https://datascience-project.eu/

Training Fiche Template

Title	**Analiza Corespondențelor (AC)**
Keywords (meta tags)	AC, Variabilele calitative, inerția explicată, valorile proprii.
Language	Română
Objectives / Goals / Learnig outcomes	**Scopul acestui modul este de a introduce și explica tehnica Analizei Corespondențelor (AC).** **La sfârșitul acestui modul veți putea:** • **Cunoaște logica AC** • **Cunoaște cerințele de aplicare a AC** • **Conduce o AC** • **Conduce o AC în R cu pachetul FactoMineR.**

Curs de Specializare:	
Competențe in Știința Datelor	
Modul de vizualizare a datelor și analiză vizuală	X
Introducere în știința datelor pentru științele umane și sociale	
Data Science for good	
Jurnalism de date și storytelling	

Descriere	În acest modul de instruire vă va fi prezentată tehnica de analiză multidimensională numită Analiză a Corespondenței, AC. Analiza Corespondenței este o formă de scalare multidimensională, care construiește în esență un fel de model spațial care arată asocierile între un set de variabile categoriale. Dacă setul include doar două variabile, metoda este de obicei numită Analiză Simplă a Corespondenței (ASC). Dacă analiza implică mai mult de două variabile, atunci este de obicei numită Analiză Multiplă a Corespondenței (AMC). În acest modul ne vom ocupa de analiza corespondențelor simple, obiectivul acestei analize este de a reduce dimensionalitatea fenomenului investigat păstrând totuși informațiile conținute de acesta. Tehnica este aplicabilă la fenomene măsurate cu variabile calitative.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	Ultima parte a modulului va fi dedicată aplicării AC cu ajutorul software-ului R.
Conținut aranjat pe 3 nivele	**1. INTRODUCERE** Analiza corespondenței, AC, este o tehnică de analiză multidimensională care poate traduce aproape orice tip de tabel format din date numerice într-o formă grafică. Obiectul AC sunt matricele de contingenta, ale căror elemente indică de câte ori caracteristicile a două cantități diferite au fost detectate împreună. Scopul principal al AC este de a analiza relațiile dintre două variabile calitative observate într-un colectiv de unități statistice. Acest lucru se realizează prin identificarea unui spațiu "optimal", adică a unei dimensiuni reduse care reprezintă sinteza informațiilor structurale conținute în datele inițiale. Scopul analizei este de a evidenția împletirea legăturilor sau corespondențelor care există între datele examinate. **2. CERINȚE PENTRU ANALIZA DE CORESPONDENȚĂ** Pentru a efectua o analiză a corespondenței, este important să analizăm variabilele care urmează să fie utilizate și să înțelegem clar unele dintre caracteristicile lor. În mod specific, variabilele trebuie să îndeplinească următoarele cerințe: - *Variabilele trebuie să fie calitative:* Variabilele calitative nu sunt reprezentate de numere, ci de modalități sau categorii. De exemplu, genul, nivelul de educație, starea civilă etc. Aceste modalități trebuie să fie exhaustive și mutual exclusive. Mutual exclusive înseamnă că modalitățile variabilelor nu trebuie să conțină același tip de informație. De exemplu, pentru variabila "culoarea părului", nu se pot introduce modalitățile "păr închis" și "păr brun", deoarece "păr închis" înseamnă și "păr brun" și viceversa. ”Exhausitv” înseamnă că modalitățile unei variabile trebuie să acopere toate posibilitățile. De exemplu, pentru variabila "nivelul de educație", se introduc modalitățile "diplomă", "licență", "masterat". Aceste trei modalități nu acoperă toate posibilitățile de nivel de educație.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

- *Variabilele trebuie să fie interdependente:*

Înainte de a efectua analiza corespondențelor, este necesar să verificăm gradul de interdependență între cele două variabile luate în considerare, deoarece, în cazul în care acestea ar fi independente, analiza corespondențelor ar putea să nu aibă sens.

Pentru a face acest lucru, se realizează testul chi-pătrat:

H_0: cele două variabile sunt independente

H_1: cele două variabile nu sunt independente

Pentru a interpreta rezultatele testului, observăm valoarea p:

valoarea p < 0,05: ipoteza nulă este respinsă și, în consecință, variabilele sunt considerate a fi interdependente într-o anumită măsură.

3. Cum să efectuați AC

După ce ați verificat cerințele pentru AC, puteți trece la analiza efectivă.

3.1) Tabele de contingență

În analiza corespondențelor lucrăm cu tabele de contingenta, care conțin frecvențele conjuncte ale modalităților celor două variabile calitative X și Y. Aceste matrici sunt întotdeauna alcătuite din numere întregi niciodată negative care reprezintă numărul de apariții, adică înregistrări simple ale ceea ce s-a întâmplat. În plus, ambele variabile categoriale au un rol simetric în care toate elementele au aceeași natură.

