

OTTEMPERANZA CONDIZIONI AMBIENTALI N. 1÷5

ALLEGATO D

**METODI DI ANALISI STATISTICA DEI DATI DI MONITORAGGIO
OFF-SHORE E ASSUNTI**

METODI DI ANALISI STATISTICA DEI DATI DI MONITORAGGIO OFF-SHORE

Sebbene le restituzioni tabellari puntuali e le procedure statistiche univariate siano tecnicamente semplici da applicare in fase di elaborazione dei dati, la complessità aumenta con l'aumentare della complessità del sistema di studio e, pertanto, diminuisce la chiarezza di lettura del quadro complessivo. Per questo, i dati presentati in forma disaggregata rendono complessa la valutazione degli elementi molteplici e concomitanti che determinano la variabilità in sistemi complessi.

Pertanto, i risultati relativi alle attività di monitoraggio *ante operam* - AO e fase di cantiere - C (fase attuale) per il prolungamento del pontile del Progetto Tempa Rossa, sono stati analizzati mediante approccio di statistica multivariata. L'analisi statistica multivariata rappresenta un potente strumento per effettuare inferenze statistiche sintetiche su database molto ampi e per permettere la valutazione simultanea del peso di molteplici variabili nel determinare l'evoluzione di un determinato ambiente secondo fattori di interesse specifico definiti *a priori* (Shepard, 1962; Chatfield e Collins, 1980; Somerfield e Clarke, 1995; Clarke e Warwick, 1998; Warwick e Clarke, 2001).

La scelta di rappresentare i dati in modo aggregato con questo approccio è finalizzata a permettere una maggiore comprensione del quadro di insieme fornito dalla valutazione del contesto ambientale e della sua evoluzione in relazione all'intervento, evidenziando tendenze complessive difficilmente rilevabili mediante l'approccio univariato basato, al contrario, su un'analisi puntuale dei singoli valori per ogni variabile.

La statistica multivariata permette di analizzare simultaneamente più caratteri definendo anche il peso relativo dei diversi fattori esterni nel determinare le variabili del sistema. Il risultato complessivo dell'analisi consente di individuare, fra le numerose variabili analizzate, quelle che danno il maggior contributo alla distribuzione multivariata complessiva di tutta la matrice dei dati.

Le elaborazioni statistiche univariate e multivariate rappresentate sono state condotte impiegando il Software Primer® v.7.0. (Plymouth Marine Laboratory, UK) in accordo con le metodiche riportate da Clarke and Warwick, (2001).

Per le rappresentazioni sono state utilizzate le tecniche univariate dei Box-Whiskers e delle correlazioni lineari di Spearman e le tecniche multivariate di analisi classica delle componenti principali (PCA), l'analisi della non-metric Multi-Dimensional Scaling (nMDS) e il test ANOSIM della varianza. Le specifiche di ognuno di questi sono riportate a seguire.

La tecnica di calcolo univariata dei Box-Whiskers è stata utilizzata per visualizzare l'andamento aggregato di ogni parametro rispetto al fattore "fase" evidenziando, in ogni rappresentazione, il valore mediano (barra centrale), il primo e terzo quartile (variazione del box), il minimo e massimo (Whiskers) e gli *outlayers* (cerchi).

L'analisi classica delle componenti principali (PCA) è stata effettuata sui dati non trattati preliminarmente per individuare le funzioni in grado di descrivere in modo sintetico la maggior parte della varianza della distribuzione congiunta di tutti i composti analizzati (Legendre & Legendre 1998). Per ogni rappresentazione sono riportati anche i vettori di correlazione degli assi rispetto alle singole variabili considerate.

I dati sono stati acquisiti nella loro integrità, in questa fase, salvo casi evidenti, non sono state effettuate valutazioni relative alla qualità del dato prodotto. È anche da evidenziare che dal punto di vista del numero delle acquisizioni, il dataset ante operam è sicuramente più abbondante rispetto al dataset relativo alle fasi di cantiere con un conseguente sbilanciamento nelle elaborazioni statistiche che deve essere tenuto in debita considerazione.

