https://datascience-project.eu/

Training Fiche Template

Title	Analisi delle Corrispondenze, AC
Keywords (meta tag)	AC, Variabili Qualitative, Inerzia Spiegata, Autovalori
Language	Italiano
Objectives / Goals / Learnig outcomes	**Obiettivo di questo modulo è introdurre e spiegare la tecnica dell'Analisi delle Componenti Principali.** **Alla fine di questo modulo sarai in grado di:** **- Conoscere la logica dell'AC** **- Conoscere i requisiti** **- Condurre un AC** **- Condurre un AC in R con il pacchetto FactoMineR**

Training course:	
Data Science Literacy	
Data Visualisation and Visual Analytics Module	X
Introduction to Data science for Human & Social Sciences	
Data Science for good	
Data Journalism and Storytelling	

Description	In questo modulo formativo verrà presentata la tecnica di analisi multidimensionale denominata Analisi delle Corrispondenze, AC. L'Analisi delle Corrispondenze è una forma di scaling multidimensionale, che essenzialmente costruisce una sorta di modello spaziale che mostra le associazioni tra un insieme di variabili categoriali. Se l'insieme include solo due variabili, il metodo è usualmente chiamato Analisi delle Corrispondenze Semplici (SCA). Se l'analisi coinvolge più di due variabili, allora è usualmente chiamata Analisi delle Corrispondenze Multiple (MCA). In questo modulo si tratterà l'analisi delle corrispondenze semplici, l'obiettivo di tale analisi è quello di ridurre la dimensionalità del fenomeno oggetto di

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.

	indagine preservando l'informazione da esso contenuta. La tecnica è applicabile a fenomeni misurati con variabili qualitative. L'ultima parte del modulo sarà dedicata all'applicazione dell'AC con il software R.
Contents arranged in 3 levels	**1. INTRODUZIONE** L'analisi delle corrispondenze, AC, è una tecnica di analisi multidimensionale che è in grado di tradurre in forma grafica quasi ogni tipo di tabella costituita da dati numerici. Oggetto dell'AC sono le matrici di contingenza, i cui elementi indicato il numero di volte che sono state rilevate congiuntamente le caratteristiche di due diverse grandezze. Obiettivo principale dell'AC è quello di analizzare le relazioni esistenti tra due variabile qualitative osservate su un collettivo di unità statistiche. Questo avviene attraverso l'identificazione di uno spazio "ottimale", cioè di una dimensione ridotta che rappresenta la sintesi dell'informazione strutturale contenuta nei dati originari. Il fine dell'analisi è quello di portare alla luce l'intreccio di legami, ovvero le corrispondenze, che esistono tra i dati in esame. **2. REQUISITI DELL'ANALISI DELLE CORRISPONDENZE** Affinché sia sensato condurre l'analisi delle corrispondenze, è importante analizzare le variabili da utilizzare per avere chiare alcune loro caratteristiche. Nello specifico le variabili devono avere i seguenti requisiti: - *Le **variabili** devono essere di **tipo Qualitativo**:* Le variabili qualitative sono delle variabili che non sono rappresentate da numeri, ma da modalità, esempio: genere, livello di istruzione, stato civile ecc. Queste modalità, dette anche categorie, devono essere esaustive e mutualmente esclusive. Per mutualmente esclusive si intende che le modalità della variabile non devono contenere lo stesso tipo di informazione. Ad esemio, per la variabile ''colore di capelli'' non si possono inserire le modalità ''capelli scuri'' e ''capelli castani'', in quanto per capelli scuri si intendono anche i capelli castani e viceversa. Per estaustive si intende che le modalità di una variabile deve tener conto di tutte le possibilità. Ad esempio per la variabile "livello di istruzione" si inseriscono le modalità "diploma", "laurea di primo livello", "laurea di secondo livello".

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.

Queste tre modalità non tengono conto di tutte i possibili liville di istruzione.

- *Le varaibile devono essere interdipendenti*:
  Prima di effettuare l'analisi delle corrispondenze è necessario verificare il grado di interdipendenza tra le due variabili considerate, in quanto se dovessero risultare indipendenti potrebbe non avere senso condurre l'analisi delle corrispondenze.
  Per fare ciò si esegue il test del chi-quadro:
  H_0: le due variabili sono indipendenti
  H_1: le due variabili non sono indipendenti
  Per interpretare i risultati del test si osserva in valore del p-value: p-value < 0.05, si rifiuta l'ipotesi nulla e di conseguenze le variabili sono considerate con un certo grado di dipendenza.