$X\backslash Y$	y_1	y_2	y_3	
x_1				
x_2		$n_{i,j}$		$n_{i.}$
x_3				
		$n_{.j}$		n

X, Y sunt variabilele categoriale.

x_1, x_2 , x_3 : sunt variantele varabilei X

y_1, y_2 , y_3 : sunt variantele varabilei Y

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

$n_{i,j}$: sunt frecvențele absolute comune, adică frecvențele perechilor, de exemplu: $n_{1,1}$: $X = x_1; Y = y_1$

$n_{i\cdot}$: sunt frecvențele marginale pe rânduri: $n_{i.} = \sum_{j=1}^{C} n_{i,j}$

$n_{\cdot j}$: sunt frecvențele marginale pe coloane: $n_{.j} = \sum_{i=1}^{R} n_{i,j}$

Acestea reprezintă suma pentru rândul (sau coloana) fixat al frecvențelor conjuncte pe modalitățile lui ***Y*** (pentru coloane pe modalitățile lui ***X***).

n = volumul eșantionului, care poate fi obținut prin adunarea frecvențelor marginale de rând sau de coloană:

$$n = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{i,j} \quad \forall\, i, j$$

Puteți trece de la frecvențele absolute la frecvențele relative prin împărțirea fiecărei frecvențe absolute la n.: $f_{i,j} = \frac{n_{i,j}}{n}$

3.2) Row Profile Matrix and Column Profile Matrix

Matricea profilurilor de rând este obținută prin împărțirea frecvențelor absolute (sau frecvențelor relative) la frecvențele marginale de rând corespunzătoare. Prin urmare:

$$\frac{n_{i,j}}{n_i} = \frac{f_{i,j}}{f_{i.}} \quad \forall\, i, j$$

Tabelul de contingență devine:

$\frac{f_{i,j}}{f_{i.}} = \frac{n_{i,j}}{n_{i.}}$	1 1 1
profilo medio	1

Pe marginea de rând avem întotdeauna 1 și aceasta reprezintă suma profilurilor de rând.

Pe marginea coloanei se găsesc profilurile medii, care sunt obținute prin adunarea frecvențelor relative pe coloană; sau prin calcularea mediei elementelor matricei profilului de rând, pe coloană. Aceasta

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

este o medie ponderată, în care ponderile sunt reprezentate de frecvențele marginale de rând $f_{i.}$.

Lucrând cu frecvențe, se pierde o dimensiune, astfel încât spațiul de rând este reprezentat de un spațiu cu C-1 dimensiuni;

Se poate construi o matrice diagonală a frecvențelor marginale de rând $\boldsymbol{D_R}$, care are profilurile de rând pe diagonala principală. Matricea diagonală a frecvențelor marginale de rând este o matrice R · R, care are dimensiuni egale cu numărul de rânduri și pe diagonala principală conține frecvențele marginale de rând a tabelului de frecvențe relative. O matrice diagonală este o matrice ale cărei elemente generice de pe diagonala principală sunt marginalii de rând, iar deasupra sau dedesubtul acesteia există doar zero-uri. Este întotdeauna o matrice simetrică și pătrată. Cu matricea diagonală a marginalilor de rând se poate construi matricea profilurilor de rând: aceasta se obține prin împărțirea frecvențelor relative la marginalii de rând $\frac{\boldsymbol{F}}{\boldsymbol{D_R}}$. Dimensiunile matricei $\boldsymbol{F}$ sunt R · C, în timp ce matricea $\boldsymbol{D_R}$ are dimensiunea R · R. Deoarece nu se poate face împărțirea directă între matrice, se calculează inversa matricei $\boldsymbol{D_R}$ și se înmulțește cu matricea $\boldsymbol{F}$, rezolvând astfel problema dimensionalității. : $\boldsymbol{D_R^{-1}} \cdot \boldsymbol{F}$.

Același lucru se aplică și pentru coloane, cu unele mici diferențe. Matricea profilurilor de coloană este construită prin împărțirea frecvențelor absolute la marginalii relative de coloană:

$$\frac{n_{i,j}}{n_{.j}} = \frac{f_{i,j}}{f_{.j}} \qquad \forall\, i,j$$

Tabelul de contingență devine:

$\frac{f_{i,j}}{f_{.j}} = \frac{n_{i,j}}{n_{.j}}$			profilo medio
1	1	1	1

În acest caz, pe marginea coloanei veți avea întotdeauna 1, iar pe marginea de rând veți avea profilul mediu al coloanei. În acest caz,

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

ponderile sunt reprezentate de marginalii de coloană f_(.j). Evident, chiar și în spațiul de coloană se lucrează cu mai puțin de o dimensiune, deci spațiul de coloană este R-1. Se poate construi o matrice diagonală a marginalilor de coloană $\boldsymbol{D_C}$, care are profilurile de coloană pe diagonala principală. Matricea diagonală a marginalilor de coloană este o matrice C·C, care are dimensiuni egale cu numărul de coloane și pe diagonala principală conține marginalii de coloană ai tabelului de frecvențe relative. O matrice diagonală este o matrice ale cărei elemente generice de pe diagonala principală sunt marginalii de coloană, iar deasupra sau dedesubtul acesteia există doar zero-uri. Este întotdeauna o matrice simetrică și pătrată.

Cu matricea diagonală a marginalilor de coloană se poate construi **matricea profilurilor de coloană**: aceasta se obține prin împărțirea frecvențelor relative la marginalii de coloană $\frac{\boldsymbol{F}}{\boldsymbol{D_R}}$. Dimensiunile matricei F sunt R · C, în timp ce matricea $\boldsymbol{D_C}$ are dimensiunea C · C. Deoarece împărțirea între matrici nu poate fi efectuată direct, se calculează inversa matricei $\boldsymbol{D_C}$ și se înmulțește cu F prin post-înmulțire, rezolvând astfel problema dimensionalității : $\boldsymbol{F \cdot D_C^{-1}}$.

