UNIVERSITÀ
DEGLI STUDI
DI MESSINA

Dipartimento di Ingegneria
C.da Di Dio - Villaggio S. Agata - 98166 Messina – Italy P.I. 00724160833 - c.f. 80004070837

AMBIENTE STATISTICO

SOFTWARE PER L'ANALISI STATISTICA DI DATI PROVENIENTI DAL MONITORAGGIO AMBIENTALE

Release 4.0 – 20/03/2018

Manuale d'uso

Ambiente Statistico è un software sviluppato nell'ambito del Progetto "*MAGINOT: Sistema integrato per il monitoraggio e la tutela dell'ambiente urbano, extraurbano e marino*" - PON Ricerca e Competitività, Asse I, Obiettivo operativo 4.1.1.1., Azione II - PON01_02309/4 - CUP B44B140000600008 - Responsabile Scientifico: *Prof. Signorino Galvagno*

Sviluppatore: *Ing. Vito Puliafito* – vpuliafito@unime.it

SOMMARIO

Direzione: tel.+39.090.3977157; fax +39.090.3977457 Segreteria: tel. +39.090.3977357; fax: +39.090.3977475
Amministrazione: tel. +39.090.3977567; fax: +39.090.3977480

Introduzione

AMBIENTE STATISTICO è un'interfaccia software per l'ANALISI STATISTICA DI DATI PROVENIENTI DAL MONITORAGGIO AMBIENTALE.

Le tecniche di analisi statistica multivariata che possono essere utilizzate sono l'analisi della componenti principali (PCA) e l'analisi a gruppi o cluster analysis (CA).

È inoltre possibile, nel caso in cui siano presenti variabili climatiche nella matrice dei dati, escludere quelle variabili dall'analisi statistica o selezionare una di esse e definire un range di suoi valori per filtrare i dati.

Il presente manuale d'uso descrive tutte le funzionalità del software, spiegando il modo corretto con cui utilizzarlo. Gli aspetti più importanti del funzionamento sono evidenziati in grassetto.
Per informazioni e aiuti sull'utilizzo del software, inoltre, è possibile consultare le finestre informative cliccando sui pulsanti di informazione *i* presenti in ogni pannello.

Per gli aspetti teorici sulle tecniche di analisi statistica e per il significato dei risultati di quelle tecniche si rimanda alla letteratura del settore e alle relazioni del progetto "Maginot".

Direzione: tel.+39.090.3977157; fax +39.090.3977457
Segreteria: tel. +39.090.3977357; fax: +39.090.3977475
Amministrazione: tel. +39.090.3977567; fax: +39.090.3977480

Installazione

Il software non richiede installazione. Il file AmbienteStatistico.exe può essere eseguito e avvia la finestra d'apertura.

PREREQUISITI PER IL FUNZIONAMENTO

Ambiente Statistico 4.0 funziona sul sistema operativo Windows® (versione 8 e successive) e richiede la corretta installazione di Matlab® (versione R2015a e successive) e di Microsoft® Excel.

Direzione: tel.+39.090.3977157; fax +39.090.3977457
Segreteria: tel. +39.090.3977357; fax: +39.090.3977475
Amministrazione: tel. +39.090.3977567; fax: +39.090.3977480

Finestra di apertura

Ambiente Statistico 4.0 si presenta come nella figura 1 che segue.

Figura 1 – La finestra di apertura di AmbienteStatistico 4.0.

Essa include un pannello laterale contenente il logo dell'Università di Messina, il nome ed i dettagli del progetto MAGINOT, il nome dello sviluppatore dell'interfaccia software, e la versione della stessa.
Nella parte principale della finestra compaiono, dall'alto in basso, i pannelli relativi ai dati da analizzare, alle opzioni nel caso di variabili climatiche, alle due tecniche di analisi statistica.

Per il corretto utilizzo dell'interfaccia software, è necessario procedere come segue:
1) Caricare la matrice dei dati;
2) Selezionare, se opportuno, le variabili climatiche da escludere e la variabile climatica che filtri i dati (in questo caso indicare il range di valori di interesse);
3) Selezionare la tecnica di analisi statistica desiderata.

Le note principali di cui sopra sono riportate anche nella finestra informativa che si apre cliccando il bottone *i* posto accanto al titolo della finestra.

CARICAMENTO DEI DATI

Nel pannello DATI DA ANALIZZARE è presente un bottone per caricare il file di dati. Come indicato sotto il bottone, **il file deve essere un foglio di lavoro** (per esempio Excel) **in cui i dati siano in forma matriciale, con le osservazioni nelle righe e le variabili nelle colonne**.

