Calcolo della Concentrazione Rappresentativa della Sorgente (CRS)

Renato Baciocchi, Emiliano Scozza

Università di Roma "Tor Vergata"

Valutazione dei Dati

Data Set di ingresso

Il data-set deve essere suddiviso in relazione ad ogni sorgente secondaria di contaminazione:

Suolo Superficiale (SS),

Suolo Profondo (SP)

Falda (GW).

Valutazione dei Dati

Applicabilità dei criteri statistici

1. Ampiezza del Data Set: Per ogni data set (SS, SP, GW) il numero dei dati disponibili non deve essere inferiore a 10;
2. Campionamento: Va verificato che il campionamento sia uniformemente distribuito sulla sorgente di contaminazione;
3. Outlier: Vanno distint i falsi outlier dai veri outlier;
4. Non Detect: Questi valori vanno identificati e posti pari al corrispondente Detection Limit (DL).

Valutazione dei Dati

Outlier

L'identificazione e la trattazione degli outlier può essere distinta in tre fasi

- Identificazione (tramite analisi visiva dei dati opportunamente graficati)
- Distinzione tra veri (da rimuovere) e falsi (da mantenere) outlier
- Studio scientifico degli outlier identificati

Valutazione dei Dati

Non Detect

Sono quei valori al di sotto del Detection Limit ma il cui valore non è nullo.

Aggiungere parte sui metodi???

Distribuzione dei Dati

Tipi di distribuzione

Le distribuzioni di probabilità più comunemente utilizzate per la loro rappresentazione sono:

- distribuzione gaussiana o normale
- distribuzione lognormale
- distribuzione gamma
- distribuzione non parametrica.

Distribuzione dei Dati

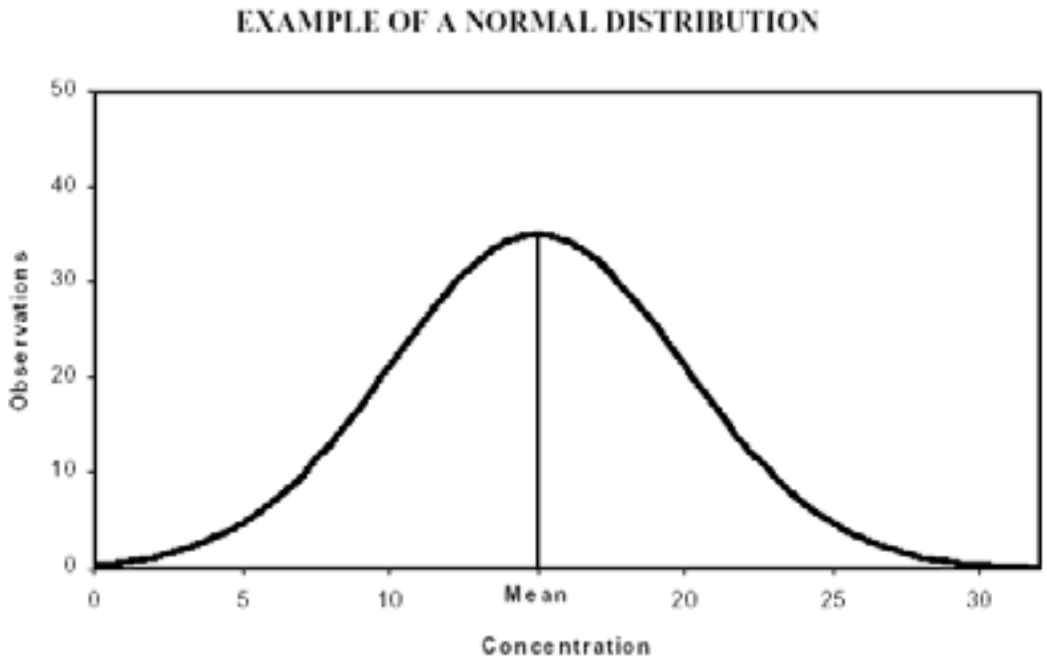
Distribuzione Normale

è una distribuzione di tipo simmetrico la cui tendenza centrale è data dal calcolo della media aritmetica dei valori $x_1, x_2, x_3, \ldots, x_n$ delle grandezze considerate.

La forma della distribuzione normale è descritta dalla funzione Densità di Probabilità, definita da due parametri: la media aritmetica e la varianza del campione, che è indice della dispersione dei dati rispetto al valor medio.