3.3) Distanțele

În analiza corespondențelor este necesar să înțelegem ce distanță există între valorile respective, în scopul de a înțelege dacă modalitățile sunt apropiate sau îndepărtate una de cealaltă și, prin urmare, dacă se aseamănă sau nu. Acest lucru poate fi realizat prin observarea frecvențelor: cu cât sunt mai mici, cu atât sunt mai apropiate, și viceversa. Există diferite metode de calcul al distanței, cum ar fi **distanța euclidiană și distanța chi-pătrat.**

Distanța euclidiană este cea mai simplă și acordă mai multă importanță diferențelor mari în defavoarea celor mici. Se calculează prin diferența dintre frecvențele relative și ridicarea lor la pătrat.

Pentru profilurile de rând:

$$d_{(i,i')} = \sqrt{\sum_{j=1}^{C} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}} \right)^2}$$

Pentru profilurile de coloană:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

$$d_{(j,j')} = \sqrt{\sum_{i=1}^{R} \left(\frac{f_{i,j}}{f_{.j}} - \frac{f_{i,j'}}{f_{.j'}} \right)^2}$$

Distanța chi-pătrat recompensează distanțele mai mici deoarece frecvențele cu valori mici sunt reponderate în raport cu rândurile, introducând în formulă inversul frecvențelor marginale de coloană (respectiv inversul frecvențelor marginale de rând). Dezavantajul distanței chi-pătrat este că valoarea reciprocă a frecvențelor marginale de coloană (sau de rând) poate tinde către zero și, prin urmare, o singură valoare poate contribui excesiv la calculul distanței.

3.4) Spațiul rândurilor și spațiul coloanelor

În spațiul rândurilor cele două componente sunt:

- Profile de rând: $\mathbf{D_R^{-1}} \cdot \mathbf{F}$
- Metrica: $\mathbf{D_C^{-1}}$

Să începem cu următoarea formulă:

$$\boldsymbol{\Psi_{n\times 1} = X_{n\times p} \cdot u_{p\times 1}}$$

După înlocuirile adecvate rezultă:

$$\boldsymbol{\Psi = D_R^{-1} \cdot F \cdot D_C^{-1} \cdot u}$$

Obiectivul analizei corespondențelor constă în găsirea setului de axe unitare care permite maximizarea distanțelor dintre proiecțiile profilurilor de rând. Trebuie, așadar, să căutăm acei vectori care maximizează proiecțiile. Deoarece vectorii ***u*** pot fi infiniți, se adaugă restricția normei unitare.

$$\boldsymbol{u^T \cdot D_C^{-1} \cdot u = 1}$$

Problema de maximizare: Se maximează inerția explicată (variația explicată), care corespunde variabilității pentru variabilele cantitative.

$$\begin{cases} \text{MAX: } \left\{ \hat{\psi}^T D_R \hat{\psi} \right\} \\ v^T D_C^{-1} v = 1 \end{cases}$$

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Pentru a rezolva problema de maximizare cu restricții, utilizați metoda multiplicatorilor Lagrange:

$$\mathcal{L}(v, \lambda) = (\hat{\psi}^T D_R \hat{\psi}) - \lambda(v^T D_C^{-1} v - 1)$$

$\boldsymbol{\lambda}$= multiplicatorul Lagrange, care este un scalar;

***u*=** vectorul ponderilor, pe care dorim să îl determinăm

Prin efectuarea înlocuirilor necesare, vom avea:

$$\mathcal{L}(v, \lambda) = (D_R^{-1} F D_C^{-1} v)^T D_R (D_R^{-1} F D_C^{-1} v) - \lambda(v^T D_C^{-1} v - 1)$$

Efectuăm operațiile de transpunere, înlocuim matricea identitate I cu o $\boldsymbol{D_R \cdot D_R^{-1}}$ și $[(-\boldsymbol{\lambda}) \cdot (-1)]$ o înlocuim cu λ. Putem apoi elimina transpusa din matricile diagonale $\mathbf{D_C^{-1}}$ și $\mathbf{D_R^{-1}}$, deoarece transpusa unei matrice diagonale nu se modifică.

Rezultă:

$$\mathcal{L}(v, \lambda) = v^T D_C^{-1} F^T D_R^{-1} F D_C^{-1} v - \lambda v^T D_C^{-1} v + \lambda$$

Calculăm derivatele parțiale, derivând Lagrange-anul în raport cu ***u*** și le egalăm cu 0:

$$\frac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow \frac{\partial \mathcal{L}}{\partial v} = 2F^T D_R^{-1} F D_C^{-1} v - 2\lambda v = 0$$

Se înmulțește ecuația cu$\mathbf{D_C^{-1}}$:

$$F^T D_R^{-1} F D_C^{-1} v = \lambda v$$

Dacă înlocuim transpunerea profilelor de rând și matricea profilelor de coloană cu ***S,*** putem scrie ecuația caracteristică sub forma:

$$Sv = \lambda v$$

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Maximizarea inerției explicate a profilelor de rânduri este echivalentă cu descompunerea acestei matrice în valori proprii și vectori proprii ai acesteia. Prima valoare proprie este asociată cu primul vector propriu care explică inerția maximă. Vectorii proprii care se extrag ulterior vor fi extrași ortogonal, aplicând constrângerea de ortogonalitate:

$$\boldsymbol{u_1^T \cdot D_C^{-1} \cdot u_2 = 0}$$

Utilizăm constrângerea de ortogonalitate pentru a putea alege cea de-a doua componentă care va explica inerția care nu este explicată de prima componentă. Evident, prima componentă extrasă explică inerția maximă, adică alungirea maximă a norului de puncte.

