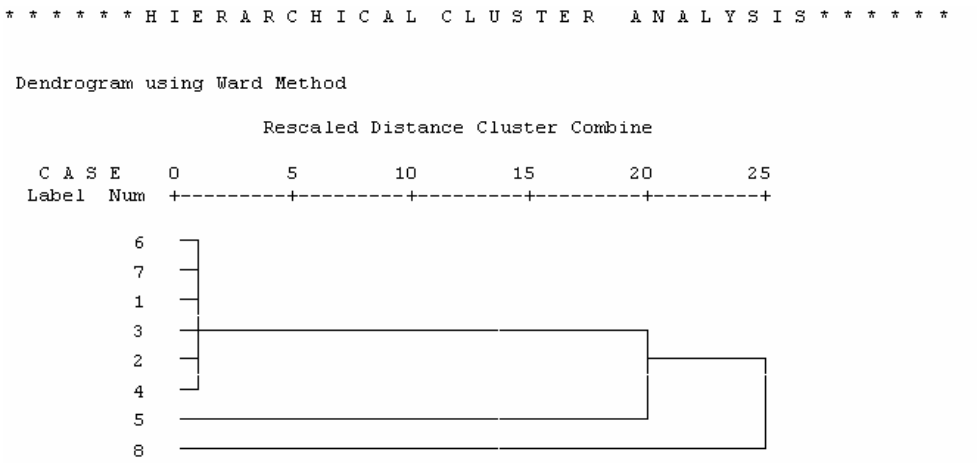

Fig. 1. Dendograma rezultată în urma utilizării algoritmului WARD

Se observă că Bucureştiul (regiunea 8 - județul Ilfov şi Municipiul București), precum şi regiunea Vest (5) sunt regiuni care la primul pas de grupare au rămas vizibil izolate fiind foarte diferite de toate celelate regiuni. Toate celelalte regiuni ale țării sunt asemănătoare formând împreună un cluster.

Fig. 2. Clasificarea regiunilor României după câştigul salarial nominal mediu lunar al populației civile ocupate în domeniile de activitate ale economiei naționale

Exceptând Bucureştiul, a cărui situație în peisajul economic al țării este complet specială, în capitală fiind atrase conform datelor statistice peste 50% din totalul investițiilor străine înregistrându-se de asemenea o substanțială migrație pozitivă şi regiunea de Vest a țării unde proximitatea piețelor vestice acționează ca factor de difuzare a creşterii

economice, celelalte regiuni ale țării sunt asemănătoare din punctul de vedere analizat. Disparități în dezvoltarea economică şi socială există şi la nivel intraregional existând încă un număr important de județe unde predomină un singur tip de activitate economică, de regulă în întreprinderi de Stat, care ocupă o pondere foarte mare a populației ocupate.

4. Concluzii

Algoritmii cluster sunt destul de complicați pentru a putea fi instrumentați manual, utilizarea pachetelor de programe specializate fac însă din analiza de cluster o metodă de clasificare la îndemâna nespecialistilor. Totuşi trebuie avut în vedere faptul că metoda de clasificare cluster va determina întotdeauna o grupare. Grupările rezultate se pot sau nu dovedi utile pentru clasificarea obiectelor. Dacă grupările fac diferență între variabilele nefolositoare grupării şi acele diferențieri sunt utile, atunci analiza de clustering este utilă.

Metodele analizei de cluster nefiind clar stabilite, existând mai multe opțiuni cu precădere atunci când se apelează la pachetele de programe statistice, aceastea fiind încă deschise criticii deseori găsindu-ne în situația de a fi nevoiți a încerca mai multe variante până când să ajungem la o structură a datelor convenabilă.

5. Bibliografie*:*

1. Anderberg, M. R*., Cluster Analysis for Applications, New York:* Academic Press, 1973;
2. Babucea, A. G., *Algoritmi de clasificare utilizând analiza de cluster şi SPSS/WIN,* Lucrările sesiunii internationalã de comunicări ştiințifice „Integrare europeană în contextul globalizării”, 2003, Piteşti, Editura AGIR, ISBN 973-8466-02-4, p.409-418;
3. Babucea, A. G., *Analiza Cluster în statistica teritorială,* Volumul Simpozionul stiintific international “Universitaria ROPET 2003”, Petrosani, pag.15-18, Editura Universitas, ISBN 973-8260-37-X;
4. Furtună, T. F., *Algoritmi de clasificare în statistica teritorială*, Revista de statistică, nr. 2/2002, p. 72-80;
5. Johnson, R. A., Wichern, D. W., *Applied Multivariate Statistical Analysis* (3rd Edition) New Jersey: Prentice Hall, 1992;

CÂŞTIGUL SALARIAL NOMINAL MEDIU NET LUNAR, PE ACTIVITĂŢI ALE ECONOMIEI NAŢIONALE - 2003

Lei/salariat

Tabelul nr. 1.

