Approfondimenti di statistica e geostatistica

Ing. Antonella Vecchio, Dr. Marco Falconi

APAT
Agenzia per la Protezione dell'Ambiente e per i Servizi Tecnici

Geostatistica

Cos'è la geostatistica?

La Geostatistica studia i **fenomeni naturali** che si sviluppano su **base spaziale** a partire dalle informazioni derivanti da un loro campionamento. In particolare studia **la variabilità spaziale dei parametri** che descrivono tali fenomeni.

Definizioni

Variabile regionalizzata (VR): è una grandezza espressa come una funzione numerica z(**x**) il cui valore dipende dalla localizzazione ovvero dal vettore **x** (x, y) delle coordinate spaziali.

Campo: è il dominio all'interno del quale si studia la variabilità della variabile z.

Supporto: è l'entità geometrica sulla quale vengono misurati i valori della variabile z. Quando le dimensioni sono molto piccole (rispetto alla scala del lavoro) il supporto può considerarsi puntuale.

Analisi spaziale – Approccio probabilistico

Perché un approccio probabilistico?

Necessità di passare da dati puntuali (misurati) ad un dato omogeneo e continuo nel dominio di studio.

Variabile aleatoria (VA): $Z(\mathbf{x_0})$ rappresenta l'insieme dei valori che può assumere la variabile regionalizzata $z(\mathbf{x})$ nel punto $\mathbf{x_0}$ (x_0, y_0) del dominio di studio, ovvero è una variabile che assume dei valori numerici appartenenti ad un certo intervallo secondo una legge di densità di probabilità $f_0(Z)$.

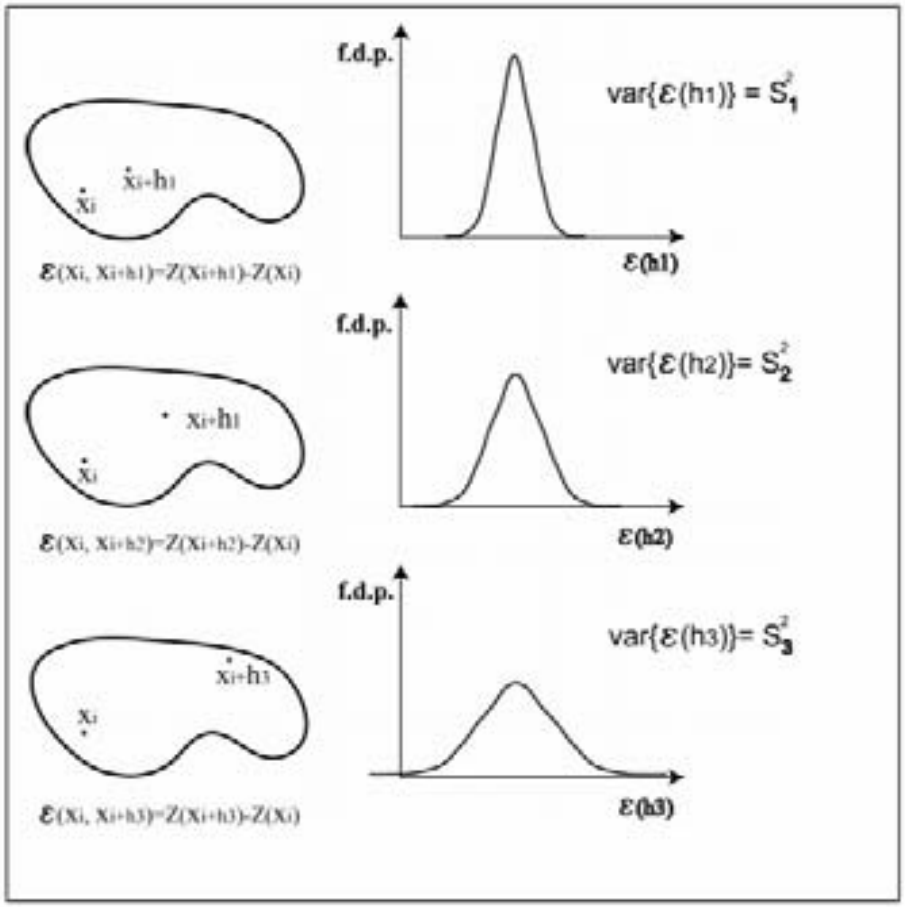
Funzione aleatoria (FA): FA $Z(\mathbf{x})$ è l'insieme di tutte le variabili aleatorie $Z(\mathbf{x})$ per ogni punto **x** (x, y) del dominio di studio, ovvero l'insieme di tutti i valori che può assumere la variabile regionalizzata $z(\mathbf{x})$ nel dominio di studio.

Approccio Probabilistico

Valori misurati di z(x) ⇒ **Funzione aleatoria FA Z(x)**

Funzione aleatoria FA Z(x) ⇒ **Stima di z(x) nell'intero dominio**

Il Variogramma



Incrementi:

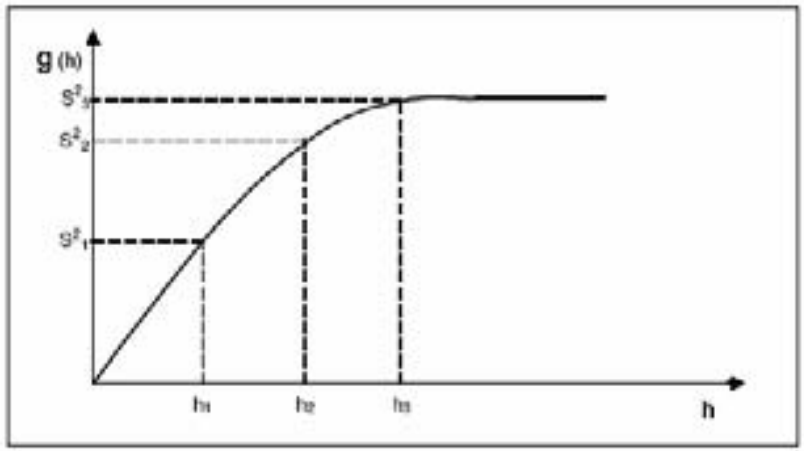
$\varepsilon(\mathbf{h}) = Z(\mathbf{x}_i+\mathbf{h}) - Z(\mathbf{x}_i)$

Esprimono la variazione della variabile spaziale con la posizione ovvero al variare del vettore **h**

Dispersione (varianza) degli incrementi : aumenta al crescere di **h** e quindi decresce con la distanza l'influenza della distanza stessa sulla variazione di Z(**x**)