În **spațiul coloanelor** sunt două componente:

- Profilul de coloana: $\mathbf{F \cdot D_C^{-1}}$
- Metrica: $\mathbf{D_R^{-1}}$

Începem cu formula:

$$\boldsymbol{\varphi_{p\times 1} = \left(X^T_{n\times p}\right)_{p\times n} \cdot v_{n\times 1}}$$

După înlocuire, se obține:

$$\boldsymbol{\varphi = D_C^{-1} F^T D_R^{-1} v}$$

Problema de maximizare care trebuie rezolvată cu multiplicatori Lagrange este:

$$\begin{cases} \text{MAX: } \{\hat{\varphi}^T D_C \hat{\varphi}\} \\ \nu^T D_R^{-1} \nu = 1 \end{cases}$$

Procedând ca în spațiul rândurilor, în final vom obține:

$$F D_C^{-1} F^T D_R^{-1} \nu = \mu \nu$$

Co-funded by the Erasmus+ Programme of the European Union

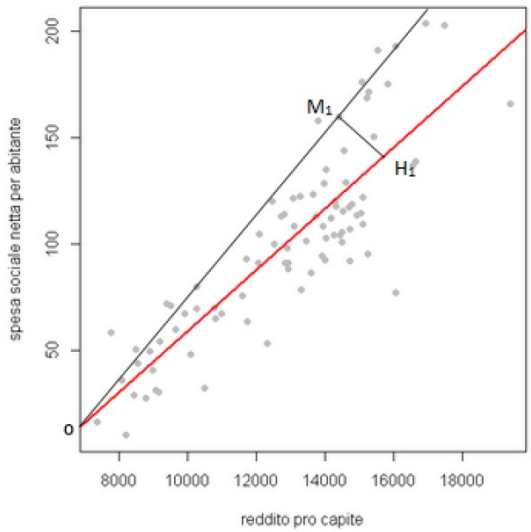
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Înlocuind matricea profilelor de coloană și metrica transpusă a profilelor de rând cu ***S**** se obține ecuația caracteristică:

$$S^*\nu = \mu\nu$$

Maximizarea geometrică a inerției explicate, și anume, realizarea unei cantități cât mai mici de informații pierdute și a unei cantități cât mai mari de informații observate, va fi următoarea: distanța M_1H_1să fie cât mai mică și distanța OH_1 cât mai mare.

Figura 1.3: Diagramma di dispersione



Prin urmare, trebuie să găsim dreapta f (în roșu) care interpolează punctele din spațiul vectorial, astfel încât distanța dintre toate punctele din spațiu și punctele proiectate ortogonal pe dreapta f să fie cea mai mică posibilă.

Valorile proprii în spațiul liniilor corespund vectorilor proprii în spațiul coloanelor, astfel încât valorile proprii ale lui ***S*** corespund celor ale lui ***S*.*** Vectorii proprii sunt egali între ei, cu excepția unei constante. Astfel, atunci când trebuie să maximizăm, nu trebuie să descompunem în valori proprii și vectori proprii **S** și ***S*,*** ci doar să o facem cu unul singur. Cantitatea de inerție explicată este egală indiferent dacă

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

calculăm ***S*** sau ***S****,* relația dintre cele două spații este reprezentată de formulele de tranziție:

$$S \to \nu = \frac{1}{\sqrt{\lambda}} F D_C^{-1} \upsilon \equiv S^* \to \upsilon = \frac{1}{\sqrt{\lambda}} F' D_R^{-1} \nu$$

Spațiul rândurilor:

$$\hat{\psi} = D_C^{-1} \upsilon$$

Cu:

$$\upsilon = \frac{1}{\sqrt{\lambda}} F' D_R^{-1} \nu$$

După efectuarea substituțiilor corespunzătoare, se obține:

$$\frac{1}{\sqrt{\lambda}} D_C^{-1} F' D_R^{-1} \upsilon \to \frac{1}{\sqrt{\lambda}} D_C^{-1} F' \hat{\psi}$$

Rezultă:

$$\sqrt{\lambda}\hat{\psi} = D_C^{-1} F' \hat{\psi} \to \hat{\psi} = \frac{1}{\sqrt{\lambda}} D_C^{-1} F \hat{\psi} \to \sqrt{\lambda}\hat{\psi} = D_C^{-1} F \hat{\psi}$$

Pentru spațiul rândurilor:

$$\sqrt{\lambda}\hat{\psi} = D_C^{-1} F \hat{\psi} \equiv \hat{\psi} = \sqrt{\lambda}\hat{\psi}$$

Spațiul coloanelor:

$$\hat{\psi} = D_R^{-1} \nu$$

Unde:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

$$\nu = \frac{1}{\sqrt{\lambda}} F D_C^{-1} v$$

După efectuarea substituțiilor corespunzătoare, se obține:

$$\frac{1}{\sqrt{\lambda}} D_R^{-1} F D_C^{-1} v \rightarrow \frac{1}{\sqrt{\lambda}} D_R^{-1} F \hat{\psi}$$

Rezultă:

$$\frac{1}{\sqrt{\lambda}} D_R^{-1} F \hat{\psi} \rightarrow \sqrt{\lambda} \hat{\psi} \rightarrow D_R^{-1} F \hat{\psi}$$

Pentru spațiul coloanelor:

$$\sqrt{\lambda} \hat{\psi} = D_R^{-1} F \hat{\psi} \equiv \hat{\psi} = \sqrt{\lambda} \hat{\psi}$$

4) Exemplu cu R

Verificarea unei posibile relații între distribuția animalelor și diferitele regiuni italiene. Datele se referă la anul 2011, colectate de către băncile disponibile pe site-ul Istat.