Distribuzione dei Dati

Distribuzione Normale



Distribuzione dei Dati

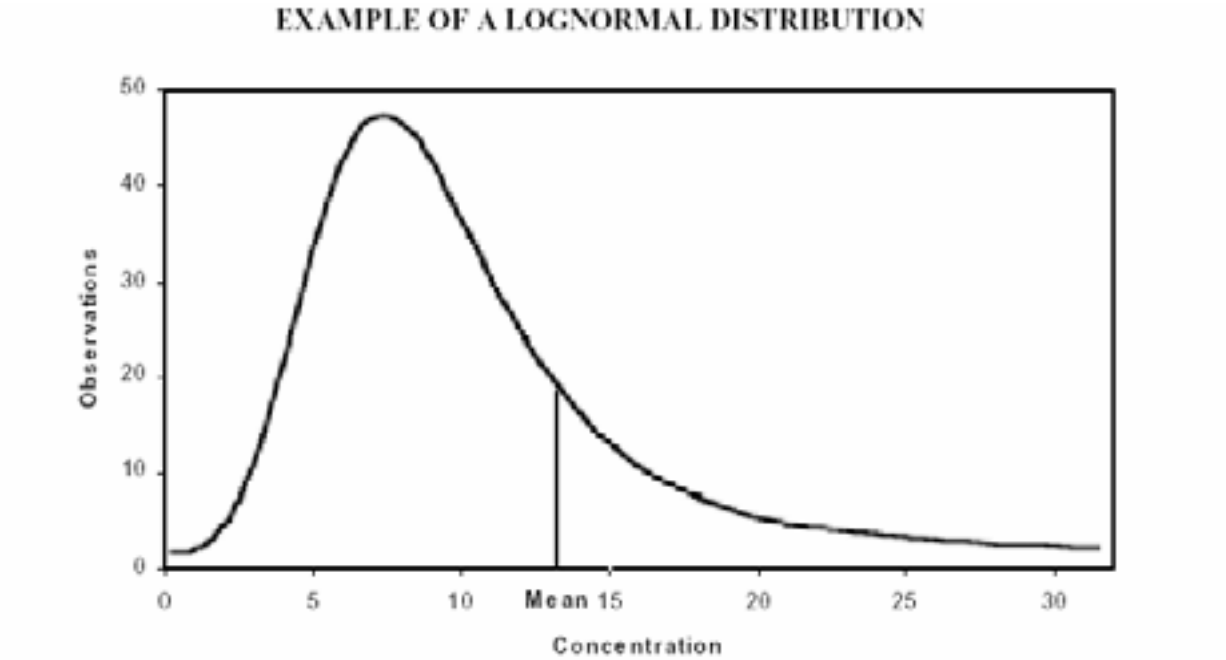
Distribuzione Lognormale

è un tipo di distribuzione asimmetrica, derivante dal calcolo della media geometrica dei valori. La sua forma è più pendente di quella di una distribuzione normale ed è delimitata a sinistra dallo zero, mentre la parte finale all'altra estremità risulta avere una specie di coda più lunga di quella normale.

La distribuzione lognormale è generalmente definita da due parametri $\overline{y}$ e σ_y^2 (media e varianza della variabile trasformata).

Distribuzione dei Dati

Distribuzione Lognormale



Distribuzione dei Dati

Distribuzione Gamma

Molti data set che presentano asimmetrie possono essere rappresentati sia mediante una distribuzione lognormale che da una distribuzione di tipo gamma, specialmente nei casi in cui il numero di campioni n è inferiore a 70-100.

La distribuzione gamma è generalmente definita da due parametri: k (parametro di forma) e θ (parametro di scala); il loro prodotto è pari alla media aritmetica

Distribuzione dei Dati

Distribuzione Non Parametrica

Nel caso in cui non sia possibile dimostrare che i valori di un data set seguano una tra le suddette distribuzioni (ad esempio a causa dello scarso numero di campioni) o qualora risulti, dalla applicazione dei test statistici, che nessuna distribuzione approssimi bene l'insieme dei dati, allora si parla di data set non parametrici.

In tal caso esistono delle procedure specifiche, per l'individuazione del valore rappresentativo dell'insieme dei dati, indipendenti dai parametri statistici e dal tipo di distribuzione dei dati.

Distribuzione dei Dati

Test per la selezione del tipo di distribuzione

TIPO DI TEST	TIPO DI DISTRIBUZIONE				Rif. Bibliografico
	NORMALE	LOG NORMALE	GAMMA	NON PARAMETRICA	
"Shapiro e Wilk test" (n < 50)	X	X	---	---	[Gilbert, 1987], [software ProUCL ver. 3.0]
"D'Agostino test" (n = 50)	X	X	---	---	[Gilbert, 1987]
"Normal Quantile-Quantile (Q-Q) Plot"	X	X	---	---	[software ProUCL ver. 3.0]
"Lilliefors Test"	X	X	---	---	
"Gamma Quantile-Quantile (Q-Q) Plot"	---	---	X	---	
"Kolmogorov-Smirnov test"	---	---	X	---	
"Anderson Darling test"	---	---	X	---	

Criteri di Calcolo

Criteri di stima della Concentrazione Rappresentativa

I criteri di calcolo per la stima della CRS si riferiscono essenzialmente alle seguenti grandezze statistiche:

1. Valore massimo;
2. Media aritmetica, per una distribuzione normale;
3. Media geometrica, per una distribuzione lognormale;
4. UCL 95%
5. Percentile 95%.