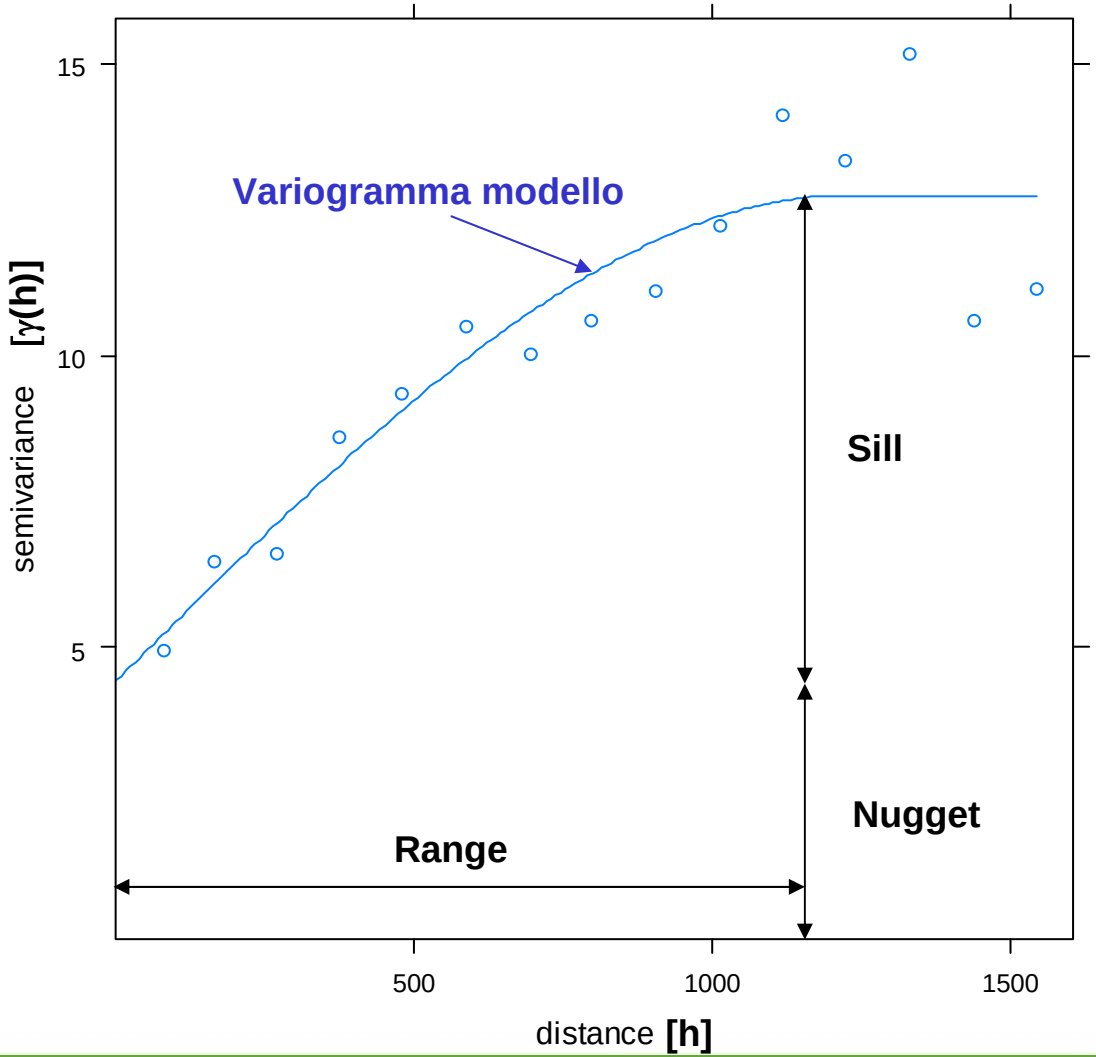
Variogramma: Esprime la correlazione (varianza) di Z(**x**) con la distanza **h**

$$\gamma(\mathbf{h}) = \tfrac{1}{2}\ \mathrm{var}\,[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})]$$



Elementi del Variogramma

Variogramma sperimentale



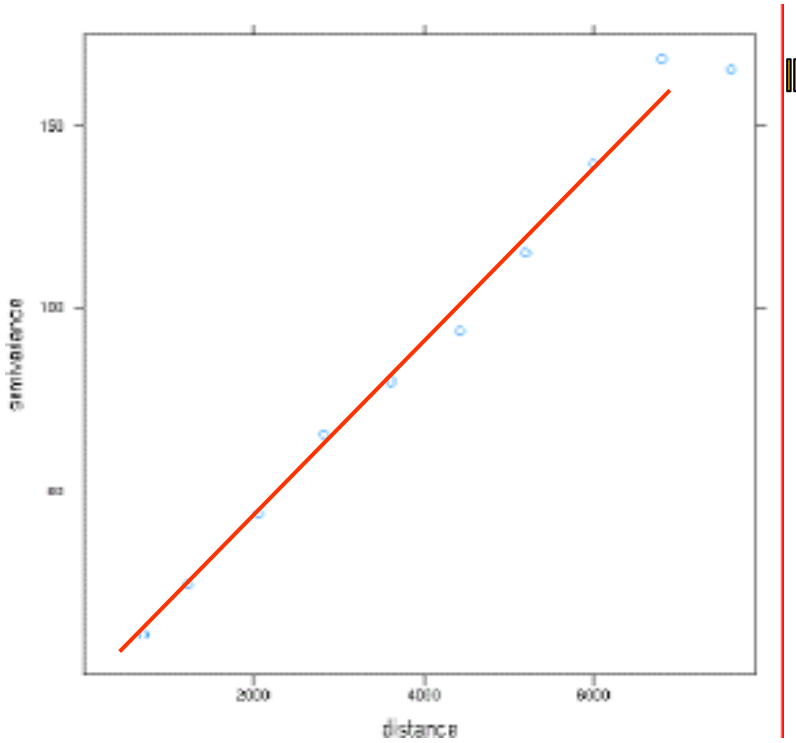
$$\gamma(h) = \sum_{i=1}^{n(h)} \frac{[z(x+h) - z(x)]^2}{n(h)}$$

Nugget: variabilità casuale non correlata alla distanza (es. errori di misura)

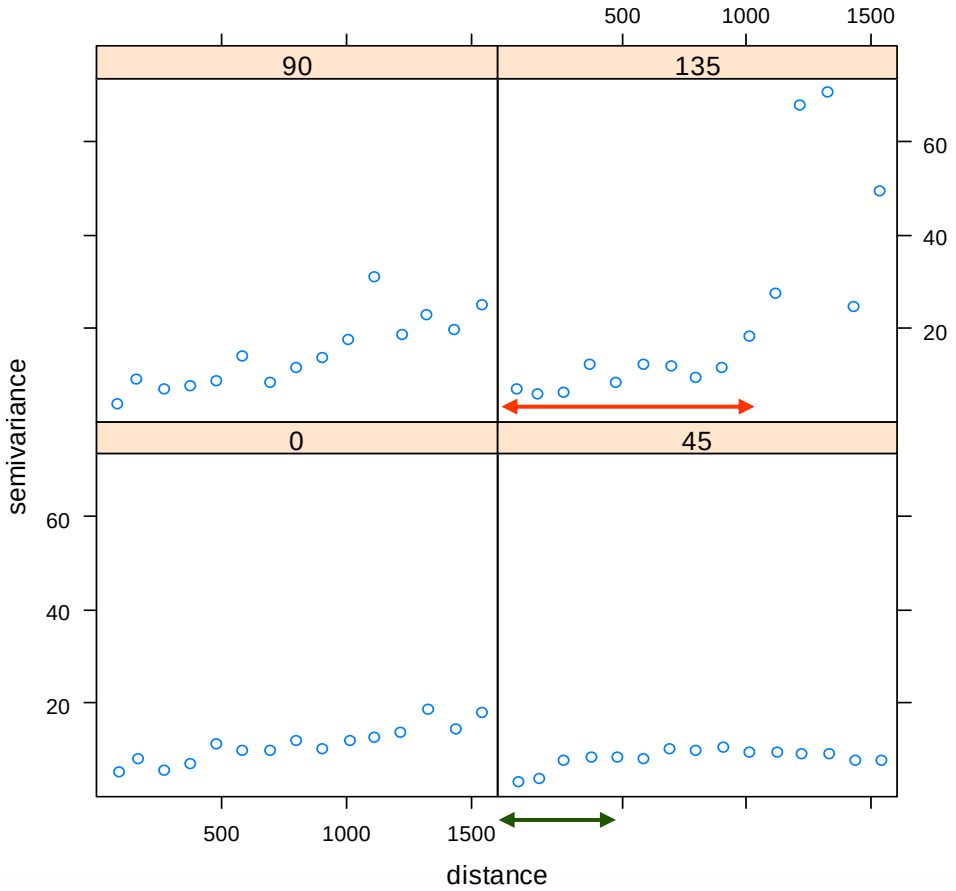
Sill: variabilità correlata alla distanza

