APPENDICE H

CONCENTRAZIONE RAPPRESENTATIVA ALLA SORGENTE

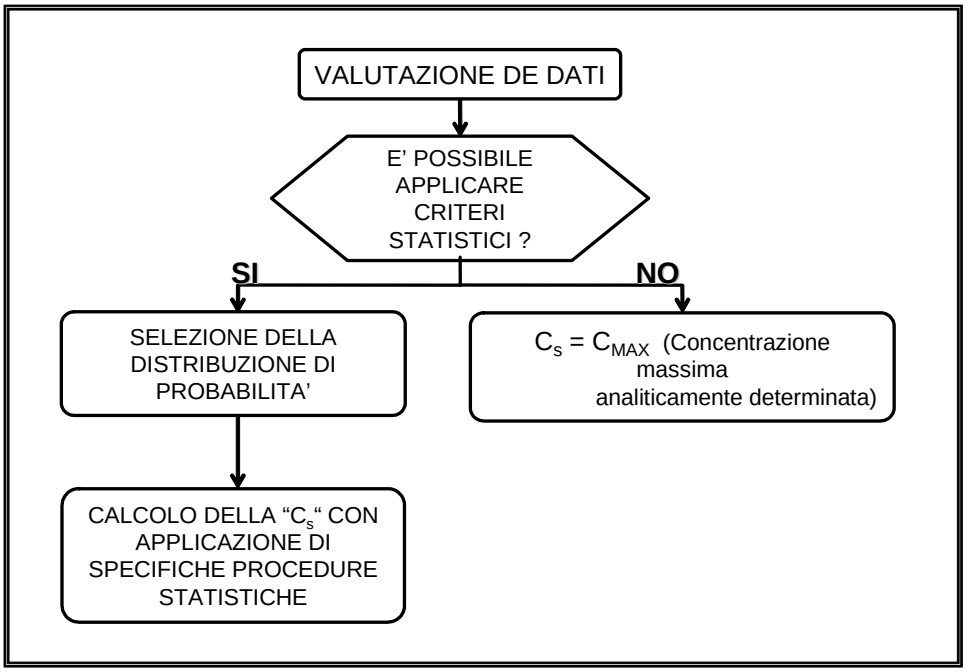
Per la definizione del criterio utile per la selezione del valore di concentrazione rappresentativo alla sorgente (C_s) in corrispondenza ad un livello 2 di analisi di rischio sanitario, sono stati criticamente valutate le informazioni e i risultati derivanti da due successive e distinte fasi di lavoro.
Nella prima fase (par. H.1 - H.3) è stato condotto un approfondito esame di tutti i testi ed i software adottati quali riferimento di base (vedi capitolo 1); nello studio è stato inoltre consultato materiale bibliografico supplementare specifico sull'argomento [OSWER 9285.6-10, EPA 2002] [D.Kofi Asante-Duah, 1993] [Gilbert, 1987] [Florida Dep. E.P.D., 2004] [Supplemental Guidance to RAGS, EPA 1992] [EPA. 2000a, QA/G-4HW] [EPA 2000b, QA/G-9] [Manuale ProUCL, EPA 2004].
Nella seconda fase (par. H.4) sono stati applicati i criteri sopra individuati a due casi studio rappresentativi.

Il principio ispiratore che ha guidato la definizione del criterio per la selezione della C_s è stato, come in tutte le fasi decisionali presenti in queste linee guida, il principio di cautela, o conservatività.

Per l'individuazione della concentrazione rappresentativa alla sorgente (C_s) è necessario innanzitutto effettuare una accurata valutazione dei dati (Paragrafo H.1), in grado di stabilire l'applicabilità di criteri statistici sui valori di concentrazione analiticamente determinati nei campioni di suolo e di falda.
Nel caso in cui si abbiano pochi dati o non possano essere applicati criteri statistici suddetti, si utilizza come C_s il valore massimo di concentrazione (C_{MAX}) riscontrato in un determinato mezzo.
Nel caso in cui sia possibile applicare i criteri statistici, la prima cosa da fare è individuare la distribuzione di probabilità che approssimi meglio l'insieme dei dati disponibili (Paragrafo H.2), per poi procedere all'applicazione della procedura statistica corrispondente al tipo di distribuzione riconosciuta (Paragrafo H.3).
Tali procedimenti statistici infatti, possono influenzare in maniera sostanziale i risultati dello studio, perciò l'individuazione della distribuzione più rispondente a rappresentare

del set di dati a disposizione, è fondamentale per ricavare stime di parametri il più possibile attinenti alla situazione reale.
Quanto detto sopra è riassunto, sotto forma di diagramma di flusso, in figura H.1.

Figura H.1 – Selezione della concentrazione rappresentativa alla sorgente (C_s)



H.1 VALUTAZIONE DEI DATI

In tale contesto, si presuppone che i dati analitici a disposizione siano stati già validati, ossia sia stata verificata la loro attendibilità.

Nel seguito verranno descritti i principali elementi che è necessario prendere in considerazione per stabilire l'applicabilità di criteri statistici atti ad individuare un valore unico di concentrazione rappresentativo (C_s) dell'insieme di dati a disposizione.

- Per ogni data-set (SS, SP, GW), il numero di dati a disposizione non può essere inferiore ad un valore minimo. L'ampiezza del data set è di particolare importanza soprattutto nei casi in cui si abbia una grande variabilità della distribuzione dei dati.

  Tutti i testi esaminati concordano nel porre tale valore minimo pari a 10 (paragrafo H.3.2).

Al di sotto di tale soglia, non essendo attendibile alcuna stima statistica e in accordo con il principio di massima conservatività, si pone la concentrazione rappresentativa alla sorgente coincidente con il valore di concentrazione massimo analiticamente determinato ($C_s = C_{MAX}$).

In tale ambito, si ritiene opportuno evidenziare che l'Allegato 2 al D.M. 471/99 fornisce indicazioni riguardanti il numero di punti da sottoporre a campionamento sulla base della estensione del sito da investigare. Tali indicazioni sono riportate in tabella H.1.

Tabella H.1 – Numero di punti di campionamento in funzione della estensione della sorgente (D.M. 471/99)

Estensione della sorgente (m^2)	Numero di punti di campionamento
< 10.000	Almeno 5 punti
10.000 – 50.000	Da 5 a 15 punti
50.000 – 250.000	Da 15 a 60 punti
250.000 – 500.000	Da 60 a 120 punti
> 500.000	Almeno 2 punti ogni 10.000 m^2

- I criteri statistici applicabili per la stima della C_s si basano sulla assunzione che il campionamento sia uniformemente distribuito su tutta la sorgente di contaminazione (campionamento random o campionamento a griglia) [OSWER 9285.6-10, EPA 2002].

  Possono comunque verificarsi casi in cui il campionamento sia più concentrato nella porzione del sito maggiormente sospetta di contaminazione. In tali situazioni il tener conto di tali aree sovra-rappresentate può comportare una sovrastima della C_s. Poiché tale approccio risulta essere conservativo e quindi protettivo per la salute umana, lo stesso può ritenersi accettabile. Non è invece ammissibile il caso in cui le aree caratterizzate da un maggiore grado di contaminazione, o sospette tali, siano sotto-rappresentate [Florida Dep. E.P.D., 2004].

- Possono, infine ma non per ultimo, verificarsi casi in cui la presenza di "outlier" e/o di un numero consistente di dati "non-detect" può comportare effetti negativi sul calcolo della C_s.

  Data la rilevanza di tali due argomenti, questi verranno trattati nel seguito con particolare dettaglio.

H.1.1 GLI "OUTLIER"

Gli outlier sono quei valori di un data set che non sono rappresentativi dell'insieme di dati nel suo complesso. Non sono rappresentativi perché, in genere, sono quantitativamente in numero estremamente ridotto e qualitativamente assumono dei valori molto grandi o molto piccoli rispetto al resto del data set.
In campo ambientale di inquinamento dei suoli, valori di concentrazione molto alti in genere corrispondono ai picchi (hot spot) locali di contaminazione.
Comunque, in generale, tali valori estremi possono costituire dei "veri outlier" o dei "falsi outlier". I **primi** possono derivare da errori di trascrizione, di codifica dei dati o da una qualsiasi inefficienza degli strumenti del sistema di rilevazione dei dati. I **secondi** sono quei valori estremi reali, spesso presenti in questo tipo di indagini soprattutto, come gia detto, in campo ambientale. La rimozione dei secondi e/o la mancata rimozione dei primi può condurre ad una visione errata del data set [EPA 2000b, QA/G-9]. Infatti è di fondamentale importanza tener conto e quindi non rimuovere i "falsi outlier" dal data set [OSWER 9285.6-10, EPA 2002].
Se il data-set a disposizione è stato già validato si esclude automatocamente la presenza di veri outlier.

L'identificazione degli outlier può essere condotta attraverso le seguenti fasi [EPA 2000b, QA/G-9]:

1. Identificazione dei valore estremi che potranno essere potenziali outlier. Questo può essere fatto mediante rappresentazione grafica dell'insieme dei valori rilevati:è possibile così individuare velocemente quei punti che corrispondono a valori più elevati o più ridotti rispetto agli altri. Una volta identificati i potenziali outliers, è necessario procedere a ulteriori indagini, applicando uno dei test statistici disponibili.

2. Applicazione di un opportuno test statistico. Esistono molti test statistici atti a verificare se un outlier statistico, cioè un potenziale vero outlier, sia tale o meno. I principali test statistici utili a tale scopo sono quattro:
   - Extreme value test (Dixon's Test)

- Discordance Test
- Rosner's test
- Walsh's test

Questi, descritti dettagliatamente nel seguito, si differenziano per le dimensioni del data set da considerare, il numero di potenziali outlier da analizzare e la necessità o meno di una distribuzione di tipo normale dei dati raccolti.

In particolare la guida raccomanda l'uso del Rosner's test quando il data set contiene un numero di elementi maggiore di 25; in caso contrario suggerisce quello dell'Extreme Value test. Se si ha un solo valore sospetto outlier il Discordance test può essere sostituito a uno di questi test. Se però i dati non seguono una distribuzione normale si deve considerare un test non parametrico, come il Walsh's test (Tabella H.2).

Tabella H.2 - Criteri di selezione del test per la identificazione degli outlier

DIMENSIONE DEL DATA SET	TEST	DISTRIBUZIONE NORMALE
$n \leq 25$	Estreme Value Test	Sì
$n \leq 50$	Discordance Test	Sì
$n \geq 25$	Rosner's Test	Sì
$n \geq 50$	Walsh's Test	No

Per la descrizione di dettaglio dei test si rimanda al documento [EPA 2000b, QA/G-9].

3. Studio scientifico degli outlier identificati per la scelta di trattamento del dato. I test statistici (fase 2) da soli non permettono di stabilire se comprendere o escludere il dato dall'insieme considerato. Le scelte possibili sono:
   - Effettuare ulteriori approfondimenti e indagini al fine di correggere il valore di outlier.
   - Utilizzare il data set comprensivo dei valori di outlier.
   - Escludere l'inserimento di tali valori dal data set. Tale scelta può avvenire solo se è possibile accompagnare i risultati dei test statistici (fase 2) con valide giustificazioni scientifiche.

4. Nel caso di esclusione degli outlier dal data set, conduzione della successiva analisi statistica dei dati sia sull'insieme dei dati comprensivo di outlier, sia su quello rivisto con l'eventuale soppressione degli outlier.

5. Documentazione dell'intero procedimento, con la descrizione di tutti i passaggi e le scelte effettuate.

H.1.2 I "NON-DETECT"

Tutte le tecniche analitiche di laboratorio hanno un " Detection Limit"(DL) (limite di rilevazione): i valori cosiddetti "non-detect" (ND) sono quelle concentrazioni realmente o virtualmente pari a zero, o comunque maggiori di zero, ma al di sotto delle possibilità di misurazione del laboratorio. Il DL dipende dalla sensibilità della metodica di estrazion ed analisi.
Un data set contenente non-detect viene definito in letteratura "censored" (o "left-censored") a indicare la sua incompletezza, che può essere più o meno grave a seconda del DL del laboratorio che ha condotto il campionamento: per questo motivo alla documentazione dello studio il laboratorio è opportuno che alleghi informazioni sul "Quantitation Limit"(limite di misura) che dipenderà dalla strumentazione di cui si è servito. Il Quantitation Limit può essere definito come il livello più basso al quale una sostanza chimica può essere misurata con precisione, generalmente pari al DL dello strumento moltiplicato per un fattore compreso fra tre e cinque, ma comunque variabile a seconda della sostanza considerata e del tipo di campione [RAGS Part A, EPA 1989].

La presenza di ND in un insieme di dati può influire pesantemente sul calcolo della media, della varianza, sull'orientamento dei dati e su vari altri parametri, pregiudicando quindi il procedimento statistico nel caso in cui questo risulti applicabile nonostante la loro presenza.
I laboratori di analisi riportano questi valori come "non detected" (ND), oppure li pongono pari a zero o come dati "less-than"(LT) cioè "minori di" una certa quantità, in genere pari proprio al DL, o ancora capita di trovarli indicati come valori pari ad una frazione del DL (in genere a ½ DL). E' comunque preferibile, qualora le tecniche di

misurazione lo consentano, che siano riportate le loro misure esatte, benché minime, per non perdere informazioni utili all'analisi dei dati.

Nel seguito è riportato quanto proposto dai testi bibliografici presi quali riferimento.

Il documento [OSWER 9285.6-10, EPA 2002] descrive quattro possibili approcci per la trattazione dei non-detect, finalizzati alla applicazione di analisi statistiche dell'insieme dei dati e alla conseguente individuazione di un valore rappresentativo.

1. Riesame del modello concettuale del sito: da questo riesame potrebbe risultare una distribuzione dei valori di concentrazione tali da permettere l'individuazione di aree a maggior grado di contaminazione e aree a minor grado di contaminazione. In tal caso, il sito oggetto di indagine potrebbe essere suddiviso in sotto-aree, alcune delle quali presenteranno una maggiore e altre una minore concentrazione di non-detect. In tale caso potrebbe risultare necessario raccogliere un maggior numero di campioni per permettere una migliore caratterizzazione del sito.
2. Metodo della sostituzione semplice ("Simple Substitution Methods"): questo metodo prevede l'assegnazione di un valore costante ai dati non-detect. Tale valore potrà essere:
    - pari a zero;
    - pari al Detection Limit;
    - pari alla metà del DL.

    L'incertezza associata a questo metodo aumenta all'aumentare del valore del DL e all'aumentare del numero di non-detect. Quindi si consiglia di scegliere, senza un preciso criterio, il valore costante da attribuire tra i tre proposti solo nel caso in cui il numero dei non-detect costituisce al massimo il 15% di tutto il data set [EPA 2000b, QA/G-9].
3. Metodo degli estremi ("Bounding Methods"): Tale metodo propone di calcolare il valore di concentrazione rappresentativo alla sorgente (generalmente l'UCL) attribuendo, di volta in volta, uno dei valori costanti elencati sopra. Questi metodi forniscono una stima del limite superiore e di quello inferiore dei vari UCL95%, calcolati sulla base dell'intero range di valori dei non-detects possibili (da 0 fino al DL).
4. Metodi della distribuzione ("Distributional Methods"): Si basano sull'ipotesi che la forma della distribuzione dei dati non-detects sia simile a quella delle concentrazioni

misurate che superano il DL. Tra questi metodi il più utilizzato è il Metodo di Cohen ("Cohen's Method").