3. Come Condurre l'AC

Dopo aver verificato la sussistenza dei requisiti dell'AC si può passare all'analisi vera e propria.

3.1) Tabelle di contingenza

Nell'analisi delle corrispondenze si lavora con le tabelle di contingenza, le quali contengono le frequenze congiunte delle modalità delle due variabili qualitative X e Y. Queste matrici sono sempre costituite da numeri interi mai negativi che sono i conteggi, vale a dire semplici registrazioni di ciò che si è verificato. Inoltre, entrambe le variabili categoriche svolgono un ruolo simmetrico in cui tutti gli elementi hanno la stessa natura.

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

X, Y sono le variabili qualitative.

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

x_1, x_2 , x_3 : sono le modalità della variabile di X

y_1, y_2 , y_3 : sono le modalità della variabile di Y

$n_{i,j}$: sono le frequenze congiunte assolute, ossia le frequenze della coppie, esempio $n_{1,1}$: $X = x_1; Y = y_1$

$n_{i\cdot}$: sono i marginali di riga: $n_{i.} = \sum_{j=1}^{C} n_{i,j}$

$n_{\cdot j}$: sono i marginali di colonna: $n_{.j} = \sum_{i=1}^{R} n_{i,j}$

Questi non sono altro che la somma per la riga fissata (o per la colonna) delle frequenze congiunte sulle modalità di Y (per le colonne sulle modalità di X).

n = è la numerosità campionaria, che si può ottenere sommando i marginali di riga o colonna: $n = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{i,j} \quad \forall\, i,j$

Si può passare dalle frequenze assolute alle frequenze relative dividendo ogni frequenza assoluta per n: $f_{i,j} = \frac{n_{i,j}}{n}$

3.2) Matrice profilo Riga e Matrice profilo Colonna

La matrice dei profili riga si ottiene dividendo le frequenze assolute (o le frequenze relative) per i rispettivi marginali di riga. Quindi:

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

La tabella di contingenza sarà:

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

Sui marginali di riga si abbiamo tutti 1 e questo rappresenta la somma dei profili riga.

Sui marginali di colonna, invece, ci sono i profili medi che si ottengono sommando le frequenze relative per colonna; oppure calcolando la media degli elementi della matrice dei profili riga, per colonna. Si tratta

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.

di una media ponderata, dove le masse sono rappresentate dai marginali di riga $f_{i.}$.

Lavorando con le frequenze si perde una dimensione, quindi lo spazio delle righe è rappresentato da uno spazio C-1 dimensioni, ossia

Si può costruire una **matrice diagonale dei marginali di riga D_R,** che ha sulla diagonale maggiore i profili riga. La matrice diagonale dei marginali di riga è una matrice **R·R,** che ha dimensioni pari alle righe e sulla diagonale maggiore contiene i marginali di riga della tabella delle frequenze relative. Una matrice diagonale è una matrice il cui elemento generico sulla diagonale maggiore è il marginale di riga, al di sopra o al di sotto di essa, ci sono tutti zero. È una matrice sempre simmetrica e quadrata. Con la matrice diagonale dei marginali di riga si può costruire la **matrice dei profili riga**: essa si ottiene dividendo le frequenze relative per i marginali di riga $\frac{F}{D_R}$**.** Le dimensioni di **F** sono R·C, mentre D_R ha come dimensione R·R, dato che la divisione tra matrici non si può fare, si calcola l'inverso di D_R e si moltiplica per **F**, risolvendo così il problema della dimensionalità: $D_R^{-1} \cdot F$.

Lo stesso discorso vale per le colonna, con qualche piccola differenza.

Si costruisce la matrice dei profili colonna dividendo le frequenze assolute per i relativi marginali di colonna:

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

La tabella di contingenza che si ottiene sarà:

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

Ovviamente in questo caso sui marginali di colonna si avranno tutti 1 e sui marginali di riga si ha il profilo colonna medio. In questo caso le masse sono rappresentate dai marginali di colonna $f_{.j}$. Ovviamente,

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.

anche nello spazio delle colonne si lavora a meno di una dimensione, quindi lo spazio delle colonne è R-1.