Range: distanza oltre la quale non si osserva più correlazione spaziale

Stima del variogramma modello



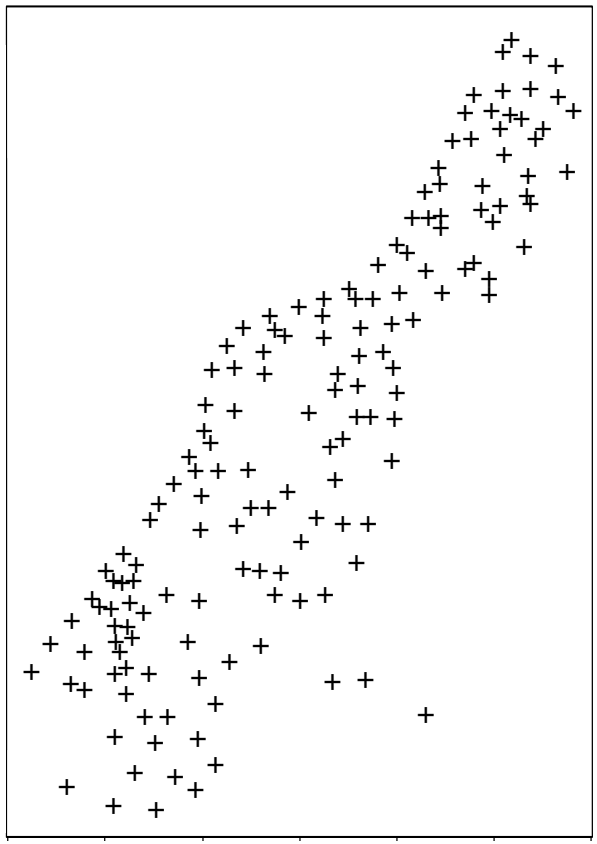
I dati presentano un trend: occorre eliminare l'effetto del trend (detrendizzare)

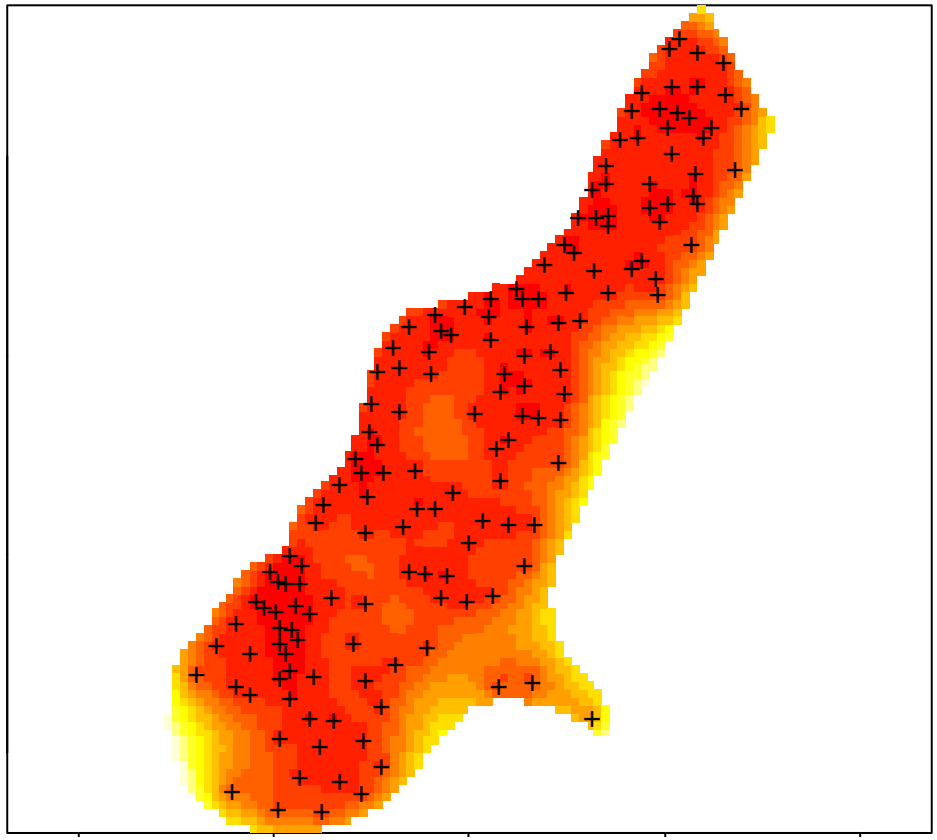


Anisotropia: diversa correlazione a seconda della direzione (45° vs 135°)

Modelli di interpolazione dei dati

Modelli di interpolazione





Dato puntuale → **Interpolazione** → **Dato continuo**

Modelli di interpolazione dei dati

Modelli di interpolazione

Per ottenere una rappresentazione spaziale continua nel dominio di studio di una VR occorre:

- stimare i valori della VR ai vertici di una griglia regolare sufficientemente fitta sulla base dei valori misurati (**operazione di stima**);
- interpolare sui lati della griglia i valori della VR sulla base dello stimatore selezionato (**operazione di interpolazione**).

Stimatori lineari

La stima della VR nei vertici della griglia $z^*(\mathbf{x_0})$ è una combinazione lineare dei valori misurati nei punti vicini $z(\mathbf{x}_\alpha)$

$$z^*(\mathbf{x_0}) = \sum_{\alpha=1}^{n} \lambda_\alpha \cdot z(\mathbf{x}_\alpha)$$

- i punti $\mathbf{x}_\alpha$ rappresentano rispetto a $\mathbf{x_0}$ il cosiddetto **vicinaggio di stima**;
- i coefficienti λ_α sono i **pesi** della combinazione lineare.

Correttezza ed accuratezza della stima

Errore della stima

Differenza nel punto $\mathbf{x_0}$ tra valore vero e valore stimato:

$$\varepsilon = z(x_0) - z^*(x_0) = z(x_0) - \sum_{\alpha=1}^{n} \lambda_\alpha \cdot z(x_\alpha)$$

Correttezza della stima

La stima è **corretta** se la **media degli errori** di stima è **nulla**.

$$E[\varepsilon] = E\left[z(x_0) - \sum_{\alpha=1}^{n} \lambda_\alpha \cdot z(x_\alpha) \right] = 0$$

Accuratezza della stima

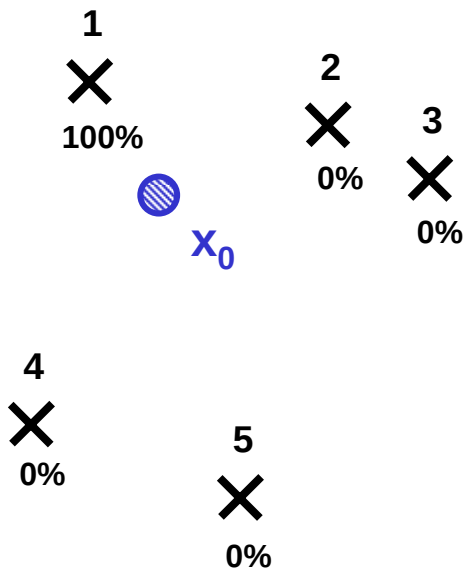
La **stima** è tanto **più accurata** quanto **più bassa** è la dispersione degli errori ovvero della **varianza di stima**:

$$\sigma_s^2 = D^2[\varepsilon] = 2 \cdot \sum_{\alpha=1}^{n} \lambda_\alpha \cdot \gamma(x_\alpha - x_0) - \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \cdot \lambda_\beta \cdot \gamma(x_\alpha - x_\beta)$$

Modelli deterministici

- I pesi λ_α della combinazione lineare sono funzioni matematiche date.
- I **valori stimati** sono quindi calcolati a partire dai dati misurati sulla base di precise leggi matematiche che **non tengono conto della legge di autocorrelazione spaziale (variogramma)**.
- I modelli più comunemente utilizzati sono:
  - **Poligoni di influenza (Nearest Neighbour Analysis – NNR)**
  - **Media mobile**
  - **Inverso delle distanze (Inverce Distance Weighted – IDW)**
  - Regressioni Polinomiali
  - Spline
- Il risultato fornito dai modelli deterministici è unicamente una **mappa delle previsioni**.
- E' possibile comunque determinare l'**accuratezza della stima** calcolando la **varianza di stima** in base al variogramma.