È previsto che la matrice includa le intestazioni con i nomi delle osservazioni nella prima colonna e i nomi delle variabili nella prima riga. I dati da elaborare saranno quindi ottenuti eliminando la prima riga e la prima colonna della matrice caricata.

Il file deve trovarsi nella stessa cartella dell'eseguibile.

Una volta caricato il file, nelle caselle a destra compariranno il numero di osservazioni e il numero di variabili della matrice dei dati da analizzare.
Nel proseguo del manuale identificheremo questi due numeri con N_{OBS} e N_{VAR}.

Le note principali di cui sopra sono riportate anche nella finestra informativa che si apre cliccando il bottone *i* posto accanto al titolo del pannello.

VARIABILI CLIMATICHE E FILTRAGGIO DEI DATI

Nel caso in cui la matrice dei dati includa variabili climatiche, il pannello centrale della finestra permette di aggiornare i dati con le seguenti due opzioni:
1) escludere una o più variabili climatiche;
2) filtrare i dati da analizzare in funzione dei valori di una variabile climatica.

L'esclusione delle variabili climatiche può fornire un'analisi statistica legata esclusivamente al tipo di sorgente inquinante.
Per escludere una variabile è necessario inserire il numero della colonna associata alla variabile nella matrice di dimensioni [N_{OBS}+1 x N_{VAR}+1] caricata come foglio di lavoro (non il numero d'ordine della variabile nella matrice dati priva delle intestazioni).

Se si vuole effettuate un filtraggio dei dati, analizzando solo quelli all'interno di un range di valori di una variabile climatica, indicare la colonna della variabile filtro, spuntare il bottone filtro, e indicare il minimo e massimo della variabile.

Cliccando sul pulsante AGGIORNA I DATI, compariranno i numeri delle osservazioni e delle variabili che saranno soggetti all'analisi statistica multivariata.

Anche in questo caso, le istruzioni d'uso del pannello sono riportate nella finestra informativa che si apre cliccando il bottone *i* posto accanto al titolo del pannello.

LA SELEZIONE DELLA TECNICA DI ANALISI

Dopo il caricamento dei dati, e l'eventuale aggiornamento degli stessi, si può procedere alla selezione della tecnica di analisi statistica multivariata, l'analisi delle componenti principali (in inglese principal component analysis, da cui PCA) e l'analisi a gruppi (cluster analysis, da cui CA).

La selezione di una delle due tecniche, cliccando il pulsante corrispondente, permette l'apertura della finestra appropriata per la visualizzazione dei risultati.

La selezione di una delle due tecniche cancella eventuali file di risultati creati precedentemente e presenti nella cartella dell'eseguibile.

Direzione: tel.+39.090.3977157; fax +39.090.3977457
Segreteria: tel. +39.090.3977357; fax: +39.090.3977475
Amministrazione: tel. +39.090.3977567; fax: +39.090.3977480

Finestra per l'analisi delle componenti principali

La finestra di Ambiente Statistico 4.0 rivolta all'analisi delle componenti principali (PCA) si presenta come nella figura 2 che segue.

Figura 2 – La finestra di AmbienteStatistico 4.0 per l'analisi PCA.

Nel primo pannello, sono riportati il numero delle osservazioni ed il numero delle variabili dei dati caricati, ed eventualmente aggiornati, nella finestra di apertura.

AmbienteStatistico 4.0 effettua l'analisi PCA sui dati standardizzati e può fornire i risultati cliccando i pulsanti corrispondenti nel pannello PCA.

RISULTATI IN TABELLE

Cliccando sul pulsante OUTPUT PCA in TABELLE verrà creato, e salvato nella cartella dell'eseguibile, il file "**resultsPCA.xlsx**" con diversi fogli di lavoro:

1) Dati analizzati - dati caricati, eventualmente aggiornati nella finestra iniziale;

2) Importanza % - importanza percentuale delle N_{VAR} componenti principali, inclusa la somma delle percentuali al crescere delle componenti principali;

3) Coefficienti delle variabili - matrice [N_{VAR} x N_{VAR}] con i coefficienti delle variabili rispetto alle componenti principali;

4) Coefficienti delle osservazioni - matrice [N_{OBS} x N_{VAR}] con i coefficienti delle osservazioni nello spazio delle componenti principali;

5) Correlazione dati nei 2 spazi - matrice [N_{VAR} x N_{VAR}] con la correlazione tra i dati nello spazio delle variabili originarie e i dati nello spazio delle componenti principali.

RISULTATI GRAFICI

Nel pannello PCA è inoltre possibile aprire dei grafici con i risultati dell'analisi statistica.
In particolare, è possibile visualizzare un grafico a barre con l'importanza delle componenti principali ed una linea a rappresentare la somma di queste percentuali al crescere delle componenti principali

considerate. Si ricorda che, di norma, le componenti principali che hanno il 70% dell'importanza percentuale possono spiegare bene statisticamente l'intero set di dati.

