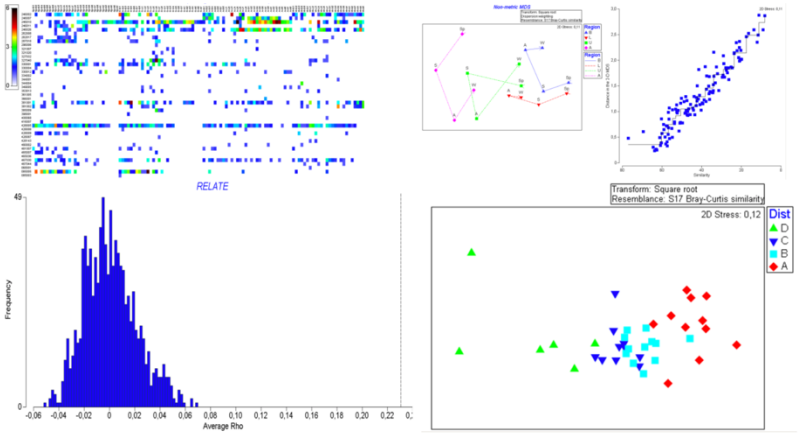
CENTRO DI RICERCA
PRODUZIONI VEGETALI
Sant'Anna
Scuola Universitaria Superiore Pisa

Corso di Analisi Multivariata in Agroecologia



Il percorso didattico del Corso prevede una fase di formazione d'aula: periodo dal 21 marzo al 5 maggio 2023, per un totale 38 ore, di cui 17 di didattica frontale e 17 di pratica e 4 ore di pratica in gruppi, così articolato:

Docenti:

Elisa Pellegrino – Scuola Superiore Sant'Anna (coordinatrice)
Marco Pittarello – Università di Torino
Emanuele Barca – CNR Bari
Andrea Onofri - Università di Perugia
Alessia Perego - Università di Milano
Anna Maria Stellacci - Università di Bari

I partecipanti alla fine del corso di formazione svilupperanno conoscenze relative al calcolo delle misure di resemblance (similarity/dissimilarity/distance) in strutture multivariate, la valutazione degli effetti del pre-trattamento dei dati (standardizzazione, trasformazione, normalizzazione), la scelta del pre-trattamento per i diversi tipi di dati; metodi di classificazione univariata e multivariata: metodi per la valutazione del numero di clusters, come Elbow method, silhouette method, gap statistic method. Metodi di clustering: K-means, e (fuzzy) C-means; analisi delle Componenti Principali; metodi di ordinazione parametrici e non nell'ambito delle analisi parametriche constrained e unconstrained, come la correspondence analysis, la detrended correspondence analysis, la redundancy analysis (RDA), e la canonical correspondence analysis; la forward selection delle variabili esplicative; il Monte Carlo Test per modelli completamente randomizzati e modelli a blocchi randomizzati; la ripartizione della varianza e tecniche per la gestione di misure ripetute; i metodi di ordinazione non parametrici; l'analisi non-parametric analysis of similarities (ANOSIM); la permutational ANOVA e MANOVA (PERMANOVA) per l'analisi multivariata di dati in disegni complessi; la partitioning variation sulla base della scelta delle misure di similarità e metodi di permutazione; il test di omogenità delle dispersioni (PERMDIPS); l'impiego di metodi di analisi multivariata per la variable/feature selection, l'analisi discriminante stepwise, e la partial least squares regression con impiego di statistiche VIP (Variable importance for projection).

Il corso fornirà una panoramica completa e approfondita dei metodi statistici di analisi parametrica e non parametrica di dati multivariati utilizzando R ed PRIMER 7 + PERMOVA, che sarà fornito e utilizzabile free per tutta la durata del corso. Il corso proposto consiste in 34 ore ed una mezza giornata di test dedicata all'analisi di data set comuni (4 ore). In tale giornata verrà programmata una consultazione online individuale separata con il/i docenti (inclusa la condivisione di dati/documenti) per gruppi di partecipanti per discutere i progetti di analisi dei dati, ed ottenere consigli e assistenza e/o porre ulteriori domande. Il corso sarà costituito da lezioni teoriche e sessioni pratiche in consultazione con i relatori. Ciascuna lezione teorica sarà seguita da una sessione pratica. La sessione al computer (il partecipante userà il proprio computer) sarà seguita dal docente che utilizzerà il proprio computer per riepilogare i punti salienti dell'interpretazione dei risultati, e rispondere alle domande. I partecipanti sono tenuti a utilizzare il proprio laptop o computer desktop, che deve essere dotato di un accesso ad Internet sicuro e affidabile, un microfono e una telecamera, che consentano la comunicazione diretta con il docente (e potenzialmente anche con gli altri partecipanti al corso). Il software PRIMER 7 + PERMOVA gira su Windows, quindi gli utilizzatori Mac dovranno eseguire le analisi in emulazione di Windows o dual boot, mentre la piattaforma R (https://cran.r-project.org/) funziona sia su Windows che su Mac. La conoscenza della statistica di base e dei principali disegni sperimentali è richiesta per la partecipazione al corso. Inoltre, una conoscenza di base di R è fortemente consigliata. Il Corso prevede che i partecipanti abbiano installato sul proprio computer i software proposti, con i packages che verranno indicati dopo l'iscrizione.