I dati ambientali (descrittori principali della colonna d'acqua e concentrazione di contaminanti nelle

diverse matrici) sono stati trasformati applicando la funzione *square root* quindi normalizzati (Clarke and Green, 1988) per procedere al calcolo della matrice Euclidea delle distanze dei campioni e, quando significativa, delle variabili. Nel caso di dati relativi ad abbondanze biologiche (es. dati di contaminazione microbica espressi come conteggi/volume), si è optato per trasformazione mediante la funzione *square root* e successivo calcolo della matrice delle similarità di Bray-Curtis considerata dalla letteratura specifica, maggiormente idonea per trattare questo genere di dati di partenza.

Le matrici delle distanze sono state utilizzate per effettuare l'analisi *non-metric Multi-Dimensional Scaling* (nMDS) condotta applicando al calcolo la Kruskal stress formula: 1 con minimum stress: 0,01 (Kruskal, 1964). La proiezione bidimensionale delle distanze ottenuta e associata al diagramma di Sheppard è stata plottata evidenziando i fattori di interesse specifico definiti *a priori* e considerati in grado di "condizionare" la varianza complessiva del sistema. Tra i fattori testati, la "fase" (sei livelli fissi: AO1, AO2, AO3, C1, C2 e C3) è stata considerata di interesse specifico ed evidenziata nelle rappresentazioni.

L'analisi della significatività della distribuzione osservata dai diagrammi bidimensionali nMDS rispetto al fattore "fase" è stata testata mediante il test ANOSIM (Analysis of Similarity) *one-way*. Questo è un test statistico non parametrico ampiamente utilizzato nel campo dell'ecologia. Il test è stato suggerito dalla letteratura citata come un test simile all'ANOVA, dove invece di operare su dati grezzi, opera su una matrice di dissimilarità classificata.

Data una matrice di differenze di rango tra un insieme di campioni, ciascuno appartenente esclusivamente a un gruppo di trattamento, l'ANOSIM verifica se è possibile rifiutare l'ipotesi nulla che la somiglianza tra i gruppi sia maggiore o uguale alla somiglianza all'interno dei gruppi. La statistica del test R viene calcolata secondo la seguente formula: $R = 2 * (rB - rW) / M$; dove rB è la media delle somiglianze di rango di coppie di campioni indicizzati con fattori diversi tra loro, rW è la media della somiglianza di rango di coppie degli stessi fattori e M è uguale a $n(n - 1) / 2$ dove n è il numero di campioni considerati. La statistica del test R è vincolata tra i valori compresi tra -1 e 1. Ai fini della verifica delle ipotesi, dove l'ipotesi è nulla la statistica R viene solitamente confrontata con un insieme di valori R' che si ottengono mediante mescolamento casuale delle etichette dei siti tra i campioni e calcolando la R' risultante, ripetuta molte volte. La percentuale di volte in cui la R effettiva ha superato i valori R' derivati dalle permutazioni è il valore p per la statistica R effettiva.

Infine, alcune premesse alla valutazione sintetica sono risultate necessarie in relazione alle specifiche caratteristiche dei dati e dei processi di analisi statistica.

Le analisi di laboratorio riportate in questo documento sono state condotte, nel corso del tempo (novembre 2015 ÷ ottobre 2021), da diversi laboratori di analisi. Questo, fermo restando l'impiego da parte del laboratorio incaricato di metodi validati e riferibili (es. ISO ecc.), può avere implicato, in alcuni casi, una differenza nel metodo di analisi adottato dal laboratorio per misurare le variabili oggetto di indagine.

