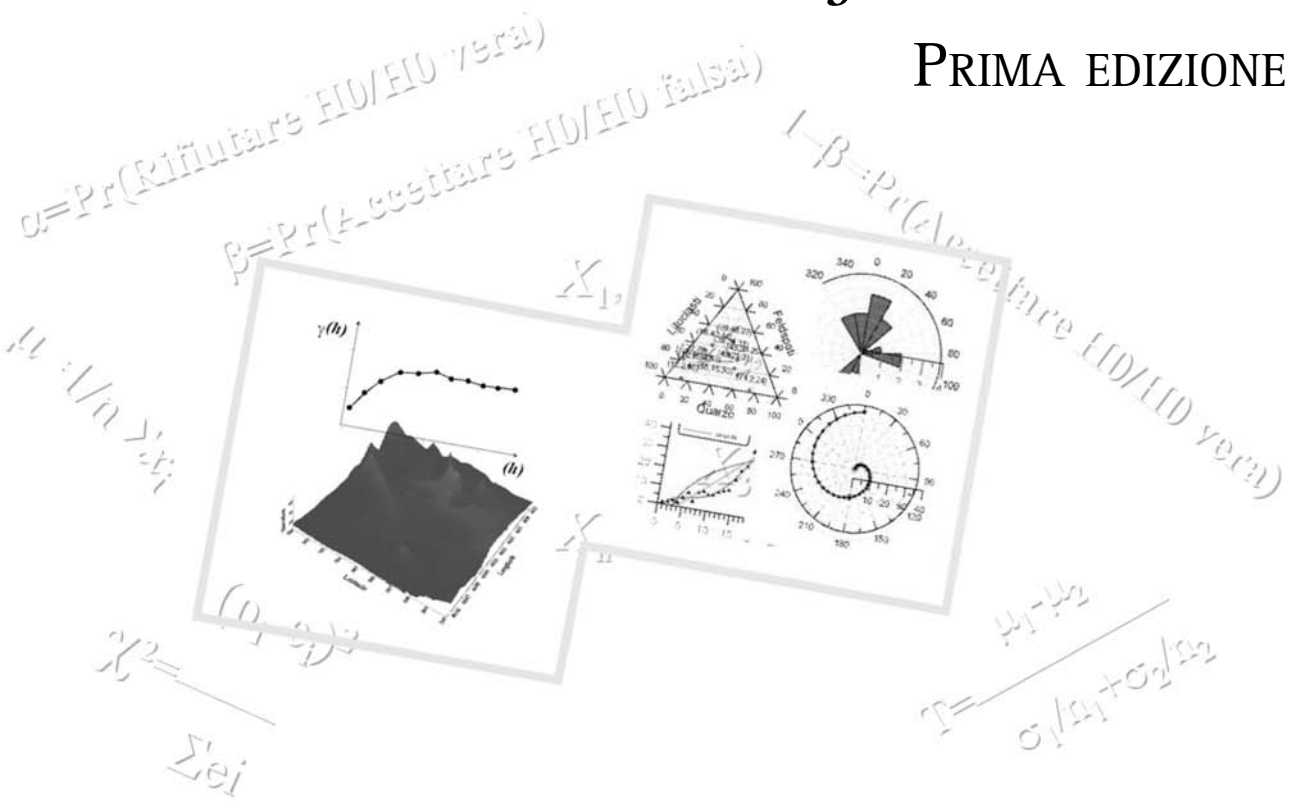
Giancarlo Ciotoli
Dipartimento di Scienze della Terra,
Università di Roma "La Sapienza"

Maria Grazia Finoia
Istituto Centrale per la Ricerca Applicata al Mare

Dalla Statistica alla Geostatistica:

Introduzione all'analisi dei Dati Geologici e Ambientali

PRIMA EDIZIONE



ARACNE

www.aracneeditrice.it
info@aracneeditrice.it

via Raffaele Garofalo, 133 A/B
00173 Roma
(06) 93781065

ISBN 88–548–0191–7

I edizione: settembre 2005

L'esperienza è il tipo di insegnante più difficile.
Prima ti fa l'esame, e poi ti spiega la lezione.

Anonimo

SOMMARIO

CAPITOLO X – La Geostatistica: studio della correlazione spaziale

CAPITOLO XI – La Geostatistica: i metodi di stima

Softwares Utilizzati:

Minitab Release 14© 1972 - 2005 Minitab Inc., www.minitab.com

STATISTICA (sistema software di analisi dei dati), versione 6. StatSoft Italia srl (2003). www.statsoft.it

Grapher – Version 5.04.21 January 19, 2005. Graphing System, Copyright© 1992-2005, Golden Software Inc. Golden (CO, USA), www.goldensoftware.com.

Surfer 8, Golden Software Inc., Golden Software Inc., Golden, CO, USA, 1994, www.goldensoftware.com.

Statgraphic Plus 5.0, Copyright 1994-2000, Statistical Graphics Corps, www.statgraphics.com

In nihil sapiendo
iucundissima vita!
Adagiorum Chiliades, 2.10.81

Desideri Erasmi Roterodami

PREMESSA

L'esigenza di raccogliere ed organizzare in un opera unica tutte le informazioni ottenute in sette anni di studi e di ricerche sull'applicazione dei metodi statistici e geostatistici nell'analisi dei dati ambientali nasce dalla consapevolezza, maturata in questo periodo, dell'esistenza di un "gap" culturale nell'ambito delle discipline che fino a qualche anno fa appartenevano alle Scienze "naturalistiche" (Geologia, Biologia, Zoologia, Botanica, Biochimica, Mineralogia, Paleontologia, ecc.), ma che ora con l'avvento delle nuove tecnologie sperimentali ed informatiche fanno parte integrante delle Scienze Applicate.

L'idea è nata nel famoso periodo "buio" del Dottorato di Ricerca (chi lo ha vissuto in prima persona lo conosce bene!!) un momento in cui, pur consapevoli di dover affrontare da soli i primi ed importanti passi nel mondo della ricerca, non si riesce a definirne le modalità (come, dove e perché!!). Fu proprio durante questo periodo che il dott. Maurizio Guerra, grande amico, nonché compagno di "sventura", lanciò l'input che mi ha fatto balenare la classica "lampadina" dei fumetti di Archimede. Il consiglio che mi fu dato riguardava il miglioramento delle tecniche d'interpretazione dei dati sperimentali mediante l'applicazione di una disciplina, allora abbastanza sconosciuta tra noi geologi, nota con il nome di ***Geostatistica****.*

Intrapresi questa strada consapevole delle limitate conoscenze di Matematica Applicata e di Statistica fornite dal vecchio corso di Laurea in Scienze Geologiche, ma incoraggiato dall'idea di dare nuovi stimoli e potenzialità al mio futuro lavoro di ricerca nel Laboratorio di Chimica dei Fluidi del Dipartimento di Scienze della Terra, dell'Università di Roma "La Sapienza" dove tuttora passo gran parte delle mie giornate.

Dedicai, inizialmente, molto tempo allo studio teorico, ma soprattutto all'applicazione sui dati sperimentali dei metodi di interpretazione afferenti alla Statistica Classica, passando di seguito allo studio ed all'applicazione delle tecniche geostatistiche per l'analisi della correlazione spaziale di variabili ambientali. Durante questo periodo ho avuto la fortuna di rincontrare un vecchio amico, il dott. Sergio Cedrone, che con la sua grande esperienza e competenza nel campo della Statistica inferenziale, mi ha fornito le basi necessarie per affrontare da solo gli studi avanzati di Statistica, nonché per cimentarmi anche con le problematiche connesse all'applicazione dei metodi dell'analisi multivariata. Per quanto riguarda lo studio della Geostatistica e le sue applicazioni sono stati fondamentali il corso e, soprattutto, i consigli e la pazienza del Prof. Giovanni Raspa docente di Geostatistica Applicata presso la Facoltà di Ingegneria dell'Università di Roma "La Sapienza".

Con la breve, ma fondamentale, esperienza maturata in quegli anni la mia ricerca di dottorato ottenne un risultato finale molto soddisfacente, e al di là di ogni aspettativa, questo lavoro fu talmente gratificante da indurmi a continuare quella strada che avevo intrapreso tra mille difficoltà. La maturazione dell'esperienza passata ed la continua voglia di migliorala mi hanno condotto a prendere in seria considerazione la possibilità di raccogliere ed uniformare le mie conoscenze riguardo l'applicazione delle tecniche principali dell'analisi statistica ai dati sperimentali legati all'ambiente ed al territorio.

Le informazioni raccolte in questo libro sono, pertanto, rivolte soprattutto a chi ha "subito" durante il Corso di Laurea il "peso didattico" di un insegnamento di Statistica fine a se stesso e poco funzionale alle esigenze di chi è impegnato giornalmente nella raccolta e nell'interpretazione dei dati ambientali e territoriali. E' mia opinione, infatti, che sia lo statistico sia il matematico sono in grado di fornire un'elaborazione completa ed anche complessa dei dati sperimentali, ma solo il geologo, il biologo, il petrografo, il biochimico, conoscitore di statistica potrà utilizzare le tecniche dell'analisi statistica e geostatistica in modo da ottenerne una corretta interpretazione.

Due sono sostanzialmente le motivazioni che mi hanno suggerito e spinto a scrivere un lavoro così ambizioso ma, secondo me, necessario. Da un lato la necessità di introdurre nel corso di Laurea in Scienze Geologiche un insegnamento di Statistica e di Geostatistica Applicata tenuto da un docente geologo che sia in grado sia di fornire agli studenti le conoscenze teoriche di base, ma soprattutto che sia in grado di fornire con la sua esperienza gli strumenti per una corretta interpretazione qualitativa e quantitativa dei dati sperimentali relativi ai fenomeni naturali. L'insegnamento dei metodi della Statistica Applicata deve necessariamente diversificarsi da sede a sede e nell'ambito dei diversi corsi di laurea, in funzione degli interessi ed in sintonia con l'autonomia didattica del nuovo ordinamento universitario.

Un secondo aspetto riguarda lo sviluppo in quest'ultimo decennio degli strumenti di calcolo e di softwares specifici. L'esplosiva proliferazione dei potenti PC di nuova generazione ha permesso anche a piccole compagnie, società e singoli professionisti di entrare in un mondo che in precedenza era ristretto alle grandi aziende o alle Università. Pertanto molti geologi (come anche i biologi ed i naturalisti) sono stati "catturati" dall'improvvisa evoluzione dei sistemi informatici, ma, educati tradizionalmente ad enfatizzare la qualità alle spese della quantità, la loro conoscenza dei metodi della Statistica Applicata è completamente inadeguata, nonostante molte delle case produttrici accompagnino i loro prodotti con una serie di manuali in cui sono messe in risalto le principali procedure geomatematiche. La tentazione di utilizzare tali programmi supera la conoscenza di base che in alcuni casi è completamente sconosciuta. Un esempio è costituito dai softwares, estremamente attuali, utilizzati per la costruzione e la gestione dei Sistemi Geografici Informativi, softwares che generalmente sono utilizzati al 10-20% delle loro reali potenzialità.

Pertanto, l'obiettivo del testo più che limitarsi ad una trattazione teorica degli argomenti, di poca utilità per chi si trova ad affrontare giornalmente la problematica dell'interpretazione dei dati sperimentali, è quello di aiutare il lettore nella scelta delle tecniche più opportune dell'analisi statistica al fine di trarre dai dati a disposizione più informazioni possibili per una buona valutazione qualitativa e quantitativa del fenomeno da essi rappresentato.

L'organizzazione degli argomenti trattati nel testo prevede una parte iniziale in cui sono fornite le informazioni fondamentali della statistica di base applicate all'analisi dei dati sperimentali, fino a giungere all'applicazione delle tecniche dell'analisi statistica multivariata. Poiché l'obiettivo del testo è l'analisi delle variabili ambientali e, quindi legate al territorio, gli ultimi capitoli sono stati dedicati all'analisi spaziale delle variabili regionalizzate per lo studio del comportamento spaziale delle variabili geologiche e naturali attraverso l'introduzione dei metodi geostatistici. A tale scopo è stata aggiunta una sezione riguardante le tecniche moderne di "contour mapping".

Nell'ambito della sezione dedicata ai metodi della statistica di base sono forniti, inoltre, i fondamenti del calcolo delle probabilità necessari in quasi tutte le procedure di analisi dei dati, soprattutto nell'applicazione dei test statistici. A tale proposito è stato introdotto un capitolo sui test non parametrici che non richiedono un livello di misurazione elevato: molti di essi, infatti,

possono essere applicati su scale ordinali e nominali generalmente utilizzate per la misura di molte variabili geologiche quali, i dati geochimichi, petrografici e granulometrici.

Al termine di ogni capitolo è presente una breve lista di lavori e/o testi specialistici riguardanti gli argomenti trattati. I dati utilizzati nelle finestre di esercizio e negli esempi riportati costituiscono una raccolta che in generale proviene dalle differenti discipline nell'ambito delle Scienze Geologiche e Naturali.

Sono stato fortunato ad avere l'aiuto e l'incoraggiamento incondizionato della dott.ssa Maria Grazia Finoia, ricercatrice dell'Istituto Centrale per la Ricerca Applicata al Mare, che ha messo a disposizione la sua grande esperienza e, soprattutto, il suo tempo per la buona riuscita di questa opera. Il suo lavoro di costruzione, elaborazione e revisione ha premesso di creare quell'"anello mancante" di congiunzione tra i metodi della Statistica e la loro applicazione per l'interpretazione delle problematiche legate alle Scienze Ambientali in genere.

Vorrei, inoltre, ringraziare il Prof. Salvatore Lombardi per l'aiuto e la fiducia incondizionata in tutti questi anni di collaborazione trascorsi presso il suo laboratorio.

In genere un libro, e soprattutto il primo, riporta sempre una dedica per le persone care, i familiari, ecc. In generale vorrei dedicare questo mio lavoro a tutti coloro che mi sono vicini e che con la loro presenza mi incoraggiano giornalmente dandomi la forza ed il sorriso per superare le mille difficoltà della vita, anche le più piccole. In particolare, desidero dedicare questo libro a un incontro speciale con una persona speciale, Marilena!

Un ultimo, ma non ultimo, pensiero è per mio fratello Fabio, forza!!!!

Giancarlo Ciotoli
Agosto 2005, Roma

CAPITOLO I

INTRODUZIONE

1.1 E' veramente necessaria l'analisi statistica dei dati?

1.2 Perché analizzare i dati ambientali e territoriali

1.3 Scopo ed organizzazione del libro

1.4 Gli obiettivi dell'analisi statistica dei dati

1.5 Il "Data Mining"

Riassunto

Il testo inizia con una domanda fondamentale, "Abbiamo davvero bisogno della Statistica?" Se il nostro primo "amore" fosse davvero stato la Statistica, probabilmente non avremmo studiato o lavorato nel campo delle Scienze Ambientali e Territoriali. In effetti, probabilmente, essere un brillante statistico non è una condizione necessaria per essere un brillante biologo, geologo, ecc. Il fatto che la maggior parte dei corsi di laurea in discipline afferenti alle scienze naturali e territoriali includa un breve, obbligatorio e, talvolta forzato, corso di statistica, non significa che esso non debba fornire le informazioni necessarie su come, quando e perché applicarne le tecniche fondamentali.

Tutto quello che riguarda la descrizione qualitativa e quantitativa dei fenomeni naturali, caratterizzati da una notevole complessità e variabilità, è effettuata mediante l'interpretazione dei dati prodotti dal lavoro di campagna, di misure e di determinazioni analitiche spesso condotte da soggetti diversi, con metodologie differenti e, in alcuni casi, in tempi diversi. Questa enorme mole di dati sperimentali deve essere opportunamente gestita ed elaborata al fine di valutare correttamente le condizioni reali dei fenomeni da essi descritti. Tutto questo dimostra l'utilità e la necessità del trattamento statistico dei dati che può essere effettuato con diversi gradi di approfondimento a seconda delle conoscenze, della disponibilità di mezzi ed informazioni e dell'importanza dei problemi.

Il testo ha quindi lo scopo di mettere a disposizione dei ricercatori, ma anche degli studenti e di tutti gli operatori di Enti pubblici e/o privati (biologi, geologi, naturalisti, ecc.), uno strumento utile per la descrizione e l'interpretazione dei dati acquisiti su fenomeni ambientali anche al fine di poter meglio valutare le situazioni di rischio causate dalle attività antropiche. Tuttavia, l'utilizzo delle tecniche statistiche e geostatistiche non costituisce uno strumento esclusivo per l'analisi dei dati ambientali e territoriali, ma concorre con le altre tecniche tradizionali di indagine nell'interpretazione dei fenomeni e delle variabili che li descrivono. E', comunque , necessario sottolineare che uno dei requisiti fondamentali rimane sempre la profonda conoscenza del fenomeno che si sta studiando.

1.1 E’ veramente necessaria l’analisi statistica dei dati?

Se qualcuno cerca di venderti un teso di statistica sicuramente non sarà la persona che ti chiederà se hai bisogno della Statistica! In ogni caso tu avrai aperto questo libro perché hai compreso l’immediato bisogno che hai di applicare le tecniche statistiche. Comunque sia la risposta è chiaramente sì, hai bisogno della Statistica! Tuttavia per sapere se la Statistica è veramente necessaria domanderai a diverse persone (studenti, ricercatori, ecc,) se effettivamente se l’applicazione delle tecniche statistica ha portato giovamento al loro lavoro. La risposta, naturalmente, sarà ambigua ed inoltre troverai diverse persone che avranno sicuramente reso più pregevole il loro lavoro di ricerca pur non essendo statistici. In effetti, probabilmente, essere un brillante statistico non è una condizione necessaria per essere un brillante biologo, geologo, ecc. Ed in ogni caso anche se non si ha bisogno immediato di avere basi di Statistica, esse sono necessarie per capire il vero significato di quasi tutte le informazioni scientifiche in nostro possesso.