Poligoni di influenza (NNR)

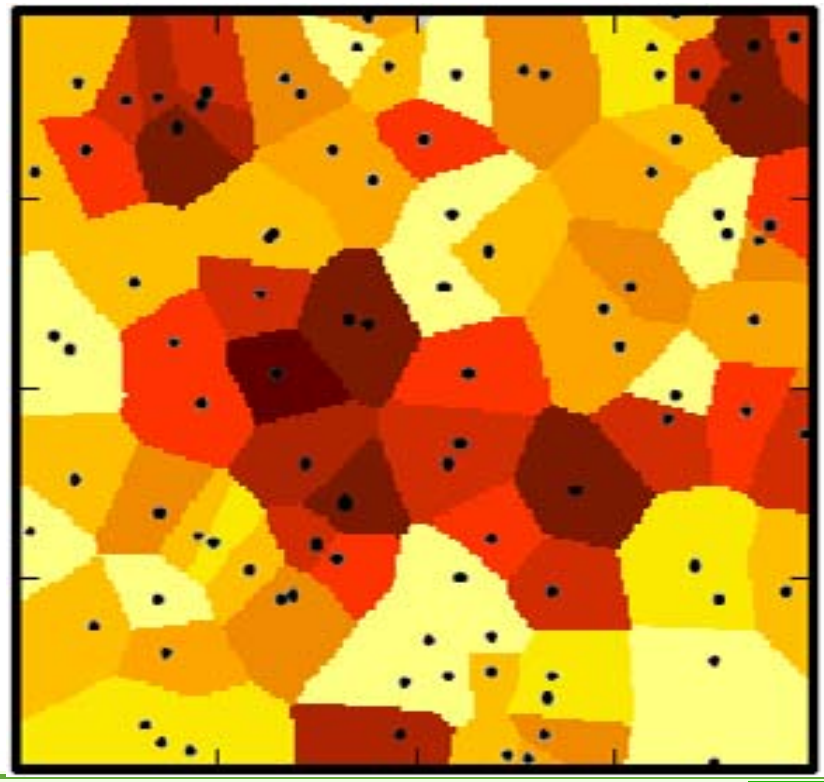


Il valore stimato è pari al valore del punto più vicino all'interno del vicinaggio di stima.

$z^*(x_0) = z_1$

$l_1 = 100\% = 1$

$l_{i \neq 1} = 0\% = 0$



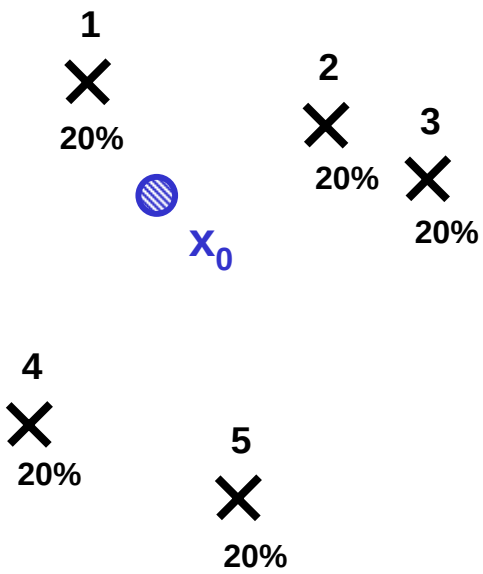
L'influenza è quindi limitata ad un solo dato ovvero al punto più vicino. Si trascurano gli altri contributi.

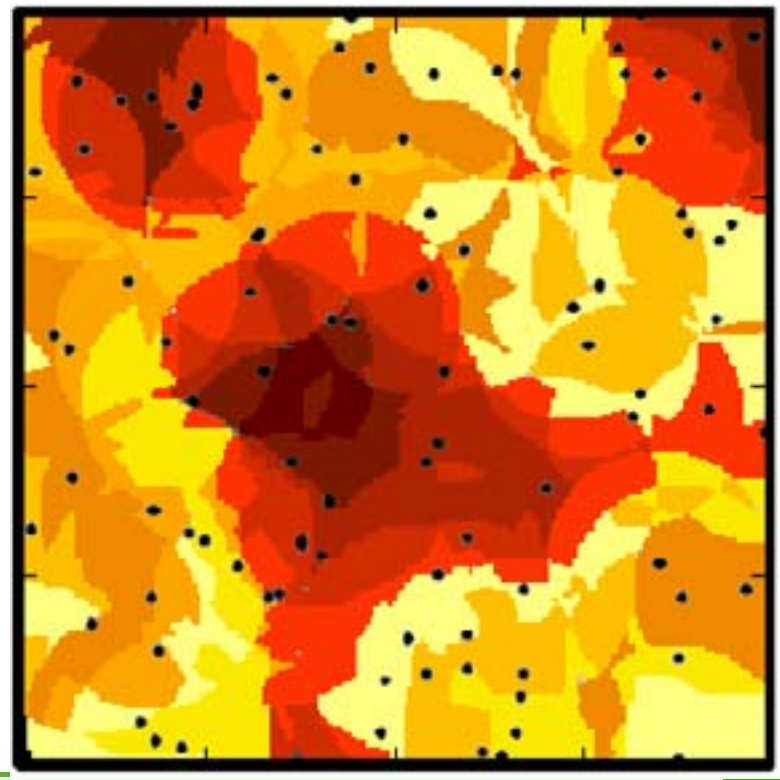
Media mobile

Il valore stimato è pari alla media dei valori dei punti all'interno del vicinaggio. I pesi sono uguali e dipendono dal numero di punti considerati.

$$z^*(x_0) = \frac{\sum_{i=1}^{n} z_i}{n}$$

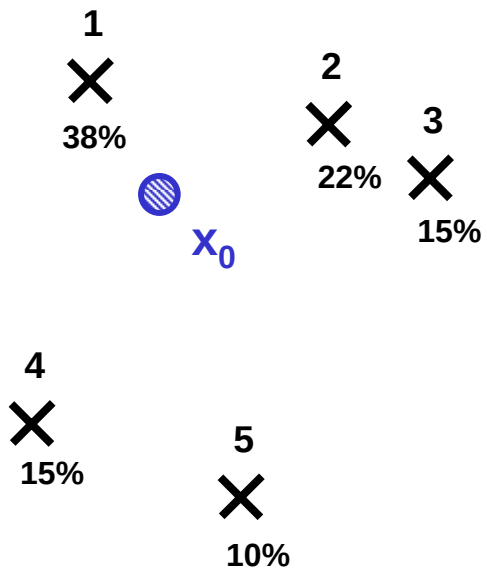
$\lambda_i = 1/n \quad \Rightarrow \quad \lambda_i = 1/5 = 20\%$





L'influenza sulla stima (pesi) non dipende né dalla distanza del punto x_0 rispetto ai punti all'interno del vicinaggio, né dai valori assunti dalla variabile nei punti stessi.