Metodo di Cohen ("Cohen's Method) [EPA 2000b, QA/G-9]: è applicabile se i non-detect costituiscono il 15-50% del data set disponibile, se la forma della distribuzione dei dati senza i valori non-detect sia di tipo normale e che il DL sia sempre lo stesso .

Questo metodo adatta la media e la deviazione standard per valori al di sotto del DL, basandosi sulla tecnica statistica della stima più probabile della media e della varianza, in modo che sarà possibile applicare i vari test statistici al data set.

Nella applicazione di questo metodo i non-detect non si assumono mai pari a zero.

Le stime derivanti dai campionamenti sono $x_1, x_2, x_3, ..., x_n$ di cui i primi m valori rappresentano i dati sopra il DL. Quelli sotto il DL saranno dunque n-m .

La scelta del metodo più appropriato dipende dal grado di incompletezza del data set, dalle sue dimensioni e dalla distribuzione più idonea a rappresentare i campioni.

Inoltre, sempre il documento [OSWER 9285.6-10, EPA 2002] fornisce cinque raccomandazioni su come trattare un insieme di dati in cui siano presenti dei non-detects:

- ✓ I Detection Limits devono sempre essere specificati e i non-detects riportati con il valore osservato se possibile.
- ✓ I non-detects non devono mai essere riportati come valori zero senza specifiche giustificazioni.
- ✓ Se un'analisi condotta con un Bounding Method rivela che gli effetti quantitativi della presenza di non-detects nel data set è trascurabile non sono necessari ulteriori esami.
- ✓ Se si vuole procedere ad ulteriori analisi è consigliabile usare un metodo per una specifica distribuzione.
- ✓ Se la quantità dei non-detects nel data set è alta (>75%) oppure se il numero di campioni è basso (n<5) nessun metodo funzionerà bene. In tal caso si può riportare la percentuale di valori al di sotto del DL, ricorrere ancora ad un Bounding Method nel quale i non-detects saranno sostituiti dal DL nel calcolo dell'UCL, che sarà riportato come un numero probabile considerevolmente maggiore della media reale.

Il documento [RAGS/HHEM, EPA 1989, Volume 1] prevede la possibilità di rianalizzare i campioni cercando di riportare i dati con il loro valore esatto, di usare concentrazioni approssimate (pari al DL o alla metà) o di eliminare alcuni dei non-detected nel caso in

cui si abbiano delle informazioni che facciano pensare all'assenza di queste sostanze dal sito. Quest'ultima possibilità deve essere valutata con particolare attenzione, in quanto il Quantitation Limit potrebbe essere maggiore della concentrazione di riferimento di alcuni contaminanti (con la quale deve essere confrontata la concentrazione rappresentativa alla sorgente) e perciò l'eliminazione di alcuni dati può comportare una lacuna nell'analisi di rischio globale del sito. Se la concentrazione di un certo elemento chimico non è stata rilevata in nessun campionamento nel mezzo indagato questa sostanza generalmente viene esclusa dal data set, in modo da avere alla fine dell'analisi dei campioni raccolti un data set comprendente solo quelle sostanze di cui si possiede un valore di concentrazione in almeno un campione per ogni mezzo (aria,acqua,suolo) dell'area di interesse.

Nell'appendice A del [Concawe Report NO 2/1997] sono previsti diversi trattamenti per i valori non-detect, a seconda dell'analisi statistica è possibile:

- Porli pari a zero, anche se questa assunzione porterebbe ad una concentrazione media calcolata aritmeticamente minore di quella reale.
- Porli pari a ½ DL, se la distribuzione dei dati è normale (gaussiana).
- Porli pari a $1/\sqrt{2}$ DL,se la distribuzione dei dati è lognormale.

Il documento [D.Kofi Asante-Duah, 1993] suggerisce l'utilizzo di ½ DL per quei valori non rilevabili, ma cita anche il metodo consigliato nella guida EPA 1989, cioè quello di utilizzare proprio il DL, nel caso in cui vi sia motivo di ritenere che la concentrazione della sostanza considerata sia più vicina al DL piuttosto che alla sua metà.

Per l'applicazione delle presenti linee guida, seguendo il principio di cautela, si ritiene opportuno porre, in ogni caso e quindi in corrispondenza a qualsiasi distribuzione dell'insieme dei dati, i Non-Detect pari al corrispondente Detection Limit (ND = DL).

H.2 DISTRIBUZIONE DEI DATI

Quando si ha a che fare con dati ambientali (in particolare, concentrazioni di specie chimiche nei comparti ambientali: suolo, acqua, aria), le distribuzioni di probabilità più comunemente utilizzate per la loro rappresentazione sono:

- distribuzione gaussiana o normale
- distribuzione lognormale

- distribuzione gamma
- distribuzione non parametrica.

Nel seguito sono descritte sinteticamente le caratteristiche delle distribuzioni suddette (paragrafo H.2.1) e i test utili per identificare quale di queste distribuzioni rappresenti al meglio l'insieme di dati in esame (paragrafo H.2.2).

H.2.1 TIPI DI DISTRIBUZIONE DEI DATI

Distribuzione Gaussiana o normale– La distribuzione Gaussiana, o normale, è una distribuzione di tipo simmetrico la cui tendenza centrale è data dal calcolo della media aritmetica dei valori $x_1, x_2, x_3, ..., x_n$ 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.

Funzione $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} \left(x - \bar{x}\right)^2 \right]$$

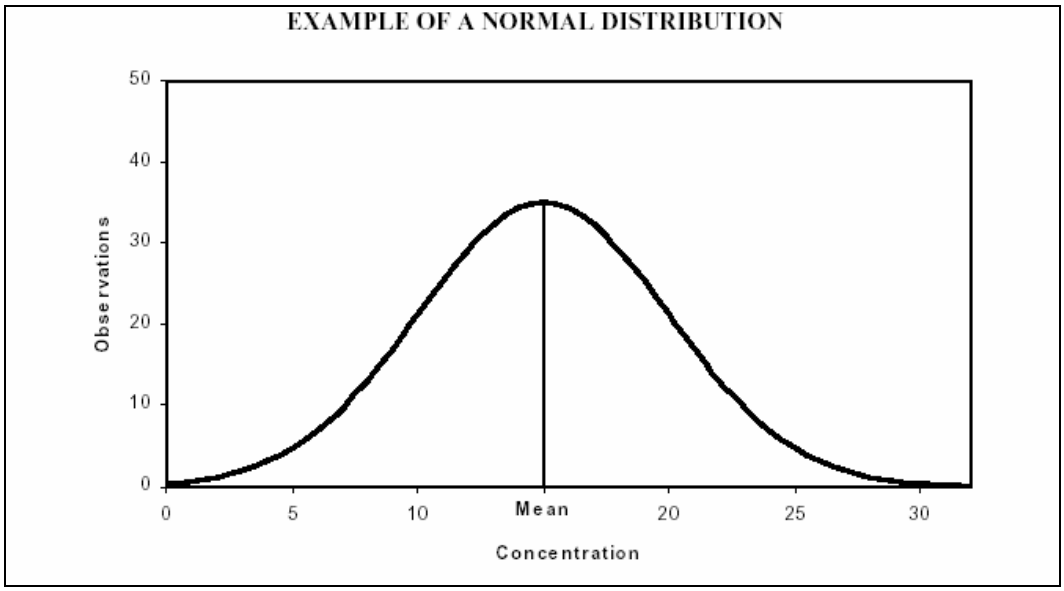
Media $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Varianza $$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

dove n è il numero di valori considerati.

In figura H.2 è riportato un esempio di distribuzione normale.

Fig. H.2 – Esempio di distribuzione normale



Distribuzione lognormale – La 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. Quindi, la distribuzione lognormale è caratterizzata da una asimmetria positiva (coda a destra) dovuta al fatto che ad un'elevata frequenza di valori bassi si associa una coda di valori molto meno frequenti ma, allo stesso tempo, molto elevati.

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

Funzione $$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma_y^2}\left(ln\, x - \bar{y}\right)^2 \right] \quad x > 0, \quad -\infty < \bar{y} < \infty,$$

$$\sigma_y > 0$$

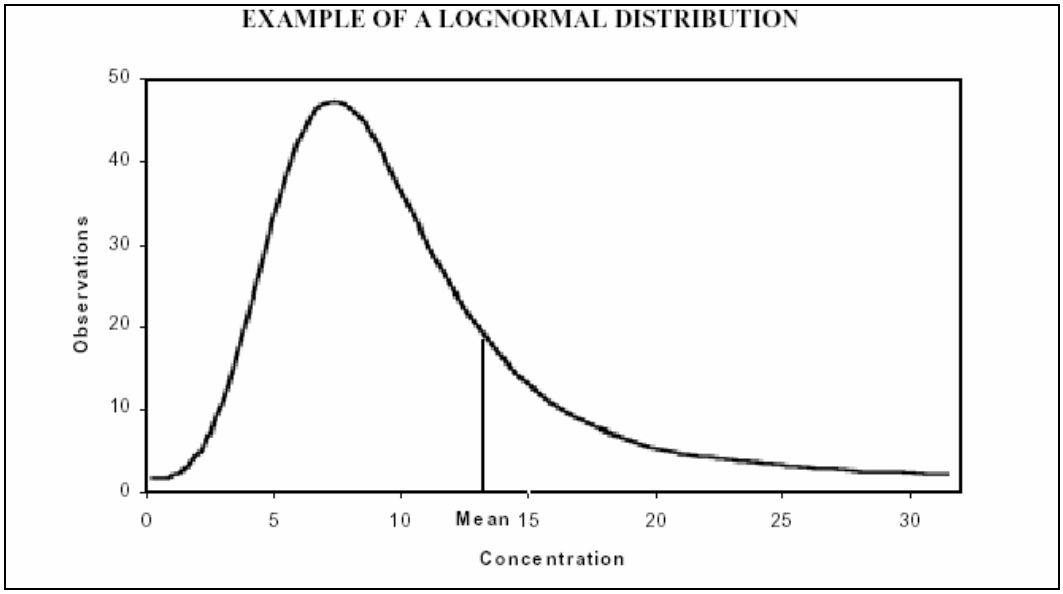
Media $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} ln\, x_i$$

Varianza $$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

dove n è il numero di valori considerati.

In figura H.3 è riportato un esempio di distribuzione lognormale.

Fig. H.3 – Esempio di distribuzione lognormale



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 $\bar{x}$.

Funzione $$f(x;k,\theta)=\frac{1}{\theta^k\Gamma(k)}x^{(k-1)}e^{(-x/\theta)}$$ x>0, k>0, θ>0

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.

Nel seguito sono descritte altre grandezze statistiche utili per lo studio del tipo di distribuzione dei dati.

Mediana - La mediana di una distribuzione è quel valore al di sopra del quale e al di sotto del quale giace metà dell'insieme dei dati.

La mediana si individua facilmente una volta ordinati in senso crescente gli n valori del data set :

- se n è pari la mediana sarà il valore $x_{[(n+1)/2]}$;
- se n è dispari la mediana sarà il valore $\frac{1}{2}\left(x_{n/2} + x_{[(n+2)/2]}\right)$.

Se la distribuzione è simmetrica, allora la mediana coincide con la media. Se la distribuzione dei dati è lognormale pendente verso destra la mediana sarà minore della media, e viceversa.

Coefficiente di skewness - Il valore di questo coefficiente fornisce una stima della asimmetria della forma di distribuzione dei dati. Si calcola secondo la seguente espressione:

$$asimmetria(\ skewness\ ) = \frac{1}{\sigma^3}\sum_i \frac{\left(x_i - \bar{x}\right)^3}{n}$$

Tale coefficiente può risultare:

- maggiore di zero: in tal caso la distribuzione avrà una coda verso destra;
- pari a zero: in tal caso la distribuzione sarà di tipo simmetrico, tipicamente gaussiana;
- minore di zero: in tal caso la distribuzione avrà una coda verso sinistra.

Il coefficiente di skewness non varia per traslazioni e cambiamenti di scala.

Coefficiente di curtosi - Il valore di questo coefficiente fornisce una stima della acutezza della curva di distribuzione dei dati. Si calcola secondo la seguente espressione:

$$curtosis = \frac{1}{\sigma^4}\sum_i \frac{\left(x_i - \bar{x}\right)^4}{n}$$

Tale coefficiente può risultare:

- maggiore di 3: in tal caso la curva avrà un picco che determinerà una forma aguzza;
- pari a 3: in tal caso la distribuzione sarà di tipo simmetrico,con la forma a campana tipicamente gaussiana;
- minore di 3: in tal caso la forma della curva sarà appiattita.

Il coefficiente di curtosi non varia per traslazioni e cambiamenti di scala.

Coefficiente di variazione – E' un indice di dispersione che permette di analizzare la dispersione dei valori attorno alla media indipendentemente dall'unità di misura, fornendo un'indicazione sulla variabilità delle osservazioni rilevate. E' definito come il rapporto tra la deviazione standard dell'insieme dei dati ed il valore assoluto della loro media aritmetica: $CV=\frac{\sigma}{|\bar{x}|}$.

In particolare:

- se CV=1 vuol dire che $\sigma = \bar{x}$ e la media $\bar{x}$ non è un indice corretto per la rappresentazione dei dati;
- se CV=0 vuol dire che $\sigma = 0$ e la media $\bar{x}$ è un indice appropriato per la rappresentazione dei dati;
- se CV>0.5 la media $\bar{x}$ non è un indice corretto;
- se $CV \leq 0.5$ la media $\bar{x}$ è un indice corretto.

H.2.2 TEST PER LA SELEZIONE DEL TIPO DI DISTRIBUZIONE

Per individuare quale distribuzione di probabilità approssimi meglio l'insieme di dati a disposizione, sono stati ideati diversi test statistici.

Nel seguito sono sinteticamente riportati i principali test statistici, per una trattazione di maggiore dettaglio si rimanda al riferimento bibliografico corrispondente.

- "Shapiro e Wilk test" ("W test")– Con questo test si può valutare se sussistono o meno le ipotesi di distribuzione normale o lognormale nei casi in cui il numero dei dati a disposizione sia inferiore a 50 (n < 50).
- "D'Agostino Test" – Con questo test si può valutare se sussistono o meno le ipotesi di distribuzione normale o lognormale nei casi in cui cui il numero dei dati a disposizione sia uguale o superiore a 50 ($n \geq 50$).
- "Normal Quantile-Quantile (Q-Q) Plot" – E' un test grafico la cui attendibilità, se non viene accompagnato da altri test più completi (come il "W test" o il "Lilliefors Test"), è piuttosto scarsa. E' tuttavia utile per avere una prima approssimativa idea sulla distribuzione che assumono i dati in caso di ipotesi di distribuzione normale o lognormale.