Ipoteză: diferitele regiuni, în funcție de caracteristicile teritoriale și de nevoile populației, aleg să crească anumite capete de bovine mai degrabă decât altele.

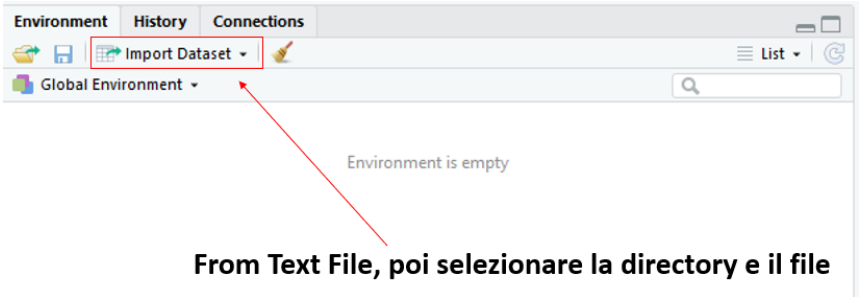
Setul de date:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Regione	Bovini	Ovini	Caprini	Equini	Suini	Conigli	Totale
Piemonte	23516	2303	3418	2370	2429	1392	35428
Valle d'Aosta	1585	347	284	53	16	11	2296
Liguria	1642	1126	549	949	258	924	5448
Lombardia	15480	2592	3175	3647	4346	1191	30431
Trentino Alto Adige	10482	2279	2424	1513	3292	266	20256
Veneto	16007	1642	1207	2429	3634	1907	26826
Friuli-Venezia Giulia	1539	83	207	280	1477	117	3703
Emilia-Romagna	8522	1315	908	3161	1541	308	15755
Toscana	4392	4918	607	2163	2046	1764	15890
Umbria	3132	2734	667	1245	4107	1924	13809
Marche	2940	1877	342	383	7103	1786	14431
Lazio	9256	8678	1624	3535	6849	4269	34211
Abruzzo	5588	6590	1710	1362	10241	2450	27941
Molise	2976	2510	610	534	3943	60	10633
Campania	10971	6248	3675	1448	15145	6708	44195
Puglia	3010	1918	826	691	759	921	8125
Basilicata	3156	7426	3562	1280	6137	2606	24167
Calabria	5496	3701	3505	1839	21522	2087	38150
Sicilia	7387	4963	1088	1930	821	63	16252
Sardegna	8200	12880	3171	3333	9324	523	37431
Totale	145277	76130	33559	34145	104990	31277	425378

Importăm setul de date:

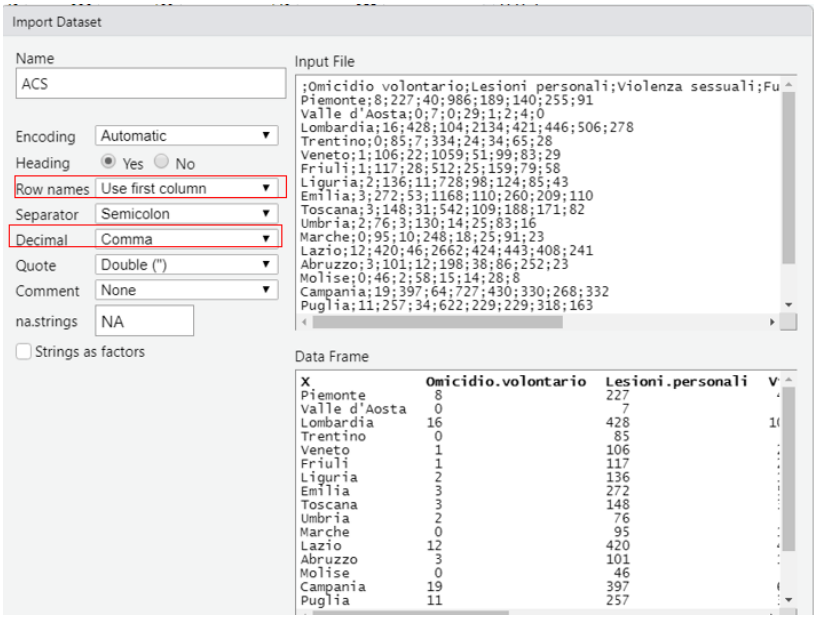


În câmpul "Row names" (Nume rânduri), selectați formularea: "use first column" (utilizați prima coloană) pentru a avea etichetele atât ale indivizilor, cât și ale variabilelor pe grafice. În câmpul zecimale,

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

selectăm " comma " (virgulă).



With the command:

X<-as.matrix(nome_del_dataset)

Atribuim lui X, ca obiect, setul de date utilizat în analiză.

Înainte de a putea efectua AC este necesar să se stabilească gradul de interdependență dintre cele două personaje luate în considerare, aceasta deoarece în cazul în care acestea sunt independente s-ar putea să nu aibă sens să se continue AC. Pentru a verifica acest lucru, efectuăm testul chi-pătrat.