Inverso delle distanze (IDW)



I pesi λ_i sono inversamente proporzionali alla distanza dell' i-esimo punto del vicinaggio rispetto al punto $\mathbf{x_0}$.

$$z^*(x_0) = \frac{\sum_{i=1}^{n} \frac{z_i}{\varphi(d_i)}}{\sum_{i=1}^{n} \frac{1}{\varphi(d_i)}} \quad \Rightarrow \quad \lambda_i = \frac{\frac{1}{\varphi(d_i)}}{\sum_{i=1}^{n} \frac{1}{\varphi(d_i)}}$$

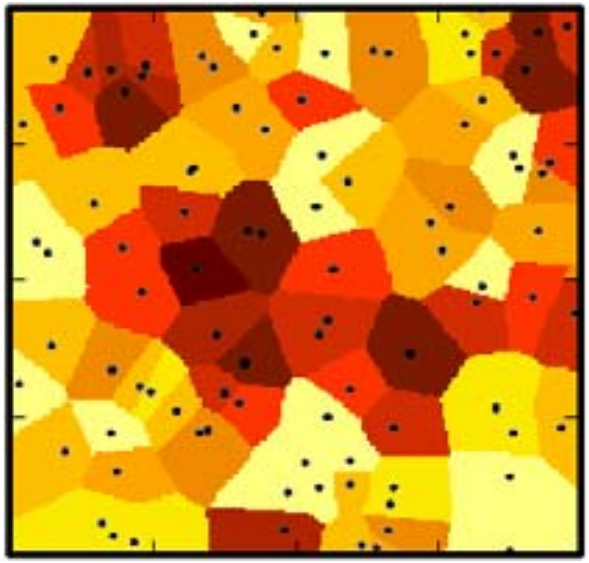
$$\varphi(d_i) = d \quad \Rightarrow \quad \textbf{Inverso della distanza}$$

$$\varphi(d_i) = d^2 \quad \Rightarrow \quad \textbf{Inverso del quadrato della distanza}$$

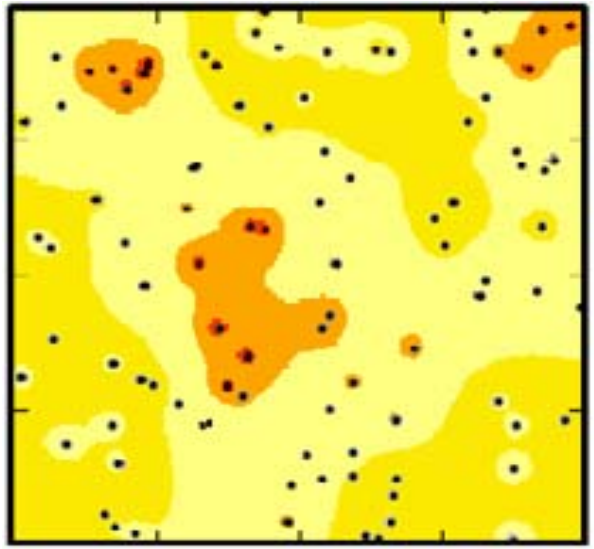
L'influenza sulla stima è data dalla distanza del punto x_0 rispetto ai punti all'interno del vicinaggio e non dai valori assunti dalla variabile nei punti stessi.

Quale modello scegliere?

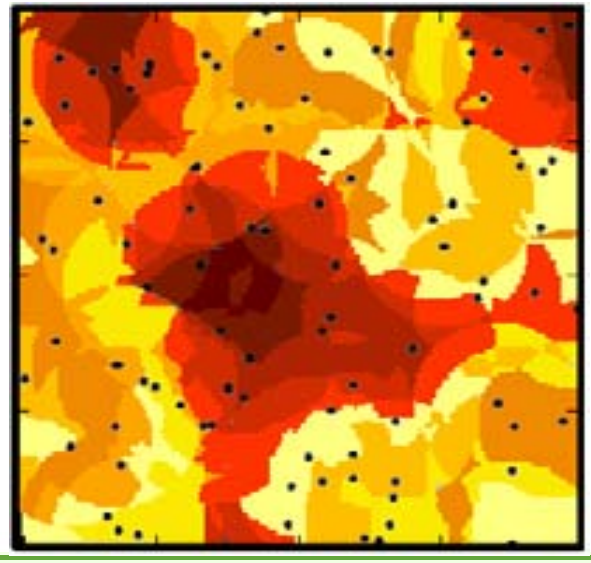
Poligoni di influenza



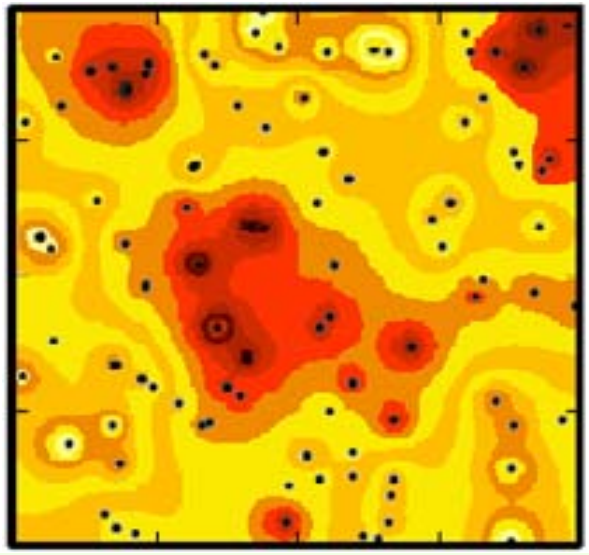
Inverso delle distanze



Media mobile



Inverso del quadrato delle distanze



Quale modello scegliere?

Il modello da scegliere è quello che minimizza la

varianza di stima

$$\sigma_s^2 = D^2[\varepsilon] = 2 \cdot \sum_{\alpha=1}^{n} \lambda_\alpha \cdot \gamma(x_\alpha - x_0) - \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \cdot \lambda_\beta \cdot \gamma(x_\alpha - x_\beta)$$

Modello	σ_s^2
Poligoni di influenza (NNR)	3,24
Media Mobile	3,01
Inverso delle distanze (IDW)	2,83
Inverso del quadrato delle distanze (IDW)	**2,81**

Modelli statistici

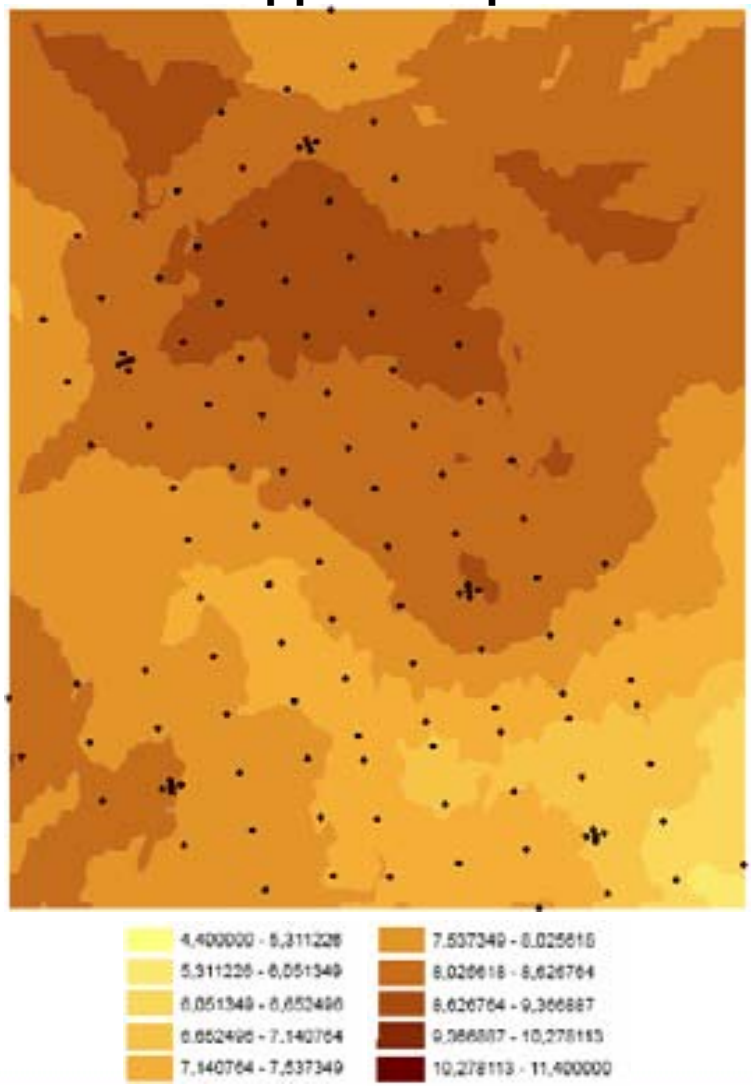
- I pesi λ_α della combinazione lineare sono determinati in modo da minimizzare la varianza di stima σ_s^2.
- I **valori stimati** quindi **tengono conto della legge di autocorrelazione spaziale (variogramma)**.
- I modelli di stima statistici sono genericamente indicati come **Kriging** ed in particolare ne esistono diverse applicazioni:
  - **Kriging Semplice**
  - **Kriging Ordinario**
  - **Universal Kriging**