- "Lilliefors Test" – Viene utilizzato, nel caso di ampi data set (n > 1000), per verificare la normalità o la lognormalità di una distribuzione di dati.
- "Quantile-Quantile (Q-Q) Plot per distribuzioni gamma" – E' un test grafico la cui attendibilità, se non viene accompagnato da altri test più completi (come l'"Anderson Darling test" o il "Kolmogorov-Smirnov test"), è piuttosto scarsa. E' tuttavia utile per avere una prima approssimativa idea sulla distribuzione che assumono i dati in caso di ipotesi di distribuzione gamma.
- "Kolmogorov-Smirnov test" - Per l'applicazione di questo test non devono essere fatte assunzioni sul tipo di distribuzione dei dati. Lo stesso viene utilizzato per dimostrare che un certo data set segue la distribuzione ipotizzata, mediante il confronto tra un determinato parametro calcolato e il corrispondente valore critico tabellato. Nel software ProUCL ver. 3.0 tale test viene utilizzato solo nel caso di ipotesi di distribuzione gamma.
- "Anderson Darling test" - Questo test è simile al Kolmogorov –Smirnov test, ma più preciso, in quanto fa uso di una distribuzione specifica per il calcolo dei valori critici (diversi dunque per ogni tipo di distribuzione), con i quali verrà confrontato il parametro calcolato.

In tabella H.3 si ripota una sintesi dei test sopra elencati e il corrispondente riferimento bibliografico, da consultare per una trattazione di maggiore dettaglio.

Tabella H.3 – 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	---	

H.3 APPLICAZIONE DELLE PROCEDURE STATISTICHE PER IL CALCOLO DELLA CONCENTRAZIONE RAPPRESENTATIVA

Nel presente paragrafo sono descritte le procedura statistiche utili per il calcolo della concentrazione rappresentativa alla sorgente (C_s).

Per chiarezza di trattazione, nel primo paragrafo (paragrafo H.3.1) sono descritte le grandezze statistiche utili per la stima della C_s più comunemente proposte dai testi adottati quali riferimento. Successivamente si riportano i criteri indicati dalla bibliografia esaminata per la scelta di tali grandezze statistiche (paragrafo H.3.2) e le opzioni previste dai diversi software esaminati (paragrafo H.3.3). In ultimo, è riportata una sintesi dei contenuti dei due documenti [OSWER 9285.6-10, EPA 2002] [Manuale ProUCL ver. 3.0, EPA 2004], che trattano l'argomento in modo più specifico e approfondito rispetto alla restante documentazione disponibile (paragrafo H.3.4).

H.3.1 PRINCIPALI METODI DI STIMA DELLA C_s

I criteri di calcolo per la stima della concentrazione rappresentativa alla sorgente C_s, più comunemente proposti da testi bibliografici adottati quali riferimento, si riferiscono essenziamente alle seguenti grandezze statistiche:

- valore massimo;
- media aritmetica, per una distribuzione normale;
- media geometrica, pre una distribuzione lognormale;
- Upper Confidence Limit 95% (UCL 95%) per una distribuzione normale;
- Upper Confidence Limit 95% (UCL 95%) per una distribuzione lognormale;
- Percentile 95%.

Nel seguito è descritto il significato e il criterio di calcolo di tali grandezze statistiche.

- VALORE MASSIMO

  Il valore massimo è il valore più alto riscontrato in un insieme di dati.

  Questa rappresenta la scelta più conservativa in quanto definisce per l'intera area interessata dalla contaminazione una concentrazione pari al valore massimo riscontrato.

- MEDIA ARITMETICA

  La media aritmetica è un valore intermedio tra l'estremo superiore e quello inferiore di un data set, pari alla somma dei valori considerati divisa per il loro numero.

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(XXX)}$$

  dove n è il numero di campioni e x_i , nello specifico, è la concentrazione dello i-simo campione.

  Tale grandezza statistica può essere utilizzata per rappresentare un insieme di dati a cui corrisponde una distribuzione normale.

- MEDIA GEOMETRICA

  La media geometrica è un valore, compreso tra il dato quantitativamente maggiore e quello minore di una popolazione, pari alla radice n del prodotto dei valori osservati:

  $$\bar{x} = \left[x_1 \times x_2 \times \ldots \times x_n\right]^{1/n} = exp\left(\frac{1}{n}\sum_{i=1}^{n} ln\, x_i\right) \qquad \text{(XXX)}$$

  dove n è il numero di campioni e x_i , nello specifico, è la concentrazione dello i-simo campione.

  Tale grandezza statistica può essere utilizzata per rappresentare un insieme di dati a cui corrisponde una distribuzione log-normale.

- UPPER CONFIDENCE LIMITS (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 C_s, 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.

  Il calcolo dell'UCL della media varia a seconda del tipo di distribuzione dei dati.

  Saranno nel seguito riportate le relazioni che permettono il calcolo del UCL 95% nel caso rispettivamente di distribuzione dei dati normale e log-normale.

Calcolo dell'UCL per distribuzione normale – "Metodo della t di Student"

Se i dati $(x_1, x_2, x_3, ..., x_n)$ sono rappresentabili da una distribuzione normale, il metodo più utilizzato per il calcolo dell'$UCL_{(1-\alpha)}$ della media è quello della t di Student, valore tabellato che dipende dal grado di approssimazione richiesto, quindi da α (nella stima dell' l'UCL95% si ha che α=0.05), e dal numero di campioni disponibili.

I passaggi da seguire per il calcolo sono i seguenti:

Calcolo dell'UCL della media aritmetica – "Metodo della t di Student"

FASE 1 Si calcola la media aritmetica degli n dati: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

FASE 2 Si calcola la deviazione standard: $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

FASE 3 Si ricerca nell'apposita tabella [Tab. A2, Gilbert 1987] il valore della t di Student per il caso particolare, univocamente determinato una volta noti il valore di α e (n-1).

FASE 4 Si calcola l'$UCL_{(1-\alpha)}$ della media:

Calcolo dell'UCL per distribuzione lognormale – "Metodo Land"

Il classico approccio per il calcolo dell'$UCL_{(1-\alpha)}$ della media in una distribuzione di tipo lognormale è il cosiddetto "Metodo Land" basato sulla H statistica, valore tabellato individuabile una volta noto il numero di campioni e la deviazione standard del data set. Il parametro α indica il grado di approssimazione richiesto (nella stima dell' l'UCL95% si ha che α=0.05).

I risultati del metodo Land sono da considerare attendibili se si hanno a disposizioni un numero di dati con $n \geq 30$.

Spesso questo metodo può portare a sovrastime dell'UCL95%. In particolare, nel caso in cui la varianza sia piuttosto elevata, è addirittura possibile che l'UCL 95% della media geometrica superi il valore massimo degli n dati disponibili.

I passaggi da seguire per il calcolo sono i seguenti:

Calcolo dell'UCL della media geometrica – "Metodo Land"

FASE 1 Si calcola la media aritmetica della variabili trasformate $y = ln\, x$:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} ln\, x_i$$

FASE 2 Si calcola la deviazione standard associata:

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

FASE 3 Si ricerca nell'apposita tabella [Tab. A10-A12, Gilbert 1987] il valore della H statistica per il caso particolare, individuabile disponendo di n e della σ_y.

FASE 4 Si calcola l'$UCL_{(1-\alpha)}$ della media:

$$UCL_{(1-\alpha)} = exp\left(\bar{y} + 0{,}5\sigma_y^2 + H_{(1-\alpha,\sigma_y)}\sigma_y / \sqrt{n-1}\right)$$

Calcolo dell'UCL con il metodo della Disuguaglianza di Chebyshev

E' possibile calcolare l'$UCL_{(1-\alpha)}$ della media con il metodo non parametrico della Disuguaglianza di Chebyshev basato sulla media e sulla deviazione standard, sia per distribuzioni di tipo normale, sia di tipo lognormale che gamma (specialmente se $\sigma_y \geq 1.5$).

Per data set distribuiti in modo lognormale è consigliato [OSWER 9285.6-10, EPA 2002] utilizzare gli MVUE (Minimum-Variance Unbiased Estimators) della media e della varianza. L' MVUE di un parametro è definito come un valore statisticamente imparziale che presenta la varianza minore rispetto a tutte le stime alternative dello stesso parametro.

In tal modo si ottiene un valore di UCL più attendibile rispetto a quello ottenuto con il metodo Land. , anche se nei casi in cui n<10-50 e $\sigma_y > 1$ l'UCL95% della media così calcolato spesso non risulta opportuno per quantificare la Cs.

Il modello da seguire, per distribuzioni lognormali, per l'approccio basato sugli MVUE con il metodo della disuguaglianza di Chebyshev è il seguente:

Nei casi in cui ci sia alta varianza e basso numero di campionamenti è più indicato considerare come valore di concentrazione rappresentativo l'UCL99% anziché l'UCL95% se si sceglie di utilizzare il Chebyshev Inequality Method dopo aver riconosciuto la distribuzione più consona a rappresentare il data set del caso.

Calcolo dell'UCL della media geometrica

"Metodo della Disuguaglianza di Chebyshev" basato sugli MVUE

FASE 1: Si calcola la media aritmetica della variabili trasformate $y = ln\, x$:

$$\overline{\ln x} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i = \bar{y}$$

FASE 2: Si calcola la varianza associata: $$\sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\ln x_i - \bar{y}\right)^2$$

FASE 3: Si calcola l'MVUE della popolazione media per una distribuzione lognormale con la seguente formula, in cui g_n è una funzione tabellata [Aitchinson and Brown 1969,Tavola A2] [Koch and Link 1980, Tavola A7]:

$$\hat{\mu}_{LN} = \exp\left(\overline{\ln x}\right) g_n\left(\sigma_y^2 / 2\right)$$

FASE 4: Si calcola l'MVUE della varianza associata alla alla media $\hat{\mu}_{LN}$ secondo la formula $$\sigma_\mu^2 = \exp(2\ln x)\left\{\left[g_n\left(\sigma_y^2/2\right)\right]^2 - g_n\left(\frac{n-2}{n-1}\sigma_y^2\right)\right\}$$

FASE 5: Si può così calcolare l'$UCL_{(1-\alpha)}$ della media: $UCL_{(1-\alpha)} = \hat{\mu}_{LN} + \sqrt{\left(\frac{1}{\alpha} - 1\right)\sigma_\mu^2}$

Il parametro α indica il grado di approssimazione richiesto (nella stima dell' l'UCL95% si ha che α=0.05).

- 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%.

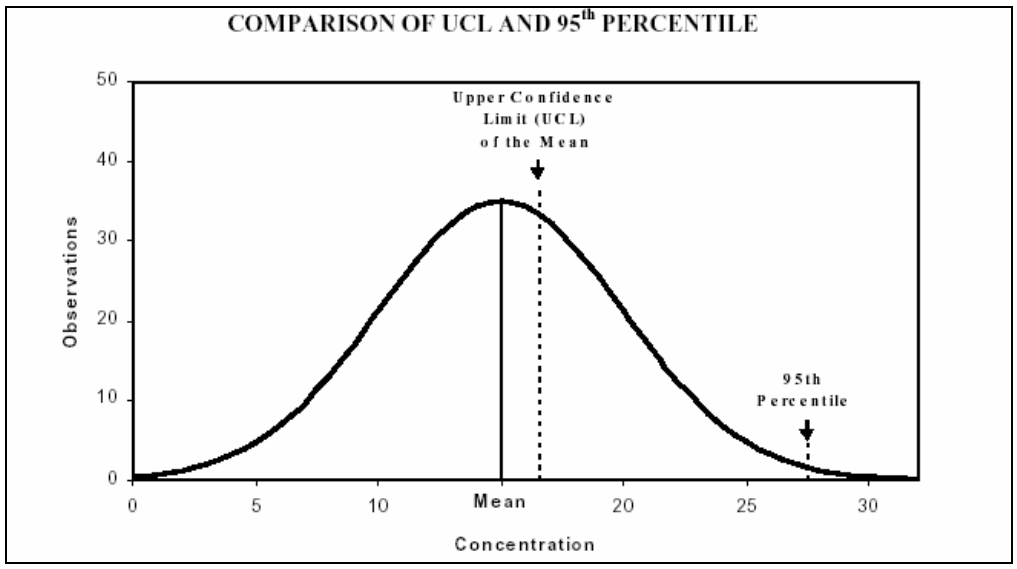
  Il percentile viene calcolato a mezzo della seguente espressione:

$\hat{x}_p = \bar{x} + Z_p \sigma$ per distribuzione normale

$\hat{x}_p = \exp\left(\bar{y} + Z_p \sigma_y\right)$ per distribuzione lognormale

dove $\bar{x}$ e σ sono rispettivamente la media aritmetica e la deviazione standard dei dati, $\bar{y}$ e σ_y quelle della variabile trasformata $y = \ln x$ e Z_p è un parametro tabellare [Tab. A1 Gilbert, 1987] che varia a seconda del percentile desiderato.
In Figura H.4 si riporta il confronto tra il valore dell'UCL e del percentile 95%. All'aumentare del numero di dati a disposizione, diminuisce l'incertezza legata alla stima di C_s , l'UCL si avvicina sempre di più alla media, mentre il percentile 95% si mantene sempre in corrispondenza della estremità inferiore della distribuzione.

Fig. H.4 – Confronto tra l'UCL e il percentile 95%



H.3.2 ANALISI DEI TESTI DI RIFERIMENTO

Nel presente paragrafo si riporta quanto proposto dalla bibliografia esaminata per la scelta delle grandezze statistiche utili per la stima della concentrazione rappresentativa alla sorgente (C_s).

Il documento [RAGS/HHEM, EPA 1989, Volume 1] raccomanda di utilizzare la concentrazione media quale valore della concentrazione rappresentativa alla sorgente (C_s).

Tale testo considera comunque, a seconda del numero di campioni rilevati, la possibilità dell'utilizzo dell'UCL per la determinazione di questo importante parametro, senza specificare però né la percentuale dell'UCL, né su quale distribuzione applicarlo.
In particolare, il numero di campionamenti deve essere stabilito, secondo la guida, prima dell'inizio delle indagini per l'analisi di rischio, in base alle dimensione dell'area di interesse, ai metodi statistici che si intendono impiegare nello studio, del livello di rappresentatività dei dati che saranno raccolti, da considerazioni pratiche di tipo logico ed economico e dal grado di errore che si è disposti ad ammettere nelle procedure statistiche. Per avere dunque bassi margini di errore le dimensioni del data set da raccogliere dovranno essere ampie, a meno che la variabilità tra i valori dei dati non sia estremamente ridotta. Se i dati a disposizione sono pochi e la loro variabilità è alta le difficoltà a cui si va incontro riguardano la definizione di un'appropriata distribuzione dei dati e l'incertezza dei risultati, per cui con grande probabilità i valori dell'UCL saranno di molto superiori a quelle della media. Per evitare che ciò si verifichi è necessario limitare la variabilità nelle aree di interesse, aumentando il numero di queste.

La guida [Supplemental Guidance to RAGS, EPA 1992] raccomanda, a causa della incertezza associata alla stima della concentrazione media effettiva di un sito, di adottare UCL 95% della media, evitando così di sottostimare tale valore. Per il calcolo dell'UCL95% della media il documento distingue due casi:

- se la distribuzione dei dati (da individuare con metodi grafici o con opportuni test, come il "W test") è di tipo normale, allora il metodo consigliato è il "Metodo della t di Student";
- se la distribuzione dei dati è di tipo lognormale, allora il metodo consigliato è quello della H statistica ("Metodo Land").