Inoltre, il fatto che la maggior parte dei corsi di laurea in discipline afferenti alle scienze naturali e territoriali includa un breve, obbligatorio e, talvolta forzato, corso di Statistica, non significa che esso non debba fornire le informazioni necessarie su come, quando e perché applicarne le tecniche fondamentali.

1.2 Perché analizzare i dati ambientali e territoriali

La ricerca nel campo delle scienze naturalistiche ed applicate necessita la conoscenza dei fenomeni naturali e della loro variabilità nello spazio e nel tempo. Pertanto, le discipline, come la Geologia, la Biologia e le Scienze Ambientali, la cui evoluzione storica vuole siano basate principalmente sull’osservazione di campagna (numero di fossili in una formazione, misure di porosità, analisi di concentrazione di inquinanti, direzione di una faglia, ecc.), dipendono molto dal modo con cui queste osservazioni sono raccolte ed elaborate,soprattutto quelle che sono caratterizzate da un grande margine di incertezza. La statistica e la geostatistica possono, pertanto, giocare un ruolo abbastanza importante nella ricerca applicata a queste discipline sebbene alcune di esse non hanno una grossa tradizione riguardo l’analisi numerica dei dati.

Il problema ha due risposte una di tipo filosofico ed una più pragmatica:

- la necessità del progresso scientifico di utilizzare procedure oggettive per la verifica delle ipotesi. Guardando una roccia non è sicuramente scientifico affermare con certezza “Questo è un granito!!”. Un geologo avrà sicuramente degli strumenti ed una serie di criteri scientifici per verificare ciò che sembra intuitivo. L’utilizzo di criteri qualitativi è inevitabilmente soggettivo, cosicché è bene affrontare il problema in termini numerici.

- in passato, ad esempio, i geologi hanno trattato le rocce sulla base dell'osservazione, approccio utile dal punto nella fase di "field mapping". Tuttavia, attualmente, si tende ad avere anche queste prime informazioni in forma numerica; in questo modo è più facile riassumere le osservazioni rilevate e comunicare i risultati ottenuti.

In effetti tutto quello che riguarda la descrizione qualitativa e quantitativa dei fenomeni naturali, in alcuni casi notevolmente complessi e variabili, è effettuata mediante l'interpretazione dei dati prodotti dal lavoro di campagna, da misure e da determinazioni analitiche spesso condotte da soggetti diversi, con metodologie differenti e, in alcuni casi, in tempi diversi. Tutto ciò grazie anche al miglioramento delle tecniche di rilevamento dati (immagini satellitari, geofisica, ecc.), alla convenienza nell'immagazzinamento dei dati in forma numerica, alla sensibilità della nuova strumentazione nell'apprezzare anche le piccole variazioni (elementi in traccia, permeabilità, ecc.). Questa enorme mole di dati sperimentali deve essere opportunamente gestita ed elaborata al fine di valutare correttamente le condizioni reali dei fenomeni da essi descritti. Quanto detto dimostra l'utilità e la necessità del trattamento statistico dei dati che può essere effettuato con diversi gradi di approfondimento a seconda delle conoscenze, della disponibilità di mezzi ed informazioni e dell'importanza dei problemi.

La raccolta dei dati sperimentali costituisce solo una informazione grezza e non la vera conoscenza del fenomeno che essi rappresentano. La sequenza logica che separa i dati dalla conoscenza è costituita dalle seguenti fasi:

- raccolta dei dati allo scopo di ottenere informazioni su un dato fenomeno;
- analisi delle informazioni raccolte e loro formalizzazione in fatti;
- definizione dei fatti e passaggio alla conoscenza del fenomeno oggetto di studio.

I dati, quindi, diventano informazioni solo quando costituiscono un supporto rilevante per le decisioni di intervento; le informazioni diventano fatti solo quando sono supportate dai dati; i fatti diventano conoscenza solo quando sono utilizzati nel completamento del processo di intervento (Fig. 1.1). Questo costituisce il motivo fondamentale che conduce alla necessità di affrontare l'analisi statistica dei dati. Essa richiede lo studio delle proprietà e delle relazioni fra i dati, nonché lo studio delle leggi di probabilità. E' auspicabile in questo modo minimizzare quella che potrebbe essere un'interpretazione soggettiva dei dati ambientali, che potrebbe di seguito avere conseguenze sull'applicazione di interventi e provvedimenti anche di tipo amministrativo, i quali potrebbero a loro volta avere risvolti negativi sugli aspetti tecnici, economici e giuridici del problema.

Tuttavia la conoscenza è molto di più della conoscenza di qualcosa di tecnico. La conoscenza è legata alla saggezza, e la saggezza all'esperienza ed all'età; la saggezza, ad esempio permette

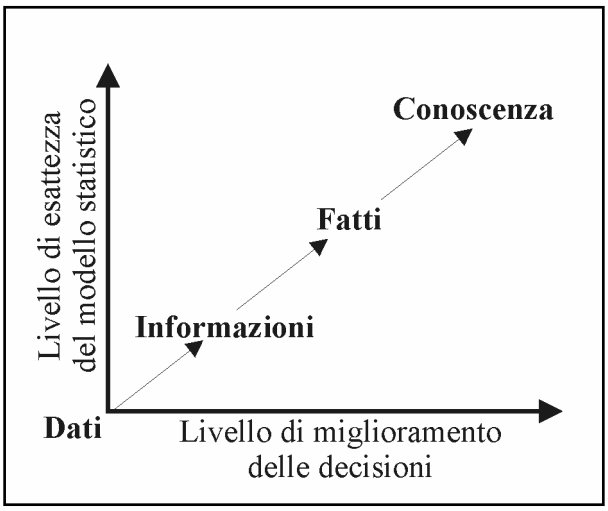


Figura 1.1 La figura illustra il processo del pensiero statistico che a partire dalla conoscenza dei dati giunge alla costruzione di un modello utilizzato nell'ambito delle operazioni di intervento nei casi di incertezza.

l'elaborazione di modelli semplici ed utili, piuttosto che di modelli complessi e tecnicamente brillanti.

La statistica è una scienza che ci consente di prendere delle decisioni nei casi di incertezza (aspetto probabilistico), decisioni che devono essere prese sulla base dei dati a disposizione non sulla base delle proprie credenze e opinioni personali.

L'aspetto probabilistico ed il relativo grado di incertezza dei risultati dipendono principalmente dal fatto che usualmente, nel campo della ricerca, lo studio di un dato fenomeno viene svolto in base ad una sua osservazione 'parziale' o campionaria. L'intento di estrapolare i risultati, in tal modo ottenuti, al fenomeno considerato in tutta la sua complessità implica, inevitabilmente, l'introduzione di un errore che, nelle aspettative del ricercatore, dovrà essere il più piccolo possibile.

Attualmente, sebbene lo studio statistico dei dati costituisca un passo fondamentale dell'attività di ricerca di molti geologi, biologi e naturalisti, esiste ancora una grande lacuna tra l'aspetto degli studi accademici e l'utilità dell'applicazione pratica delle tecniche statistiche di elaborazione dei dati. La maggior parte dei testi di statistica utilizza ancora "l'estrazione di palline bianche e nere dall'urna", mentre le studio, l'interpretazione e la comprensione delle relazioni spazio-temporali che costituiscono il "cuore" dei fenomeni naturali sono raramente trattati. Per questo motivo è difficile per i geologi, i biologi e i naturalisti tradurre la teoria statistica in una soluzione pratica ed immediata di tali problematiche.

L'obiettivo principale di questo lavoro consiste nel promuovere l'uso del pensiero statistico e delle principali tecniche statistiche nello studio dei fenomeni naturali per produrre informazioni più consistenti migliorando anche quelle azioni che costituiranno le decisioni di intervento. In particolare, questo corso mira all'applicazione delle tecniche della Statistica Classica e della Statistica dei Dati Spaziali (Geostatistica) nella pratica sperimentale, pur cercando di fornire, senza dilungarsi in pesanti dissertazioni matematico-probabilistiche, gli elementi teorici che sono alla base di ogni singola tecnica. Pertanto, questo testo può essere considerato come un breve corso di statistica che, attraverso informazioni teoriche e applicazioni sperimentali mette a disposizione dei ricercatori, ma anche degli studenti e di tutti gli operatori di Enti pubblici e/o privati (biologi, geologi, naturalisti, ecc.), uno strumento utile per la descrizione e l'interpretazione dei dati acquisiti

su fenomeni ambientali anche al fine di poter meglio valutare le situazioni di rischio causate dalle attività antropiche.

1.3 Scopo ed organizzazione del libro

Poiché l'intenzione è stata quella di scrivere a coloro che desiderano applicare le tecniche statistiche a i dati ambientali e territoriali, si presuppone che i lettori conoscano i fenomeni naturali oggetto di studio, ma non necessariamente conoscano la Statistica.

Il testo è organizzato in due parti, nella prima parte sono descritti i principi statistici elementari (Cap. II e III) che sono sviluppati in maniera tale da essere direttamente inseriti ed utilizzati nella pratica dell'analisi dei dati ambientali. Tuttavia dai principi di base si passa alla descrizione ed utilizzazione pratica di tecniche più specializzate, quali l'applicazione dei test statistici parametrici e non parametrici (Cap. IV e V) e le tecniche di base della regressione (Cap. VI), che generalmente non è trattata nei testi di statistica elementare. La prima parte del testo termina con la descrizione delle tecniche che sono alla base dello studio delle serie storiche (Cap. VII) allo scopo di identificare la natura del fenomeno rappresentato da una sequenza di osservazioni svolte nel corso di un certo intervallo di tempo e di predirne i valori futuri.

La seconda parte del testo è dedicata allo studio dei fenomeni spaziali, a partire dall'analisi dei dati direzionali (Cap. VIII), per arrivare attraverso l'analisi delle distribuzioni di punti nello spazio (Cap. IX), alla descrizione delle tecniche della geostatistica (Cap. X) e dei metodi di interpolazione dei dati spaziali (Cap. XI).

La seconda parte si conclude con un capitolo dedicato alla descrizione di base delle tecniche dell'analisi statistica multivariata più comunemente applicate nel campo dello studio dei fenomeni naturali e territoriali (Cap. XII).

Nel testo di è cercato di riportare soltanto quelle tecniche, di semplice applicazione e strettamente necessarie, che consentono un miglioramento della fase interpretativa dei dati ambientali. L'applicazione di tali tecniche di analisi è spiegata attraverso molti esempi pratici utilizzando dati ambientali e territoriali reali. Inoltre, la trattazione teorica degli argomenti richiede semplici conoscenze algebriche in quanto si è cercato di omettere il formalismo matematico richiesto da dimostrazioni di calcolo più complesse.

1.4 Gli obiettivi dell'analisi statistica dei dati

Come è stato accennato i dati non costituiscono le informazioni! Per definire in cosa consiste l'analisi statistica dei dati per prima cosa dobbiamo definire la Statistica. **La Statistica è una scienza costituita da una serie di metodologie utilizzate per la raccolta, per l'analisi, per la**

presentazione, per l'interpretazione dei dati e per la scelta degli interventi in situazioni di incertezza. L'avvento dei computers ha sicuramente facilitato il lavoro interpretativo del ricercatore, anche se la nuova tecnologia ha sopravanzato l'abilità da parte dei tecnici di utilizzare a pieno le sue potenzialità.

Come gli altri aspetti delle Scienze Naturali, della Geologia, della Biologia anche l'analisi dei dati rappresenta una disciplina caratterizzata da diversi livelli di specializzazione.
Esistono tre livelli di abilità:

1. la piena conoscenza dell'aspetto matematico;
2. la conoscenza dell'aspetto concettuale delle tecniche principali di analisi e la loro applicazione;
3. l'approccio pragmatico, cioè una conoscenza basilare, ma una buona abilità nell'utilizzo delle procedure e nel trarre le giuste conclusioni.

Naturalmente l'obiettivo principale è quello di raggiungere il livello 2 per poter avere le conoscenze fondamentali a supportare le operazioni previste nel livello 3.

Il contenuto di questo libro è disegnato per le conoscenze di uno studente di un secondo-terzo anno di una laurea in Scienze Geologiche (Ambientali, Biologiche, Naturali, ecc.), ma include informazioni dedicate anche a corsi post-laurea (Master, Dottorato di Ricerca).

Il fatto di poter trarre delle conclusioni su un numero elevato di osservazioni attraverso lo studio di una piccola parte di esse costituisce il motivo principale per utilizzare la Statistica nelle Scienze Ambientali e Territoriali.

Supponiamo di voler studiare un piccolo campione di osservazioni selezionato. Le domande a cui cercheremo di rispondere sono essenzialmente due:

- se assumiamo che il campione di osservazioni che dobbiamo studiare è rappresentativo del gruppo da cui proviene, cosa possiamo dire riguardo al gruppo?
- con quale probabilità posso affermare che il gruppo di osservazioni che sto studiando è simile al gruppo originario?

Queste domande costituiscono il punto centrale dei metodi statistici descritti in questo testo che nella maggior parte dei casi sono utilizzati nelle Scienze Ambientali e Territoriali applicative.

In genere, affrontare lo studio dei fenomeni naturali (geologici) attraverso le tecniche statistiche di analisi dei dati prevede cinque fasi operative: la definizione del problema (l'individuazione del fenomeno), il piano di campionamento, la raccolta dei dati, l'analisi dei dati, la restituzione dei risultati.

La definizione del problema, e quindi l'individuazione del fenomeno oggetto di studio, è fondamentale per assicurarsi l'accuratezza dei dati da rilevare. E' estremamente difficoltoso

ottenere dei dati senza avere una chiara definizione del problema da affrontare. Attualmente la moderna tecnologia consente rende estremamente facile sia la fase di raccolta dei dati sia la fase dell'elaborazione statistica. Paradossalmente nei testi di statistica non è mai stata data la giusta importanza alla strategia di campionamento sostituendo a questa necessità dietro una più spinta analisi numerica. Al contrario, è molto importante definire numericamente e geometricamente la distribuzione dei dati da raccogliere (popolazione) su cui poi effettuare l'inferenza statistica (cfr. par. 2.7). Due aspetti importanti di uno studio statistico sono:

- la popolazione, la serie di tutti gli elementi di interesse in uno studio;
- il campione, un sottoinsieme della popolazione.

Da questi due aspetti nasce l'inferenza statistica, cioè l'estrapolazione e l'estensione delle informazioni contenute nel campione all'intera popolazione. Nel linguaggio matematico questo concetto viene definito come Ragionamento Induttivo, ossia la conoscenza del "tutto" a partire dal "particolare". I dati possono essere qualitativi, etichette o nomi che identificano ogni singolo elemento, e quantitativi, ossia numeri che esprimono una quantità in senso di numerosità o di grandezza. I dati possono essere ottenuti da fonti esistenti o mediante osservazioni e studi sperimentali (cfr. par. 2.5).

L'analisi dei dati si divide in due categorie: i metodi esplorativi (statistica descrittiva) ed i metodi confermativi (statistica inferenziale). I metodi esplorativi sono utilizzati per individuare aspetti significativi dei dati mediante semplici procedure aritmetiche e/o grafiche che li riassumono. I metodi confermativi si basano sulla teoria delle probabilità per rispondere a dei quesiti specifici misurando ed analizzando le incertezze associate ad eventi futuri.

I risultati ottenuti dall'analisi inferenziale sono generalmente presentati in forma di tabelle numeriche e/o grafici e, poiché è stato esaminato solo una parte dell'intera popolazione (campione), i risultati dovranno rifletter l'incertezza attraverso tabelle di probabilità ed intervalli numerici.

Per quanto concerne i softwares utilizzati per l'analisi statistica dei dati, attualmente esistono diversi prodotti anche molto specialistici. Tuttavia le caratteristiche che essi devono avere sono:

1. una capacità di immagazzinamento e facilità nella manipolazione dei dati (operazioni fra righe e colonne);
2. la possibilità di elaborare grafici fondamentali, quali istogrammi, diagrammi a dispersione, ecc.;
3. la possibilità di effettuare test statistici standard (t-test, ANOVA, correlazione, regressione, ecc.)
4. la capacità di elaborare ed analizzare i dati temporali (analisi di Fourier e analisi spettrale);
5. una serie di differenti algoritmi di stima per l'analisi delle variabili spaziali (IDW, triangolazione, interpolazione polinomiale, kriging, ecc.)

6. la possibilità di effettuare operazioni tra matrici e di analisi multivariata (analisi dei gruppi, delle componenti principali, discriminante).

Poiché, in generale i softwares in commercio non posseggono tutti queste caratteristiche è fondamentale la capacità di trasferimento dei dati in modo che i vari programmi possano interagire gli uni con gli altri.

1.5 Il "Data Mining

Il "Data Mining" è una procedura analitica utilizzata per esplorare ed estrarre informazioni da grandi database alla ricerca di relazioni sistematiche fra le variabili oggetto di studio. La presenza di computer veloci, di un facile immagazzinamento dei dati e di una migliore comunicazione informatica ha reso più semplice l'applicazione di tale tecnica statistica. La procedura consiste di tre fasi: l'esplorazione dei dati, la costruzione di un modello, la convalida/verifica del modello.

Ciò che distingue il Data Mining dalla Statistica Classica consiste nel fatto che il primo pone come obiettivo un'analisi secondaria dei dati al fine di individuare relazioni inaspettate rispetto a quelle per le quali i dati sono stati raccolti.