Comanda este următoarea:

chiquadro<-chisq.test(X)

```
        Pearson's Chi-squared test

data:  X
X-squared = 126691.2, df = 95, p-value < 2.2e-16
```

Se poate observa că valoarea p este mai mică decât nivelul de semnificație cel mai frecvent utilizat, și anume 0,05. Prin urmare, putem respinge ipoteza nulă de independență statistică între cele două variabile și putem continua analiza.

Co-funded by the Erasmus+ Programme of the European Union
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Acum dorim să creăm o matrice de frecvențe relative F.

Calculăm numărul de eșantioane, cu comanda:

n<-sum(X)

și apoi împărțind matricea de pornire (deci toate frecvențele comune) la numărul de eșantioane se obține matricea F. Comandă:

F<-X/n

Următorul pas este obținerea tabelelor de profiluri de rânduri și coloane. Pentru a face acest lucru, în primul rând, este necesar să se calculeze marginalele rândului și ale coloanei. Respectiv, comenzile sunt:
sumrow<-apply(F,1,sum)
sumcol<-apply(F,2,sum)

Apoi calculăm matricea diagonală a frecvențelor marginale de rând și inversa acesteia cu ajutorul comenzilor::
Dr<-diag(sumrow)
Dr_inv<-solve(Dr)

Acum putem calcula profilurile rândurilor. În termeni matriciali, înmulțim în prealabil inversa matricei diagonale a rândului marginal cu matricea frecvențelor relative. Comanda care trebuie folosită este:
Pr<-Dr_inv%*%F

Același lucru pentru profilurile coloanelor, ținând cont de faptul că, în acest caz, inversa matricei coloanelor trebuie să fie post-multiplicată la matricea frecvențelor relative.

Dc<-diag(sumcol)
Dc_inv<-solve(Dc)
Pc<-F%*%Dc_inv

Acum putem calcula distanțele dintre puncte. După cum am menționat deja, există două tipuri de distanțe: Euclidiană și Chi-pătrat.

Profilele rândurilor de distanțe euclidiene sunt:

d_euc_r<-dist(rbind(Pr[1,],Pr[2,]))

Profilele coloanelor de distanțe euclidiene sunt:

d_euc_c<-dist(rbind(Pr[,1],Pr[,2]))

Co-funded by the Erasmus+ Programme of the European Union

Profilele rândurilor de distanțe chi-pătrat sunt:

```
d_r<-pr[1,]-pr[2,]
d<-d_r^2/sumcol
d_chi_r<-sqrt(sum(d))
```

Profilele coloanelor de distanțe chi-pătrat sunt:

```
dc<-Pr[,1]-Pr[,2]
dc<-dc^2/sumrow
d_chi_c<-sqrt(sum(dc))
```

Ecuația caracteristică a matricei profilului de rânduri:

```
S<-t(Pr)%*%Pc
```

Deoarece matricea ***S*** nu este simetrică, este necesar să o diagonalizăm pentru a obține **S_tilde:**

```
A<-t(F)%*%Dr_inv%*%F #simmetria

Dc_12<-diag(sumcol^(-1/2))

S_tilde<-Dc_12%*%A%*%Dc_12
```

Acum trebuie să maximizăm inerția explicată prin descompunerea matricei în valori proprii și vectori proprii:

```
AC<-eigen(S_tilde)

lambda<-as.matrix(AC$values)

lambda<-lambda[-1,]

w<-AC$vectors

u<-Dc^(1/2)%*%w

u<-u[,-1]
```

Ecuația caracteristică a matricei profilului de coloane:

```
S_star<-F%*%Dc_inv%*%t(F)%*%Dr_inv
```

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Pentru a ne deplasa de la u la v, folosim formule de tranziție (deoarece cantitatea de inerție explicată este egală atât în spațiul rândurilor, cât și în cel al coloanelor).

sq_lambda<-diag((sqrt(lambda))^(-1))

v<-F%*%Dc_inv%*%u%*%sq_lambda

Calculăm factorii și coordonatele, mai întâi spațiul rândurilor și apoi al coloanelor:

fp_r<-Dc_inv%*%u

fp_c<-Dr_inv%*%v

PHI_coord<-Dc_inv%*%t(F)%*%fp_c

PSI_coord<-Dr_inv%*%F%*%fp_r

Se afișează graficul coordonatelor principale:

PRINCOORD<-rbind(PSI_coord,PHI_coord)

rows<-row.names(X);columns<-colnames(X)

plot(PRINCOORD[,1],PRINCOORD[,2],type="n",main="Main Coordinates",xlab="Axis1",ylab="Axis2")+ text(PRINCOORD[1:20,1],PRINCOORD[1:20,2],labels=rows,col="spring green4")