L'utilizzo dell'UCL95% della media per distribuzioni normali o lognormali viene sconsigliato nei casi in cui i dati siano pochi o la varianza elevata, in quanto il valore ottenuto potrebbe essere maggiore del massimo rilevato; in tal caso si assume quest'ultimo come concentrazione rappresentativa alla sorgente. In particolare, quando il numero di campioni è minore di 10 ($n<10$) la differenza tra la media effettiva e l'UCL95% della media potrebbe essere troppo grande; se $10 \leq n \leq 20$ il risultato sarà più attendibile; se infine $n > 20$ l'UCL95% tenderà quasi a coincidere con la media.

La guida prevede inoltre la possibilità di utilizzare anche la media aritmetica, mentre sconsiglia l'utilizzo della media geometrica, poiché tale grandezza statistica spesso comporta una sottostima della C_s.

La guida [EPA. 2000a, QA/G-4HW] propone il metodo cosiddetto DQO (finalizzato cioè a sviluppare Data Qualità Objectives, ossia gli obiettivi di qualità dei dati). La sua applicazione è utile per riuscire a raccogliere dati del giusto tipo, della opportuna qualità e nella giusta quantità, per supportare le decisioni sugli interventi più indicati per un sito contaminato. La finalità del DQO è dunque quella di permettere un approccio sistematico alla raccolta di dati sito-specifici, evidenziando i criteri che un data set deve soddisfare, quelli che debbono essere seguiti per effettuare il campionamento (periodi di tempo necessari, luoghi idonei…), le sue dimensioni e gli errori ammessi.
In riferimento alla selezione della C_s, la guida elenca varie opzioni di scelta disponibili, indicandone i limiti:

- La concentrazione massima, per concentrazioni nei punti cosiddetti "hot spots".
- La media aritmetica nei casi in cui non vi siano disomogeneità di rilievo all'interno del data set, né grandi quantità di non-detect.
- La mediana, più attendibile della media aritmetica se vi è presenza di dispersione dei dati e se le quantità di non-detect sia considerevole. Per utilizzare però la mediana nei test statistici è necessario un ampio data set.
- Il percentile 95% è consigliato in casi in cui sia presente un alto numero di non-detect e adottabile per test statistici solo se il numero di campioni è alto.

Il documento [D.Kofi Asante-Duah, 1993] ritiene che, nonostante sia spesso utilizzata la distribuzione Gaussiana o normale per descrivere l'ambiente, i dati di concentrazione dei composti chimici nei comparti aria, acqua e suolo, risultano meglio rappresentati considerando una distribuzione log-normale. Propone quindi per il calcolo della C_s l'utilizzo dell'UCL95% di una distribuzione lognormale. Il testo considera anche l'uso dell'UCL95% di una distribuzione normale, ma ritiene tale metodo il più delle volte esageratamente conservativo.

Il manuale UNICHIM 196/1 (2002) non effettua raccomandazioni specifiche ma, rifacendosi principalmente al documento [D.Kofi Asante-Duah,1993], porta in rassegna i possibili criteri statistici per il calcolo di C_s. In particolare riporta il possibile utilizzo di:

- Valore massimo, se si ha a disposizione un ristretto numero di dati (n<10).
- Media aritmetica, se la distribuzione è di tipo normale.
- Media geometrica, se la distribuzione è di tipo lognormale (sottolineando che questo tipo di distribuzione è più indicata quando si parla di concentrazioni di specie chimiche nel suolo).
- UCL95% ,considerando così il fattore di incertezza legato al calcolo di C_s e attenendosi al principio di conservatività. Nel dettaglio suggerisce per questo calcolo il "Metodo della t di Student" per distribuzioni normali e quello della H statistica ("Metodo Land") per distribuzioni lognormali.

Gli standard [ASTM RBCA-1739-95] e [PS-104-98] suggeriscono l'utilizzo della concentrazione massima come valore rappresentativo o dell'UCL se i dati sono in misura sufficiente per tale calcolo. Senza nessune ulteriori specifiche sull'argomento.

L'Appendice A del [Concawe Report NO 2/1997] suggerisce l'utilizzo dei seguenti valori:

- La media aritmetica, se la contaminazione risulta abbastanza omogenea nell'area di interesse.
- L'UCL 95% della media, a seguito dell'individuazione di una appropriato modello di distribuzione dei dati, se la variabilità dei campioni non è troppo accentuata e il data set ha dimensioni tali per cui l'UCL 95% non superi il valore massimo di concentrazione riscontrato, rendendo altrimenti necessari ulteriori campionamenti.
- Il valore massimo, in corrispondenza agli "hot spot" di contaminazione, poiché possono comportare fenomeni di tossicità acuta; quindi considerando come C_s la media, gli alti valori di concentrazione negli "hot spot" verrebbero attenuati nella valutazione globale.

La guida [Florida Dep. E.P.D., 2004] stabilisce dei criteri utili per la identificazione della C_s finalizzati al confronto della stessa con dei limiti di accettabilità tabellati. Secondo tale guida, nel caso in cui sia possibile applicare criteri statistici al set di dati a disposizione, il valore di concentrazione rappresentativo alla sorgente C_s deve essere assunto pari all'UCL 95% della media aritmetica o geometrica, a seconda della distribuzione dei dati.
In particolare, per il calcolo dell'UCL 95% devono essere soddisfatte due condizioni:

- il numero di valori di concentrazione a disposizione deve essere uguale o superiore a 10 ($n \geq 10$);
- almeno sette del numero totale di valori di concentrazione devono essere superiori al limite di rilevazione (detection limit) ($n.TOT - n.ND. \geq 7$).

Se non sono soddisfatte le due suddette condizioni, non risulta possibile individuare alcun tipo di distribuzione, quindi si assume C_s pari al valore massimo di concentrazione analiticamente rilevato (C_{MAX}).

Infine, il testo [Coleman e Steele, 1998] definisce il numero di misure di una variabile che vanno effettuate affinché il campione possa considerarsi rappresentativo, ovvero sufficiente per la determinazione della incertezza random di una singola variabile. Tale definizione viene fornita, con riferimento ad una distribuzione normale, verificando la numerosità del campione per la quale il valore della t di Student per un assegnato limite di confidenza può essere posto pari al valore della t di Student corrispondente ad un numero infinito di campioni. A seguito di simulazioni Monte Carlo, il testo raccomanda che tale numerosità sia pari a $N \geq 10$.

In tabella H.4 si riporta una sintesi dei criteri proposti dalla bibliografia esaminata per la stima della concentrazione rappresentativa alla sorgente (C_s).

Tabella H.4 – Testi di riferimento: calcolo della concentrazione rappresentativa alla sorgente

	RAGS/HHEM (EPA 1989)	Supplemental Guidance to RAGS (EPA 1992)	[EPA. 2000a, QA/G-4HW]	[D.Kofi Asante-Duah, 1993]	UNICHIM n.196/1 2002	ASTM E-1739-95 e PS 104-98	Concawe report n.2/97	[Florida Dep. E.P.D., 2004]
MASSIMO	x	x	x		x	x	x	x
MEDIA ARITMETICA	x	x	x		x		x	
MEDIA GEOMETRICA					x			
UCL95% DELLA MEDIA	x	x		x	x	x	x	x
PERCENTILE 95%			x					
MEDIANA			x					

A conclusione della analisi dei testi bibliografici di riferimento, è possibile osservare che per la stima di C_s vengono prese in considerazione due possibili distribuzioni dei dati, quella normale e quella lognormale, non contemplando quindi la possibilità di distribuzioni di tipo gamma o non parametriche. Tali due distribuzioni vengono invece ritenute utili per rappresentare dati di concentrazione dai documenti [OSWER 9285.6-10,

EPA 2002] e [Manuale ProUCL ver. 3.0, EPA 2004], oggetto quindi di una trattazione specifica (paragrafo H.3.4).

H.3.3 ANALISI DEI SOFTWARE

Nel presente paragrafo si riportano le possibili opzioni, previste dai quattro software esaminati, per la scelta delle grandezze statistiche utili per la stima della concentrazione rappresentativa alla sorgente (C_s).

Il software Giuditta ver. 3.0 prevede due possibili casi, uno semplice ed uno complesso, per la valutazione del rischio legato a siti contaminati.
Nel primo caso la concentrazione da utilizzare per tale calcolo corrisponderà al massimo, alla media aritmetica o al percentile 95% dei dati disponibili relativi alla sorgente di contaminazione.
Nel secondo caso potranno essere eseguite procedure statistiche per tutte le aree definite nelle diverse matrici ambientali: verranno segnalate, accanto al numero di campioni, le stime del valore massimo, del percentile 95%, della media aritmetica e dell'UCL della media aritmetica.
Il software consiglia:

- nel caso in cui il numero di campioni sia inferiore a 20 di considerare quale concentrazione rappresentativa alla sorgente il valore massimo osservato;
- se invece i dati a disposizione sono più di 20 il percentile 95% o l'UCL della media aritmetica, se questi seguono una distribuzione normale.

Il software Rome ver. 2.1 propone di utilizzare, in presenza di un numero esiguo di dati, la concentrazione massima di ogni contaminante riscontrata in sito. Altrimenti, propone di utilizzare le grandezze statistiche previste nel Manuale Unichim n. 196/1. Il software, comunque, non premette di effettuare automaticamente il calcolo di tali grandezze statistiche.

Il software RBCA Tool Kit ver. 1.2 permette di selezionare il criterio di stima della concentrazione rappresentativa alla sorgente. L'utente può selezionare il calcolo della media, sia aritmetica che geometrica, del massimo (consigliato nel caso di numero di campioni inferiore a 5) o dell'UCL con la percentuale richiesta.
Il software permette inoltre di determinare il tipo di distribuzione dei dati, se normale o lognormale, attraverso un test relativo al coefficiente di variazione.

Il software RISC ver. 4.0 prevede, nel caso sia disponibile più di un campione, l'utilizzo dei seguenti metodi per la stima della concentrazione rappresentativa alla sorgente da utilizzare come input nel processo di analisi:

- La media aritmetica, se la distribuzione dei dati è di tipo normale.
- La media geometrica, se la distribuzione dei dati è di tipo lognormale. Questa non può essere usata se sono presenti valori di concentrazione pari a zero, o se si sceglie di porre come tali i non-detect.
- L'UCL95% della media, sia per distribuzioni normali (in tal caso èconsigliato per il calcolo il "Metodo della t di Student"), sia per distribuzioni lognormali se il data set non contiene valori nulli (in tal caso è consigliato il "Metodo Land"). In particolare, l'utilizzo dell'UCL95% è suggerito nei casi in cui il numero di dati non sia esiguo, preferibilmente maggiore o uguale a 20-30, ma comunque non inferiore a 10.
- L'approccio "Weighting" che consiste nell'attribuire ad ogni campione inserito nel database del software un "fattore di peso"(weighting factor). Tale valore può corrispondere all'area del sito a cui il campione fa riferimento oppure al numero di campioni per cui è stata rilevata una determinata concentrazione. Nel caso in cui tutti i fattori di peso saranno pari a uno la concentrazione stimata utilizzando l'approccio "Weighting" corrisponderà alla media aritmetica.

Per i non-detect il software contempla la possibilità per l'utente di scegliere se porli pari al DL, a ½ DL o pari a zero.

In tabella H.5 si riporta una sintesi delle opzioni previste dai quattro software esaminati per la stima della concentrazione rappresentativa alla sorgente (C_s).

Tabella H.5 – Software esaminati: calcolo della concentrazione rappresentativa alla sorgente

	RBCA Tool Kit ver. 1.2	BP-RISC ver. 4.0	ROME ver. 2.1	GIUDITTA ver.3.0
MASSIMO	x		x	x
MEDIA ARITMETICA	x	x	x	x
MEDIA GEOMETRICA	x	x	x	
UCL95% DELLA MEDIA	x	x	x	x
PERCENTILE 95%				x
MEDIANA				

A conclusione della analisi dei quattro software, è possibile osservare che, come nella analisi dei testi bibliografici di riferimento, anche in questo caso per la stima di C_s vengono prese in considerazione due possibili distribuzioni dei dati, quella normale e quella lognormale, non contemplando quindi la possibilità di distribuzioni di tipo gamma o non parametriche.

H.3.4 DOCUMENTI "OSWER 9285.6-10 (EPA 2002)" E "MANUALE ProUCL ver. 3.0" (EPA 2004)

Come detto in precedenza, i due documenti, [OSWER 9285.6-10, EPA 2002] e [Manuale ProUCL ver. 3.0, EPA 2004], sono trattati separatamente rispetto alla restante documentazione bibliografica poiché forniscono una trattazione particolarmente specifica e dettagliata del calcolo della concentrazione rappresentativa alla sorgente (C_s).

Questi documenti sono scaricabili gratuitamente rispettivamente agli indirizzi web www.epa.gov/superfund/programs/risk/ragsa/ucl.pdf e www.epa.gov/esd/tsc/form.htm; in quest'ultimo sito è possibile anche scaricare gratuitamente il software ProUCL Version 3.0.

Entrambe i documenti contemplano la possibilità che il data set dei dati di concentrazione in input possa avere una distribuzione di tipo:

- normale
- lognormale

- gamma
- non-parametrica.

Quindi prevedono la necessità di effettuare test statistici che permettano di individuare la distribuzione di probabilità che approssimi meglio l'insieme dei dati disponibili e ne descrivono le modalità di applicazione.

Successivamente propongono una ampia gamma di procedure statistiche utili per il calcolo dell'UCL della media, che si differenziano in funzione del tipo di distribuzione individuata.

Per quanto riguarda i test statistici utili per la individuazione della distribuzione di probabilità, in tabella H.6 è riportato l'elenco di quelli contenuti nel software ProUCL ver. 3.0. Per una trattazione di maggiore dettaglio è possibile consultare il manuale d'uso del software.

Tabella H.6 – Test del software ProUCL ver. 3.0 per la selezione della distribuzione

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	---

In riferimento ai criteri di calcolo dell'UCL della media, il software ProUCL ver. 3.0 permette l'applicazione di 15 metodi di calcolo statistici, di cui 5 sono parametrici:

- Student's t UCL
- Approximate gamma UCL utilizzando l'approssimazione del chi-quadro
- Adjusted gamma UCL
- Land's H UCL
- UCL basato sulla disuguaglianza di Chebyshev, utilizzando i parametri MVUE di una distribuzione lognormale

e 10 sono non parametrici:

- UCL basato sul teorema del Limite Centrale

- UCL basato sulla modified-t
- UCL basato sul teorema del Limite Centrale "Adjusted"
- UCL basato sulla disuguaglianza di Chebyshev, utilizzando la media e la deviazione standard
- UCL basato su standard bootstrap
- UCL basato su percentile bootstrap
- UCL basato su bias-corrected accelerated bootstrap
- UCL basato sulla bootstrap-t
- UCL basato su Hall's bootstrap.