Pertanto il "data mining" consiste nell'estrazione di informazioni predittive nascoste in grandi database. Essa è una nuova tecnica analitica dalle grandi potenzialità utilizzata sopratutto nel campo delle indagini di mercato per la previsione di tendenze e comportamenti futuri, ma che può trovare importanti applicazioni nell'analisi dei dati territoriali legati a grosse banche dati quali quelli dei GIS (Geographic Information Systems).

L'estrazione della conoscenza di un fenomeno naturale a partire dai dati sperimentali può essere vista come un processo contenente diversi fasi alcune delle quali ben conosciute, mentre altre dipendono strettamente dal giudizio dell'esperto. In generale tutte le fasi coinvolgono lo sviluppo di un modello definito da:

1. Raccolta dati iniziale e formulazione della problematica.
2. Scelta del software più adatto per la costruzione del modello e la simulazione.
3. Costruzione del modello concettuale.
4. Rappresentazione del modello mediante equazioni, grafici, diagramma a blocchi, ecc.
5. Verifica del modello. Il modello è verificato solo per vedere se risponde alle ipotesi fatte dall'esperto.
6. Inizializzazione del modello. Consiste nel calcolo di valori iniziali ragionevoli da cui il modello deve partire.
7. Validazione del modello. I risultati della simulazione sono confrontati con I dati sperimentali.

8. Documentazione. Descrizione scritta del processo di modellizzazione, del modello, dei risultati delle simulazioni e delle applicazioni del modello.
9. Applicazione del modello nella risoluzione di problemi.

Bibliografia

Koch G.S.Jr., Link R.F., 2002, Statistical Analysis of Geological Data, Dover Publications, Inc., Mineola, NewYork, pp. 438.

Swan A.R.H., Sandilands M., 1995, Introduction to Geological Data Analysis. Blackwell Science Ldt, pp.431

Townend J., 2004, Practical Statistics for Environmental and Biological Scientists, John Wiley & Sons, ltd, England, pp.276.

Westphal Ch., T. Blaxton, Data Mining Solutions: Methods and Tools for Solving Real-World Problems, John Wiley, 1998.

CAPITOLO II

RACCOLTA ED ORGANIZZAZIONE DEI DATI

Riassunto
Come in tutta la ricerca scientifica sperimentale, anche nelle scienze ambientali, in quelle geologiche e in quelle biologiche è indispensabile la conoscenza dei concetti e dei metodi statistici, sia per i problemi di gestione sia per quelli di indagine. Affinché i risultati di una ricerca siano facilmente comprensibili si richiede che la presentazione dei dati e la loro elaborazione seguano criteri ritenuti validi universalmente. Tali criteri prevedono una fase di raccolta dei dati, la loro descrizione, le analisi e il riepilogo. Una raccolta di dati non corretta, una loro presentazione inadeguata o un'analisi statistica non appropriata rendono impossibile la verifica dei risultati da parte di altri studiosi e il confronto con altre ricerche e analisi del settore. Per condurre in modo corretto una ricerca scientifica, una volta raccolto un campione con un numero sufficiente di dati occorre seguire alcuni passaggi metodologici, riassumibili nelle seguenti fasi: il controllo della qualità dei dati ed i tipi possibili di errore, i tipi di dati, la preparazione dei dati, il disegno sperimentale, la descrizione dei dati spaziali, i tipi di analisi e la scelta dei test per l'inferenza, le strategie di campionamento.
La procedura dell'inferenza statistica è semplice, nelle linee logiche generali. Tuttavia, le analisi e le conclusioni possono trovare complicazioni per l'elevata variabilità dei dati, a motivo soprattutto di tre cause principali : gli errori di misurazione, generati dagli strumenti e dalle differenze nell'abilità dei ricercatori; l'operare su campioni, per cui i dati utilizzati in una ricerca non sono mai identici a quelli rilevati in qualsiasi altra; la presenza di vari fattori contingenti di disturbo che, come il tempo e la località, possono incidere diversamente sul fenomeno in osservazione, con intensità e direzioni ignote. Schematicamente, esistono due tipi di variabili: le variabili qualitative o categoriali che sono quantificate con conteggi, ossia con numeri interi e discreti; le variabili quantitative che sono espresse su una scala continua. I dati che si raccolgono per analisi statistiche possono quindi essere discreti o continui e le misure possono essere raggruppate in 4 tipi di scale, che godono di proprietà formali differenti. Le scale di misura dei fenomeni territoriali ed ambientali possono essere: nominale; ordinale o per ranghi, ad intervalli, e di rapporti.

2.1 La qualità dei dati

Nell'analisi dei fenomeni ambientali e territoriali è indispensabile la conoscenza dei concetti e dei metodi statistici, sia per i problemi legati alle indagini di campionamento sia per quelli connessi con la preparazione e la gestione dei dati. E' necessario, infatti, che la presentazione dei dati e la loro elaborazione seguano criteri ritenuti validi universalmente, e che le fasi di raccolta dei dati, di descrizione, di analisi e infine di riepilogo siano in buona parte codificati, in modo dettagliato. Una raccolta di dati non corretta, una loro presentazione inadeguata o la scelta di una tecnica statistica non appropriata rendono impossibile la verifica dei risultati da parte di altri studiosi, nonché confronto con altre ricerche e analisi del settore.

La qualità dei dati richiede una buona conoscenza dei procedimenti di misura e di registrazione. Sfortunatamente nessun numero di dati è sufficiente per provare la bontà del dato stesso. Per esempio la casualità dei dati in un risultato analitico può essere reale, cioè legata ad un preciso fenomeno, oppure dovuta ad una errata procedura di misura.

E' necessario innanzitutto assicurarsi che i dati ottenuti siano in grado di produrre il tipo di risultati che ci si aspetta da essi. Normalmente un geologo formulerà delle ipotesi che dovranno essere verificate, ad esempio (vedi finestra di esercizio 2.1):

- Esempio 1. E' significativa la differenza tra la composizione di graniti appartenenti a due rocce intrusive differenti? Non è possibile rispondere a questo quesito analizzando solo un campione da ogni intrusione in quanto potrebbe non essere rappresentativo e quindi non tenere conto delle possibili variazioni all'interno della roccia;
- Esempio 2. Esiste una ciclicità nella stratificazione di una sequenza carbonatica? L'analisi può essere invalidata dalla incompletezza dei dati nella sequenza.

Un progetto di analisi dei dati deve essere, inoltre, affrontato con attenzione pianificando al meglio le differenti fasi operative. L'operatore deve conoscere a fondo la natura dei dati a disposizione, le tecniche statistiche e le regole mediante le quali possibile unire il dato alle tecniche di analisi. L'utilizzo di qualsiasi strumento di misura deve essere accompagnato dalla conoscenza due caratteristiche fondamentali dei dati sperimentali: la precisione e l'accuratezza (vedi finestra di esercizio 2.2).

Al fine di facilitare a chi legge la corretta comprensione dei risultati è necessario seguire uno schema specifico basto sullo sviluppo di quattro fasi:

1) un'**introduzione** accurata dell'argomento affrontato, e delle sue le finalità

2) una descrizione dei **materiali e metodi**, nella quale devono essere definiti il tipo di scala utilizzato, le tecniche di campionamento, e le misure sintetiche delle caratteristiche più importanti della distribuzione dei dati (media, mediana, deviazione standard, ecc.).

3) la descrizione dei **risultati,** allo scopo di permettere alla comunità scientifica di valutare: la correttezza delle ipotesi da verificare, il tipo di scala utilizzata per la misura delle variabili e le caratteristiche statistiche della distribuzione dei dati.

4) una **discussione**, in cui deve essere riportata l'interpretazione dei risultati ottenuti. L'interpretazione deve non solo comprendere l'analisi statistica, ma essere estesa al significato ecologico, ambientale o biologico dei risultati ottenuti. Non sempre un risultato statisticamente rilevante assume anche un significato importante nella disciplina specifica. Ne deriva un aspetto di estrema importanza per l'analisi statistica: per impostare correttamente una ricerca, per formulare ipotesi scientificamente valide, per raccogliere e analizzare i dati, infine per interpretarne i risultati, non è possibile scindere le analisi statistiche dalla loro interpretazione disciplinare.

Le osservazioni sperimentali contengono sempre qualche inesattezza, per cui, nell'utilizzo dei numeri che provengono da esse è quasi sempre necessario conoscere l'entità di queste incertezze. Se si utilizzano diverse osservazioni per raggiungere un certo obiettivo è necessario conoscere in quale maniera gli errori si ripercuotono sull'errore del risultato.

Quando si confronta un numero basato su di una previsione teorica con un numero sperimentale e ci si pone il problema se essi siano concordi o meno, è necessario disporre di un criterio per valutare la precisione dell'uno rispetto all'altro. La buona conoscenza del comportamento statistico degli errori i misura permette spesso la riduzione dell'effetto delle incertezze sul risultato finale.

La qualità dei dati è importante in quanto molte decisioni sono effettuate sulla base dei valori misurati. D'altra parte chi deve prendere delle decisioni è interessato a conoscere lo stato reale del fenomeno sul territorio investigato che è tuttavia influenzato dalla possibile presenza di errori di misura e di campionamento.

La possibilità di effettuare errori durante le decisioni non può essere eliminata completamente, ma può essere controllata verificando l'errore di campionamento mediate il prelievo di un elevato numero di campioni, e l'errore di misura ripetendo le analisi sui campioni e utilizzando metodi sempre più precisi.

Sulla base di queste considerazioni devono quindi essere definite le tollerabilità degli errori decisionali al fine di dimensionare i campionamento e le misure. Queste attività costituiscono quello

che viene definito il processo degli "obiettivi della qualità di dati"; esse consistono nelle seguenti fasi:

1. Definizione del problema.

- Individuazione del problema, ad esempio la costruzione di un futuro centro residenziale in una zona caratterizzata da elevate concentrazioni di radon nel suolo;
- individuazione delle risorse disponibili (strutture pubbliche, private, laboratori, ecc.)
- limiti temporali per la conclusione dei lavori.

2. Identificazione delle decisioni

In particolare, quelle legate ai valori limiti da utilizzare. Se tali valori sono superati si stabiliranno ulteriori azioni: acquisizioni di altri dati, progettazione della bonifica ed analisi del rischio.

3. Identificazioni dei dati fondamentali per le decisioni

Si definiscono le informazioni necessari per procedere ad una valutazione della situazione:

- dati pregressi (geologia, uso del suolo, ecc)
- parametri fondamentali da misurare (profondità delle indagini, tipo di campionamento, ecc.)

4. Definizione dei limiti spazio- temporali dello studio

Individuazione dell'area da investigare delle condizioni di prelievo dei campioni e della durata di campionamenti.

5. Regole decisionali

Definizione degli indicatori statistici dei campioni studiati ed i valori di riferimento, ad esempio se le concentrazioni di radon sono superiori ai limiti di legge (500 Bq/m^3);
se le concentrazioni di radon sono inferiori ai 500 Bq/m^3.

I dati devono, inoltre, permettere di:

- **determinare il range dei parametri studiati.**

I dati storici ed analitici devono essere utilizzati per stabilire i limiti dei parametri di interesse.

- **identificare gli errori e permettere la scelta dell'ipotesi nulla.**

La situazione reale che determina le maggiori potenziali conseguenze è considerata come ipotesi nulla, H_0, e l'altra come ipotesi alternativa, H_1; nel nostro caso il radon supera e non supera la concentrazione limite di 500 Bq/m^3.

F+: falso positivo. Corrisponde all'errore decisionale dove il reale stato è l'ipotesi nulla, e quindi essa viene rigettata quando è vera;

F-: falso negativo. Corrisponde all'errore decisionale dove il reale stato è l'ipotesi alternativa, e quindi l'ipotesi nulla non è rigettata quando è invece è falsa.

- **specificare il range di valori in cui le conseguenze delle decisioni sono minori.**

  La zona di incertezza indica un'area in cui le conseguenze di

- **assegnare dei valori di probabilità tollerabili**

  Ottimizzazione del campionamento per rispondere ai requisiti di obiettivi dio qualità dei dati

- **assicurare un numero di campioni sufficienti e di qualità**

  Per legare il meccanismo decisionale alla qualità dei dati esistenti è possibile adottare una procedura basata sul t-test. Nel caso di un campionamento casuale il numero di campioni necessari è dato dalla:

- $$n = \frac{\sigma^2 (z_{1-\alpha} + z_{1-\beta})^2}{\Delta} + \frac{z^2{}_{1-\alpha}}{2} \qquad 2.1$$

Dove σ^2 = la varianza stimata, z_p il percentile della distribuzione normale standard e Δ è la differenza tra il valore limite di intervento ed un valore adottato dagli organismi decisionali.

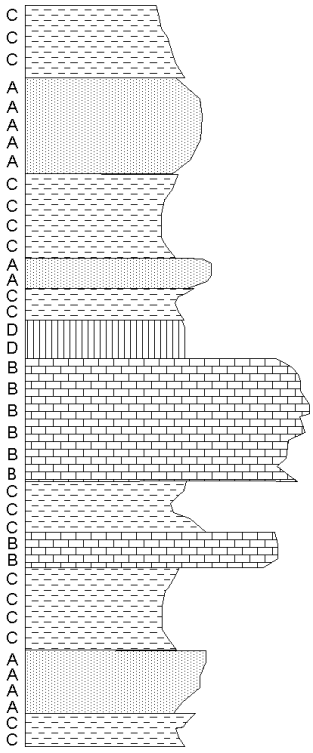
Finestra di esercizio 2.1
Esempio di ipotesi

Viene riportata la percentuale di quarzo contenuta in due rocce intrusive. I valori del contenuto in SiO_2 sono stati ottenuti prelevando per ogni roccia otto campioni (T2.1.1). La tabella riporta i valori numerici. Lo scopo è quello di verificare se esiste una differenza significativa tra le due serie di campioni al fine di classificare i due campioni di roccia. L'obiettivo viene raggiunto attraverso l'applicazione di alcuni test statistici che verificano l'uguaglianza delle medie e della variabilità delle due distribuzioni (nell'ordine: t-test, F-test).

T2.1.1. Percentuale di quarzo contenuta in due rocce intrusive.

N	Roccia 1	Roccia 2
1	23.5	20.4
2	18.6	19.7
3	25.7	22.2
4	22.4	21.4
5	26.2	16.4
6	22.8	18.6
7	20.5	21.1
8	19.3	17.3

Questo esempio mostra una tipica analisi di dati sequenziali (T2.1.2). Attraverso la costruzione di una matrice 4x4 (*matrice di frequenza delle transizioni*) è possibile analizzare la natura delle transizioni da uno strato all'altro, piuttosto che le posizioni relative degli strati nella sequenza (T2.1.3). Tale tipo di matrice costituisce un modo rapido per esprimere l'incidenza di uno strato. Il totale delle righe e delle colonne è lo stesso, ciò ad indicare che la sequenza inizia e finisce con lo stesso strato. La tendenza di uno strato a succedere ad un altro può essere enfatizzata trasformando le frequenze in percentuali ottenendo una stima della probabilità che lo strato *j* possa essere seguito dallo strato *i*.



T2.1.2. Tabella riassuntiva dei codici relativi alla successione stratigrafica rappresentata nella figura. Arenaria (A), Calcari (B), Argille (C), Torba (D).

Tetto					
C	C	C	B	C	A
C	C	D	B	C	C
C	C	D	C	C	C
A	C	B	C	C	Letto
A	A	B	C	A	
A	A	B	B	A	
A	C	B	B	A	

T2.1.3. Matrice delle frequenze delle transizioni.

		A				Tot. Riga
		A	**B**	**C**	**D**	
Da	**A**	7	0	2	0	9
	B	0	6	2	0	8
	C	3	1	12	1	17
	D	0	1	0	1	2
Tot. Col.		10	8	16	2	36

2.2 L'approccio sperimentale: il campionamento e l'inferenza

Per condurre in modo corretto una ricerca scientifica, cioè per raccogliere un campione con un numero sufficiente di dati in grado di fornire informazioni qualitativamente affidabili, occorre seguire alcuni passaggi metodologici, riassumibili nelle seguenti fasi:

1. la conoscenze del territorio
2. la pianificazione delle attività conoscitive
3. il campionamento o supporto
4. le procedure di trattamento dei dati
5. l'analisi statistica preliminare
6. la scelta dei test per l'inferenza
7. l'analisi della distribuzione spaziale
8. l'elaborazione finale di mappe tematiche.

1. La conoscenza del territorio è fondamentale per l'inquadramento generale del fenomeno oggetto di studio nelle condizioni naturali al fine di essere sicuri che i dati sperimentali rilevati siano riferibili al fenomeno che si intende studiare. Questa fase è anche definita investigazione preliminare del sito per la ricostruzione di un modello concettuale del fenomeno studiato, ad esempio nel caso di una possibile contaminazione di un sito tale fase corrisponde all'individuazione della sorgente di contaminazione, dei percorsi dell'agente contaminante e dei possibili bersagli. In questa fase è pertanto estremamente necessario l'analisi della cartografia esistente a scala regionale (1:10000 – 1:100000) e di dettaglio (1:2000 – 1:10000) per l'identificazione delle condizioni al contorno.
2. In questa fase si stabiliscono le attività da svolgere, in termini qualitativi e quantitativi, al fine di descrivere con una certa precisione le condizioni ambientali in cui si verifica il fenomeno oggetto di studio. Sin definiscono, inoltre, le tecniche di indagine che possono essere utilizzate, il tipo di campionamento e le procedure analitiche.
3. Il campionamento permette di raccogliere i dati in funzione dello scopo della ricerca. Uno dei problemi fondamentali della statistica è come raccogliere solamente un numero limitato di dati (per motivi economici, di tempo, di oggetti effettivamente disponibili, cioè per limiti oggettivi che quasi sempre esistono in qualsiasi ricerca sperimentale), ma che permetta comunque attraverso la loro analisi di raggiungere ugualmente delle conclusioni generali. Il campionamento e le indagini di

misura sono eseguite seguendo generalmente dei protocolli ben stabiliti che consentano di acquisire dati qualitativamente attendibili. In particolare il campionamento potrà subire degli aggiustamenti nel corso delle indagini a seconda dei primi risultati o delle difficoltà incontrate (in genere di tipo logistico).