  $\Rightarrow$ **geostatistica univariata**
  - **CoKriging**

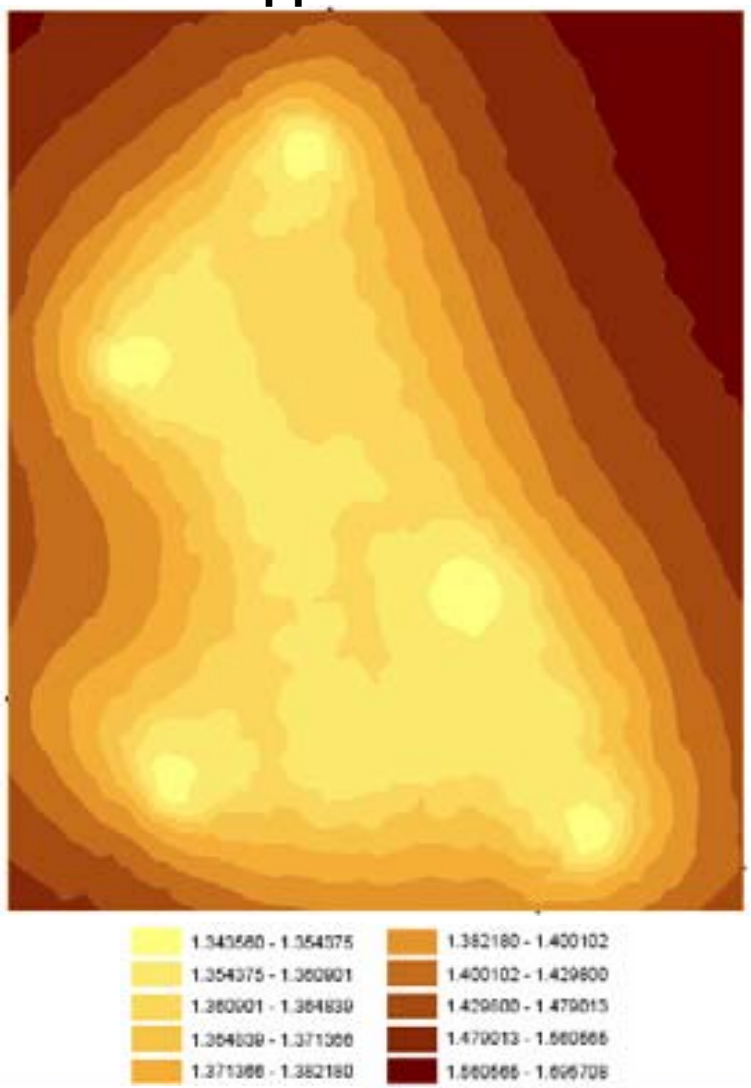
  $\Rightarrow$ **geostatistica multivariata**
- Il risultato fornito dai modelli deterministici è una **mappa delle previsioni** insieme ad una **mappa delle incertezze**.
- La mappa delle incertezze mostra la **varianza di stima** determinata in base al variogramma.

Modelli statistici

Foc - Mappa delle previsioni



Foc - Mappa delle incertezze



Kriging ordinario e Kriging semplice

Kriging ordinario (KO)

- Si applica nel caso di **funzioni aleatorie stazionarie**, ossia nel caso in cui la media dei residui sia costante in tutto il dominio di studio
- Minimizzazione della varianza di stima:

$$\frac{\partial \sigma_s^2}{\partial \lambda_\alpha} = 0 \quad \Rightarrow \sum_{\beta=1}^{n} \lambda_\beta \cdot \gamma(x_\alpha - x_\beta) + \mu = \gamma(x_\alpha - x_0) \quad \forall \alpha = 1,...,n$$

sistema di n equazioni in n+1 incognite (pesi λ_α e lagrangiano μ)

- L'ultima equazione per risolvere il sistema si ottiene dalla condizione di correttezza della stima in caso di funzioni aleatorie stazionarie:

$$E[\varepsilon] = E\left[z(x_0) - \sum_{\beta=1}^{n} \lambda_\beta \cdot z(x_\beta) \right] = 0 \quad \Rightarrow \quad \sum_{\beta=1}^{n} \lambda_\beta = 1$$

Kriging semplice (KS)

- Si applica nel caso di **funzioni aleatorie stazionarie** con **media** dei residui **costante e nota**. Necessita di un elevato numero di dati misurati.
- E' più preciso del KO nel caso di un elevato numero di misurazioni.

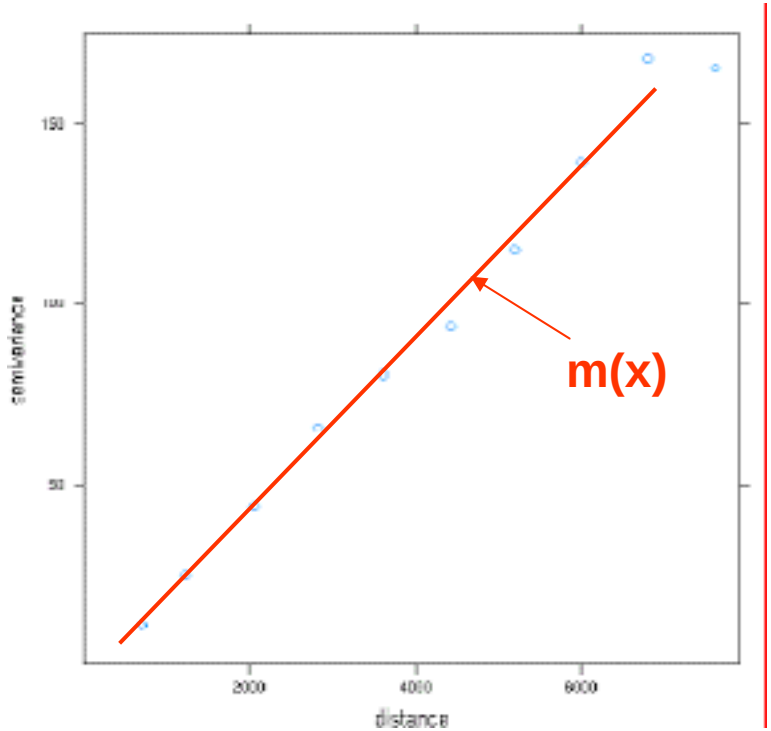
$$\sigma_s^2(KS) < \sigma_s^2(KO)$$

Universal Kriging

- Si applica nel caso di **funzioni aleatorie non stazionarie intrinseche**, ossia nel caso in cui la media dei residui non è costante e la legge di autocorrelazione presenta un trend.

- La funzione aleatoria **Z(x)** può essere considerata in ogni punto **x** del dominio come la sovrapposizione di due componenti:
  - **Il trend m(x)** che rappresenta la parte deterministica
  - **Il residuo Y(x)** che rappresenta la parte aleatoria

$$Z(x) = Y(x) + m(x)$$



- Se il residuo **Y(x)** è una funzione stazionaria e non è correlata al trend allora è possibile applicare la procedura di Kriging al residuo e quindi effettuare lo **Universal Kriging (UK)**.