In questo contesto si è assunto che le eventuali differenze tra metodi di analisi adottati per il rilievo della stessa variabile siano trascurabili e i dati sono stati gestiti come un insieme omogeneo. Le analisi statistiche, quindi, sono state condotte su tutto il *database* acquisito fino a questo momento, indipendentemente dalla metodica di analisi applicata dal laboratorio incaricato di eseguire le prove. Le unità di misura, in caso di difformità nell'espressione del dato, sono state convertite, ove possibile, per renderle uniformi.

Si è scelto di trattare separatamente le variabili per le quali le acquisizioni non risultavano effettuate su tutti i *record* di monitoraggio.

Nel caso in cui *outlayers* puntuali abbiano avuto una diversità tale da condizionare l'elaborazione

statistica multivariata agendo in modo da appiattare gli andamenti generali degli altri rilievi condotti, ove necessario per questioni di chiarezza di restituzione, l'analisi è stata effettuata escludendo la riga di registrazione specifica.

Le molecole che non sono mai state quantificate (sempre <Limit Of Quantification, LOQ) in nessun campione oggetto di indagine sono state eliminate dall'elaborazione statistica multivariata. In caso di valori rilevati occasionalmente a concentrazioni <LOQ del metodo, utilizzando l'approccio più conservativo per l'ambiente, si è inserito nel database il valore del limite di rilevabilità stesso.

Ulteriori assunti specifici preliminari all'analisi del database sono di seguito riportati per la singola matrice oggetto di analisi e d'interesse per il caso in esame.

Riferimenti bibliografici specifici

- Chatfield C. and Collins A.J., 1980. Introduction to multivariate analysis. Chapman and Hall, London.
- Clarke K.R. and Green R.H., 1988. Statistical design and analysis for a "biological effects" study. Mar. Ecol. Prog. Ser. 46: 213-226.
- Clarke K.R. and Warwick R.M., 2001. Change in marine communities: an approach to statistical analysis and interpretation, 2nd Edition. Primer-E: Plymouth
- Clarke K.R. and Warwick R.M., 1998. A taxonomic distinctness index and its statistical properties. J. appl. Ecol. 35: 523-531.
- Shepard R.N., 1962. The analysis of proximities: multidimensional scaling with an unknown distances function. Psychometrika, 27: 125-140.
- Somerfield P.J. and Clarke K.R., 1995. A comparison of some methods commonly used for the collection of sublittoral sediments and their associated fauna. Mar. Env. Res. 43: 143-156.
- Kruskal J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika. 29:1-27.
- Warwick R.M. and Clarke K.R., 2001. Practical measures of marine biodiversity based on relatedness of species. Oceanogr. Mar. Biol. Ann. Rev. 39: 207-231.
- Legendre P. & Legendre L., 1998. Numerical Ecology, 2nd Engl. Edition, Elsevier, Amsterdam.

ASSUNTI PRELIMINARI RELATIVI ALL'ELABORAZIONE STATISTICA EFFETTUATA SUL DATABASE RELATIVO AI RILEVI CON SONDA MULTIPARAMETRICA

I dati riferiti a queste elaborazioni sono relativi ai livelli registrati per conducibilità ($\mu\text{S}/\text{cm}$), ossigeno disciolto (% saturazione e mg/L), pH (udpH), potenziale redox (mV), torbidità (NTU) e clorofilla-*a* (mg/m^3) nelle fasi *ante operam* (AO1, AO2 e AO3) e cantiere (C1, C2 e C3). Le stazioni in cui non sono state registrati record completamente o in parte sono state eliminate dalla elaborazione statistica, i valori non misurati in modo sporadico sono stati trattati come "*missing value*". Gli *outlayers* evidentemente riferibili a problemi di acquisizione del dato (es. valori di torbidità negativi) sono stati eliminati dall'elaborazione statistica. Nel caso dei rilievi condotti nelle stazioni denominate ST01-ST08, i dati analizzati sono riferiti ai monitoraggi condotti dal novembre 2015 all'ottobre 2021 a diverse profondità della colonna d'acqua. I fattori d'interesse specifico sui quali testare la significatività delle segregazioni osservate sono stati la "fase" del monitoraggio (ante operam, cantiere; sei livelli, fissi), la profondità di campionamento (livelli 14). I dati relativi a questi rilievi presentavano valori positivi e negativi (es. potenziale red-ox), non è stata applicata la funzione di pretrattamento "*square root*" ma solo la normalizzazione prima del calcolo della matrice Euclidea delle distanze.