In particolare, a seconda del tipo di distribuzione corrispondente al data set esaminato, il software suggerisce il metodo più appropriato per il calcolo dell'UCL della media:

- Per distribuzioni normali consiglia il metodo della t di Student.

- Per distribuzioni lognormali distingue vari casi in funzione del numero di campioni e della deviazione standard della variabile trasformata $y = ln\, x$, secondo la tabella H.7.

Tabella H.7 – Calcolo dell'UCL per distribuzioni lognormali [software ProUCL ver. 3.0]

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

Se i valori di UCL calcolati con il metodo Hall's bootstrap risultano troppo alti è consigliato stimare l'UCL della media con il metodo della disuguaglianza di Chebyshev.

- Per distribuzioni gamma le varie possibilità di calcolo dell'UCL95%, individuate in funzione del fattore di forma k e del numero di campioni n, sono schematizzate nella tabella H.8.

Tabella H.8 – Calcolo dell'UCL per distribuzioni gamma [software ProUCL ver. 3.0]

k	Numero di campioni (n)	UCL consigliato
$k \geq 0.5$	per ogni n	Approximate gamma 95%UCL
$0.1 \leq k < 0.5$	per ogni n	Adjusted gamma 95%UCL
k<0.1	n<15	95%UCL basato sul metodo Bootstrap-t o Hall's Bootstrap
k<0.1	$n \geq 15$	Adjusted gamma 95%UCL se possibile, oppure Approximate gamma 95%UCL

E' necessario tuttavia sottolineare che i valori calcolati con i metodi Bootstrap-t o Hall's Bootstrap possono spesso risultare errati specialmente se il data set contiene degli outlier. In tali casi è preferibile utilizzare l'adjusted gamma UCL.

In riferimento alla distribuzione gamma, si ritiene opportuno sottolineare che i data set rappresentabili a mezzo di una distribuzione lognormale sono, in genere, rappresentabili anche a mezzo di una distribuzione gamma. Inoltre, calcolare l'UCL95% della media adottando una distribuzione gamma permette di ottenere valori spesso più rispondenti alla realtà di quelli ottenibili assumendo una distribuzione lognormale.

In particolare, nel caso in cui l'UCL95% calcolato con il metodo Land per un modello lognormale sia eccessivamente elevato e quindi inutilizzabile, è consigliato calcolarlo per il modello di tipo gamma. Mentre, se i dati non risultano rappresentabili da una distribuzione gamma, è più indicato calcolare un UCL95% con il metodo della Disuguaglianza di Chebyshev.

- Per distribuzioni non parametriche distingue vari casi in funzione del numero di campioni e della deviazione standard della variabile trasformata $y = ln\, x$, secondo la tabella H.9.

Tabella H.9 – Calcolo dell'UCL per distribuzioni non parametriche [software ProUCL ver. 3.0]

σ_y	**Numero di campioni (n)**	**UCL consigliato**
$\sigma_y \leq 0.5$	per ogni n	UCL calcolato con Student's t oppure Modified-t Statistic
$0.5 < \sigma_y \leq 1$	per ogni n	95% Chebyshev(Media,Dev.Standard)UCL
$1 < \sigma_y \leq 2$	n<50	99% Chebyshev(Media,Dev.Standard)UCL
	$n \geq 50$	97.5% Chebyshev(Media,Dev.Standard) UCL
$2 < \sigma_y \leq 3$	n<10	UCL calcolato con metodo Hall's bootstrap
	$n \geq 10$	99% Chebyshev(Media,Dev.Standard) UCL
$3 < \sigma_y \leq 3.5$	n<30	UCL calcolato con metodo Hall's bootstrap
	$n \geq 30$	99% Chebyshev(Media,Dev.Standard) UCL
$\sigma_y > 3.5$	n<100	UCL calcolato con metodo Hall's bootstrap
	$n \geq 100$	99% Chebyshev(Media,Dev.Standard) UCL

Se i valori di UCL calcolati con il metodo Hall's bootstrap risultano troppo alti è consigliato stimare l'UCL della media con il metodo della disuguaglianza di Chebyshev (UCL99% basato su media e deviazione standard).

Per una trattazione di maggiore dettaglio riguardo i criteri di calcolo dell'UCL per le diverse distribuzioni di dati è possibile consultare il manuale d'uso del software.

Infine, per quanto riguarda l'utilizzo del valore massimo della popolazione quale concentrazione rappresentativa, i due documenti ne sconsigliano l'utilizzo, a meno che la misura dell'UCL non superi il massimo stesso (ciò accade di frequente se ci si avvale del metodo Land), in quanto con questa scelta si ignorerebbero molte delle informazioni contenute nel data set.

H.4 APPLICAZIONE A DUE CASI STUDIO

I criteri individuati dalla analisi dei software e dei documenti utilizzati come riferimento bibliografico, riguardo la stima del valore di concentrazione rappresentativa alla sorgente di contaminazione, sono stati applicati a due casi studio. In particolare, dalla documentazione relativa al piano di caratterizzazione sono stati estrapolati i data set relativi ai valori di concentrazione analiticamente determinati a seguito della campagna di indagine diretta e per essi sono state applicate le seguenti fasi di analisi:

1 Suddividere il data-set di valori di concentrazione in funzione di ogni sorgente secondaria di contaminazione (SS, SP e GW).

2 Effettuare una valutazione dei dati

- 2.1 Esaminare l'ampiezza del data-set.
- 2.2 Verificare che il campionamento sia uniformemente distribuito su tutta la sorgente di contaminazione
- 2.3 Identificare gli outlier e distinguere i "veri outlier" dai "falsi outlier".
- 2.4 Identificare i Non-Detect.

3 Individuare la distribuzione di probabilità che approssimi meglio l'insieme dei dati disponibili.

4 Applicare la procedura statistica corrispondente al tipo di distribuzione riconosciuta

Nei seguenti paragrafi, rispettivamente per il caso 1 e per il caso 2, sono descritte le principali caratteristiche del data set e i risultati ottenuti dalla applicazione della suddetta procedura .

H.4.1 STIMA DELLA C_S : APPLICAZIONE AL CASO 1

1 Suddividere il data-set in funzione di ogni sorgente secondaria di contaminazione.

Il caso 1 è costituito da campionamenti appartenenti al solo comparto ambientale: suolo profondo. Tali campionamenti sono stati effettuati a diverse profondità che variano tra i 4,5 e i 23 m rispetto al p.c.. Le specie chimiche esaminate sono 42 e sono tutte di natura organica (fenoli non clorurati, una serie di idrocarburi policiclici aromatici, di idrocarburi leggeri e pesanti); ad ognuna di esse, per comodità, è stato fatto corrispondere un numero, riportato in tabella H.10.

Tabella H.10 – Caso 1: elenco delle specie chimiche contaminanti

SOSTANZA	NUMERO CORRISPONDENTE	SOSTANZA	NUMERO CORRISPONDENTE
Fenolo #35	1	**Idrocarburi C16#576**	24
Naphthalene#270	2	**Idrocarburi C17#650**	25
Methylnaphthalene #374	3	**Idrocarburi C18#721**	26
Acenaphthylene #518	4	**Idrocarburi C19#722**	27
Acenaphthene#546	5	**Idrocarburi C20#852**	28
Fluorene #625	6	**Idrocarburi C21#852**	29
Phenanthrene #785	7	**Idrocarburi C22#1080**	30
Anthracene #791	8	**Idrocarburi C23#1083**	31
Fluoranthene#985	9	**Idrocarburi C24#1081**	32
Pyrene#985	10	**Idrocarburi C25#1081**	33
Benzo(a)anthracene #1215	11	**Idrocarburi C26#1183**	34
Chrysene#1222	12	**Idrocarburi C27#1183**	35
Benzo(b or k)fluoranthene #1391	13	**Idrocarburi C28#1278**	36
Benzo(a)pyrene#1436	14	**Idrocarburi 29#1367**	37
Dibenzo(a,h)anthracene#1579	15	**Idrocarburi C30#1367**	38
Benzo(ghi)perylene #1588	16	**Idrocarburi C31#1367**	39
Indeno(1,2,3-cd)pyrene#1587	17	**Idrocarburi C32#1450**	40
Idrocarburi C10#57	18	**Idrocarburi C33#1531**	41
Idrocarburi C11#151	19	**Idrocarburi C34#1530**	42
Idrocarburi C12#242	20		
Idrocarburi C13#417	21		
Idrocarburi C14#417	22		
Idrocarburi C15#576	23		

2 Effettuare una valutazione dei dati

2.1 Esaminare l'ampiezza del data-set.

Per ogni specie chimica sono stati effettuati 37 campionamenti appartenenti, come gia detto, al comparto ambientale: suolo profondo.

2.2 Verificare che il campionamento sia uniformemente distribuito su tutta la sorgente di contaminazione

La campagna di indagine diretta è stata condotta adottando una modalità di campionamneto a griglia. I campioni prelevati sono quindi distribuiti uniformemente su tutta la sorgente di contaminazione.

2.3 Idetificare gli outlier e distinguere i "veri outlier" dai "falsi outlier".

Nel caso specifico, è stata omessa la fase di identificazione degli outlier. Tale scelta è dovuta al fatto che il data set a disposizione è stato già sottoposto a procedura di validazione, quindi è stata automaticamente esclusa la presenza di "veri outlier". I "falsi

outlier", di cui è stata constatata la presenza, come da procedura, non sono stati rimossi dal data set.

2.4 Idetificare i Non-Detect.

Dalla analisi dei valori di concentrazione delle diverse sostanze è stata accertata la presenza di no-detect. Gli stessi sono stati identificati e il loro numero, a seconda del contaminante, varia in parcentuale tra lo 0% (sostanze n. 23,24,25) e l'89% (sostanza n.4).

Il Detection Limit varia a seconda della specie chimica: in particolare è pari a 100 μg/Kg per il Fenolo, a 10 μg/Kg per gli Idrocarburi pesanti e leggeri e a 60 μg/Kg per i rimanenti contaminanti.

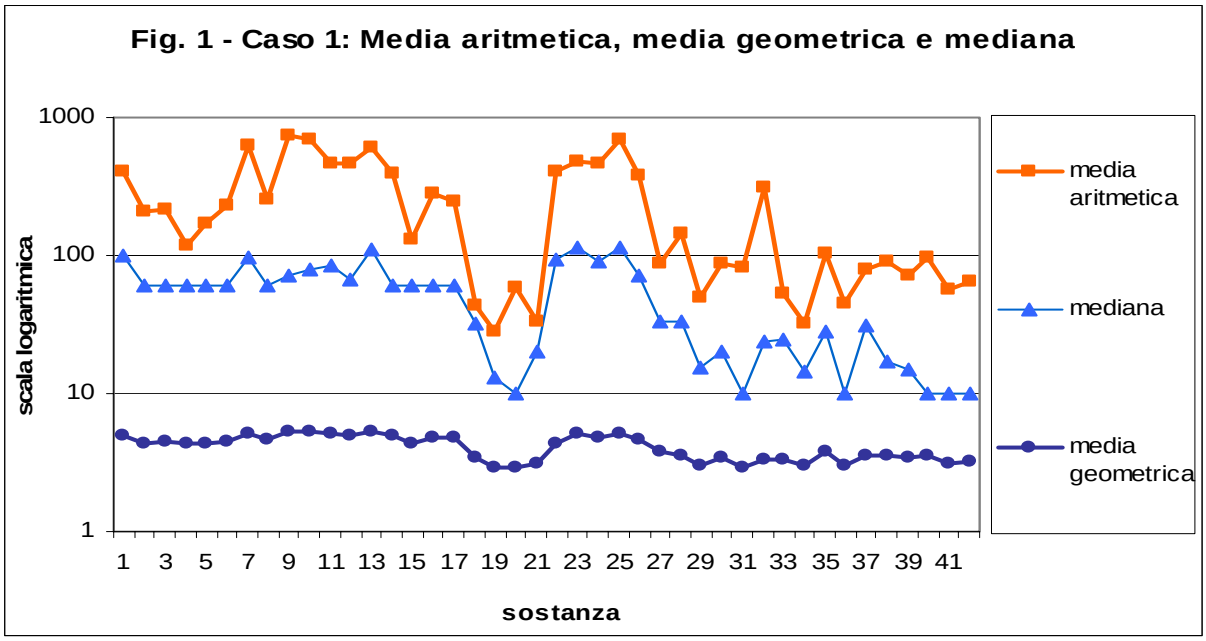
Applicando quanto previsto dalla procedura, per i no-detect è stato posto un valore pari al loro corrispondente detection limit.

3 Individuare la distribuzione di probabilità che approssimi meglio l'insieme dei dati disponibili.

In questa fase è stata individuata la distribuzione di probabilità corrispondente al data set in esame in corrispondenza ad ogni specie chimica inquinante mediante l'applicazione del software ProUCL ver. 3.0. Nel seguito si riportano i risultati ottenuti.

In figura H.5 si riporta un grafico che mostra l'andamento dei valori di media aritmetica, media geometrica e mediana per le 42 sostanze chimiche del data set del caso 1.

Figura H.5 – Caso 1: media aritmetica, media geometrica e mediana



E' possibile notare come la media aritmetica sia sempre maggiore della mediana, a sua volta nettamente superiore rispetto alla media geometrica. La selezione della media geomentrica, come anche della mediana, rappresenterebbe quindi una stima non affatto conservativa della concentrazione rappresentativa alla sorgente. Inoltre, la non coincidenza tra la media aritmetica e quella geometrica permette di affermare che la distribuzione dei dati, per tutte le specie chimiche, non è di tipo normale.

Dalla applicazione dei test statistici per la selezione della distribuzione dei dati, condotta mediante l'utilizzo del software ProUCL ver. 3.0, è emerso che:

- per 41 delle 42 sostanze esaminate, la distribuzione dei dati è di tipo non parametrico;
- per l'Idrocarburo C10 (sostanza n. 18), la distribuzione dei dati è di tipo gamma.

4 Applicare la procedura statistica corrispondente al tipo di distribuzione riconosciuta.

Per la selezione della procedura statistica utile per la stima del valore di concentrazione rappresentativa alla sorgente è stato applicato il software ProUCL ver. 3.0. Tale software fornisce come output i valori di concentrazione rappresentativi ottenuti mediante l'applicazione di tutti i metodi in esso implementati (cinque

parametrici e dieci non parametrici) e pone in evidenza quello ritenuto più indicato in funzione delle caratteristiche del caso specifico.