4. Le procedure di trattamento dei dati devono essere selezionate in funzione dell'obiettivo che si vuole raggiungere sulla base delle seguenti condizioni: una buona formulazione del problema, l'individuazione delle decisioni conseguenti al risultato delle indagini, l'individuazione delle informazioni necessarie per risolvere i problemi connessi alle decisioni (valori di riferimento e metodi di analisi), la definizione dei limiti spaziali e temporali dello studio, la definizione degli errori tollerabili nell'ambito delle decisioni finali, ottimizzazione del piano dio campionamento *in itinere*.

5. La descrizione statistica preliminare è effettuata riportando i dati di campagna su supporto informatico ed utilizzando i numerosi softwares disponibili sul mercato per calcolare gli indici statistici di base (media, deviazione standard, varianza, skewness, ecc.), per rappresentare i dati mediante grafici (istogrammi, diagrammi a scatola, a dispersione, ecc.), per individuare gli errori analitici e/o di misura, per applicare i test statistici, per verificale la presenza di un trend spaziale, per rappresentare i risultati finali mediante mappe tematiche.

6. I test devono essere già programmati nella fase di pianificazione delle attività conoscitive, poiché è da essi che può dipendere anche il tipo di campionamento utilizzato. Il test è un processo logico-matematico che porta alla conclusione di non poter rifiutare oppure di poter rifiutare l'ipotesi della casualità, attraverso il calcolo della probabilità di commettere un errore con queste affermazioni. L'ipotesi che il risultato ottenuto con i dati sperimentali raccolti sia dovuto solo al caso è chiamata ipotesi nulla e è indicata con H_0 (cfr. Cap. IV). In generale, essa afferma che le differenze tra due o più gruppi, quelle tra un gruppo e il valore atteso, oppure le tendenze riscontrate siano imputabili essenzialmente al caso. Per giungere a queste conclusioni si deve ricorrere all'inferenza, che può essere definita come la capacità di trarre conclusioni generali (sulla popolazione) utilizzando solo un numero limitato di dati variabili (campione). Nell'apprendimento e nell'uso della statistica, il primo passo è comprendere come solamente una corretta applicazione del campionamento e una scelta appropriata dei test permettano di rispondere alla domanda inferenziale di verifica dell'ipotesi nulla. Se tale probabilità risulta alta, convenzionalmente uguale o superiore al 5%, si imputeranno le differenze a fattori puramente casuali. Al contrario, se la probabilità è

inferiore al valore prefissato, si accetta come che le differenze siano dovute a fattori non casuali, rientranti tra i criteri che possono distinguere i gruppi di dati.

7. l'analisi della distribuzione spaziale dei dati sperimentali permette di studiare le relazioni spaziali tra le osservazioni e, quindi, permette di caratterizzare il fenomeno studiato in termini di porzione di territorio su cui agisce ("dominio spaziale"). La distribuzione spaziale delle osservazioni sperimentali permette inoltre di individuare la presenza di trend e quindi di una possibile anisotropia del fenomeno. L'analisi spaziale si basa sul concetto di georeferenziazione dei dati dove a ciascun valore della variabile è assegnata una precisa posizione nello spazio attraverso un sistema di coordinate appartenenti ad un qualsiasi sistema geografico (coordinate geografiche, chilometriche, ecc.) o geometrico (coordinate di disegno)

8. l'elaborazione di mappe tematiche sfrutta il concetto di georeferenziazione delle osservazioni di una variabile per poterne rappresentare la distribuzione in 2D o 3D mediante l'utilizzo di diversi metodi di interpolazione deterministici (triangolazione, regressione polinomiale, inverso della distanza, ecc.) e probabilistici (kriging).

2.3 I tipi di errore

Trattando gli errori di misura si è soliti distinguere tra **errori sistematici** ed **errori casuali** (accidentali).

Gli errori sistematici dipendono dagli strumenti e dalle tecniche di misura utilizzati. Supponiamo di avere un brachiopode lungo 10 cm e di misurarne la lunghezza facendo coincidere l'estremità di un righello con una delle estremità dell'esemplare. Se il primo centimetro del righello è stato tagliato la misura darà un valore pari a 11cm. Questo è un errore sistematico. Se un termometro immerso in acqua distillata in ebollizione a pressione atmosferica misura 102°C, esso è stato tarato in maniera sbagliata; se le misure effettuate con questo termometro vengono inserite in un risultato sperimentale, questo sarà affetto da un errore sistematico.

Gli errori casuali sono dovuti ad un gran numero di variazioni incontrollabili e sconosciute delle condizioni sperimentali. Essi possono scaturire da piccoli errori nella valutazione dello sperimentatore, come ad esempio nell'apprezzare i decimi della minima divisione su una scala.

Altre cause di errore possono essere dovute a fluttuazioni imprevedibili delle condizioni di lavoro, ad esempio variazioni della temperatura, dell'illuminazione, del voltaggio od ogni altro tipo di parametri che possono influire sugli apparati strumentali. E' stato rilevato sperimentalmente che questo tipo di errori accidentali sono distribuiti secondo una legge semplice e quindi sono

facilmente trattabili mediante l'uso di tecniche statistiche. Inoltre, è da osservare che errori di natura accidentale tendono, comunque, ad annullarsi in media.

Molto spesso nella sperimentazione gli errori sistematici hanno più peso di quelli accidentali. Essi sono, inoltre più difficili da trattare in quanto non ci sono regole per evitare gli errori sistematici; soltanto uno sperimentatore esperto può individuarli, prevenirli e/o correggerli. Tuttavia, in quanto sistematici, tali errori si ripetono costantemente in corrispondenza di ogni misurazione essi contribuiscono a rendere più variabili le osservazioni e a diminuire la precisione delle misure effettuate.

Esiste in terzo tipo di errori tra i quali si trovano quelli che a volte sono chiamati errori, ma che in effetti no lo sono. Essi includono gli sbagli che si commettono trascrivendo i dati, le sviste nel leggere gli strumenti e le distrazioni nell'eseguire i calcoli. Questi errori con l'avvento dei computers ed in un esperimento svolto con accuratezza e scrupolosità possono essere completamente eliminati.

2.4 Accuratezza e precisione

I termini accuratezza e precisione sono spesso usati per distinguere fra errori sistematici e casuali: si dice che una misura è accurata quando contiene piccoli errori sistematici, e molto precisa una misura con pochi errori casuali. Un conteggio di poche unità può fornire una misura precisa, cioè con grande probabilità la sua ripetizione determina lo stesso valore. Un conteggio ripetuto di un campione numeroso difficilmente conduce allo stesso risultato per la frequenza con la quale si possono commettere errori. Inoltre, a causa della grande variabilità dei dati ambientali e territoriali queste misure non conducono mai a risultati identici. Quando si dispone di misure ripetute, la distribuzione dei valori può essere rappresentata e quantificata mediante gli indici della statistica descrittiva. Essi servono per rispondere a due domande: Quale è il valore reale del fenomeno? Come descrivere la variabilità del fenomeno o l'errore commesso nella sua misura?

Al momento della raccolta dei dati, occorre quindi tenere presente che i valori devono essere misurati con la precisione utile per fornire una risposta accurata alle due domande precedenti. E' ovvio che quando la misura è approssimata, come il peso di una persona che sia arrotondato al chilogrammo, è impossibile valutare l'errore di una bilancia, se esso è limitato a uno o al massimo due ettogrammi. Nello stesso modo è assurdo pretendere misure precise al grammo con una bilancia pesa – persone. Tale concetto è insito nell'unità di misura con la quale sono espressi i dati. Se si afferma che un individuo pesa 68 Kg., non si intende dire che è esattamente Kg. 68.000 (cioè 68,000

gr), ma un valore compreso tra Kg. 67.5 e 68.5 o, se si vuole una misura più precisa, tra Kg. 67.50 e 68.49.

Se si vuole fornire una misura estremamente precisa, come può essere un peso fornito in grammi o un'altezza in millimetri, si dovrebbe acquistare uno strumento più sofisticato di quello abitualmente in commercio. E' un'operazione che ha un costo, che richiede maggiore attenzione, e potrebbe rappresentare un dispendio inutile di tempo e risorse. Anche le operazioni statistiche avrebbero poi un appesantimento inutile, quando non dannoso, per calcolare gli indici di tendenza centrale, dispersione e forma. Da queste osservazioni fondate sulla pratica quotidiana deriva che, al momento della raccolta dei dati, si pone un problema pratico non banale. E' necessario definire quale è il numero di cifre significative che è utile raccogliere; ricordando che una approssimazione troppo grande conduce a valori identici e fa perdere una parte importante dell'informazione; una rilevazione troppo fine aumenta inutilmente i costi e i tempi della ricerca, senza accrescere realmente l'informazione sul fenomeno studiato.

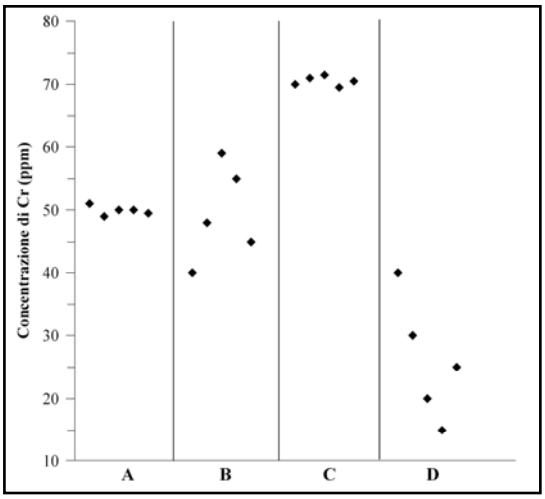


Figura 2.1. Grafico della concentrazione di Cr in campioni di suolo. Il valore corretto della misura è di 50 ppm. Come si può notare dalla figura solo nel settore A i valori misurati presentano una elevata precisione ed accuratezza. Negli altri settori si passa a diverse combinazioni tra le due qualità del dato.

Le soluzione di questo dilemma dipende dall'errore che si accetta di commettere. Esso è legato ai concetti di precisione e di accuratezza di una misura, che nel linguaggio comune sono sinonimi, ma nel linguaggio statistico hanno significati differenti.

L'accuratezza è la vicinanza di un valore misurato al suo valore reale e in buona parte dipende dallo strumento. Uno strumento che fornisce una risposta sbagliata spesso risulta essere tarato in modo non corretto e, pertanto, è definito inaccurato; nella misura dei valori si commette un errore sistematico, chiamato *bias*. Esso rappresenta un problema importante e ricorrente, in molte tecniche di stima di una quantità. In varie discipline, il progresso è spesso collegato alla ricerca di metodi di misurazione più accurati.

La precisione è la vicinanza di misure ripetute al medesimo valore. Spesso dipende dalla capacità del tecnico di ripetere la misurazione con le stesse modalità e ha origine dalla sua esperienza o abilità.

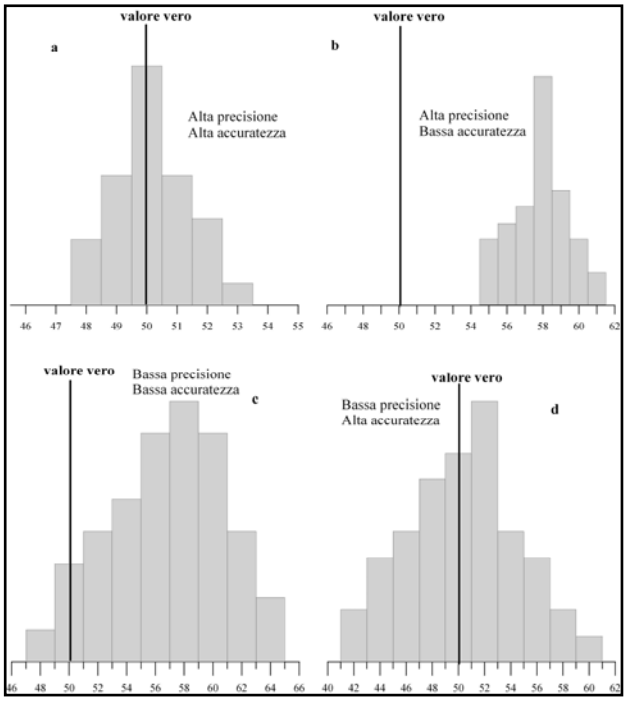
La figura 2.1 mostra le principali differenze numeriche tra misure accurate e misure precise. Considerando che il valore vero è 50 ppm, il settore A presenta una elevata accuratezza e precisione della misura; nel settore B le misure sono accurate, ma non precise; nel settore C precise ma non accurate; nel settore D ne precise ne accurate.

Finestra di esercizio 2.2

Precisione e Accuratezza

Precisione

Una misura è precisa se misurazioni ripetute della stessa entità sono simili. Le repliche della misura devono essere effettuate nelle stesse condizioni e con lo stesso strumento. Mantenendo la stessa procedura di misura l'errore coinvolto nel risultato sarà casuale. In termini quantitativi la precisione può essere espressa mediante il calcolo del coefficiente di variazione ossia del rapporto fra la deviazione standard e la media moltiplicato 100. Tanto più piccolo sarà il valore in tal modo ottenuto quanto più le misure risulteranno poco variabili.



F.2.2.1. Possibili combinazioni tra misure accurate e precise.

Accuratezza

Una misura si dice accurata se è molto vicina al valore vero. In geologia e nelle scienze ambientali il valore vero è generalmente sconosciuto (in geochimica esistono comunque dei campioni standard utilizzati per calibrare gli strumenti analitici).

Sono possibili tutte le combinazioni di alta/bassa precisione ed alta/bassa accuratezza (Fig. F2.2.1). Nell'esempio viene riportata la misura ripetuta dell'orientazione di un piano di stratificazione fatta con la bussola dalla stessa persona. Si deve supporre che la persona non ricordi la misura precedente al fine di non inficiare il risultato.

Nella figura A le misure sono accurate, vicine al valore vero, e molto precise. Nella figura B le misure non sono accurate, ma sono molto precise.

Nella figura C le misure non sono ne accurate ne precise. Nella figura D le misure sono accurate ma sono poco precise, cioè differenti tra loro.

L'operatore che ottiene una alta precisione risulta una persona competente nelle operazioni di misura; una bassa precisione indica un uso inappropriato dello strumento di misura.

L'operatore che ricade nella categoria dell'alta accuratezza indica che la bussola è tarata correttamente; una bassa accuratezza indica una cattiva taratura dello strumento che darà risultati affetti da errore sistematico.

Un valore può essere preciso ma non accurato. Un esempio didattico è il tiro ad un bersaglio, dove la media delle varie prove permette di misurarne l'accuratezza e la loro variabilità la precisione. Se tutti i colpi centrano esattamente il bersaglio o sono molto vicini a esso, con media esattamente sul centro, si ha accuratezza (il fucile è tarato esattamente per le caratteristiche visive di chi spara) e precisione (il tiratore è abile). Se i colpi sono tutti nello stesso posto, ma lontani dal centro del bersaglio, si ha una scarsa accuratezza, o misure *biased*, ma una buona precisione. Il fucile è tarato male, ma il tiratore sa sparare. Se i colpi sono molto dispersi intorno al bersaglio e la loro media coincide con quella del centro, si ha accuratezza ma bassa precisione: il fucile è tarato esattamente, ma il tiratore non sa sparare con precisione. Se i colpi formano una rosa molto ampia e la loro media è distante dal centro, si ha una scarsa accuratezza e una bassa precisione: lo strumento è *biased* e l'individuo non sa usarlo correttamente.

2.5 Numeri di riferimento, etichette, codici e formato dei dati

Come i valori misurati delle variabili, un set di dati deve necessariamente includere le informazioni necessarie per identificare le singole righe (corrispondenti alle osservazioni) e le colonne (corrispondenti alle variabili). Questa operazione può speso causare problemi di formattazione del foglio dati, problemi che possono essere evitati aggiungendo altre colonne in cui ogni campione viene identificato da un codice numerico.

Qualsiasi informazione interessante e/o necessaria relativa ad un campione, come ad esempio la località di provenienza, il tipo di preparazione del campione, può essere trattata separatamente mediante l'aggiunta di una nuova colonna corredata da un codice numerico opportuno. In particolare, la località di provenienza del campione costituisce un parametro speciale che può essere usato sia come identificatore, ma soprattutto come variabile nel caso si affronti un'analisi spaziale. In quest'ultimo caso la variabile località può essere separata in due variabili corrispondenti alle coordinate geografiche longitudine e latitudine (x e y).