CoKriging

- La stima della la **variabile principale (target)** non si basa solo sui valori della variabile esaminata ma prende in considerazione anche altre **variabili ausiliarie**.
- La **condizione necessaria** per l'applicazione del CoKriging è che la variabile target $\mathbf{z_1(x)}$ e le variabili ausiliarie $\mathbf{z_2(x)}$ siano **spazialmente correlate**.
- Lo stimatore della variabile target è dato da:

$$z_1^*(x_0) = \sum_{\alpha=1}^{n_1} \lambda_\alpha \cdot z_1(x_\alpha) + \sum_{\alpha=1}^{n_2} \omega_\alpha \cdot z_2(x_\alpha)$$

- La condizione di correttezza della stima è data da:

$$\sum_{\alpha=1}^{n_1} \lambda_\alpha = 1 \qquad \sum_{\alpha=1}^{n_2} \omega_\alpha = 0$$

- Un possibile campo di applicazione è quello a dati di caratterizzazione/monitoraggio riferiti a periodi diversi. E' possibile quindi utilizzare un maggior numero di dati anche se non riferiti alla stessa campagna di indagine.

Quale modello scegliere?

Modelli deterministici

Vantaggi
- Semplicità di utilizzo
- Possono essere utilizzati anche con pochi dati misurati
- Non richiedono ipotesi sulla distribuzione spaziale dei dati
- Non dipendono dal modello scelto per il variogramma

Svantaggi
- Non tengono conto della variabilità spaziale
- Non forniscono una mappa delle incertezze
- Spesso non tengono conto dei valori assunti dalla variabile nei punti di misura

Modelli statistici

Vantaggi
- Tengono conto della variabilità spaziale del dato
- Forniscono una mappa delle incertezze associate alla stima puntuale del dato
- Danno stime più precise con un numero discreto di misure

Svantaggi
- Sono più complessi
- Richiedono ipotesi sulla distribuzione spaziale dei dati (stazionarietà, trend, ecc.)
- Dipendono dal modello scelto per il variogramma
- Hanno poco potere previsionale se si dispone di pochi dati

Esperienze applicative

Nell'ambito della bonifica dei siti contaminati la geostatistica viene applicata soprattutto per:

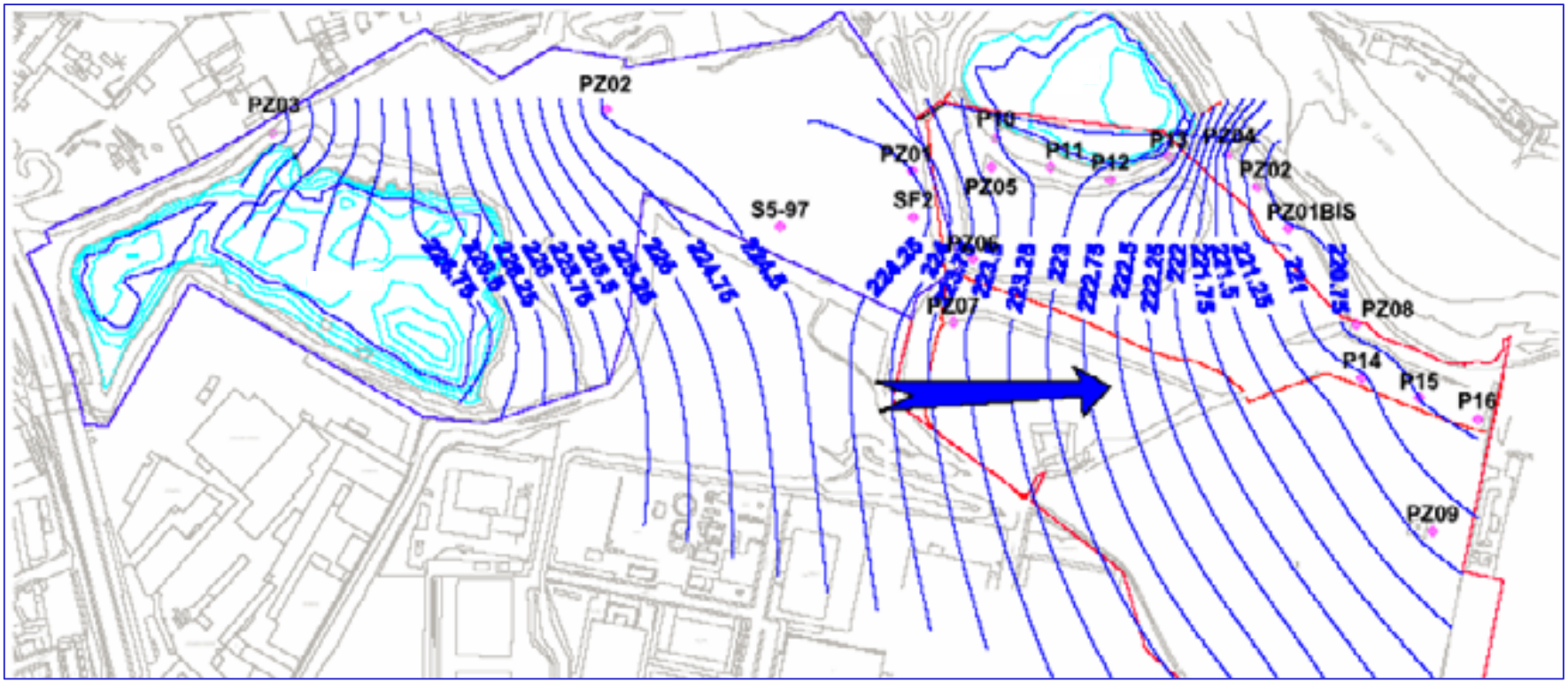
- Carta delle isopieze
- Carte di isoconcentrazione (soprattutto per la falda)

In genere i risultati ottenuti risultano carenti in quanto:

- La geostatistica viene spesso applicata con pochi dati a disposizione.
- Raramente viene indicato il modello di stima e di interpolazione utilizzato.
- Spesso non vengono indicati i punti utilizzati per l'interpolazione e/o i valori che la variabile assume in tali punti, né il dominio di studio.
- A volte vengono utilizzate condizioni al contorno nell'area (punti fittizi) che non rispecchiano dati reali misurati.
- Raramente vengono applicati più modelli di stima allo stesso set di dati ed effettuato il confronto fra le varianze di stima.
- Quando vengono utilizzati modelli statistici (kriging), non viene riportata la determinazione del variogramma sperimentale e del variogramma modello utilizzato, né viene prodotta la mappa delle incertezze.

Esempi di errori frequenti (1)