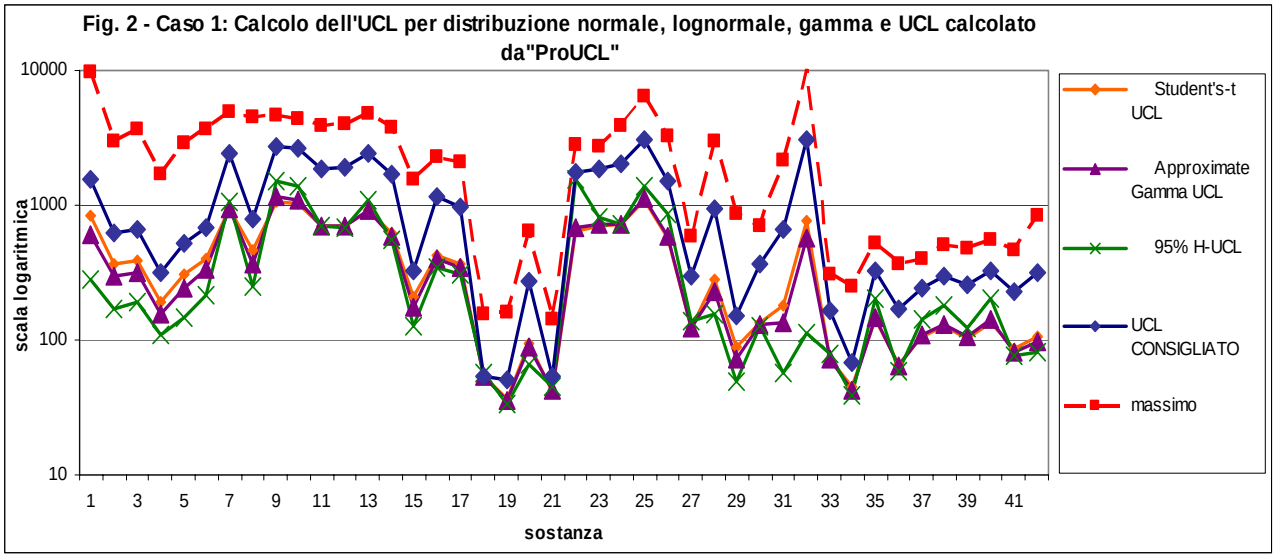
Da tale applicazione è emerso che:

- per le 41 sostanze alle quali corrisponde una distribuzione non parametrica, è stato selezionato come valore di concentrazione rappresentativa:
  - il 95% Chebyshev(Media,Dev.Standard)UCL, per $0.5 < \sigma_y \leq 1$;
  - il 99% Chebyshev(Media,Dev.Standard)UCL, per $1 < \sigma_y \leq 2$.
- Per la sostanza n. 18, alla quale corrisponde una distribuzione gamma, è stato selezionato l'Approximate gamma 95%UCL, poiché ad essa corrisponde un valore di k pari a 1,78 e quindi maggiore di 0,5.

Nel seguito si riportano i risultati ottenuti a seguito di una analisi comparativa condotta tra le possibili grandezze statistiche utilizzabili per stimare la concentrazione rappresentativa alla sorgente.

In figura H.6 è riportato il confronto tra i valori di UCL consigliati dal software ProUCL, il valore massimo e gli UCL corrispondenti ad altri tipi di distribuzione dei dati, in particolare: Student's t UCL per distribuzione normale, H UCL per distribuzione lognormale, Approximate Gamma UCL per una distribuzione di tipo gamma.

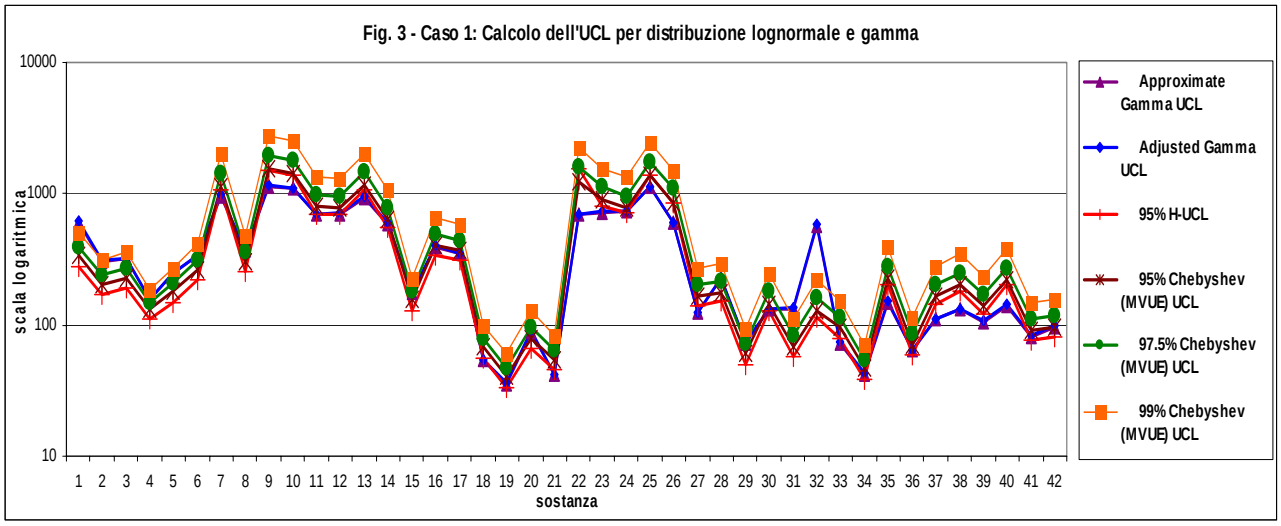
Figura H.6 – Caso 1: confronto tra i valori di UCL consigliati dal software ProUCl e quelli corrispondenti a una distribuzione normale, lognormanle e gamma.



Dalla analisi del grafico è possibile notare che l'UCL selezionato dal software segue l'andamento del massimo, senza mai superarlo, e risulta più conservativo sia rispetto all'UCL di una distribuzione normale, sia a quelli di distribuzioni lognormale o gamma. Inoltre, i valori dell'H UCL, per questo caso specifico, non risultano mai estremamente elevati (ciò avviene di frequente per alti valori di skewness e dimensioni del data set ridotte), bensì sono quasi sempre molto minori dell'UCL consigliato quale concentrazione rappresentativa.

Questa regolarità dei valori dell'H UCL è rintracciabile anche nella figura H.7, nella quale si può vedere come l'H UCL si sovrapponga quasi perfettamente all'UCL95% calcolato con il metodo di Chebyshev (MVUE).

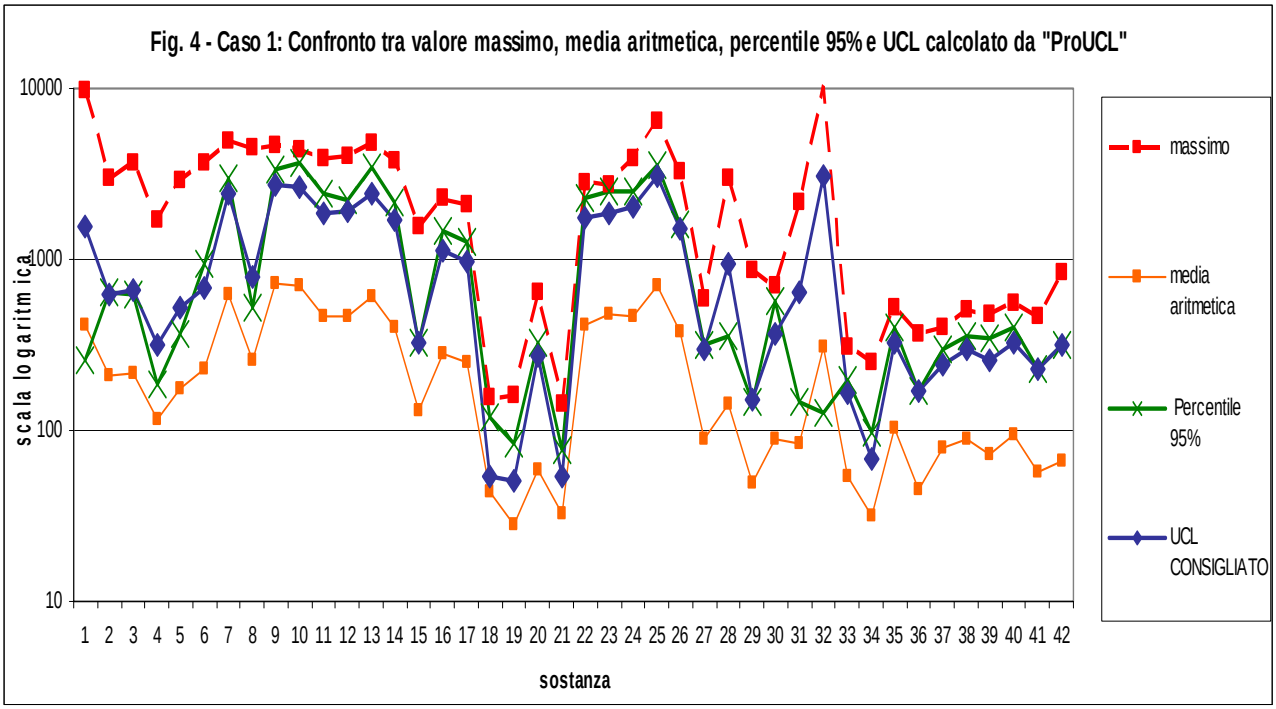
Figura H.7 – Caso 1: calcolo dell'UCL per distribuzioni lognormale e gamma



Il grafico mostra inoltre come gli UCL calcolati con i due metodi sopra citati disponibili per distribuzioni gamma (spesso addirittura coincidenti) e i quattro diversi UCL di distribuzioni lognormale abbiano un andamento molto simile.

In figura H.8 è riportato il confronto tra i valori del massimo, della media aritmetica del percentile 95% e dell'UCL suggerito dal software.

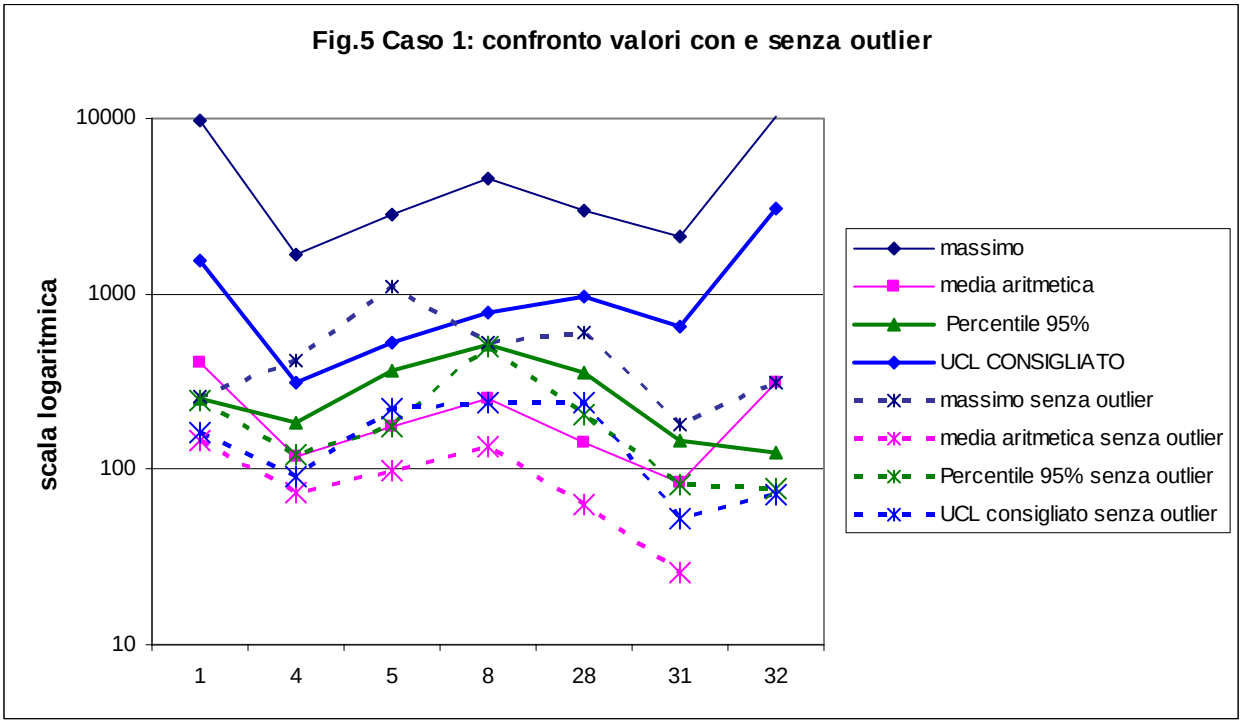
Figura H.8 – Caso 1: confronto tra valore massimo, media aritmetica, percentile 95% e UCL consigliato dal software ProUCL.



Dalla analisi del grafico è possibile evidenziare che l'UCL consigliato è sempre compreso tra il valore massimo e la media aritmetica ed in generale il suo andamento è analogo a quello del percentile 95% seppur spesso leggermente inferiore.

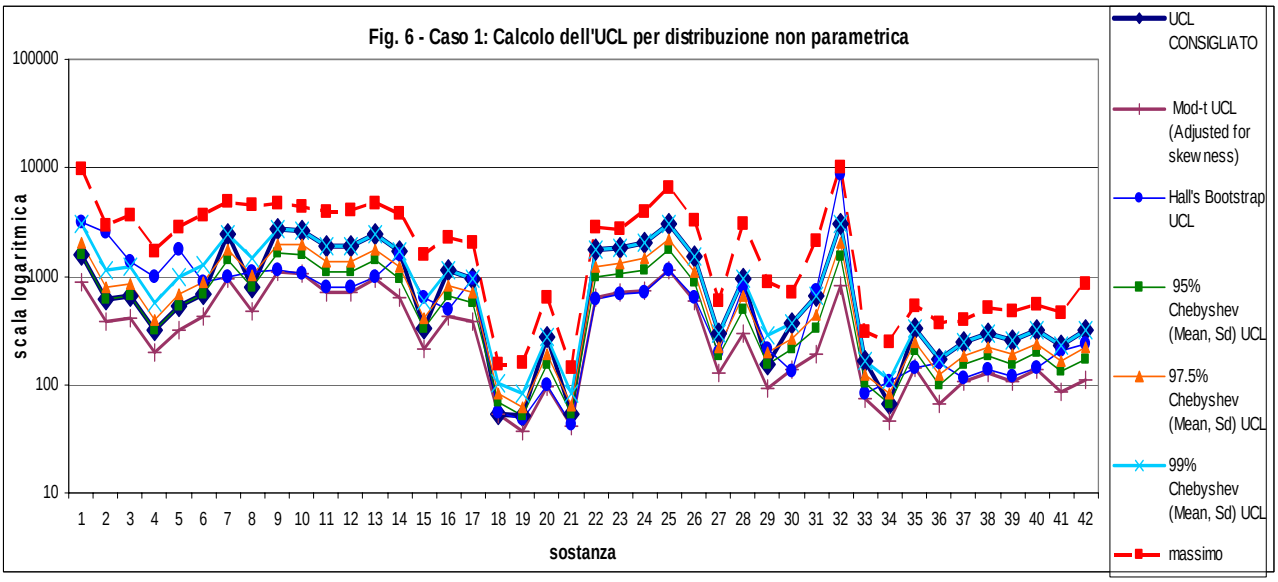
Questo però non risulta vero in corrispondenza ad alcuni punti, in particolare a quelli relativi alle sostanze numero 1, 31 e 32 e, seppur in misura minore, anche per le sostanze numero 4, 5, 8 e28. Per tali sostanze, infatti, i valori dell'UCL risultano essere superiori al percentile 95%. Analizzando i dati di concentrazione è emerso che solo in corrispondenza ad esse si ha la presenza di un unico valore identificabile come un outlier. Proprio la unicità dell'outlier comporta l'abbassamento del percentile 95% . Eseguendo i calcoli sul data set con e senza l'outlier relativo alle sostanze n.1, 4, 5, 28, 31 e 32 si vede come l'UCL consigliato calcolato sulla base del data set rivisto, risulti ancora una volta più o meno concorde con il nuovo percentile 95% (figura H.9).