Si può costruire una **matrice diagonale dei marginali di colonna** $\boldsymbol{D_C}$ che ha sulla diagonale maggiore i profili colonna. La matrice diagonale dei marginali di colonna è una matrice **C·C** , che ha dimensioni pari alle colonne e sulla diagonale maggiore contiene i marginali di colonna della tabella delle frequenze relative. Una matrice diagonale è una matrice il cui elemento generico sulla diagonale maggiore è il marginale di colonna, al di sopra o al di sotto di essa, ci sono tutti zero. È una matrice sempre simmetrica e quadrata. Con la matrice diagonale dei marginali di colonna si può costruire la **matrice dei profili colonna**: essa si ottiene dividendo le frequenze relative per i marginali di colonna $\frac{\boldsymbol{F}}{\boldsymbol{D_R}}$**.** Le dimensioni di **F** sono R·C, mentre $\boldsymbol{D_C}$ ha come dimensione C·C, dato che la divisione tra matrici non si può fare, si calcola l'inverso di $\boldsymbol{D_C}$ e si post-moltiplica a **F**, risolvendo così il problema della dimensionalità: $\boldsymbol{F \cdot D_C^{-1}}$.

3.3) Distanze

Nell'analisi delle corrispondenze è necessario capire che distanza c'è tra i valori, questo al fine di capire se le modalità sono lontane o vicini tra loro e quindi se si somigliano o meno. Si può fare questo osservando le frequenze: tanto più sono basse, tanto più sono vicine e viceversa. Ci sono vari metodi per calcolare la distanza: **distanza euclidea** e la **distanza del chi-quadro**.

La **distanza euclidea** è la più semplice e premia le distanze più alte a discapito di quelle più basse. Si calcola facendo la differenza delle frequenze relative elevandole al quadrato.

Per i profili riga:

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

Per i profili colonna:

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.

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

La **distanza del chi-quadro** premia le distanze più basse perché vengono riponderate le frequenze a bassa numerosità rispetto alle righe, inserendo nella formula l'inverso del marginale di colonna; e rispetto alle colonne, inserendo nella formula l'inverso del marginale di riga. Lo svantaggio della distanza del chi-quadro è che il reciproco dei marginali di colonna (o di riga) può tendere a zero e quindi una singola risposta può contribuire eccessivamente del calcolo della distanza.

3.4) Spazio delle Righe e Spazio delle Colonne

Nello **spazio delle righe** le due componenti sono:

- Profilo riga: $\mathbf{D_R^{-1}} \cdot \mathbf{F}$
- Metrica: $\mathbf{D_C^{-1}}$

Partiamo dalla formula:

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

Apportando le opportune sostituzioni:

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

L'obiettivo dell'analisi delle corrispondenze l'insieme degli assi unitari che consentono di massimizzare le distanze tra le proiezioni dei profili riga. Si devono, dunque, ricercare quei vettori che massimizzano le proiezioni. Dato che i vettori $\boldsymbol{u}$ possono essere infiniti si aggiunge il vincolo di norma unitaria

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

Problema di massimizzazione: massimizzare l'inerzia spiegata (variazione spiegata), che corrisponde alla variabilità per le variabili quantitative.

$$\begin{cases} \text{MAX: } \left\{ \hat{\psi}^T \boldsymbol{D}_R \hat{\psi} \right\} \\ \upsilon^T \boldsymbol{D}_C^{-1} \upsilon = 1 \end{cases}$$

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.

Per risolvere il problema di massimizzazione vincolata si utilizza il metodo dei moltiplicatori di Lagrange:

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

$\boldsymbol{\lambda}$= moltiplicatore di Lagrange, che è uno scalare;

***u*=** vettore dei pesi che stiamo cercando

Apportando le dovute sostituzioni, avremo:

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

Effettuiamo le operazioni di trasposizione, sostituiamo a $\boldsymbol{D_R \cdot D_R^{-1}}$ la matrice identità $\boldsymbol{I}$ e $[(-\boldsymbol{\lambda}) \cdot (-1)]$ lo sostituiamo con $\boldsymbol{\lambda}$**.** Possiamo poi togliere il trasposto dalle matrici diagonali $\mathbf{D_C^{-1}}$ e $\mathbf{D_R^{-1}}$, in quanto la trasposta di una matrice diagonale non cambia. Otteniamo:

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

Calcoliamo le derivate parziali, derivando la lagrangiana rispetto a $\boldsymbol{u}$ e poniamole uguali a 0:

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

Moltiplichiamo l'equazione per $\mathbf{D_C^{-1}}$**:**

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

Se sostituiamo la trasposta dei profili riga e la matrice dei profili colonna con ***S***, possiamo scrivere l'equazione caratteristica come:

$$Sv = \lambda v$$

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.