Il corso sarà così articolato:

- **I Lezione (3 ore totali: Elisa Pellegrino)** - Teoria di 1.5 ora su misure di resemblance (similarity/dissimilarity/distance) in una struttura multivariata, valutazione degli effetti del pre-trattamento dei dati (standardizzazione, trasformazione, normalizzazione), e guida nelle scelte per i diversi tipi di dati. Pratica di 1.5 ora su gestione dei dati e dei fattori e sul pre-trattamento dei dati in Primer 7 + PERMANOVA. Orario 14:30-17:30 (21 marzo 2023).
- **II Lezione II (3 ore totali: Andrea Onofri)** - Lezione di 3 ore tra teoria e pratica per la gestione, costruzione del data set e pre-trattamento dei dati in una struttura multivariata utilizzando R. Orario 14:30-17:30 (24 marzo 2023).
- **III Lezione (5 ore totali: Emanuele Barca) -** Teoria di 2.5 ore su metodi di classificazione univariata e multivariata; metodi per la valutazione del numero di clusters (Elbow method, silhouette method, gap statistic method); metodi di clustering: K-means, e (fuzzy) C-means. Pratica di 2.5 ore: applicazione di metodi di classificazione con data set messi a disposizione, gestione del workspace e plotting con R. Orario 9:30 -12.00; 14:30-17:00 (28 marzo 2023).
- **IV Lezione (5 ore totali: 2 ore Andrea Onofri e 3 ore Alessia Perego)** - Teoria di 2.5 ore di Andrea Onofri sull'Analisi delle Componenti Principali (PCA): questione di 'punti di vista'. Possiamo 'semplificare' un dataset senza perdere informazioni? Ordinare e classificare le osservazioni: la PCA come metodo di ordinamento. Come nasce e come si interpreta un biplot e riconoscimento dei limiti della PCA. Pratica di Alessia Perego di 2.5 ore: la PCA con R pratica su data set messi a disposizione. Orario 14:30-19:30 (31 marzo 2023).
- **V – VI Lezione (8 ore totali Marco Pittarello) -** Teoria di 4 ore su metodi di ordinazione parametrici e non, e nell'ambito delle analisi parametriche constrained e unconstrained, analisi come la correspondence analysis CA, la detrended correspondence analysis (DCA) la redundancy analysis (RDA), la canonical correspondence analysis (CCA) e la non metric multi-dimensional scaling (nMDS). Per i metodi parametrici constrained sarà introdotta la forward selection delle variabili esplicative, il Monte Carlo Test per i modelli completamente randomizzati ed i modelli a blocchi randomizzati, la ripartizione della varianza e le tecniche per la gestione di misure ripetute. Mentre per l'analisi nMDS

saranno forniti i concetti di base. Pratica (4 h): pratica con R su data set messi a disposizione. Due giornate dalle 14:30-18:30 (4 e 12 aprile 2023)

- **VII Lezione (4 ore: totali Elisa Pellegrino)** - Teoria di 2 ore su metodi di ordinazione non parametrici (nMDS il concetto di diagramma di Shepard), metric multi-dimensional scaling (mMDS) e combinazione delle due analisi, minimum spanning tree, e cluster overlay. Test multivariato per le differenze tra gruppi di campioni specificati a priori utilizzando la non-parametric analysis of similarities (ANOSIM unidirezionale, test globali e pairwise). Plots di ordinazione per esaminare le medie multivariate. Introduzione al calcolo del bootstrap delle stime approssimative per le medie nei grafici di mMDS. Test ANOSIM per fattori a diversi livelli e per disegni a più vie (fino a 3 fattori). Pratica di 2 ore su ANOSIM con PRIMER 7 + PERMAOVA. Orario 14:30-18:30 (14 aprile 2023).
- **VIII Lezione teorica (4 ore totali: Elisa Pellegrino)** - Teoria di 2 ore su Permutational ANOVA e MANOVA (PERMANOVA) per l'analisi multivariata di dati in disegni complessi, partitioning variation sulla base della scelta delle misure di similarità e metodi di permutazione. Analisi e stima delle componenti di variazione in disegni sperimentali complessi, incluse interazioni, utilizzo di variabili covariate, contrasti, fixed o random effects, crossed o nested models, disegni non bilanciati, blocchi radomizzati e misure ripetute. Test di omogenità delle dispersioni (PERMDIPS). Pratica di 2 ore: applicazione della PERMANOVA in disegni più o meno complessi e di applicazione della PERMDIPS con PRIMER 7 + PERMAOVA. Orario 14:30-18:30 (18 aprile 2023).
- **IX-X Lezione (6 ore totali: Anna Maria Stellacci e Emanuele Barca)** - Teoria di 3 ore di Anna Maria Stellacci e Emanuele Barca sull'impiego di metodi di analisi multivariata per la variable/feature selection: descrizione del problema e sua importanza. Illustrazione dei principali metodi allo stato dell'arte con particolare riguardo all'analisi delle componenti principali, analisi discriminante stepwise, partial least squares regression con impiego di statistiche VIP (Variable importance for projection). Pratica di 3 ore di Emanuele Barca e Anna Maria Stellacci su applicazione di metodi di variable selection con data set messi a disposizione (soil data e hyperspectral data). Orario 14:30-17:30 da distribuire su 2 pomeriggi (21 – 28 aprile 2023)

La discussione del Project Work avverrà alla fine del corso (4 ore; data prevista 5 maggio 2023).

Docente di riferimento:
Elisa Pellegrino
elisa.pellegrino@santannapisa.it

Con il patrocinio della **Società Italiana di Agronomia**

Società Italiana di Agronomia