Inoltre, dopo aver selezionato due variabili principali rispetto cui visualizzare gli altri risultati (**se non vengono selezionate, sono considerate di default le prime due variabili principali**, cioè le due con maggiore importanza percentuale), si possono visualizzare altri due grafici:

- le variabili nel piano delle componenti principali selezionate;
- le osservazioni nel piano delle componenti principali selezionate.

Guardando insieme questi due grafici, è possibile interpretare al meglio le caratteristiche statistiche dei dati.

Direzione: tel.+39.090.3977157; fax +39.090.3977457
Segreteria: tel. +39.090.3977357; fax: +39.090.3977475
Amministrazione: tel. +39.090.3977567; fax: +39.090.3977480

Finestra per l'analisi a gruppi

La finestra di Ambiente Statistico 4.0 rivolta all'analisi a gruppi (CA) si presenta come nella figura 2 che segue.

Figura 2 – La finestra di AmbienteStatistico 4.0 per l'analisi a gruppi.

Nel primo pannello, sono riportati il numero delle osservazioni ed il numero delle variabili dei dati caricati, ed eventualmente aggiornati, nella finestra di apertura.

AmbienteStatistico 4.0 permette di effettuare l'analisi a gruppi utilizzando due tipici algoritmi, uno non gerarchico, il K-means, e uno gerarchico aggregativo.

L'ALGORITMO NON-GERARCHICO K-MEANS

Per effettuare l'analisi statistica a gruppi tramite l'algoritmo K-means, è necessario indicare il numero K di gruppi (cluster) in cui si vogliono dividere i dati e cliccare sul pulsante che genera gli output.

In uscita, viene creato, e salvato nella cartella principale, il file "**resultsCAkmeans.xlsx**" con diversi fogli di lavoro:

1) Dati analizzati - matrice [N_{OBS} x N_{VAR}] dei dati caricati ed eventualmente aggiornati nella finestra iniziale;

2) Cluster - numero di cluster scelto (K);

3) Indice delle osservazioni - matrice [N_{OBS} x 1] contenente gli indici del cluster associato a ciascuna delle Nobs osservazioni;

4) Centroidi - matrice [K x N_{VAR}] corrispondente alle N_{VAR} coordinate dei centroidi dei K cluster;

5) Somma distanze - matrice [K x 1] contenente la somma delle distanze di tutti i punti di un cluster rispetto al centroide dello stesso, a misurare la compattezza del cluster stesso;

6) Distanze - matrice [N_{OBS} x K] corrispondente alle distanze di tutte le osservazioni dai K centroidi.

Direzione: tel.+39.090.3977157; fax +39.090.3977457
Segreteria: tel. +39.090.3977357; fax: +39.090.3977475
Amministrazione: tel. +39.090.3977567; fax: +39.090.3977480

L'ALGORITMO GERARCHICO AGGREGATIVO

Per effettuare l'analisi statistica a gruppi tramite l'algoritmo gerarchico aggregativo, è necessario selezionare una delle due metriche possibili, euclidea o Chebychev, indicare il numero di gruppi in cui si vogliono dividere i dati, e cliccare sul pulsante che genera gli output.

In uscita, viene creato, e salvato nella cartella principale, un file "**resultsCAgerarchico.xlsx**" con diversi fogli di lavoro:

1) Dati analizzati - matrice [N_{OBS} x N_{VAR}] dei dati caricati ed eventualmente aggiornati nella finestra iniziale;

2) Cluster - numero [K] dei cluster;

3) Distanze - matrice [N_{OBS} x N_{OBS}] delle distanze tra le N_{OBS} osservazioni;

4) Gerarchia - tabella con la gerarchia dei raggruppamenti, in cui saranno indicati passo per passo i cluster aggregati (le prime due colonne sono i cluster aggregati, la terza colonna la loro distanza, le righe sono i passi di aggregazione - i numeri fino a N_{OBS} rappresentano i cluster formati dalla singola osservazione).

Viene inoltre visualizzata una figura a rappresentare il dendrogramma dell'aggregazione.

Le istruzioni sull'uso di questi pannelli sono, ancora una volta, riportate nelle finestre informative che possono essere aperte cliccando sui bottoni *i* presenti nei diversi pannelli.

Direzione: tel.+39.090.3977157; fax +39.090.3977457
Segreteria: tel. +39.090.3977357; fax: +39.090.3977475
Amministrazione: tel. +39.090.3977567; fax: +39.090.3977480