I dati possono essere immagazzinati in un computer utilizzando un editor di testo (Notebook, Word, ecc..), un foglio elettronico (Excel, Access, ecc.), oppure un qualsiasi altro software. In genere, ogni software deve essere in grado di leggere i dati nel formato in cui sono stati immagazzinati. Tuttavia se si ha la necessità di leggere i dati utilizzando diversi pacchetti informatici la scelta del formato in cui essi dovranno essere salvati nel computer diventa molto importante. Uno dei formati più facilmente trasferibile e leggibile dai softwares di uso più comune è il formato ASCII (American Standard Code for Information Interchange) in cui i valori

I dati possono essere immagazzinati in un computer utilizzando un editor di testo (Notebook, Word, ecc..), un foglio elettronico (Excel, Access, ecc.), oppure un qualsiasi altro software. In

genere, ogni software deve essere in grado di leggere i dati nel formato in cui sono stati immagazzinati. Tuttavia se si ha la necessità di leggere i dati utilizzando diversi pacchetti informatici la scelta del formato in cui essi dovranno essere salvati nel computer diventa molto importante. Uno dei formati più facilmente trasferibile e leggibile dai softwares di uso più comune è il formato ASCII (*American Standard Code for Information Interchange*) in cui i valori numerici sono immagazzinati come caratteri, e le righe e le colonne sono separate da virgole o tabulatori.

La finestra di esercizio 2.3 mostra un esempio di foglio dati.

Finestra di esercizio 2.3
Organizzazione e formattazione di un foglio dati

ID	X (m)	Y (m)	NOME	COMUNE	PROV	T(°C)	pH	COND	Ca	Mg	Na	K	SO4	Cl	NO3	HCO3
1	342303	5073722	COURMAYEUR	Morgex	AO	6,40	7,40	2010	517,0	67,0	1,0	2,00	1371,00	1,00	2,00	168,00
2	346009	4905581	CIME BIANCHE	Vinadio	CN	**-9999**	7,60	71	12,0	**-9999**	1,5	**-9999**	8,70	**-9999**	**-9999**	36,00
3	346422	4905967	SANT'ANNA	Vinadio	CN	8,50	7,70	51	11,0	0,5	1,0	0,30	7,00	0,20	**-9999**	27,00
4	394275	4909309	ABRAU	Chiusa di Pesio	CN	11,80	7,41	200	26,0	12,5	1,1	0,20	2,10	2,10	1,50	140,00
5	436587	5108225	AUSONIA	Bognanco	VB	11,60	5,85	710	53,5	63,6	28,4	5,90	70,00	9,60	1,10	460,00
6	378355	4909573	CAMOREI	S.Dalmazzo	CN	9,00	7,60	418	82,1	6,7	4,9	1,10	15,30	6,50	19,60	284,80
7	447941	5119332	CRODO-LISIEL	Crodo	NO	9,60	7,75	357	60,0	7,7	5,7	3,10	104,00	1,70	3,60	103,10
8	396913	4906407	GARBARINO	Mondovì	CN	9,00	7,05	107	21,9	1,3	12,0	1,50	7,00	10,20	1,60	75,00
9	399987	4908249	GAREISA	Mondovì	CN	11,00	7,50	180	27,0	2,1	9,5	2,50	19,00	1,00	3,10	100,00
10	436527	5108121	GAUDENZIANA	Bognanco	VB	10,20	7,88	139	25,4	1,9	1,7	1,70	12,70	0,50	3,20	73,70

L'esempio mostra come può essere organizzato un set di dati in un foglio elettronico per l'immagazzinamento in un computer. La tabella riporta i dati relativi ad alcuni parametri chimico-fisici ed alla composizione chimica di alcune acque minerali italiane. Ogni colonna contiene le informazioni relative alle singole variabili espresse sotto forma numerica, di codice o di etichetta. Le righe contengono le informazioni relative ai singoli campioni di acqua.

Le colonne rappresentano le seguenti variabili:

ID	numero progressivo del campione
X(m)	longitudine geografica in metri
Y(m)	latitudine geografica in metri
Nome	denominazione dell'acqua
Comune	località in cui è situata la sorgente
Prov	provincia
T °C	temperatura dell'acqua alla sorgente
pH	concentrazioni degli ioni H^+
Cond	conducibilità elettrica in ohm m-1
Ca^{2+}, Mg^{2+}, Na^+, K^+	contenuto in cationi maggiori
SO_4^{2-}, Cl^-, HCO_3^-, NO_3^-	contenuto in anioni maggiori

Il numero –9999 rappresenta il dato mancante.

2.6 I tipi di dati

Nell'analisi statistica, occorre porre sempre molta attenzione alle caratteristiche dei dati. Schematicamente ai dati ambientali e territoriali appartengono due tipi di variabili casuali: **le variabili qualitative e le variabili quantitative**.

Le variabili qualitative o categoriali sono quantificate con punteggi, (numeri interi e discreti) e sono riferibili ad ogni specifica caratteristica delle unità oggetto di misura Ad esempio, per valutare gli effetti di un elemento tossico è possibile contare quante cavie muoiono o sopravvivono.

Le variabili quantitative richiedono risposte numeriche, espresse su una scala continua, esse sono riferibili ad una quantità numerica che rappresenta una specifica caratteristica delle unità oggetto di misura, quali ad esempio il peso e l'altezza di ogni individuo.

I dati che si raccolgono per analisi statistiche possono quindi essere discreti o continui. Questa suddivisione, ormai storica nella presentazione ed elaborazione dei dati, è stata resa più chiara e funzionale dalla classificazione delle scale di misurazione proposta dallo psicologo Stevens (1946). Tale classificazione è stata divulgata da Siegel (1956).

I dati qualitativi possono essere a loro volta suddivisi in:

Nominali $\Rightarrow$ dove non esiste un ordine significativo delle categorie, ad esempio il tipo di roccia;

La scala nominale è il livello più basso di misurazione. E' utilizzata quando i risultati possono essere classificati o raggruppati in categorie qualitative, dette anche nominali ed eventualmente identificate con simboli. I caratteri nominali, detti anche "sconnessi", costituiscono variabili le cui modalità o attributi non assumono alcun ordine precostituito. In una popolazione animale si possono distinguere gli individui in maschi e femmine, contando quanti appartengono ai due gruppi; con una classificazione a più voci, possono essere suddivisi e contati secondo la loro specie.

Nella scala nominale o qualitativa, esiste una sola relazione, quella di identità: gli individui attribuiti a classi diverse sono tra loro differenti, mentre tutti quelli della stessa classe sono tra loro equivalenti, rispetto alla proprietà utilizzata nella classificazione. Un caso particolare è quello dei caratteri dicotomici o boleani che possono assumere solo due modalità, spesso indicate in modo convenzionale con 0 e 1 oppure + (più) e – (meno). L'operazione ammessa è il conteggio degli individui o dei dati presenti in ogni categoria. I quesiti statistici che possono essere posti correttamente riguardano le frequenze, sia assolute che relative.

Ordinali ⇒ La scala ordinale o per ranghi rappresenta una misurazione che contiene una quantità di informazione immediatamente superiore a quella nominale. In questa scala le categorie possono essere arrangiate secondo un ordine significativo, in genere sequenziale, non importa se in ordine crescente o decrescente. La scala non è regolare e gli intervalli non sono costanti, per cui l'obiettivo della scala è solo quello di assegnare un ordine relativo alle osservazioni. Non è possibile fare operazioni (somma, moltiplicazione, ecc.) su questo tipo di dati, ad esempio il rischio di un'eruzione vulcanica (alto, medio, basso), l'appartenenza di un fossile ad una determinata specie (completa, parziale, nessuna), la scala di durezza di Mohs, ecc. Quindi alla proprietà precedente di equivalenza tra gli individui della stessa classe, si aggiunge una graduazione tra le classi o tra individui di classi differenti.

Questa tipo di misura ha un limite fondamentale: non è possibile quantificare le differenze di intensità tra le osservazioni. Forniscono l'informazione di una scala ordinale anche misure che sono rappresentate con simboli, come --, -, =, +, ++; raggruppamenti convenzionali o soggettivi in classi di frequenza variabili come 0, 1-2, 3-10, 11-50, 51-100, 101-1.000, >1.000. Resta l'impossibilità di valutare quanto sia la distanza tra insufficiente e sufficiente; oppure se essa sia inferiore o superiore alla distanza tra buono ed ottimo. La scala ordinale o per ranghi è pertanto una scala monotonica.

Alle variabili così misurate è possibile applicare una serie di test non parametrici (CFR. Cap V); ma non quelli parametrici. In questi casi, non sarebbe possibile utilizzare quei test che fanno riferimento alla distribuzione normale, i cui parametri essenziali sono la media e la varianza, poiché non si possono definire le distanze tra i valori. Tuttavia questa indicazione di massima sulla utilizzazione della statistica non parametrica è spesso superata dall'osservazione che variabili discrete o nominali tendono a distribuirsi in modo approssimativamente normale, quando il numero di dati è sufficientemente elevato.

Anche i dati quantitativi possono essere a loro volta suddivisi in due categorie:

Discreti ⇒ quando la variabile può assumere soltanto valori numerici finiti, generalmente numeri interi, ad esempio il numero di microfossili per cm^2 in una sezione stratigrafica (x =0,1,2,3…..);

Continui ⇒ quando la variabile può assumere un qualsiasi valore numerico associato ad un determinato intervallo di valori reali, quali ad esempio misure di lunghezza o peso, misure in forma percentuale (ppm, parti per milione) o qualsiasi altro modo di esprimere una proporzione di un totale fisso (dati geochimici), dati di direzione; ad esempio il contenuto in carbonato di calcio di una roccia calcarea (mg/100g), la conducibilità elettrica di un'acqua ($0<x<\infty$)In genere, è preferibile

lavorare sempre con dati di tipo quantitativo (numerici), ma talvolta le informazioni a disposizione sono rappresentate da nomi (tipo di roccia, specie fossili, ecc.). Tuttavia anche avendo a disposizione tale tipo di dati è possibile effettuare analisi statistiche rigorose trasformando la variabile qualitativa in forma numerica. Nella maggior parte dei casi tale trasformazione viene effettuata utilizzando un codice binario, per esempio 1/0 ad indicare presenza/assenza.

Tra i dati di tipo quantitativo continui è possibile, inoltre, distinguere altre quattro categorie (Fig. 2.2):

Dati relativi o ad intervalli ⇒ differiscono dai precedenti in quanto il punto zero non costituisce un limite fondamentale della scala, ad esempio la temperatura in °C o °F. Al contrario, la temperatura misurata in °K appartiene ai dati assoluti in quanto 0 °K rappresenta la mancanza assoluta di calore. La scala ad intervalli aggiunge la proprietà di misurare le distanze o differenze tra tutte le coppie di valori. La scala di intervalli si fonda su una misura oggettiva e costante, anche se il punto di origine e l'unità di misura sono arbitrari. I valori di temperatura, oltre a poter essere facilmente ordinati secondo l'intensità del fenomeno, godono della proprietà che le differenze tra loro sono direttamente confrontabili e quantificabili.

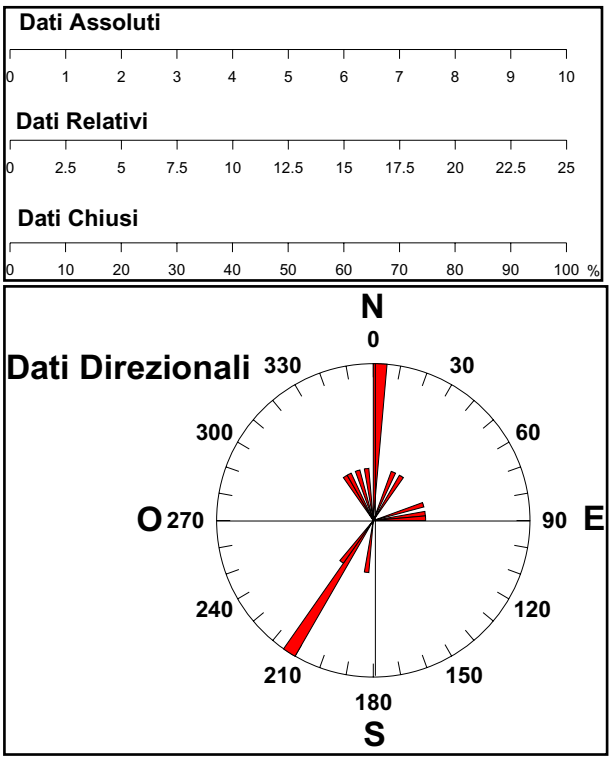


Figura 2.2. Alcuni categorie di dati quantitativi e loro scale di misura.

Tuttavia la scala ad intervalli ha un limite, non gode di un'altra proprietà importante nella elaborazione statistica dei dati, quella del rapporto tra coppie di misure. Ad esempio, una temperatura di 80 gradi Celsius non è il doppio di una di 40 gradi. Se una persona ponesse la mano destra in una bacinella con acqua a 80 gradi e la mano sinistra in un'altra con acqua a 10 gradi, non direbbe certamente che la prima scotta 8 volte più della seconda, ma solo che la prima è molto calda e la seconda fredda.

In una scala ad intervalli solo per le differenze sono permesse tutte le operazioni: possono essere tra loro sommate, elevate a potenza oppure divise, determinando le quantità che stanno alla base della statistica parametrica.

Dati chiusi ⇒ appartengono a questa categoria i dati in forma di percentuale, ppm, o altri modi di esprimere la proporzione di un totale; sono spesso utilizzati in geochimica. Si richiede molta attenzione nel trattare tale tipo di dati soprattutto nella statistica bivariata e multivariata, poiché le variabili sono in genere interdipendenti.

Dati di direzione ⇒ sono generalmente espressi in angoli, come ad esempio la direzione rispetto al nord del piano di una stratificazione o di un discontinuità tettonica. Il tipo e la qualità dei dati sono sicuramente influenzati dalla sorgente di provenienza del dato stesso.

Dati assoluti ⇒ che costituiscono la maggior parte dei dati come le misure di peso o lunghezza.

La scala dei dati assoluti ha il vantaggio di avere un'origine reale. Sono tipiche scale assolute l'altezza, la distanza, la velocità, l'età, il peso, il reddito, la temperatura in gradi Kelvin; più in generale, tutte quelle misure in cui 0 (zero) significa quantità nulla. Non solo le differenze, ma gli stessi valori possono essere moltiplicati o divisi per quantità costanti, senza che l'informazione di maggiore importanza, il rapporto tra essi, ne risulti alterata. A tale tipo di dati può essere applicato qualsiasi test statistico. Possono essere utilizzati anche la media geometrica ed il coefficiente di variazione, i quali richiedono che il punto 0 (zero) sia reale e non convenzionale. Anche una scala assoluta è possibile trasformarla in una scala di rango o addirittura qualitativa con perdita rilevante della quantità d'informazione, che essa fornisce.

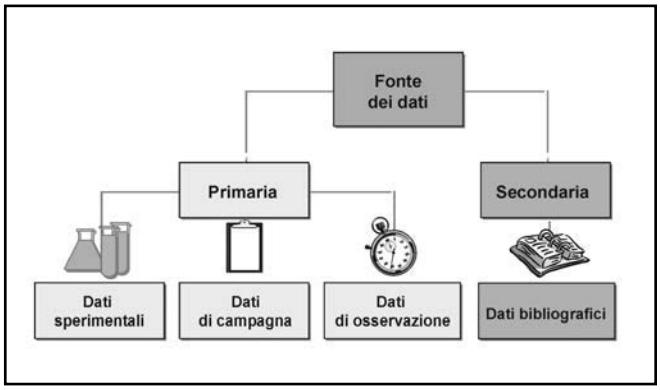


Figura 2.3 Principali sorgenti di dati.

E' importante considerare anche la sorgente dei dati che può essere **primaria o secondaria**. I dati che hanno origine da una sorgente primaria sono quelli derivati da esperimenti, osservazioni sperimentali e di campagna; i dati che provengono da una sorgente secondaria sono quelli bibliografici (Fig. 2.3).

2.7 Campioni e popolazioni

Quando ad esempio facciamo una serie di misure su pezzi di roccia e ci accingiamo ad effettuare un'analisi statistica, in genere non siamo interessati alle proprietà delle unità che abbiamo misurato, ma speriamo che le misure fatte siano rappresentative dell'intero corpo roccioso. Se i dati che abbiamo raccolto rappresentano l'intero set di misure che ci interessano, allora si parlerà di popolazione (l'intero corpo roccioso). Qualora i dati raccolti rappresentano una parte dei dati che ci interessano, si parlerà di campioni (i pezzi di roccia). Nella pratica delle analisi statistiche nel caso di una popolazione per trarre le opportune conclusioni basta applicare le tecniche dell'analisi statistica descrittiva, mentre nel caso si abbia a disposizione un sottogruppo della popolazione totale dei dati (campione) è necessario utilizzare le tecniche dell'analisi statistica inferenziale per trarre dal campione le corrette conclusioni sulla popolazione.

La **popolazione** è rappresentata dal tutte le misure ipoteticamente possibili che possono essere effettuate sull'entità oggetto di studio. I limiti della popolazione (in Geologia di solito sono di tipo spaziale) in genere sono definiti dall'operatore nella fase iniziale di un progetto, ad esempio tutte le misure di permeabilità che è possibile fare su una formazione sabbiosa interessata da un giacimento petrolifero; la composizione di tutti i graniti del mondo; le dimensioni di tutti gli esemplari di una specie di trilobite; il contenuto in elementi in traccia di tutti i sedimenti fluviali entro un raggio di 5 km da una miniera di ferro.

In questo caso la confusione tra il significato geologico e statistico di campione è inevitabile. In Geologia per campione si intende un pezzo di roccia una certa quantità di sedimento, un esemplare di una specie fossile, o un minerale. In statistica il **campione** è costituito da una serie di oggetti o misure che costituiscono una parte dell'intera popolazione di interesse e sono raccolti per rappresentarla. Tali oggetti o misure costituiscono i dati disponibili per le analisi.