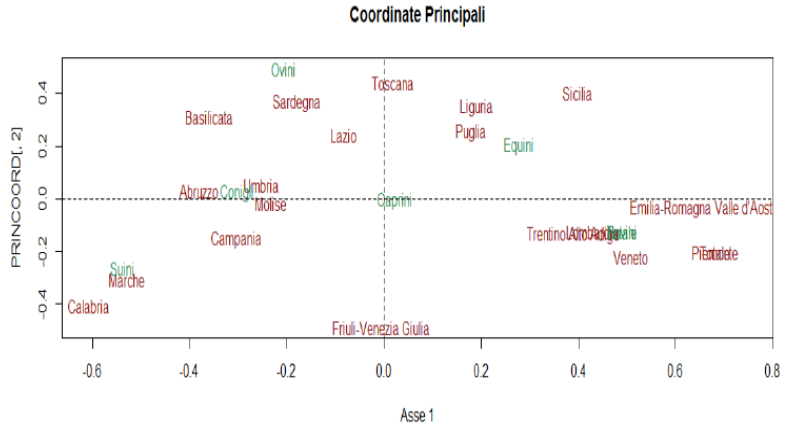
text(PRINCOORD[21:29,1],PRINCOORD[21:29,2],labels=columns,col="violetred")

abline(h=0,v=0,lty=2,lwd=1.5)

Rezultă:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Dacă ne uităm la acest grafic, putem spune, de exemplu, că în regiuni precum Abruzzo, Molise, Umbria se cresc în principal iepuri.

Alegem componentele:

inertia<-sum(diag(S))-1

sum(lambda)

in_exp<-lambda/inertia

in_exp_<-cumsum(in_exp)

Vizualizăm rezultatele obținute:

```
> inerzia
[1] 0.2978321
> in_exp
[1] 0.58571295 0.23305781 0.10382933 0.04875445 0.02864546
> in_exp_cum
[1] 0.5857130 0.8187708 0.9226001 0.9713545 1.0000000
```

Prima dimensiune explică singură 58,57% din variabilitate, iar primele trei împreună explică 92,26% din variabilitatea globală a datelor.

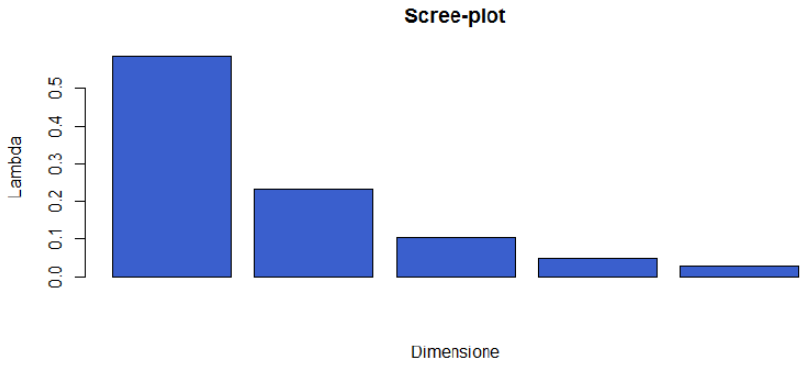
Rezultatele obținute pot fi afișate grafic cu ajutorul graficului **scree-plot** a inerției explicate:

screeplot<-barplot(in_exp,main="Scree-plot inertia", xlab="Size", ylab="Lambda", col="lightblue")

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Figura 1.10: Scree-plot dell'inerzia spiegata



Pentru calitatea reprezentării:

- pentru a evalua cât de mult influențează sau participă un mod la axa factorială, se calculează **contribuțiile absolute, CA**, atât pentru rânduri, cât și pentru coloane:

ca_r<-Dr%*%fp_c^2

ca_c<-DC%*%fp_r^2

- -Pentru a evalua calitatea reprezentării, calculăm **contribuțiile relative, CR**. Acestea oferă o măsură mai bună a reprezentării punctelor pe axe și este dată de cosinusul unghiului format de vectorul de proiecție al punctului și vectorul relativ i (sau j) în punctul i (sau j) din spațiul său original:

G<-matrix(sumcol,20,9,byrow=T)

di<-(Pr-G)^2%*%Dc_inv

d_ig<-apply(di,1,sum)

cos2r<-PSI_coord^2/d_ig

H<-matrix(sumrow,20,9)

dj<-Dr_inv%*%(Pc-H)^2

d_jh<-apply(dj,2,sum)

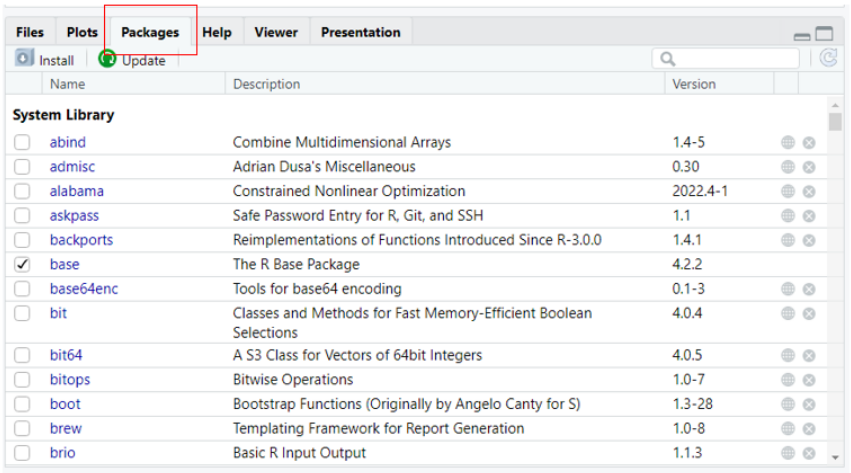
Co-funded by the Erasmus+ Programme of the European Union