ASSUNTI PRELIMINARI RELATIVI ALL'ELABORAZIONE STATISTICA EFFETTUATA SUL DATABASE RELATIVO AL MONITORAGGIO DELLA COLONNA D'ACQUA

Le stazioni in cui non sono state registrati record completamente o in parte sono state eliminate dalla elaborazione statistica, i valori sporadici non misurati sono stati trattati come "*missing value*". Gli *outlayers* evidentemente riferibili a problemi di acquisizione del dato oppure tali da determinare effetti significativi di "schiacciamento" sugli andamenti generali, se presenti, sono stati eliminati dall'elaborazione statistica.

ASSUNTI PRELIMINARI RELATIVI ALL'ELABORAZIONE STATISTICA EFFETTUATA SUL DATABASE RELATIVO AL MONITORAGGIO DEI SEDIMENTI

Questi dati sono riferiti ai livelli registrati per contaminanti ambientali, ecotossicologia e microrganismi nelle fasi *ante operam* (AO1, AO2 e AO3) e cantiere (C1, C2 e C3).

Relativamente alle determinazioni chimico-fisiche sul sedimento, l'elaborazione dei risultati dei sedimenti è stata condotta su due livelli separando le stazioni per le quali si ha il *dataset* completo su tutti i parametri dalle stazioni per le quali si ha un parziale di analiti determinati su una percentuale del totale di campioni prelevati. Gli Xileni sono stati considerati nell'elaborazione statistica solo come totale; anche DDD, DDT, DDE sono stati considerati solo come totale. In alcuni casi è riportata separatamente la quantificazione della singola molecola Benzo(j)Fluorantene e della molecola Benzo(k)Fluorantene; in altri casi le due molecole sono quantificate dal laboratorio come Benzo(j+k)Fluorantene; si è scelto, per uniformità, di sommare sempre le concentrazioni delle due molecole di IPA analizzando il contributo di Benzo(j+k)Fluorantene.

Sono state eliminate dall'elaborazione statistica tutte quelle molecole che hanno mostrato sempre concentrazione inferiore al limite di rilevabilità. Le stazioni in cui non sono stati registrati record completamente o in parte sono state eliminate dalla elaborazione statistica, i valori sporadici non misurati sono stati trattati come "*missing value*".

Gli *outlayers* evidentemente riferibili a problemi di acquisizione del dato, se presenti, sono stati eliminati dall'elaborazione statistica. Per le molecole che occasionalmente hanno mostrato valori <LOQ, a scopo cautelativo, si è scelto di riportare il valore del limite di quantificazione come valore rilevato.

Per il dataset ecotossicologico N.D. è stato interpretato come non determinato (*missing*) e N.C. come effetto non calcolabile (tossicità assente). La fase di cantiere non è presente nelle elaborazioni box-plot in quanto i risultati sono sempre stati "missing value".

Anche i dati microbiologici sono stati trattati separatamente come dati di abbondanze per volume di campione testato. Per la trattazione dei dati microbiologici valgono le indicazioni fornite in precedenza. I dati sono stati convertiti in abbondanze numeriche espresse come numero a notazione non esponenziale. Le differenze metodologiche relative all'espressione MPN (Major Probable Number) e UFC (Unità Formanti Colonia) sono state considerate trascurabili per questa finalità e la popolazione dei dati è stata considerata omogenea; tuttavia, i risultati statistici relativamente al fattore "fase" devono essere letti tenendo in considerazione questa importante premessa alla elaborazione dei dati.