Figura H.9 – Confronto dei valori con e senza outlier per le sostanze n.1, 4, 5, 28, 31 e 32



Infine, in figura H.10 sono riportati i valori dell'UCL calcolati con i principali metodi non parametrici e gli stessi sono posti a confronto con l'UCL consigliato.

Figura H.10 – Caso 1: confronto tra l'UCL consigliato e l'UCL applicando i restanti 9 metodi di stima per distribuzioni non parametriche



Dal grafico emerge che il ProUCL consiglia sempre, tra tutti i metodi non parametrici per il calcolo dell'UCL, il più conservativo, ed in particolare quello di Chebyshev per l'UCL

95% e lo stesso per l'UCL99%, a seconda che la deviazione standard della variabile trasformata $y = \ln x$ sia minore o maggiore di 1. L'unico metodo che fornisce in alcuni casi (ad esempio per le sostanze n. 2, 3, 4, 5, 15, 31, 32) valori maggiori a quelli dell'UCL consigliato è l'Hall's bootstrap: l'eventualità di ottenere da questo metodo misure inadatte per la stima di concentrazioni rappresentative (specialmente in presenza di outlier) è però presa in considerazione dalla guida del software ProUCL, che consiglia come alternativa l'UCL 99% calcolato con la disuguaglianza di Chebyshev per distribuzioni non parametriche.

A conclusione della analisi dei dati sopra esposta, è possibile affermare che i valori di UCL proposti dal software ProUCL ver. 3.0 rappresentano una stima attendibile e adeguatamente conservativa della concentrazione rappresentativa alla sorgente.

H.4.2 STIMA DELLA C_S : APPLICAZIONE AL CASO 2

1 Suddividere il data-set in funzione di ogni sorgente secondaria di contaminazione.

Il caso 2 è costituito da 21 campionamenti, suddivisibili in 6 per il suolo superficiale (campionamenti eseguiti fino a 1 metro di profondità dal p.c.), e 15 per il suolo profondo (campionamenti eseguiti da 5 a 20 metri di profondità dal p.c.). Le specie chimiche esaminate sono 7, in particolare:TPH, Benzene, Toluene, Etilbenzene, Xilene, Piombo e Benzine.

Il data set è stato quindi suddiviso in corrispondenza ai due comparti ambientali, e le successive fasi di analisi sono state condotte separatamente per i due sotto insiemi.

2 Effettuare una valutazione dei dati

2.1 Esaminare l'ampiezza del data-set.

Come gia detto in precedenza, per ognuna delle 7 specie chimiche sono stati effettuati 21 campionamenti, di cui:

- n. 6 per il suolo superficiale;
- n. 15 per il suolo profondo.

In corrispondenza al suolo superficiale il numero di dati a disposizione è quindi inferiore a 10. Poiché al di sotto di tale soglia, non è attendibile il risultato di alcuna stima

statistica, in accordo con la procedura proposta il valore di concentrazione rappresentativa alla sorgente è stato posto coincidente con il valore di concentrazione massimo analiticamente determinato ($C_s = C_{MAX}$).

2.2 Verificare che il campionamento sia uniformemente distribuito su tutta la sorgente di contaminazione

La campagna di indagine diretta è stata condotta adottando una modalità di campionamneto a griglia. I campioni prelevati sono quindi distribuiti uniformemente su tutta la sorgente di contaminazione.

2.3 Idetificare gli outlier e distinguere i "veri outlier" dai "falsi outlier".

Nel caso specifico, è stata omessa la fase di identificazione degli outlier. Tale scelta è dovuta al fatto che il data set a disposizione è stato già sottoposto a procedura di validazione, quindi è stata automatocamente esclusa la presenza di "veri outlier". I "falsi outlier", di cui è stata constatata la presenza, come da procedura, non sono stati rimossi dal data set.

2.4 Idetificare i Non-Detect.

Caso 2 - Suolo superficiale (SS)

Dalla analisi dei dati di concentrazione emerge che per le sostanze:

- Benzene, Toluene, Etilbenzene e Xilene, tutti i 6 campioni sono dei no-detect, tali composti sono quindi stati esclusi dalla analisi;
- Benzine, 5 dei 6 campioni sono dei no-detect;
- TPH e Piombo, non sono presenti valori no-detect.

Caso 2 - Suolo profondo (SP)

Dalla analisi dei dati di concentrazione emerge che per le sostanze:

- Benzene, Toluene, Etilbenzene e Xilene, 13 dei 15 campioni sono dei no-detect;
- Benzine, 11 dei 15 campioni sono dei no-detect;
- TPH e Piombo, non sono presenti valori no-detect.

In tabella H.11 si riportano, per le sostanze Benzene, Toluene, Etilbenzene, Xilene e Benzine, il corrispondente valore del Detection Limit, e la percentuale di valori no-detect per il suolo superficiale e profondo.

Tabella H.11 – Caso 2: analisi dei no-detect.

	DL (µg/Kg)	ND in SS+SP (%)	ND in SS (%)	ND in SP (%)
Benzene	0.05	90.5	100	86.66
Toluene	0.1	90.5	100	86.66
Etilbenzene	0.1	90.5	100	86.66
Xilene	0.1	90.5	100	86.66
Benzine	1	76.2	83	73.33

Applicando quanto previsto dalla procedura, per i no-detect è stato posto un valore pari al loro corrispondente detection limit.

3 Individuare la distribuzione di probabilità che approssimi meglio l'insieme dei dati disponibili.

Dalla applicazione dei test statistici per la selezione della distribuzione dei dati, condotta mediante l'utilizzo del software ProUCL ver. 3.0, è emerso che:

Caso 2 - Suolo superficiale (SS)

- per le Benzine, la distribuzione dei dati è di tipo non parametrico;
- per i TPH, la distribuzione dei dati è di tipo normale;
- per il Piombo, la distribuzione dei dati è di tipo gamma.

Caso 2 - Suolo profondo (SP)

- per i TPH, Benzene, Toluene, Etilbenzene, Xilene e Benzine, la distribuzione dei dati è di tipo non parametrico;
- per il Piombo, la distribuzione dei dati è di tipo normale.

4 Applicare la procedura statistica corrispondente al tipo di distribuzione riconosciuta.

Per la selezione della procedura statistica utile per la stima del valore di concentrazione rappresentativa alla sorgente è stato applicato il software ProUCL ver. 3.0. Tale software fornisce come output i valori di concentrazione rappresentativi ottenuti mediante l'applicazione di tutti i metodi in esso implementati (cinque parametrici e dieci non parametrici) e pone in evidenza quello ritenuto più indicato in funzione delle caratteristiche del caso specifico.

Da tale applicazione è è stato selezionato come valore di concentrazione rappresentativa:

Caso 2 - Suolo superficiale (SS)

- per le Benzine (distr. non parametrica), il 95% Chebyshev (Media,Dev.Standard) UCL;
- per i TPH (distr. normale), il t di Student UCL;
- per il Piombo (distr. gamma), l'Approximate gamma 95%UCL.

Caso 2 - Suolo profondo (SP)

- per i TPH, Benzene, Toluene, Etilbenzene, Xilene e Benzine (distr. non parametrica), il 99% Chebyshev(Media,Dev.Standard)UCL;
- per il Piombo (distr. normale), il t di Student UCL.

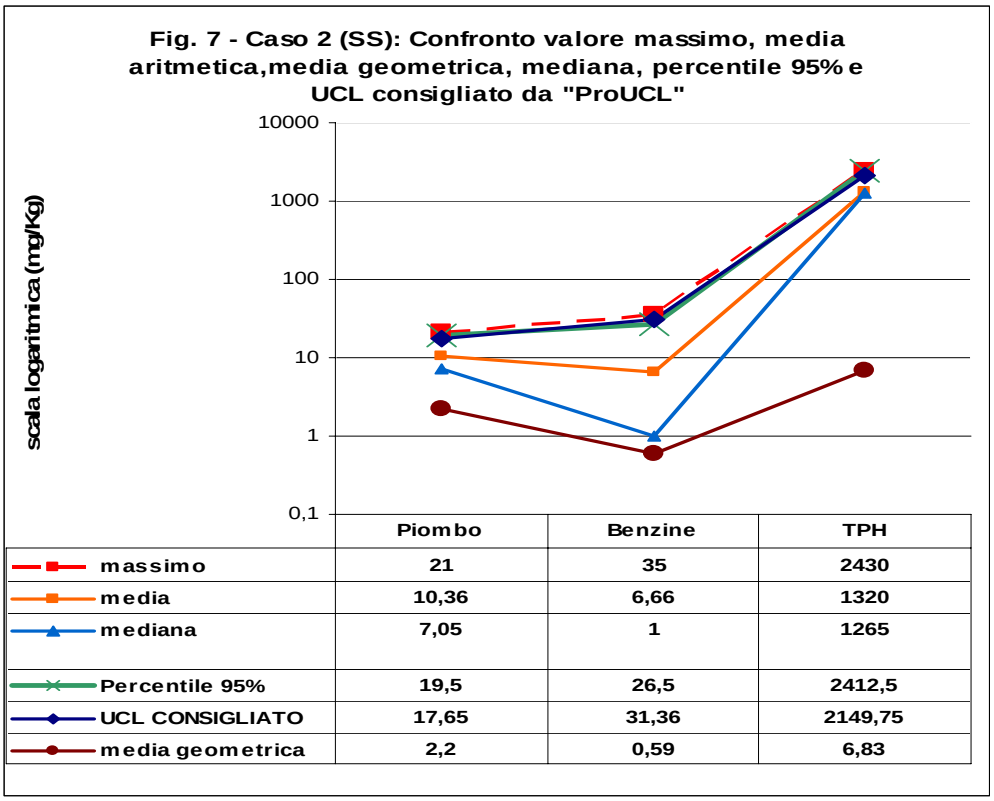
Nel seguito si riportano i risultati ottenuti a seguito di una analisi comparativa condotta tra le possibili grandezze statistiche utilizzabili per stimare la concentrazione rappresentativa alla sorgente.

Caso 2 - Suolo superficiale (SS)

Nonostante sia stato posto il valore di concentrazione rappresentativa alla sorgente coincidente con il valore di concentrazione massimo analiticamente determinato ($C_s = C_{MAX}$), è stata comunque effettuata una analisi dei dati di concentrazione assumibili come rappresentativi.

In figura H.11 si riporta un grafico che mostra l'andamento del valore massimo, della media aritmetica, del percentile 95% e dell'UCL consigliato dal software ProUCL per Piombo, Benzine e TPH.

Figura H.11 – Caso 2 (SS): confronto tra valore massimo, media aritmetica, percentile 95% e UCL consigliato dal software ProUCL



	Piombo	Benzine	TPH
massimo	21	35	2430
media	10,36	6,66	1320
mediana	7,05	1	1265
Percentile 95%	19,5	26,5	2412,5
UCL CONSIGLIATO	17,65	31,36	2149,75
media geometrica	2,2	0,59	6,83

Si può notare come l'UCL suggerito corrisponda più o meno precisamente al valore massimo e al percentile 95% e nettamente superiore alla media, sia aritmetica che geometrica e alla mediana, a dimostrazione del fatto che per n<10 (in questo caso n=6) il massimo sia il valore più appropriato da considerare per la concentrazione rappresentativa.

I seguenti grafici (figura H.12, H.13 e H.14) mostrano l'inadeguatezza dei valori elaborati attraverso criteri statistici in presenza di data set limitati.

Figura H.12 – Caso 2 (SS): calcolo dell'UCL per distribuzione normale e gamma

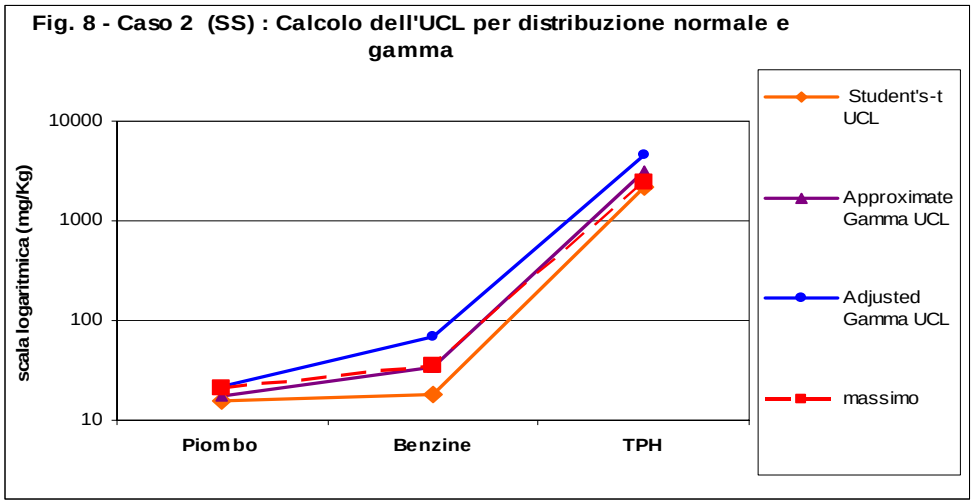


Figura H.13 – Caso 2 (SS): calcolo dell'UCL per distribuzione lognormale

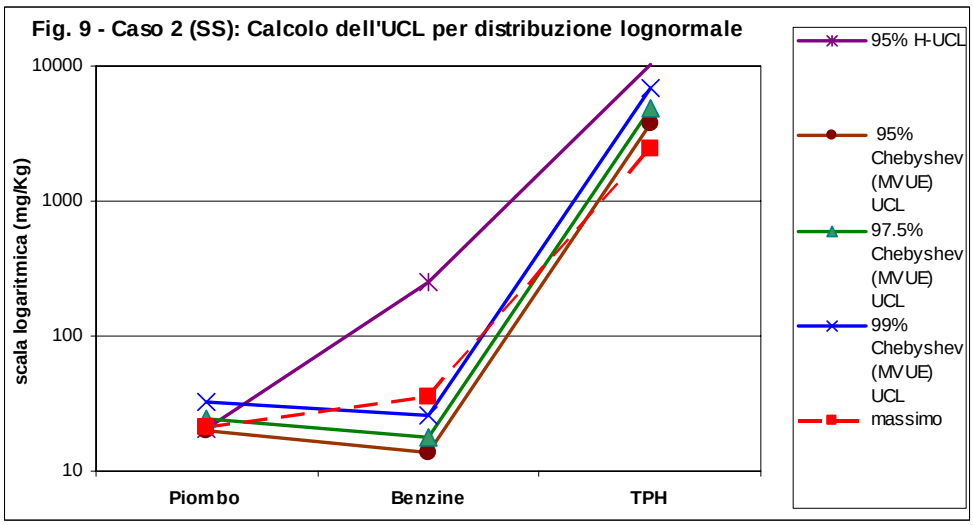
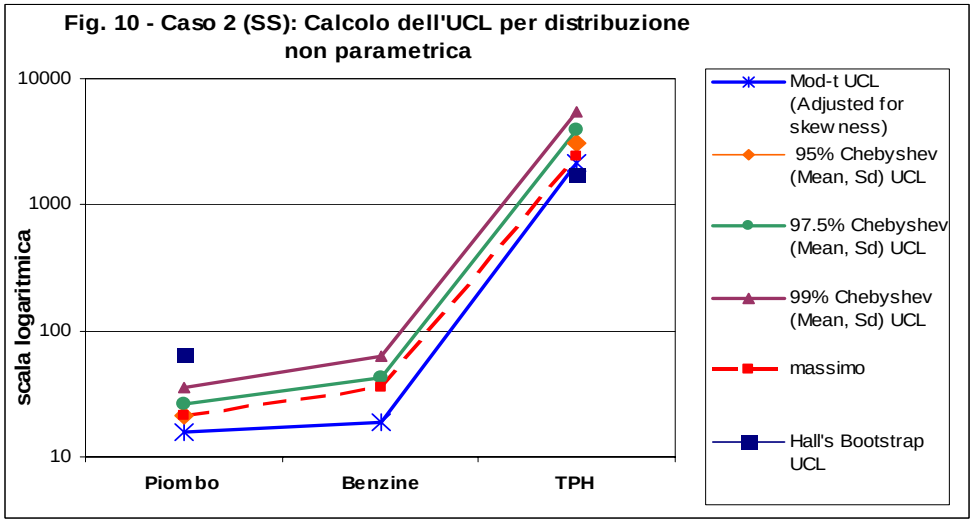


Figura H.14 – Caso 2 (SS): calcolo dell'UCL per distribuzione non parametrica



Nel caso di distribuzione ipotizzata normale i valori dell'UCL risultano inferiori a quelli del massimo, per gli altri tipi di distribuzione accade che, per il TPH, l'Approximate gamma UCL, l'Adjusted gamma UCL, l'H UCL e gli UCL calcolati con il metodo della disuguaglianza di Chebyshev per distribuzione lognormale e non parametrica (sia l'UCL al 95%, che quelli al 97.5% e al 99%) superino il valore massimo, a causa dello scarso numero di dati e dell'alta deviazione standard degli stessi.