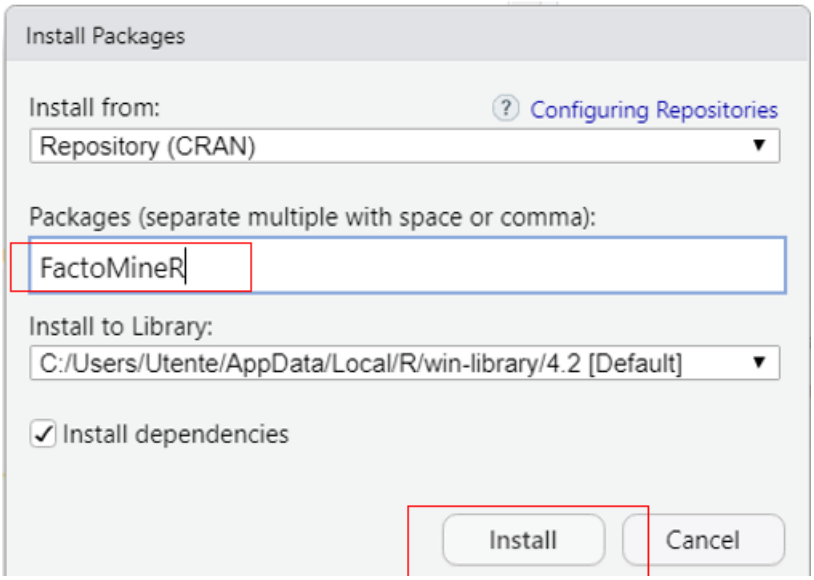
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

COS2C<-PHI_coord^2/d_jh

R oferă un pachet numit **FactoMineR** pentru analiza corespondențelor, care adaugă informații despre indivizi și variabile și vă permite să creați un grafic bidimensional comun al indivizilor și variabilelor.

Pe R pentru a putea utiliza acest pachet trebuie mai întâi să îl descărcați:



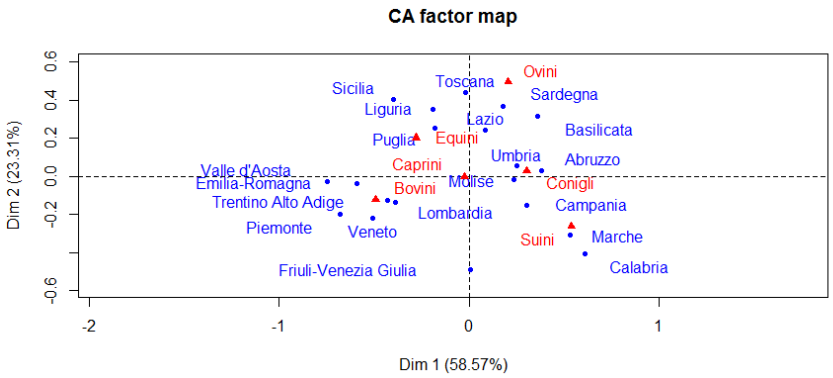


După ce l-ați instalat, trebuie să îl apelați cu comanda

library(FactoMineR)

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	Să trecem la crearea graficului bidimensional Persoane și variabile: **CA(X, ncp = 5, row.sup = NULL, col.sup = NULL, quanti.sup=NULL, quali.sup = NULL, graph = TRUE, axes = c(1,2), row.w = NULL)** Grafic, obținem:  Interpretarea rezultatelor: Putem spune că ipoteza inițială este confirmată. În special, regiunile cele mai dedicate creșterii ovinelor par a fi Toscana, Sardinia și Basilicata, iar acest lucru poate fi explicat prin faptul că aceste regiuni sunt zone de munte și de transhumanță. Caii sunt crescuți mai ales în Puglia, Liguria și Sicilia, deoarece aceste animale au fost întotdeauna folosite pentru muncă în mediul rural. Bovinele sunt prezente în Trentino Alto-Adige, Veneto, Piemont, Lombardia și Emilia-Romagna; de fapt, aceste regiuni au o tradiție de creștere mai dezvoltată pentru uz alimentar. Iepurii apar mai ales în Umbria, Abruzzo și Molise. În schimb, porcii par să fie crescuți mai mult în Marche, Campania și Molise; Aceste regiuni au, de asemenea, o tradiție de creștere mai dezvoltată pentru uz alimentar. Caprele, pe de altă parte, sunt plasate la mijlocul axei, probabil pentru că nu există regiuni care să prefere creșterea lor.
Autoevaluare (întrebări și răspunsuri cu alegere multiplă)	1. Ce fac formulele de tranziție? A) Comută între spații **B) Trec de la reprezentarea contribuțiilor absolute la cea a contribuțiilor conexe** C) Trec de la matricea frecvențelor relative la cea a profilurilor

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	2. De ce se face testul chi pătrat înainte de a implementa AC? **A) Pentru a verifica dacă variabilele sunt cantitative** B) Pentru a evalua dacă variabilele sunt calitative C) Pentru a analiza existența unei interdependențe între cele două variabile 3. Care este scopul analizei corespondenței? A) Maximizarea variabilității explicate B) Maximizarea inerției explicate **C) Minimizarea inerției explicate**
Resurse (videoclipuri, link de referință)	
Materiale conexe	
În legătură cu PPT	
Bibliografie	van der Heijden, P. G. M. & de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis, Psychometrika, 50, pp. 429-447. Le, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. 25(1). pp. 1-18. Mineo, A. M. (2003). Una Guida all'utilizzo dell'Ambiente Statistico R, http://cran.r-project.org/doc/contrib/Mineo-dispensaR.pdf.
Furnizat de	[Unisalento]

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.