Massimizzare l'inerzia spiegata dei profili riga equivale a scomporre questa matrice in autovalori e autovettori della stessa. Il primo autovalore è associato al primo autovettore che spiega la massima inerzia. Gli autovettori che vengono estratti successivamente saranno estratti in modo ortogonale, ponendo il vincolo di ortogonalità

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

Utilizziamo il vincolo di ortogonalità per poter scegliere la seconda componente che andrà a spiegare l'inerzia che non viene spiegata dalla prima componente. Ovviamente, la prima componente estratta spiega la massima inerzia, cioè il massimo allungamento della nuba dei punti.

Nello **spazio delle colonne** le due componenti sono:

- Profilo colonna: $\mathbf{F \cdot D_C^{-1}}$
- Metrica: $\mathbf{D_R^{-1}}$

Partiamo dalla formula:

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

Sostituiamo e otteniamo

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

Il problema di massimizzazione da risolvere con i moltiplicatori di Lagrange è:

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

Procedendo come nello spazio delle righe, infine otterremo:

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.

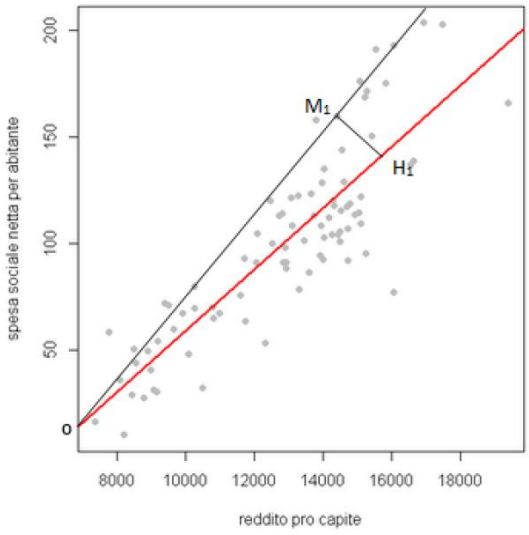
$$FD_C^{-1}F^T D_R^{-1}\nu = \mu\nu$$

Sostituendo la matrice dei profili colonna e la metrica trasposta dei profili riga con S^* otteniamo l'equazione caratteristica:

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

Massimizzare l'inerzia spiegata, ossia rendere l'informazione persa la più piccola possibile e l'informazione osservata la più grande possibili, dal punto di vista geometrico vuol dire rendere il più piccola possibile la distanza M_1H_1 e il più grande possibile la distanza OH_1.

Figura 1.3: Diagramma di dispersione



Si deve trovare dunque quella retta f (in rosso) interpolante i punti dello spazio vettoriale in modo che la distanza tra tutti i punti dello spazio e i punti proiettati ortogonalmente sulla retta f sia la minima possibile.

Gli autovalori nello spazio delle righe corrispondono agli autovettori nello spazio delle colonne, quindi gli autovalori di **S** corrispondono a quelli di S^*. Gli autovettori sono uguali tra loro a meno di una costante. Quindi quando dobbiamo massimizzare non serve scomporre in autovalori e autovettori **S** e S^*, basta farlo solo con una. La quantità di

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.

inerzia spiegata è uguale sia che si calcoli **S** o $\boldsymbol{S}^*$, la relazione trai due spazi è rappresentata dalle **formule di transizione**:

$$\boldsymbol{S} \rightarrow \boldsymbol{\nu} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F} \boldsymbol{D}_C^{-1} \boldsymbol{\upsilon} \equiv \boldsymbol{S}^* \rightarrow \boldsymbol{\upsilon} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\nu}$$

Spazio delle righe:

$$\hat{\boldsymbol{\psi}} = \boldsymbol{D}_C^{-1} \boldsymbol{\upsilon}$$

dove:

$$\boldsymbol{\upsilon} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\nu}$$

Applicando le opportune sostituzioni:

$$\frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\upsilon} \rightarrow \frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F}' \hat{\boldsymbol{\psi}}$$

Otteniamo:

$$\sqrt{\lambda} \hat{\boldsymbol{\psi}} = \boldsymbol{D}_C^{-1} \boldsymbol{F}' \hat{\boldsymbol{\psi}} \rightarrow \hat{\boldsymbol{\psi}} = \frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\boldsymbol{\psi}} \rightarrow \sqrt{\lambda} \hat{\boldsymbol{\psi}} = \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\boldsymbol{\psi}}$$

Per lo spazio delle righe, dunque:

$$\sqrt{\lambda} \hat{\boldsymbol{\psi}} = \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\boldsymbol{\psi}} \equiv \hat{\boldsymbol{\psi}} = \sqrt{\lambda} \hat{\boldsymbol{\psi}}$$

Spazio delle colonne:

$$\hat{\boldsymbol{\psi}} = \boldsymbol{D}_R^{-1} \boldsymbol{\nu}$$

Dove:

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.