In genere, nelle scienze ambientali e territoriali è possibile distinguere tre categorie di popolazioni:

- la popolazione ipotetica, che comprende l'intera entità del fenomeno studiato esistente,ad esempio le misure granulometriche di un orizzonte sabbioso che affiora in superficie e che è stato parzialmente eroso;
- la popolazione esistente, costituita dalla porzione rimanente dell'intera entità (i.e orizzonte sabbioso) che rappresenta la popolazione oggetto di studio;
- la popolazione disponibile, costituita dalla popolazione esistente da cui è possibile, dal punto di vista logistico, estrarre un campione rappresentativo. Laddove la popolazione disponibile

presenta dei limiti di campionamento, non è possibile estendere le conclusioni derivate dall'analisi statistica a tutta la popolazione esistente.

2.8 I tipi di analisi

Idealmente un progetto di analisi statistica può essere schematizzato come riportato nella figura 2.4. In genere si tende ad iniziare ponendosi un obiettivo oppure una domanda:

- "è il grado del giacimento tale da permettere uno sfruttamento economico?"
- "è la roccia sufficientemente permeabile da costituire un buon serbatoio?"
- "che forma ha il corpo sabbioso che costituisce il serbatoio?"
- "è la roccia in esame una tholeite?"

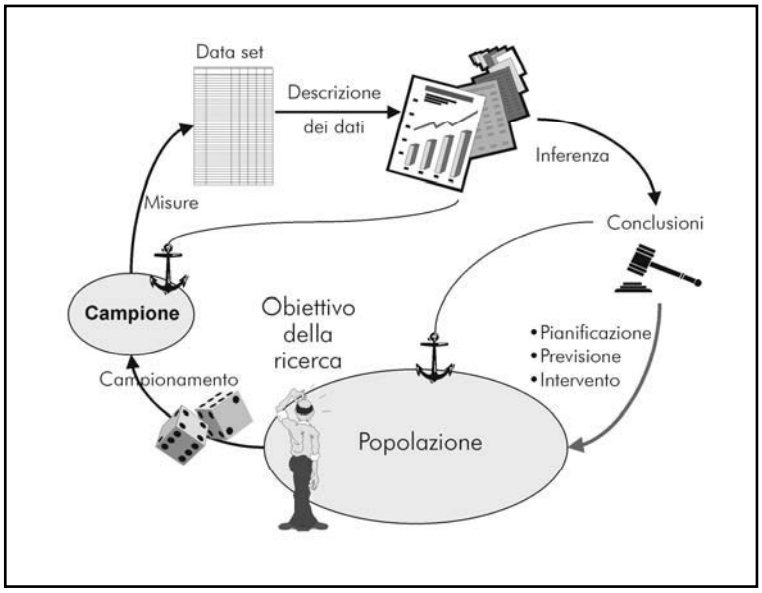


Figura 2.4. Schema di un progetto di analisi statistica dei dati sperimentali.

La statistica in senso stretto è una speciale disciplina in cui le risposte alle domande di cui sopra vengono date attraverso il calcolo della stima del valore reale. La veridicità della stima è spesso espressa in termini probabilistici, per esempio *siamo sicuri al 95% che quel giacimento è economicamente sfruttabile.* Questo approccio, molto formale e rigoroso, non è spesso necessario. Esiste una grande varietà di tecniche analitiche del dato disponibile in maniera da poter esprimere giudizi per la stima delle variabili, tali tecniche includono: metodi grafici, EDA (*Exploratory Data Analysis*), stima delle superfici, ecc. Questi metodi naturalmente non daranno delle risposte definitive, ma permettono tuttavia di costruire alcune ipotesi plausibili.

Le tecniche descritte in questo libro comprendono sia i metodi di analisi sperimentale sia un approccio teorico più rigoroso. Il lettore dovrà tenere conto del fatto che in generale i metodi sperimentali non danno comunque risposte definitive, ma risultati che si prestano all'interpretazione dell'operatore. L'approccio teorico di contro impone una certa forma e rigore alle esercitazioni sperimentali.

La statistica moderna può essere distinta in tre parti: descrittiva, matematica, inferenziale.

1 - La **statistica descrittiva** spiega come i dati raccolti devono essere riportati in tabella, rappresentati in grafici e sintetizzati in indici matematici, allo scopo di individuare le caratteristiche fondamentali del campione.

2 - La **statistica matematica** presenta le distribuzioni teoriche sia per misure discrete sia per misure continue, allo scopo di illustrarne le caratteristiche fondamentali, le relazioni che esistono tra esse, gli usi possibili;

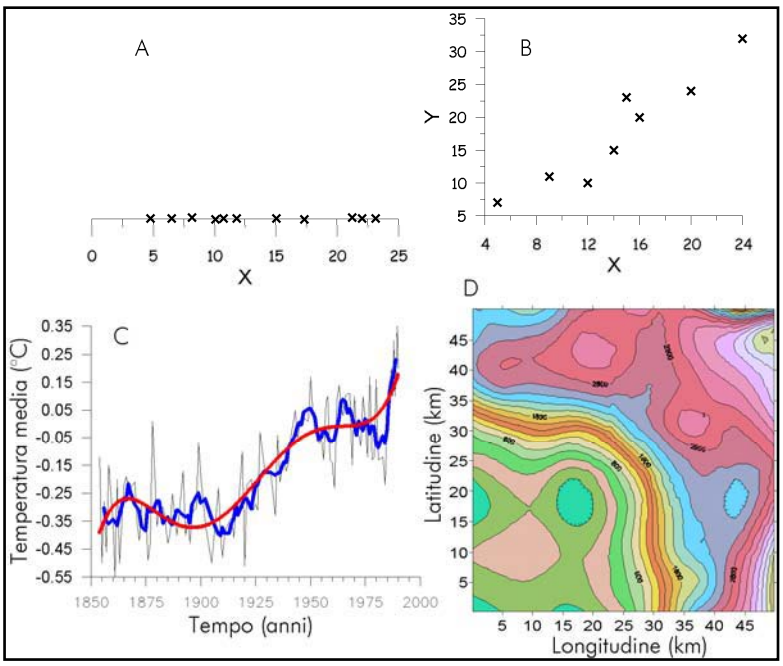


Figura 2.5. Alcuni risultati grafici derivati da differenti tipi di analisi: analisi statistica monovariata (A); analisi statistica bivariata (B); analisi delle serie temporali (C); analisi spaziale (D).

3 – L**a statistica inferenziale**, serve per la verifica delle ipotesi, coinvolgendo anche i processi di stima e la formulazione di ipotesi per l'elaborazione di un modello probabilistico che permetta di prendere delle decisioni di intervento.

Quest'ultima può essere distinta in diverse tecniche a seconda delle caratteristiche dei dati (se permettono o meno il ricorso alla distribuzione normale: **statistica parametrica e non parametrica**, oppure a seconda del numero di variabili (se una, due o più: **statistica univariata, bivariata, multivariata**) (Fig. 2.5)..

Tecniche della statistica monovariata. Ogni variabile viene analizzata singolarmente; i dati possono essere rappresentati come una serie di punti su una retta che rappresenta la scala appropriata per quella variabile. L'osservazione della distribuzione di punti lungo la retta permette la descrizione dei dati (Fig.2.5a).

Tecniche della statistica bivariata. Le variabili sono analizzate a coppie. Le misure fatte su un oggetto costituiscono le coordinate di un punto in uno spazio a due dimensioni (2D), ed una serie di misure possono essere rappresentate come un grafico a dispersione in 2D. I metodi della statistica bivariata descrivono ed analizzano la forma dei diagrammi a dispersione al fine di investigare le relazioni tra i punti e/o le relazioni tra le variabili (Fig. 2.5b).

Analisi delle serie temporali. Sono effettuate su sequenze di dati nel tempo (o nello spazio) che in genere possono essere trattate come dati a due variabili, con una delle variabili rappresentata dal tempo. I dati formano una sequenza, e la posizione dei punti nella sequenza costituisce il fattore di studio più importante. Dati di questo tipo sono abbastanza comuni in Geologia e comprendono misure di successioni stratigrafiche, misure geochimiche o mineralogiche lungo profili o perforazioni, logs elettrici in pozzi per l'estrazione petrolifera, ecc. (Fig. 2.5c).

Tecniche di analisi spaziale. Viene effettata considerando tre o quattro variabili contemporaneamente, due o tre delle quali rappresentano le coordinate spaziali, quali valori di griglia o latitudine/longitudine con o senza altitudine o profondità. La quarta variabile rappresenta la variabile oggetto di studio. I dati devono essere immaginati come punti in uno spazio a tre dimensioni, e l'analisi spesso si pone come obiettivo la costruzione di superfici che descrivono le variazioni spaziali delle variabili stesse (Fig. 2.5d).

2.9 Le strategie di campionamento

La scelta dello schema di campionamento è fondamentale per il raggiungimento di risultati significativi scientifici e statistici, soprattutto in funzione del numero di campioni necessari per un determinato livello di significatività.

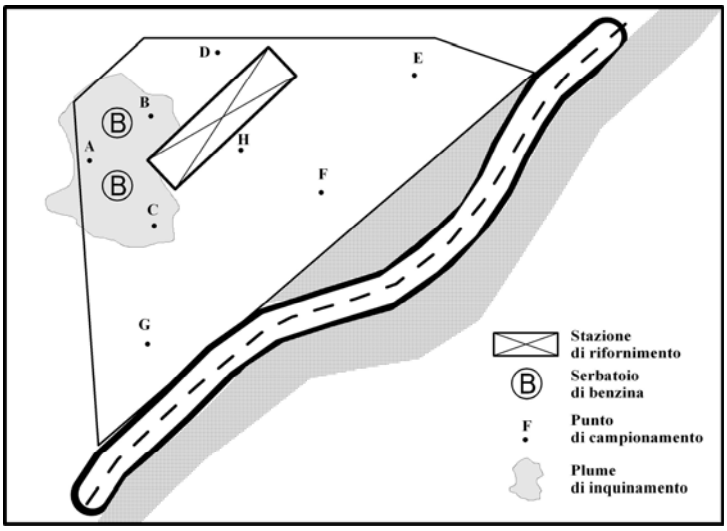


Figura 2.6. Scelta dello schema di distribuzione dei campioni in funzione di una buona qualità del campionamento.

Uno schema di campionamento ben organizzato deve consentire di:

1. supportare le decisioni relative ai livelli di una certa variabile ambientale laddove questa sia superiore alla soglia di rischio;
2. determinare se le caratteristiche di due popolazioni differiscono di una certa quantità;
3. stimare gli indici statistici caratteristici di una popolazione;
4. localizzare e delimitare le aree anomale, e determinarne la loro natura;
5. identificare e controllare le tendenze nei dati ambientali.

La figura 2.6 Mostra un'area extraurbana in cui è ubicata una stazione di rifornimento per auto. L'obiettivo che si vuole raggiungere è la verifica di un inquinamento nell'area occupata dalla stazione in cui il serbatoio interrato di gasolio riversa nel sottosuolo una parte del contenuto a causa di una falla. Assumendo che la raccolta dei campioni di suolo sia stata effettuata correttamente e che le analisi abbiano una adeguata qualità, si vuole sapere se i campioni di suolo prelevati nei siti A, B e C siano rappresentativi del problema legato all'inquinamento del suolo. Tuttavia, il valore medio dei campioni raccolti non possono essere rappresentativi delle concentrazioni medie di inquinanti in tutta l'area della stazione; infatti a questo scopo è necessario prelevare altri campioni in tutta l'area della stazione di servizio (D, C, E e F). Se il campionamento non è rappresentativo, nonostante l'elevata qualità delle analisi di laboratorio non è possibile compensare la mancanza di campioni rappresentativi.

Nel predisporre un piano di campionamento le strategie che possono essere seguite appartengono a due gruppi:

- campionamento ragionato nel quale la posizione di punti di misura è determinata in base alle conoscenze pregresse sull'area di indagine;
- campionamento probabilistico dove la posizione dei punti è basata su concetti statistici che determinano tipi differenti di campionamento quali, casuale, stratificato, sistematico, pesato, raggruppato, adattato e composito.

Per la descrizione delle differenti strategie di campionamento ci si riferisce a USEPA (1992) per eventuali approfondimenti.

Il concetto di popolazione costituisce un concetto importante nella definizione di un piano di campionamento (cfr. par. 2.7). La popolazione comprende tutte le unità che rappresentano l'obiettivo della ricerca, è proprio sulla popolazione che si devono trarre le conclusioni. La popolazione campionata è quella parte della popolazione che è accessibile e disponibile per il campionamento come l'area della stazione di servizio riportata nella figura 2.6.

Il campione è un componente della popolazione che può essere selezionato per il campionamento, quale ad esempio un albero, u volume specifico dio acqua o gas, ecc. E' importante per chi pianifica un campionamento definire molto bene l'unità del campionamento stesso. La definizione fisica di un campione in termini di grandezza, forma e orientazione è definita come supporto di campionamento (Starks, 1986). Esso rappresenta la porzione dell'unità di campionamento come l'area, il volume, la massa, ecc. che viene prelevata e misurata. Ad esempio

se l'unità di campionamento è di 10 g di suolo prelevato in un sito di coordinate x e y, il supporto può essere dato da 1 g di suolo dopo l'omogeneizzazione.

Una volta selezionato il campione si dovrà applicare un protocollo di misura che ha l'obiettivo di ottenere informazioni sulle caratteristiche di interesse del campione mediante osservazioni o analisi. Il protocollo comprende sia la procedura di campionamento, che la preparazione del campione per le analisi.

Lo schema di campionamento deve specificare il numero, il tipo e la localizzazione (spaziale e/o temporale) delle unità di campionamento (campioni) che dovranno essere selezionate (prelevate) per le misure. L'unica considerazione che si può fare sulle dimensioni numeriche del campione riguardano il grado di precisione che vogliamo assegnare alla ricerca che stiamo affrontando, e nella maggior parte dei casi è un fattore che non possiamo prevedere. Supponiamo di voler sapere se la porosità di una formazione sabbiosa è più grande del 5%, se i cinque campioni di sabbia che abbiamo prelevato hanno dato valori di 15%, 18%, 16%, 20% e 18% siamo abbastanza sicuri sulla risposta. Se tuttavia i cinque campioni hanno dato valori di 8%, 4%, 5%, 9% e 5% non possiamo essere sicuri della nostra ipotesi e quindi sono necessari altri campioni per giungere ad una qualsiasi conclusione. In genere è buona pratica studiare le proprietà di un piccolo campione pilota per poi decidere il campione da prelevare.

Nello studio dei fenomeni naturali le variabili in gioco sono spesso georeferenziate, cioè corredate dall'informazione sulla loro ubicazione geografica mediante le due variabili x (longitudine) ed y (latitudine). Poiché l'obiettivo principale nello studio dei fenomeni naturali è l'autocorrelazione spaziale tra le osservazioni, prima di effettuare qualsiasi analisi statistica è opportuno ricostruire una mappa simbolica relativa alla distribuzioni delle osservazioni sul territorio. Tale rappresentazione è importante perché consente un primo esame sull'uniformità della copertura di campionamento, sulla densità del campionamento e sulle possibili relazioni tra un punto e l'altro.

Di seguito sono riportati alcune strategie di campionamento per i quali sarà data una breve sintesi allo scopo di indicarne le caratteristiche generali ed i limiti di utilizzo.

Campionamento ragionato

In questo tipo di campionamento la selezione dei punti di prelievo è effettuata sulla base delle conoscenze pregresse sull'area investigata. Questo tipo di campionamento è molto utile quando si ha una buona conoscenza sulle caratteristiche del sito, esso è consigliato nelle seguenti situazioni:

- lo studio è condotto a piccola scala;
- il numero di campioni è piccolo;le conoscenze pregresse del sito sono buone;
- l'obiettivo è l'individuazione delle aree anomale ed eventualmente approfondire le indagini con campionamenti mirati;
- si ha poco tempo a disposizione per organizzare una strategia di campionamento completa.

I vantaggi di questi tipo di campionamento sono la rapidità di esecuzione ed il baso costo che consentono contemporaneamente il raggiungimento di buoni obiettivi senza eccessivo bisogno di risorse.

Campionamento casuale

E' la strategia di campionamento più semplice dove la selezione dei punti viene effettuata in modo tale che tutti i campioni abbiano la stessa probabilità di essere selezionati evitando così gli errori dovuti alla soggettività. In questo tipo di campionamento la popolazione da campionare deve essere più omogenea possibile.

Il vantaggio principale è che in questo caso, avendo un numero sufficiente di campioni, è garantita la selezione di un campione rappresentativo della situazione generale. Inoltre, l'analisi statistica è molto agevolata soprattutto nel caso dell'applicazione dei test statistici (USEPA, 2002). Tuttavia questa strategia di campionamento ha due limitazioni: i campioni potrebbero non essere distribuiti uniformemente nello spazio e/o nel tempo; il campionamento casuale tende a non tenere conto di ogni tipo di informazione pregressa relativa all'area investigata.

Campionamento Stratificato

In questo campionamento le informazioni relative alla popolazione vengono utilizzate per determinare dei gruppi (strati) che sono poi campionati separatamente. Ogni unità di campionamento deve appartenere ad un solo strato, così facendo si possono ottenere risultati più precisi soprattutto se gli strati sono scelti in modo che ciascuno di essi abbia caratteristiche più omogenee in rapporto al totale della popolazione. L'applicazione del campionamento stratificato è utile ad esempio nel caso di un monitoraggio temporale. Il criterio di selezione dei punti di campionamento all'interno di ogni strato può essere una qualunque delle strategie di campionamento.

I vantaggi sono generalmente connessi con l'ottimizzazione delle risorse finanziarie, infatti gli strati possono essere selezionati in modo da minimizzare i costi associati al campionamento in aree

differenti unendo per esempio aree vicine in un singolo strato. Tale scelta necessita, tuttavia, di buone conoscenze pregresse dell'area investigata che non sono sempre disponibili.