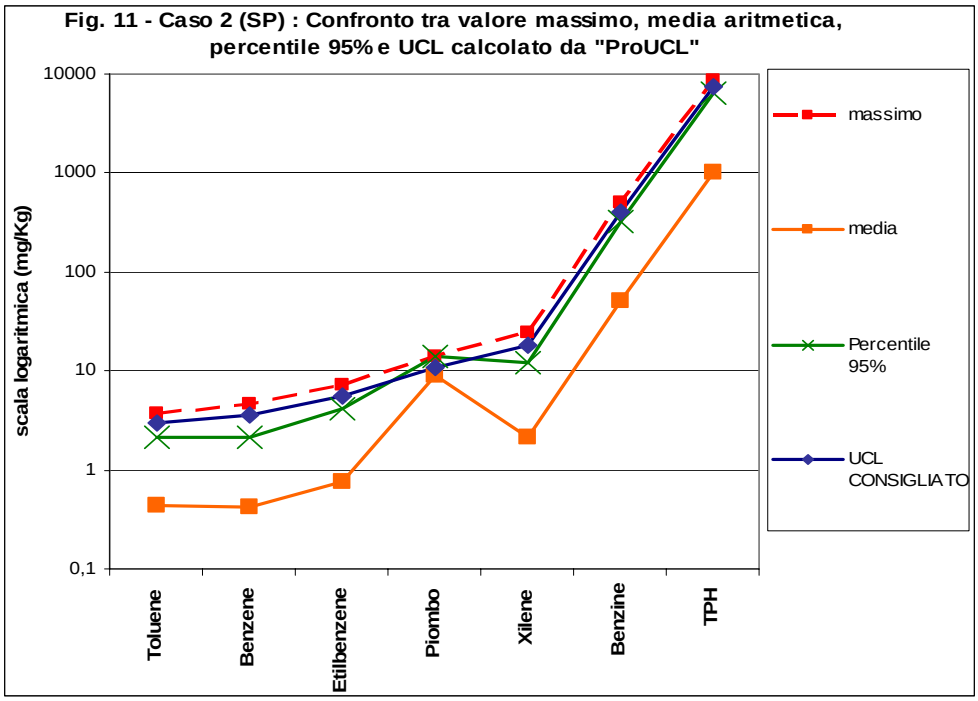
Anche per il Piombo e le Benzine si verificano casi in cui gli UCL siano maggiori del valore massimo, e quindi senza significato statistico, soprattutto se calcolati con i metodi della H statistica (metodo Land) e con quello non parametrico degli Hall's Bootstrap.

Risulta dunque inappropriata, a causa dei risultati poco o affatto significativi, l'applicazione di qualsiasi procedura statistica per la stima di parametri rappresentativi per data set con $n \leq 10$, che devono essere presi con il valore più conservativo tra quelli riscontrati, nel caso di concentrazione alla sorgente pari al valore massimo.

Caso 2 - Suolo profondo (SP)

In figura H.15 sono posti a confronto il valore massimo, la media aritmetica, il percentile 95% e l'UCL consigliato dal software ProUCL, per le sostanze rilevate, nel suolo profondo nei 15 campionamenti effettuati.

Figura H.15 - Caso 2 (SP): confronto tra valore massimo, media aritmetica, percentile 95% e UCL consigliato dal software ProUCL

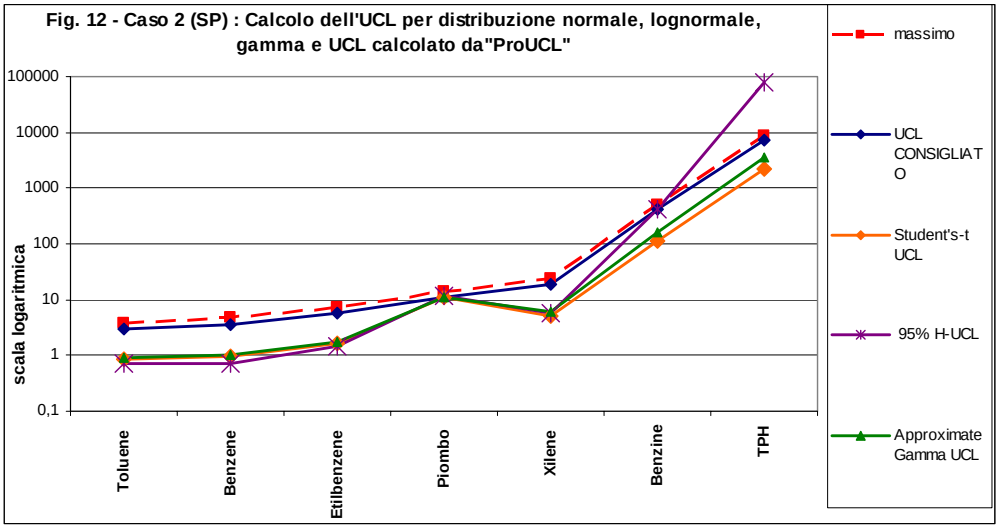


Si può osservare che, per il piombo, il valore massimo, la media aritmetica, il percentile 95% e l'UCL consigliato dal software sono molto più vicini tra loro di quelli degli altri contaminanti.

Per questi ultimi, l'UCL suggerito (il 99%UCL calcolato con la disuguaglianza di Chebyshev) risulta più conservativo del percentile 95%, maggiore della media aritmetica e molto prossimo, ma mai superiore, al valore massimo.

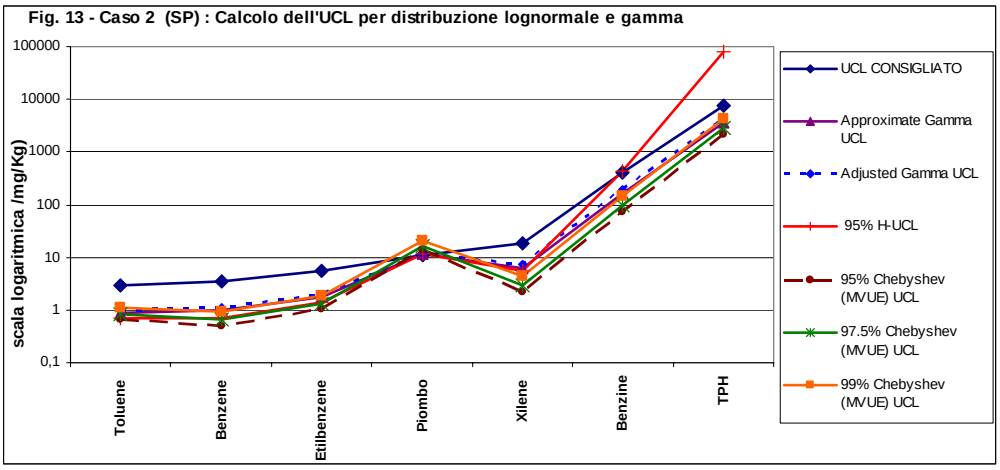
In figura H.16 è riportato l'andamento degli UCL calcolati assumendo tutte le distribuzioni citate dai principali testi esaminati e i metodi di calcolo più rappresentativi: quello della t di Student per distribuzione normale, il metodo Land per distribuzioni lognormale e quello per il calcolo dell'Approximate gamma UCL.

Figura H.16 – Caso 2 (SP): calcolo dell'UCL per distribuzione normale, lognormale e gamma



Si vede come, anche per il suolo profondo, lo scarso numero di dati, su cui basarsi per la stima della concentrazione rappresentativa, porti a valori di H UCL non realistici, superiori al valore massimo nel caso del TPH, o che comunque a volte non seguono l'andamento generale degli altri UCL (quello degli UCL per distribuzioni normale e gamma invece sono molto simili per tutte le sostanze esaminate). Questo risulta anche dalla figura successiva (figura H.17), che riporta gli UCL calcolati con tutti i metodi disponibili assumendo distribuzioni lognormale e gamma.

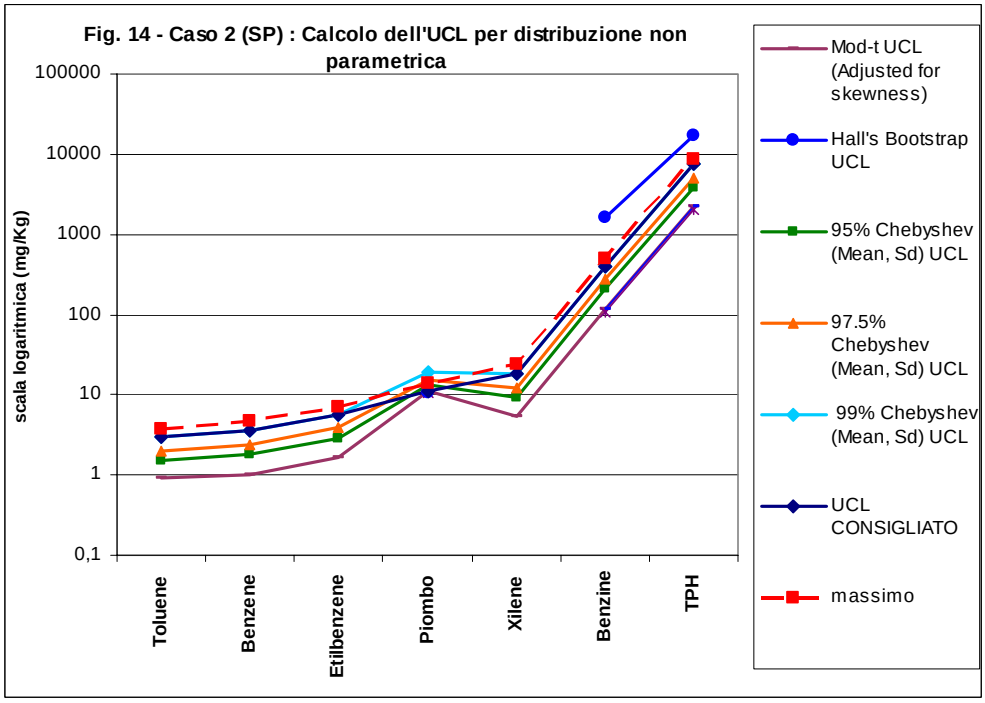
Figura H.17 – Caso 2 (SP): calcolo dell'UCL per distribuzione normale e gamma



Si vede inoltre che per il piombo, i cui dati seguono una distribuzione di tipo normale, gli UCL calcolati con la disuguaglianza di Chebyshev ipotizzando una distribuzione lognormale, sono più conservativi dell'UCL basato sulla t di Student, consigliato dal ProUCL e suggerito da tutti i testi esaminati per distribuzioni di questo tipo.

Infine si riporta il grafico (figura H.18) relativo ai metodi di calcolo dell'UCL per distribuzioni non parametriche, osservando ancora una volta che l'Hall's bootstrap UCL non risulta idoneo a rappresentare la concentrazione alla sorgente, in quanto maggiore del valore massimo (Benzine, TPH).

Figura H.18 – Caso 2 (SP): calcolo dell'UCL per distribuzione non parametrica



H.4.3 ANALISI DEI NO-DETECT : APPLICAZIONE AL CASO 2 (SP)

Per quanto riguarda i no-detect presenti nel data set relativo al Caso 2 – Suolo Profondo, si riportano i risultati ottenuti ponendo il valore dei no-detect rispettivamente pari al DL, a ½ DL o a $\frac{1}{\sqrt{2}}$ DL.

In sintesi, l'assunzione di questi valori non comporta sostanziali variazioni nei parametri stimati nell'analisi dei campioni. Nel seguito si riportano gli andamenti della media aritmetica e dell'UCL consigliato, in quanto potenzialmente sensibili alla variazione dei valori no-detect, contrariamente al massimo e al percentile 95%. (figura H.19 e H.20).

Figura H.19 – Caso 2 (SP): calcolo della media al variare del valore dei no-detect

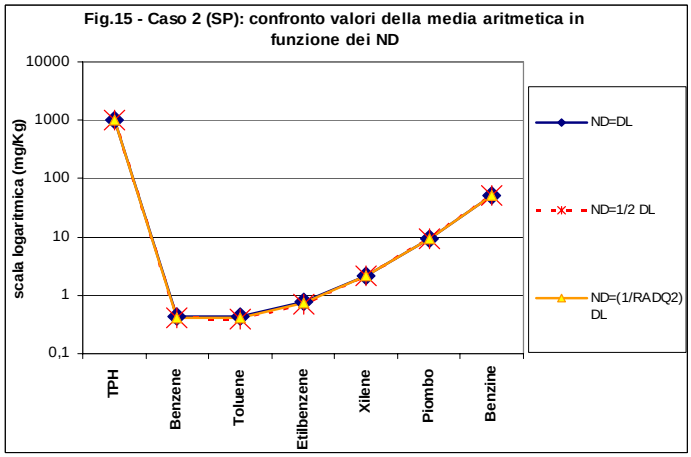


Figura H.20 – Caso 2 (SP): calcolo dell'UCL al variare del valore dei no-detect