Nr. crt.	REGIUNEA Judeţul	Agricultură	Silvicultură, exploatare forestieră şi ec. vânatului	Industrie	Construcţii	Comerţ	Hoteluri şi restaurante	Transport şi depozitare	Poştă şi telecomunicaţii	Activităţi financiare, bancare şi de asigurări	Tranzacţii imobiliare şi alte servicii	Administraţie publică	Învăţământ	Sănătate şi asistenţă socială	Alte activităţi
1	**NORD-EST** Bacău, Botoşani, Iaşi, Neamţ, Suceava, Vaslui	4089955	2925432	4272847	3850006	2998849	2797509	6215158	10421624	3691188	6815313	4688286	4251360	3449052	4089955
2	**SUD-EST** Brăila, Buzău, Constanţa, Galaţi, Tulcea, Vrancea	3538806	2689193	5208600	4588863	3078410	3165376	6738499	9378970	3551002	6407259	4867400	3920127	3454903	3538806
3	**SUD** Argeş, Călăraşi, Dâmboviţa, Giurgiu, Ialomiţa, Prahova, Teleorman	3618340	3307289	5020467	3967495	3317994	2773906	6381718	10248871	4669265	6130259	4614418	3894763	3203082	3618340
4	**SUD-VEST** Dolj, Gorj, Mehedinţi, Olt, Vâlcea	3622616	3597827	5810986	4229799	3104724	2769780	6087832	9712919	3806872	6323225	5073307	4234666	3487388	3622616
5	**VEST** Arad, Caraş-Severin, Hunedoara, Timiş	4792965	3892431	3768107	5166499	4288496	3284106	2759832	6461385	9568871	4423103	6519912	4621452	3847251	4792965
6	**NORD-VEST** Bihor, Bistriţa-Năsăud, Cluj, Maramureş, Satu Mare, Sălaj	4068533	3818752	4343310	4213999	3404110	3127015	5797448	10562129	3931141	6698627	5028731	4120212	3820539	4068533
7	**CENTRU** Alba, Braşov, Covasna, Harghita, Mureş, Sibiu	3790137	3918079	4451884	4096393	3370432	3129332	6233418	10376759	4311670	6280891	4599443	4060908	3642012	3790137
8	**BUCUREŞTI** Ilfov, Municipiul Bucureşti	4086195	3254639	5169593	4514791	5236276	4662267	8252807	17223889	5530684	8835984	4680659	4624007	6015093	4086195

SURSA: Anuarul statistic 2004

Matricea distanţelor dintre regiuni (similitudinilor) - Proximity Matrix

Tabelul nr. 2.

Squared Euclidean Distance								
Case	**1 Nord-Est**	**2 Sud-Est**	**3 Sud**	**4 Sud-Vest**	**5 Vest**	**6 Nord-Vest**	**7 Centru**	**8 Bucureşti**
1 Nord-Est		3611778088960,000	2720023969792,000	4116611072000,000	76860176203776,000	1745267720192,000	2170529775616,000	74451051872256,000
2 Sud-Est	3611778088960,000		3359828344832,000	2231432642560,000	74778090143744,000	5267910557696,000	4446954192896,000	88262064472064,000
3 Sud	2720023969792,000	3359828344832,000		2385169612800,000	64887896145920,000	3026782781440,000	1280957743104,000	76447574130688,000
4 Sud-Vest	4116611072000,000	2231432642560,000	2385169612800,000		69004043485184,000	3707147386880,000	3192598036480,000	86032632512512,000
5 Vest	76860176203776,000	74778090143744,000	64887896145920,000	69004043485184,000		68055958814720,000	66015819988992,000	195959946477568,000
6 Nord-Vest	1745267720192,000	5267910557696,000	3026782781440,000	3707147386880,000	68055958814720,000		877543882752,000	69531284275200,000
7 Centru	2170529775616,000	4446954192896,000	1280957743104,000	3192598036480,000	66015819988992,000	877543882752,000		71980036390912,000
8 Bucureşti	74451051872256,000	88262064472064,000	76447574130688,000	86032632512512,000	195959946477568,000	69531284275200,000	71980036390912,000	