Campionamento sistematico

Il campionamento sistematico consiste nella selezione dei punti nel tempo e/o nello spaio secondo uno schema predefinito. Ad esempio, i campioni possono essere prelevati secondo una griglia a maglia quadrata oppure ad intervalli regolari nel tempo. Affrontare un campionamento secondo schemi regolare permette di ottenere una copertura uniforme che garantisce di non trascurare parti importanti della popolazione. Inoltre, campioni equispaziati permettono una stima migliore della correlazione spaziale o temporale tra di essi, e l'identificazione della presenza di un trend.

E' consigliabile utilizzare questi tipo di campionamento quando si vogliono formulare ipotesi di parametri statistici della popolazione; quando si vuole stimare una tendenza o la correlazione spaziale fra i campioni; quando si vuole determinare la dimensione di un'anomalia che può sfuggire a seconda della dimensione della maglia di campionamento. Esistono diversi schemi spaziali e/o temporali per organizzare un campionamento sistematico; i principali sono riportati di seguito nel paragrafo.

I vantaggi principali sono i seguenti:

- si ottiene una copertura uniforme e nota dell'area investigata, a parità di numero di campioni la copertura garantita da questo tipo di campionamento è maggiore rispetto a quella dio altri schemi;
- la progettazione della griglia e semplice;
- è possibile procedere per fasi successive incrementando il dettaglio in aree in cui inizialmente non sono state riscontrate anomalie, oppure utilizzando griglie a scala differente;
- i campioni raccolti con uno schema regolare consentono di calcolare le correlazioni spaziali e/o temporali;
- non sono necessarie informazioni pregresse sul sito, che se, comunque, sono disponibili permettono una ottimizzazione del campionamento.

Campionamento pesato

Questo tipo di campionamento combina il campionamento casuale con le conoscenze iniziali dl sito che permettono un campionamento ragionato. Un'alternativa è costituita dall'esecuzione di

misure in situ che consentono di effettuare il campionamento pesato. La scelta dei punti di campionamento si basa sull'individuazione, mediante tecniche di misure in situ, di caratteristiche che indichino la presenza di valori anomali della variabile studiata, quali ad esempio l'utilizzo di foto aeree e da satellite, oppure indagini speditive e a basso costo come le prospezioni geochimiche o geofisiche superficiali.

Il campionamento pesato è consigliabile in presenza di un costo elevato delle misure di laboratorio; a seguito del giudizio di un esperto; laddove è necessaria una stima più precisa degli indici statistici.

Uno dei vantaggi di questa strategia è che essa fornisce una stima statisticamente esatta della media reale della popolazione rispetto al campionamento casuale, e consente di individuare le differenze tra le medie o le mediane di due popolazioni (valori anomali e valori di fondo). Tuttavia la precisione di questa stima prevede il prelievo di un maggior numero di campioni con conseguente aumento dei costi, che sono comunque relativi rispetto al costo delle analisi di laboratorio.

2.10 Schemi di campionamento

Abbiamo visto come il temine campionamento identifica una serie di metodi per selezionare ed analizzare una parte di "universo", allo scopo di effettuare inferenze sull'intero universo, assicurando economia, velocità ed, in certe circostanze, qualità ed accuratezza (Cochran, 1977). In particolare, possiamo dunque definire come campionamento (spaziale e/o temporale), il processo secondo il quale si selezionano alcuni elementi di un'entità (in un'area o nel tempo) oggetto di studio, sui quali sono disponibili alcune informazioni riguardanti la distribuzione degli elementi stessi.

In geologia e nelle scienze ambientali è molto importante dove e come campionare, infatti i geologi sono molto spesso interessati a come sono distribuiti i campioni su una superficie bidimensionale o su una mappa al fine di individuare:

- localizzazioni che presentano valori molto elevati delle misure effettuare (valori anomali) per identificare eventuali zone a rischio;
- direzioni privilegiate che presentano un aumento (o diminuzione) sistematico delle misure di concentrazione (trend)
- correlazioni tra i valori osservati (campioni molto vicini caratterizzati da valori molto simili, rispetto ai valori osservati in campioni più distanti).A questo punto risulta fondamentale definire la distribuzione sul territorio di n punti campione u_1, u_2, u_3, ……u_n la quale può essere effettuata secondo cinque schemi principali: regolare, uniforme, casuale, in raggruppamenti, e secondo profili (Fig. 2.7).Un campionamento è detto **regolare o sistematico** quando i punti sono distribuiti su un reticolo più o meno regolare, generalmente quadrato, ma anche rettangolare, triangolare, esagonale, ecc. In questo tipo di distribuzione la distanza tra il generico punto *i* e *j* lungo una specifica direzione è la stessa per tutte le coppie dei punti *i* e *j* presenti sulla mappa (Fig. 2.7a).

Un campionamento è detto **uniforme** quando ogni punto viene scelto indipendentemente ed uniformemente all'interno del dominio oggetto di studio. A prima vista una distribuzione dei punti uniforme potrebbe sembrare casuale, in realtà dividendo la mappa in celle regolari si può notare come in ogni cella ricada in genere lo stesso numero di punti, pertanto la distribuzione presenta un aspetto casuale ed un aspetto regolare. Questo tipo di distribuzione viene detto uniforme, in quanto la densità dei punti è uguale per ogni cella elementare (Fig. 2.7b).

Nell'analisi spaziale l'uniformità della distribuzione dei punti sul territorio è importante, per esempio l'attendibilità di una carta ad isolinee è direttamente dipendente dalla densità e dall'uniformità dei punti. Nonostante ciò il grado di uniformità di una distribuzione di punti è

raramente misurato. Se la distribuzione dei punti è uniforme, ci aspettiamo che in ogni cella in cui è possibile suddividere un territorio ricada lo stesso numero di punti. Questa ipotesi può essere verificata usando il metodo del "chi quadrato" (χ^2); essa non dipende dalla forma e dall'orientamento delle celle. Il numero di punti atteso per ogni cella elementare è dato da: $E = N_{tot} / N_C$ dove $Ntot$ = numero di campioni ed Nc = numero di celle. Il risultato del test di uniformità della distribuzione è dato da: $\chi^2 = \Sigma (N_{tot} - E)^2/E$. Il test ha $\nu=(T-2)$ gradi di libertà, dove T è il numero di celle elementari. Se il valore calcolato supera il valore critico χ^2 per un livello di significatività del 5% la distribuzione dei punti non deve considerarsi uniforme.

Una distribuzione di punti è detta **casuale** quando la posizione di un punto non presenta alcuna relazione con la posizione di un altro punto (Fig 2.7c).

In un campionamento **a raggruppamenti (clusters)**, ogni punto viene scelto indipendentemente ed all'interno del dominio oggetto di studio laddove è possibile prelevare il campione. Sebbene in alcuni casi la distribuzione dei campioni possa sembrare uniforme, il numero dei campioni in ogni area elementare non è sempre lo stesso. Generalmente tale tipo di campionamento è giustificato dalla presenza nell'area investigata di zone impraticabili, od inaccessibili, pertanto i campioni vengono prelevati in numero maggiore nelle zone facilmente raggiungibili(Fig 2.7d).

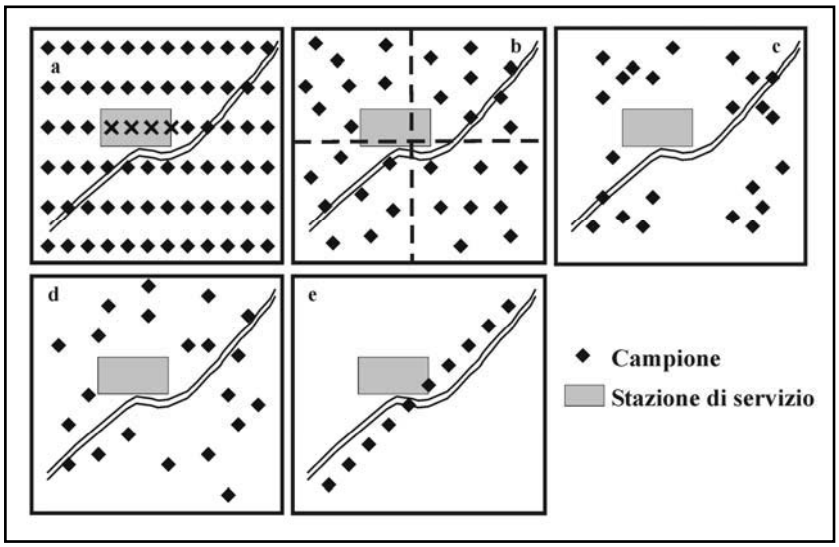


Figura 2.7. Tipi di distribuzione dei campioni: regolare (a), uniforme (b), a raggruppamenti (c), casuale (d), lungo profili (e). I campionamenti casuale, a raggruppamenti e lungo profili sono generalmente condizionati dalle possibilità di accesso, dalla mancanza di affioramenti, dalla necessità di investigare fenomeni direzionali o localizzati.

Nella distribuzione dei campioni lungo **profili** ogni punto viene scelto sulla stessa direzione a distanza fissa o variabile dagli altri punti. Generalmente, questo tipo di distribuzione viene utilizzata per investigare le caratteristiche direzionali (anisotropia) di un fenomeno (attraverso questo tipo di distribuzione è possibile mettere in evidenza fenomeni trasversali alla direzione di campionamento),

ovviare a difficoltà di accesso ai siti di indagine, e per facilità logistiche, come la presenza di una strada (Fig 2.7e).

Nel capitolo VIII sono riportati alcuni esempi relativi ai diversi modi di rappresentare graficamente la distribuzione dei punti di campionamento al fine di mettere in evidenza le caratteristiche spaziali delle variabili oggetto di studio.

La distribuzione dei campioni sul territorio oltre a considerare la posizione reciproca dei punti deve necessariamente tenere conto degli obiettivi delle indagini che si andranno ad effettuare. Quanti campioni devono essere raccolti affinché si abbia un numero sufficiente di dati per trarre delle conclusioni esaustive? Per rispondere a questa domanda è necessario considerare tre aspetti fondamentali: **l'estensione dell'area** da investigare e, di conseguenza, il tipo di fenomeno da rappresentare (regionale o locale), il **tempo** a disposizione e, soprattutto, le i **risorse economiche** a disposizione. In generale, ogni indagine di campagna (prospezione) può essere organizzata secondo lo schema e le caratteristiche riportate nella tabella 2.1. La **prospezione regionale** (angl. *Regional Survey*) viene utilizzata quando si deve indagare la distribuzione spaziale di un fenomeno poco conosciuto su aree molto vaste in tempi limitati. Le dimensioni delle porzioni elementari in cui si suddivide il territorio indagato variano da 1 a 5km con una densità di campionamento di 1-10 campioni per kmq.

Tabella 2.1. Tipi di campionamento

Tipo di prospezione	**Obiettivo**	**Maglia di campionamento**	**Densità di campionamento**
Regionale	Indagini preliminari su aree vaste e poco note	1-5 km	1-10 kmq
Dettaglio	Individuazione di zone anomale in aree già indiziate	0.5-1 km	20-100 kmq
Alta Risoluzione	Scoperta, delimitazione ed origine delle zone anomale	10-100 m	> 100kmq

La **prospezione di dettaglio** (angl. *Detailed Survey*) viene utilizzata sia per investigare fenomeni molto localizzati (già noti), quali ad esempio la dispersione di inquinanti nel suolo e nel sottosuolo da serbatoi interrati di carburanti, sia aree anomale evidenziate dalla prospezione regionale, ad esempio le zone caratterizzate da una concentrazione più elevata in elementi metallici che potrebbero indicare la presenza di un giacimento minerario. Vista l'estensione limitata di queste aree la maglia di campionamento può variare da 0.5 ad 1km di lato con una densità di 20-100 campioni a kmq.

La **prospezione ad alta risoluzione** (angl. *High Resolution Survey*) viene utilizzata per investigare fenomeni molto localizzati quale la precisa delimitazione delle zone anomale ed, come

nel caso dei giacimenti, per la caratterizzazione economica dei siti ai fini di un eventuale sfruttamento. In tali zone la maglia di campionamento può avere dimensioni variabili da 10 a 100m e la densità dei campioni può superare le 100 unità per kmq.

2.11 La presentazione grafica dei dati

In Statistica l'elaborazione di alcuni tipi di grafici costituisce una fase importante del processo di conoscenza e di controllo del fenomeno rappresentato dai dati raccolti durante la fase di campionamento. Gli obiettivi principali di questo paragrafo sono:

- definire il contributo fornito dall'elaborazione grafica quale rappresentazione sintetica dei dati per una migliore comprensione ed interpretazione del fenomeno oggetto di studio;
- descrivere la "bontà" del grafico, definita da Edward Tufte "*graphical integrity*", data dalla combinazione di un buon risultato visivo congiuntamente ad una buona informazione quantitativa del dato;
- elencare e distinguere le rappresentazioni grafiche più importanti.

I metodi grafici insieme alle informazioni fornite dai parametri numerici riassuntivi (indici statistici), costituiscono la "prima linea d'attacco" per l'interpretazione dei dati. E' indubbio che dei chiari grafici possono migliorare significativamente la comprensione di complessi *data sets*.

Ciò che costituisce l'efficacia di un grafico nel rappresentare i dati è ancora argomento di dibattito nella comunità scientifica. L'avvento di softwares per l'elaborazione di grafici statistici sempre più sofisticati ha sicuramente migliorato l'aspetto visivo, creando grafici di effetto che catturano l'attenzione, ma che in alcuni casi non hanno un reale valore interpretativo. Pur tuttavia, attraverso la conoscenza dei principali tipi di grafici e dei loro limiti rappresentativi, è possibile elaborare chiare rappresentazioni che si fondano su solidi principi statistici.

Tabella 2.2. Dati numerici relativi a 4 campioni, ogni campione è caratterizzato da due variabili x ed y.

DATA SET							
1		**2**		**3**		**4**	
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Il primo approccio alla costruzione di un grafico viene effettuato dall'osservazione delle variabili (colonne) nella tabella dei dati (Tab. 2.2) e dall'analisi dei principali indici statistici (Tab. 2.3). Ad un esame preliminare delle variabili i 4 campioni mostrano tutti gli stessi indici statistici, ma se rappresentiamo le variabili x ed y con dei grafici a dispersione vediamo come l'interpretazione può sensibilmente cambiare a seconda del tipo di correlazione che si evidenzia per ogni coppia di variabili (Fig. 2.8).

Tabella. 2.3. Principali indici statistici per le variabili x ed y della tabella 2.2.

Numero di punti (n)	11
Media di X	9.0
Media di Y	7.5
Regressione	y = 3 + 0.5x
Coefficiente di correlazione (r)	0.82
Livello di correlazione (r^2)	67%

I dati relativi alla prima coppia di variabili X1 e Y1 costituiscono un classico esempio di punti che presentano una dispersione più o meno equidistante attorno ad una linea retta (Fig. 2.8a) suggerendo l'esistenza di una relazione lineare, positiva tra le due variabili. Anche le due variabili del secondo campione sono fortemente correlate, in questo caso il tipo di relazione non è lineare, ma di tipo polinomiale (Fig. 2.8b). Le variabili del campione 3 mostrano una correlazione lineare molto forte (quasi perfetta), ma la relazione lineare è complicata dalla presenza di un singolo valore anomalo (outlier), la sua rimozione potrà condurre ad un grafico più chiaro (Fig. 2.8c). Nel caso della figura 2.7 d i valori di Y sono indipendenti dalla X (costante) ad eccezione di un outlier.

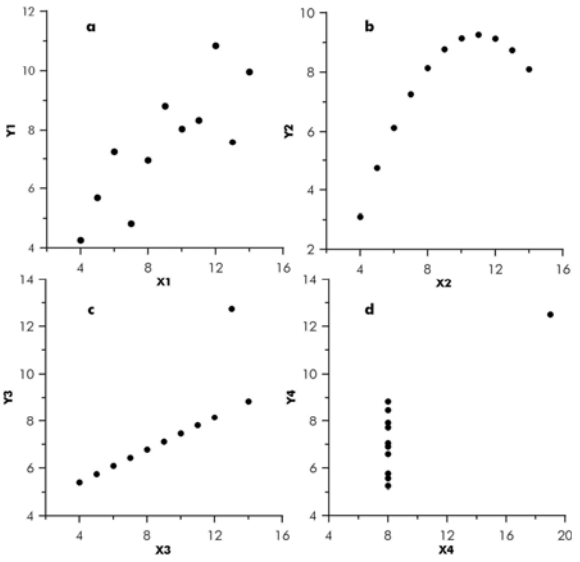


Figura 2.8. Grafici di correlazione x-y per le variabili della tabella 2.2.

2.12 I tipi di grafici

In questo paragrafo saranno presentati i tipi di grafici più utilizzati e per ognuno saranno mostrati uno o due esempi con alcuni commenti per il loro buon uso. Una descrizione più accurate sarà presentata nei capitoli successivi.