Criteri di Calcolo

UCL 95%

Statisticamente l'UCL 95% di una media è definito come un valore che, quando calcolato ripetutamente per un sottoinsieme di dati scelti a caso, eguaglia o supera il valore vero della media il 95% delle volte.

Tale valore rappresenta una stima altamente conservativa del valore vero della media. Viene comunque utilizzato nel calcolo della concentrazione rappresentativa alla sorgente Cs, poichè tiene conto dell'incertezza legata al calcolo della media che non detto fornisca sempre una stima realmente rappresentativa, dato il numero finito di campioni a disposizione.

Criteri di Calcolo

Percentile 95%

Il percentile rappresenta la condizione in cui una percentuale *x* della distribuzione è minore o pari al valore del percentile. In particolare, quindi, il percentile al 95% è quel valore che eguaglia o supera il 95% dei valori di concentrazione che costituiscono l'insieme dei dati. Tale valore rappresenta quindi, in genere, una stima più conservativa rispetto all'UCL 95%.

Criteri di Calcolo

Documento APAT 2006

1. Individuare la distribuzione di probabilità che approssimi meglio l'insieme dei dati disponibili (software ProUCL ver. 3.0).
2. Applicare la procedura statistica corrispondente al tipo di distribuzione riconosciuta. A seconda del tipo di distribuzione, selezionata come maggiormente rappresentativa del data set in esame, è possibile individuare il più appropriato criterio per il calcolo dell'UCL. (software ProUCL ver. 3.0).

Criteri di Calcolo

Software Pro-UCL

Test Statistici

TIPO DI TEST [software ProUCL ver. 3.0]	TIPO DI DISTRIBUZIONE			
	NORMALE	LOG NORMALE	GAMMA	NON PARAMETRICA
"Normal Quantile-Quantile (Q-Q) Plot"	X	X	---	---
"Shapiro e Wilk test" (n < 50)	X	X	---	---
"Lilliefors Test"	X	X	---	---
"Gamma Quantile-Quantile (Q-Q) Plot"	---	---	X	---
"Kolmogorov-Smirnov test"	---	---	X	---
"Anderson Darling test"	---	---	X	---

Criteri di Calcolo

Software Pro-UCL

Distribuzione lognormale

σ_y	Numero di campioni (n)	UCL consigliato
$\sigma_y < 0.5$	per ogni n	Student's t, Modified-t, H-UCL(metodo Land)
$0.5 \le \sigma_y < 1$	per ogni n	H-UCL
$1 \le \sigma_y < 1.5$	n<25	95% Chebyshev (MVUE) UCL
	$n \ge 25$	H-UCL
$1.5 \le \sigma_y < 2$	n<20	99% Chebyshev (MVUE) UCL
	$20 \le n < 50$	95% Chebyshev (MVUE) UCL
	$n \ge 50$	H-UCL
$1.5 \le \sigma_y < 2$	n<20	99% Chebyshev (MVUE) UCL
	$20 \le n < 50$	97.5% Chebyshev (MVUE) UCL
	$50 \le n < 70$	95% Chebyshev (MVUE) UCL
	$n \ge 70$	H-UCL
$2.5 \le \sigma_y < 3$	n<30	Il maggiore tra 99% Chebyshev (MVUE) UCL e 99% Chebyshev(Media,Dev.Standard)
	$30 \le n < 70$	97.5% Chebyshev (MVUE) UCL
	$70 \le n < 100$	95% Chebyshev (MVUE) UCL
	$n \ge 100$	H-UCL
$3 \le \sigma_y < 3.5$	n<15	UCL calcolato con metodo Hall's bootstrap
	$15 \le n < 50$	Il maggiore tra 99% Chebyshev (MVUE) UCL e 99% Chebyshev(Media,Dev.Standard) UCL
	$50 \le n < 100$	97.5% Chebyshev (MVUE) UCL
	$100 \le n < 150$	95% Chebyshev (MVUE) UCL
	$n \ge 150$	H-UCL
$\sigma_y > 3.5$	per ogni n	Utilizzare UCL calcolato con metodi non parametrici