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

Applicando le opportune sostituzioni:

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

Otteniamo:

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

Per lo spazio delle colonne:

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

4) Esempio con il software R

Verificare una possibile relazione tra le distribuzioni dei capi di allevamento e le diverse regioni italiane. I dati sono relativi all'anno 2011, raccolti dalle banche disponibili sul sito dell' Istat.

Ipotesi: le varie regioni, a seconda delle caratteristiche territoriali e delle esigenze della popolazione, scegli di allevare alcuni capi di bestiame piuttosto che altri.

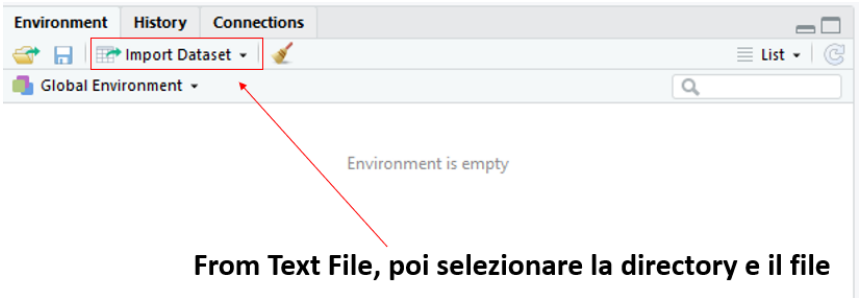
Dataset:

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

Importiamo il dataset:



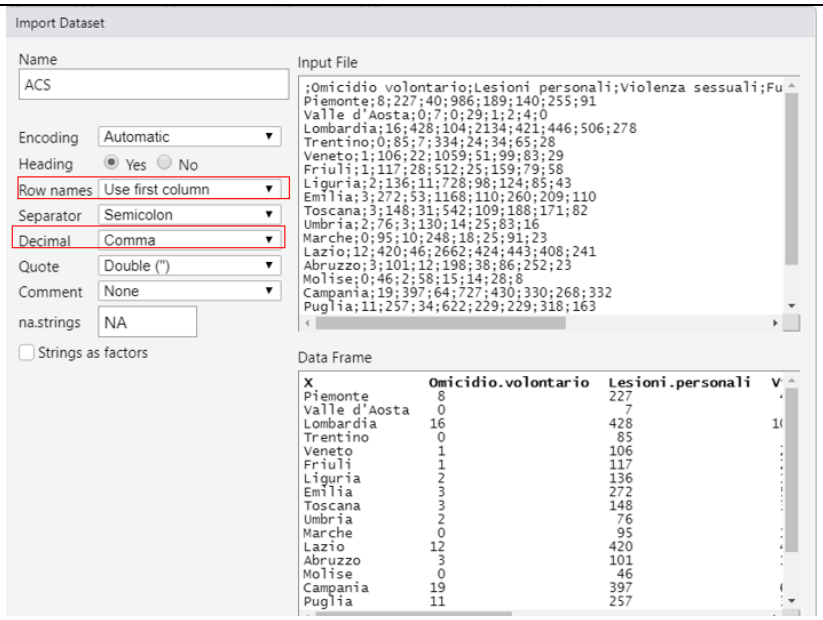
From Text File, poi selezionare la directory e il file

Nel campo ***row names*** selezionare la dicitura: "***use first column***" in modo da avere sui grafici le etichette sia degli individui sia delle variabili.

Nel campo ***decimal*** selezioniamo la dicitura "***comma***".

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.



Con il comando:

X<-as.matrix(nome_del_dataset)

Attribuiamo ad **X**, come oggetto, il dataset utilizzato nell'analisi.

Prima di poter effettuare l'AC è necessario stabilire il grado di interdipendenza tra i due caratteri considerati, questo perché nel caso in cui essi risultassero indipendenti potrebbe non aver senso proseguire AC. Per verificare questo si effettua il test del chi-quadro.