2.12.1 Grafici a dispersione (angl. Scatterplot)

I grafici a dispersione sono largamente utilizzati in Geologia e nelle Scienze Naturali in genere per presentare dati relativi a due o più variabili che possono essere in qualche modo relazionate fra

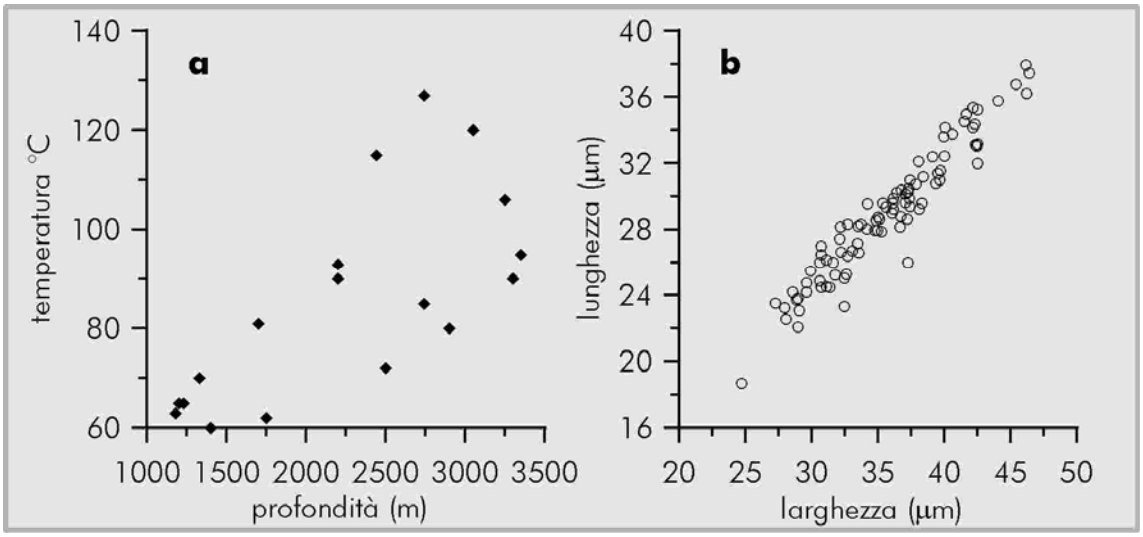


Figura 2.8. Esempi di diagramma a dispersione che mostrano una relazione lineare positiva fra le variabili. Nel primo caso (a) il grafico mostra l'esistenza di una relazione lineare (anche se non molto chiara) tra la profondità di alcuni giacimenti di olio e la temperatura del sottosuolo; nel secondo caso (b) si può notare una forte la relazione lineare tra le dimensioni di alcune specie di foraminiferi.

loro. La variabile che viene rappresentata sull'asse delle ordinate viene detta variabile dipendente, mentre la variabile rappresentata sull'asse delle ascisse è detta variabile indipendente. Se la dispersione numerica di una delle variabili è maggiore di due o più ordini di grandezza, è possibile convertire la variabile con i valori più elevati a scala logaritmica al fine di ottenere un grafico leggibile. La figura 2.10 riporta un esempio di grafico a dispersione con due variabili.

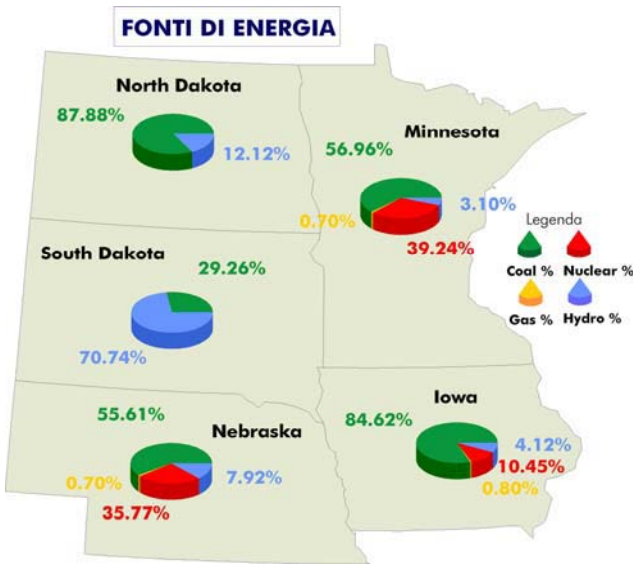


Figura 2.10. Grafici a torta posizionati su una mappa. I grafici presentano le fonti di energia ed il loro utilizzo (%) in alcuni Stati degli USA.

*2.12.2 Diagrammi a torta (*angl. *Pie Charts*)

Il diagramma a torta è una delle forme

grafiche principali per presentare i dati (anche se non molto utilizzato). Se utilizzato con criterio può costituire un valido aiuto per rappresentare gruppi di dati non numerosi, con le seguenti limitazioni:

- quando la somma dei valori è costante, come ad esempio il 100%;
- quando i singoli valori mostrano variazioni significative;
- quando il numero di categorie è abbastanza risotto, generalmente tra 3 e 10. La figura 2.10 mostra un esempio di tale grafico.

2.12.3 Diagrammi a barre e a colonna (angl. bar and charts column)

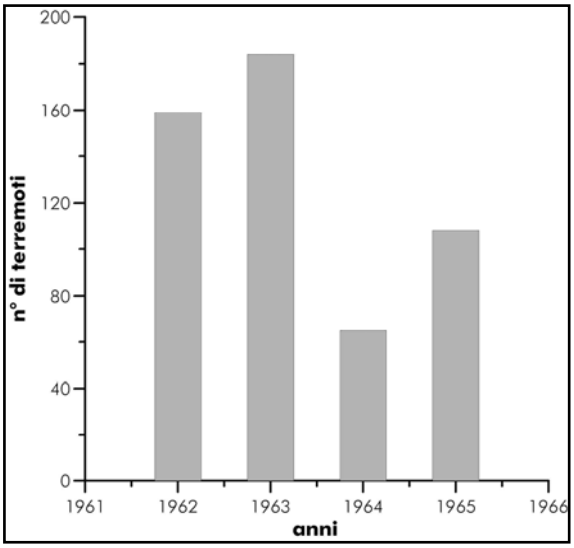


Figura 2.11. Grafico a colonna in cui è rappresentato il numero di terremoti avvenuti dal 1962 al 1965.

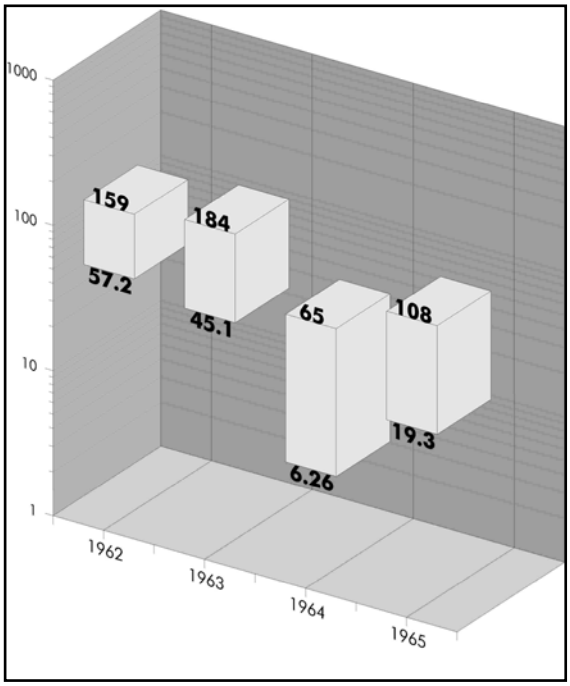


Figura 2.12. Grafico a colonna in 3D. Il grafico rappresenta due variabili: i microsismi avvenuti dal 1962 al 1965, indicati dal numero in grassetto sulla base superiore dei blocchi, ed la quantità di materiale iniettato nel sottosuolo che li ha generati (numero in grassetto alla base del blocco). Poiché i range delle due variabili sono di un ordine di grandezza differenti, l'asse delle ordinate è riportato a scala logaritmica.

Come i diagrammi a torta, anche i grafici a barre (Fig. 2.11) e a colonna (Fig. 2.12) possono essere elaborati per rappresentare dati raggruppati. L'uso corretto di tali grafici è limitato da alcune regole:

- possono essere utilizzati per rappresentare gruppi di dati discretizzate sia a scala ordinale che nominale;
- possono essere utilizzati, generalmente, per dati che mostrano una sequenza nel tempo delle categorie in cui sono rappresentati, come le serie temporali, ad es. January, February, March…1987, 1988, 1989…oppure serie di dati suddivisi in classi 0-10, 11-20, 21-30;
- i grafici a barre orizzontali(fig. 2.13) sono in genere utilizzati per rappresentare dati che non mostrano una successione naturale delle categorie in cui sono suddivisi, come strade, ferrovie, aria, mare, ecc.

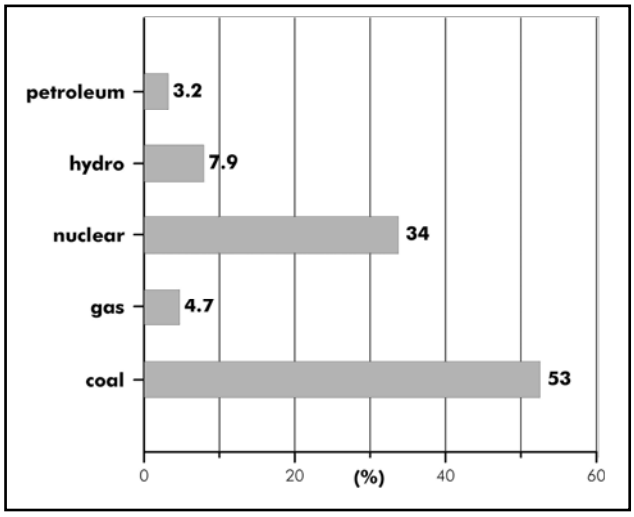


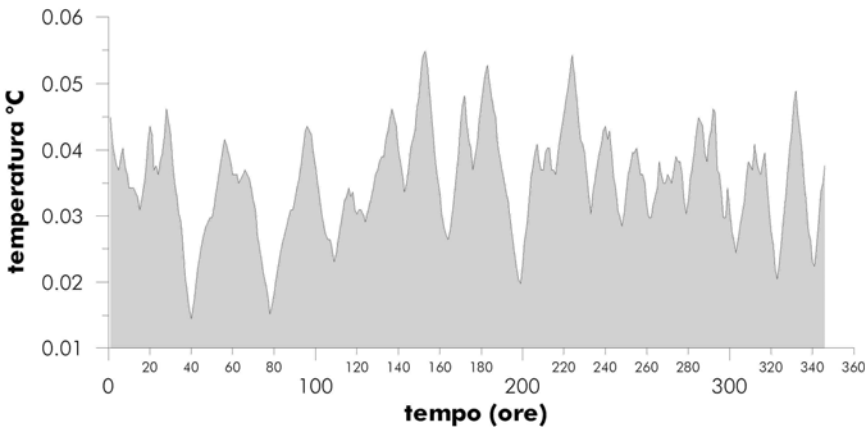
Figura 2.13. Grafico a barre orizzontali che riporta i consumi percentuali relativi al tipo di energia utilizzata.

- è ragionevole utilizzare una scala logaritmica per l'asse delle ordinate se il range dei valori è più grande di due ordini di grandezza (es. 0-200).

2.12.4 Diagrammi a linea (angl. line charts)

I diagrammi a linea (*angl. line charts*) sono molto simili ai diagrammi a dispersione con la differenza che I valori della variabile indipendente (x) sono caratterizzate da una sequenza propria. Inoltre, tali campioni in genere provengono da serie continue, come la temperatura, la pressione, i prezzi di mercato, ecc. Con tali grafici impossibile rappresentare differenti variabili nello stesso grafico che, se necessario, possono anche avere differenti scale. La figura 2.14 mostra un esempio di line chart.

Figura 2.14. Il grafico a linee rappresenta la variazione di temperatura in un esperimento della durata di 346 ore.



2.12.5 Diagrammi polari (angl. polar charts)

Mentre i diagrammi a dispersione, a torta ed a barre sono utilizzati per diversi tipi di dati, esistono alcuni tipi di grafici particolari utilizzati per rappresentare dati che possono apparire insoliti, ma che in Geologia sono molto comuni. Uno di questi grafici è il grafico polare (*angl. polar chart*, fig. 2.15) che viene utilizzato per rappresentare dati discreti in cui ogni campione ha un valore che esprime la direzione rispetto ad un punto fisso (0 - 360 gradi) ed un valore che esprime una quantità come la pendenza, la forza di un campo come quello magnetico, ecc. Questo grafici servono a rappresentare in genere dati vettoriali.

2.12.6 Diagrammi ternari (angl. ternary plot)

I diagrammi triangolari (Fig. 2.16) sono utilizzati per rappresentare dati discretizzate in cui ogni campione possiede tre valori che in genere sono espressi come percentuale. Esempi di questo tipo di grafici sono quelli relativi alla composizione percentuale di un suolo rispetto alle componenti sabbia, argilla e silt, oppure la composizione di una roccia rispetto ai minerali principali di cui è costituita.

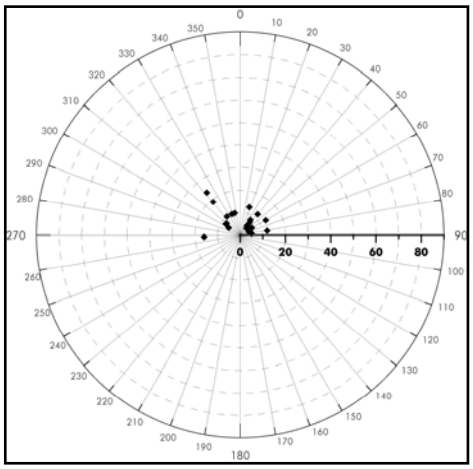


Figura 2.15. Grafico polare che rappresenta la direzione e l'immersione di uno strato sabbioso in una formazione marina pliocenica.

2.12.7 Diagrammi 3D

Finora tutti i grafici presentati mostrano una o due variabili contemporaneamente. Tuttavia in alcuni casi è necessario rappresentare tre variabili che possono essere tra loro interdipendenti. Se i dati sono raggruppati è possibile utilizzare un grafico a barre in 3D (Fig. 2.12) dove gli assi di base rappresentano due variabili indipendenti. Al contrario se le variabili sono tra loro dipendenti e sono costituiti da valori continui è possibile rappresentarle con una superficie in tre dimensioni, detto *angl. surface plot* (Fig. 2.17). La superficie può essere rappresentata in forma continua o con una serie di isolinee.

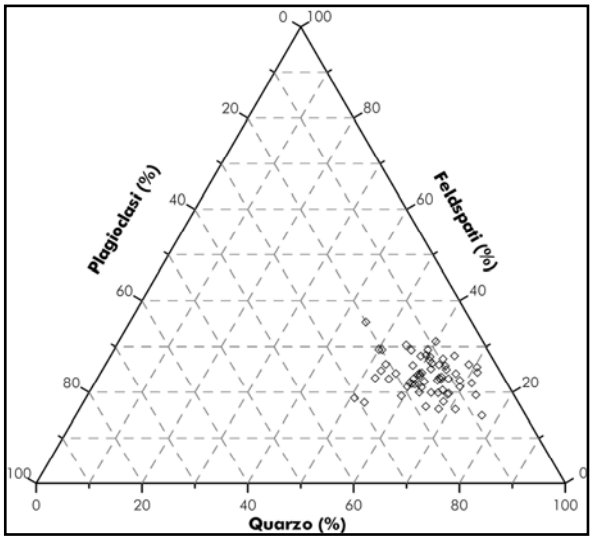


Figura 2.16. Diagramma ternario relativo alla composizione percentuale di una roccia intrusiva.

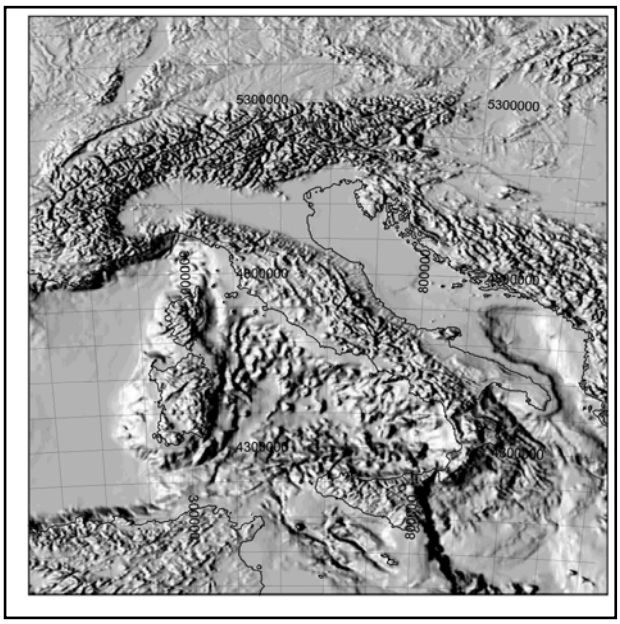


Figura 2.17. Mappa della superficie topografica d'Italia (Modello Digitale del Terreno, DTM), rappresentata con effetto tridimensionale.

Bibliografia

Cochran W.G., 1977, Sampling Techniques, third ed. John Wiley & Sons Inc., New York, pp. 428..

Koch G.S.Jr., Link R.F., 2002, Statistical Analysis of Geological Data, Dover Publications, Inc., Mineola, NewYork, pp. 438.

Soliani L., 2004, Manuale di statistica per la ricerca e la professione, UNI. Nova di Pietro Lia, Parma, pp. 550.

Starks T.H., 1988, Evaluation of control chart methodologies for RCRA waste sites. U.S. Environmental Protection Agency Technical Report CR814241-01-03, Washington D.C.

Swan A.R.H., Sandilands M., 1995, Introduction to Geological Data Analysis. Blackwell Science Ldt, pp.431

Townend J., 2004, Practical Statistics for Environmental and Biological Scientists, John Wiley & Sons, ltd, England, pp.276.

USEPA, 2002, Guidance on choosing a sampling design for environmental data collection. EPA QA/G-5S, Office of Environmetal Information, Washington D.C.

Westphal Ch., T. Blaxton, Data Mining Solutions: Methods and Tools for Solving Real-World Problems, John Wiley, 1998.