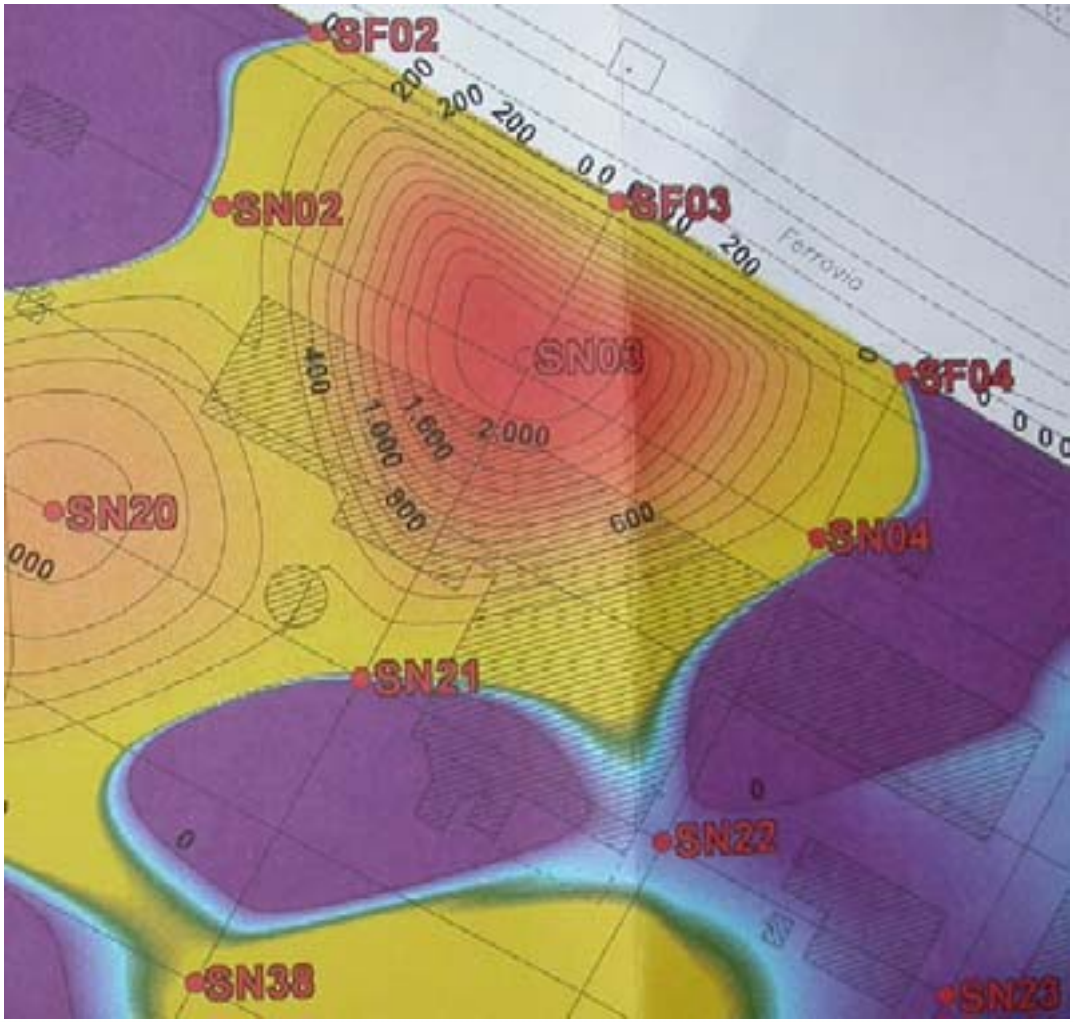
Carta delle isopieze



- Qual è il dominio di studio? E' il sito o un'area più ampia?
- Quali sono i valori riscontrati nei piezometri?
- Non è stato riportato il modello di stima (… forse IDW?)
- Incertezza della stima?

Esempi di errori frequenti (2)

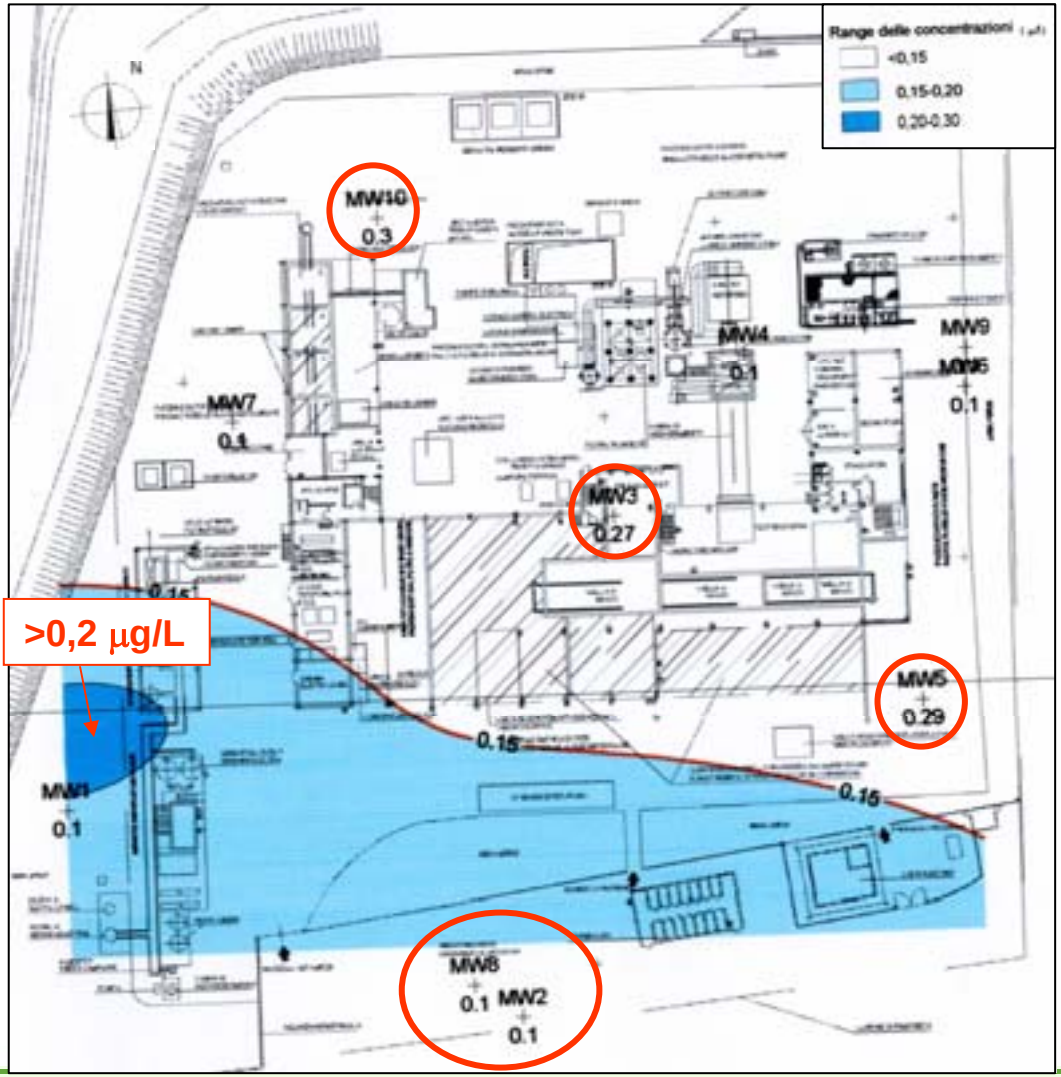
Carta di isoconcentrazione nei suoli



- Modello utilizzato: Kriging
- Al confine del sito (dominio di studio) sono stati introdotti dei punti fittizi (SF_n) ai quali è stata assegnata una concentrazione arbitraria non misurata.
- La concentrazione nei punti fittizi è stata posta a volte pari al valore del punto più vicino e a volte pari a zero.
- Le aree in viola indicano le zone a concentrazione inferiore a zero!
- Variogramma?
- Incertezza della stima?

Esempi di errori frequenti (3)

Carta di isoconcentrazione in falda



- Qual è il dominio di studio?
- Perché i piezometri MW8 e MW2 non sono stati considerati nell'interpolazione?
- Le curve di isoconcentrazione non rispecchiano i dati misurati!

Considerazioni conclusive

- La geostatistica è uno strumento utile e potente e viene utilizzata spesso nel campo dei siti contaminati.
- L'esperienza nella valutazione degli elaborati progettuali relativi ai siti contaminati di interesse nazionale purtroppo mostra che il livello di utilizzo di questo strumento è ancora carente in quanto:
  - spesso ci si limita ad una mera applicazione di un software senza adeguati controlli e valutazioni dei risultati;
  - alcuni modelli (es. poligoni di influenza) sono ritenuti più conservativi di altri, ma alla fine possono sottostimare la situazione reale del sito;
  - spesso si vogliono usare modelli complessi (es. kriging) non applicabili ai dati disponibili o con limitati set di dati.
- Occorre soprattutto non limitarsi solo a fornire delle "mappe", ma indicare tutte le valutazioni che le hanno prodotte oltre all'incertezza dei risultati.