Il comando è:

chiquadro<-chisq.test(X)

```
        Pearson's Chi-squared test

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

Si può osservare che il ***p-value*** è inferiore al livello di significatività più comunemente utilizzato ossia 0.05. Possiamo dunque rigettare l'ipotesi nulla di indipendenza statistica tra le due variabili e si può proseguire con l' analisi.

Vogliamo ora creare una matrice delle frequenze relative **F**.

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.

Per fare questo innanzitutto calcoliamo la numerosità campionaria, con il comando:

n<-sum(X)

e successivamente dividendo la matrice di partenza (quindi tutte le frequenze congiunte) per la numerosità campionaria otteniamo la matrice **F**. Comando:

F<-X/n

Il passo successivo è quello di ottenere le tabelle dei **profili riga e colonna**. Per poter fare questo, innanzitutto, è necessario calcolare i marginali di riga e di colonna. Rispettivamente i comandi sono:

sumrow<-apply(F,1,sum)
sumcol<-apply(F,2,sum)

Successivamente si calcola la matrice diagonale dei marginali di riga e la sua inversa con i comandi:

Dr<-diag(sumrow)
Dr_inv<-solve(Dr)

Ora possiamo calcolare i profili riga. In termini matriciali pre-moltiplichiamo l'inverso della matrice diagonale dei marginali di riga alla matrice delle frequenze relative. Il comando da utilizzare è:

Pr<-Dr_inv%*%F

La stessa cosa per i profili colonna, ricordando che in questo caso l'inverso della matrice colonna deve essere post-moltiplicato alla matrice delle frequenze relative.

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

Ora possiamo calcolare le distanze tra i punti. Come già detto esistono due tipi distanze: **euclidea** e del **chi-quadro**.

Distanza **euclidea profili riga**:

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

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

Distanza **euclidea profili colonna**:

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

Distanza del **chi-quadro profili riga**:

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

Distanza del **chi-quadro profili colonna:**

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

L'equazione caratteristica della matrice dei **profili riga**:

S<-t(Pr)%*%Pc

Dato che la matrice **S** non è simmetrica, è necessario diagonalizzarla per ottenere **S_tilde:**

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

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

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

Ora dobbiamo massimizzare l'inerzia spiegata scomponendo la matrice in autovalori e autovettori:

AC<-eigen(S_tilde)

lambda<-as.matrix(AC$values)

lambda<-lambda[-1,]

w<-AC$vectors

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

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.

u<-u[,-1]

L'equazione caratteristica della matrice dei **profili colonna:**

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

Per passare da ***u*** a ***v***, utilizziamo le formule di transizione (dato che la quantità di inerzia spiegata è uguale sia nello spazio delle righe che in quello delle colonne).

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

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

Calcoliamo fattori e coordinate, prima spazio delle righe e poi colonne:

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

Visualizziamo il grafico delle coordinate principali:

PRINCOORD<-rbind(PSI_coord,PHI_coord)

righe<-row.names(X);colonne<-colnames(X)

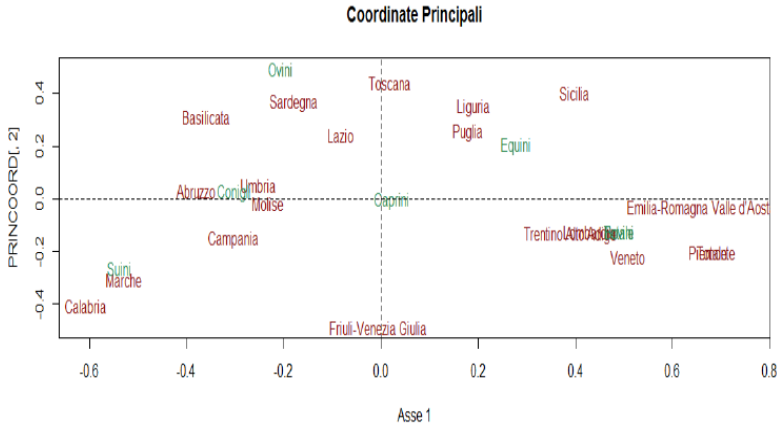
plot(PRINCOORD[,1],PRINCOORD[,2],type="n",main="Coordinate Principali",xlab="Asse1",ylab="Asse2")+ text(PRINCOORD[1:20,1],PRINCOORD[1:20,2],labels=righe,col="spring green4")

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

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

Avremo:

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



Osservando questo grafico possiamo dire, ad esempio, che nelle regioni come Abruzzo, Molise, Umbria prevalentemente si allevano conigli.

Scegliamo le componenti:

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

sum(lambda)

in_exp<-lambda/inerzia

in_exp_cum<-cumsum(in_exp)

Visualizziamo i risultati ottenuti:

```
> 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
```

La prima dimensione da sola spiega il 58.57% della variabilità e le prime tre insieme spiegano il 92.26% della variabilità complessiva dei dati.

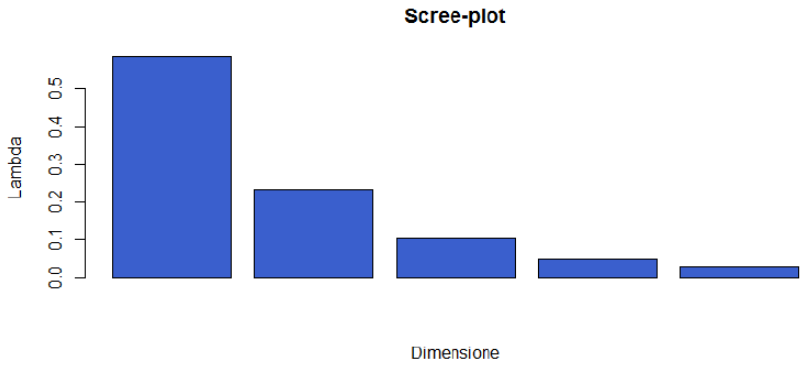
I risultati ottenuti si possono visualizzare graficamente con lo **scree-plot dell'inerzia spiegata**:

screeplot<-barplot(in_exp,main="Scree-plot inerzia", xlab="Dimensione", 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



Per la qualità della rappresentazione:

- per valutare quanto una modalità influenza o è partecipe all'asse fattoriale calcoliamo i **contributi assoluti**, **CA**, sia per le righe che per le colonne:

ca_r<-Dr%*%fp_c^2

ca_c<-Dc%*%fp_r^2

- per valutare la qualità della rappresentazione calcoliamo i **contributi relativi, CR**, che forniscono una misura migliore della rappresentazione stessa dei punti sugli assi ed è data dal coseno dell'angolo formato dal vettore di proiezione del punto e il vettore relativo *i* (o *j*) al punto *i* (o *j*) nel suo spazio originario:

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)

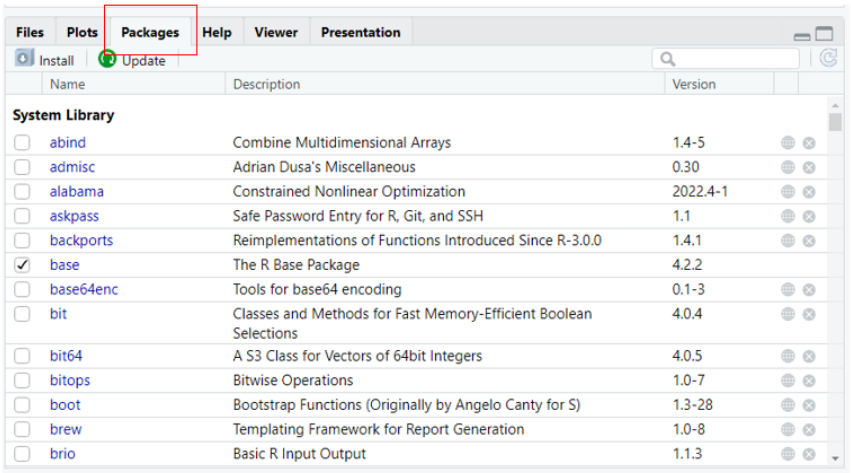
cos2c<-PHI_coord^2/d_jh

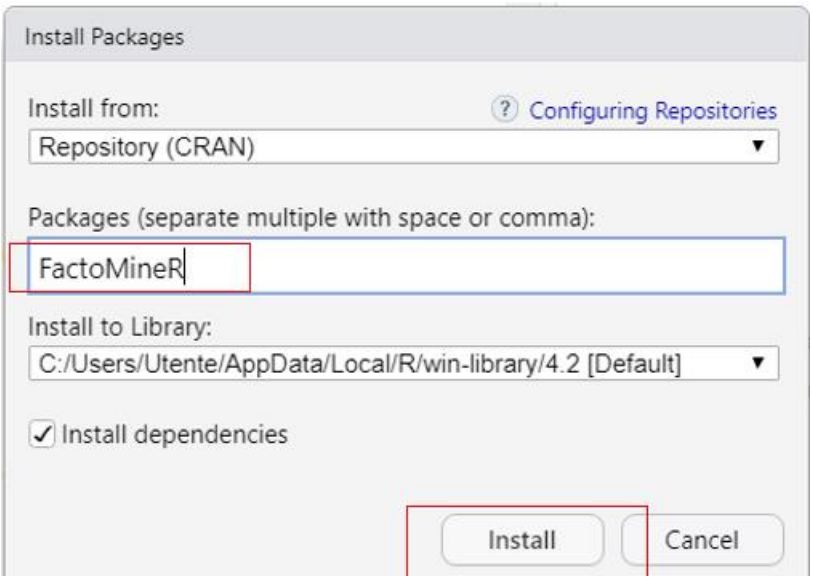
Co-funded by the Erasmus+ Programme of the European Union

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.

R per l'analisi delle corrispondenze mette a disposizione un pacchetto chiamato **FactoMineR**, che aggiunge informazioni su individui e variabili e permette di creare un grafico congiunto bidimensionale degli individui e delle variabili.

Su R per poter utilizzare tale pacchetto innanzitutto è necessario scaricarlo:





Dopo averlo istallato è necessario richiamarlo con il comando

library(FactoMineR)

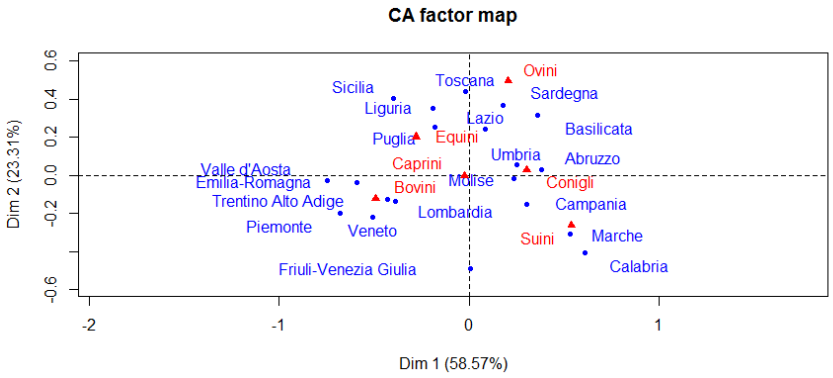
passiamo alla creazione del grafico bidimensionale individui e variabili:

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.

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)

Graficamente avremo:



Interpretazione dei risultati:

possiamo dire che l'ipotesi iniziale viene confermata. In particolare, le regioni più dedite all' allevamento degli ovini sembrano essere Toscana, Sardegna e Basilicata e ciò si può spiegare per il fatto che queste regioni sono zone di montagna e di transumanza. Gli equini sono maggiormente allevati in Puglia, Liguria e Sicilia poiché questi animali da sempre sono utilizzati per i lavori nelle campagne. I bovini sono presenti in Trentino Alto-Adige, Veneto, Piemonte, Lombardia ed Emilia-Romagna; infatti queste regioni hanno una tradizione di allevamento più sviluppata ad uso alimentare. I conigli compaiono principalmente in Umbria, Abruzzo e Molise. Invece i suini sembrano essere maggiormente allevati nelle Marche, Campania e Molise; anche queste regioni hanno una tradizione di allevamento più sviluppata ad uso alimentare.
I caprini, invece, si collocano al centro degli assi, probabilmente perché non ci sono regioni che prediligono il loro allevamento.

Self-assessment (multiple choice queries and answers)

1. Che cosa consentono di fare le formule di transizione?

   A) Passare da uno spazio all'altro
   B) Passare dalla rappresentazione dei contributi assoluti a quella dei contributi relativi
   C) Passare dalla matrice delle frequenze relative a quelle dei profili

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. Per quale motivo si effettua il test del chi quadro prima di implementare l'AC? A) Per verificare se le variabili sono quantitative B) Per valutare se le variabili sono qualitative C) Per analizzare l'esistenza di interdipendenza tra le due variabili 3. Quale obiettivo ha l'Analisi delle Corrispondenze? A) Massimizzare la variabilità spiegata B) Massimizzare l'inerzia spiegata C) Minimizzare l'inerzia spiegata
Resources (videos, reference link)	
Related material	
Related PPT	
Bibliography	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.
Provided by	[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.