UNIVERSITÀ
DEGLI STUDI
DI MILANO

DiSAA
DIPARTIMENTO
di SCIENZE
AGRARIE e
AMBIENTALI

Elementi di Geostatistica

Metodi statistici per la ricerca ambientale 2019/2020 **Prof. MARCO ACUTIS**

Aula C22 [Agraria Edificio 6]

UNIVERSITÀ
DEGLI STUDI
DI MILANO

DiSAA
DIPARTIMENTO
di SCIENZE
AGRARIE e
AMBIENTALI

Obiettivi

Definizioni di Geostatistica e storia

Motivazioni (agricoltura)

Flusso di lavoro

Inverse Distance Weighted

Kriging

Variogramma

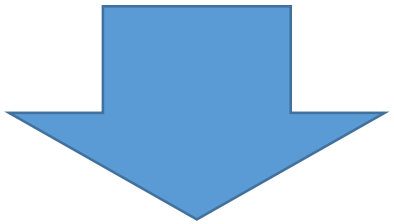
Implicazioni della stima delle variabili spaziali in agronomia

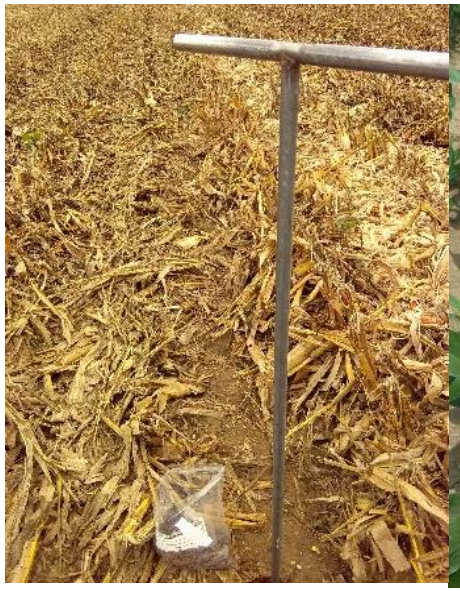
Esempio 1.

Vengono presi dei campioni di suolo in un campo per determinarne le proprietà del suolo. Dopo averle determinate viene applicato un fertilizzante omogeneamente, la resa totale verrà registrata. Agricoltore soddisfatto.

Esempio 2.

Dopo 3 anni lo stesso agricoltore vuole determinare la fertilità dei suoi campi (ad esempio il contenuto in fosforo). Ma non gli basta il valore medio per il campo. Vuole più dettaglio e soprattutto vuole fertilizzare laddove è necessario.






201
202
212
203
211
204
205
206
301
210
302
333
207
310
209
309
303
208
308
304
307
305
306
0 25 50 100
m


TOPCON

Considerazioni

In ogni momento è necessario considerare l'equilibrio tra il costo della fornitura delle informazioni e i guadagni finanziari che verranno accumulati dall'agricoltore mediante fertilizzazione a rateo variabile.

In un contesto più ampio potrebbe esserci un ulteriore vantaggio se si può aiutare a evitare l'eccessiva fertilizzazione e quindi proteggere l'ambiente dall'inquinamento provocato dall'eccesso di fosforo.

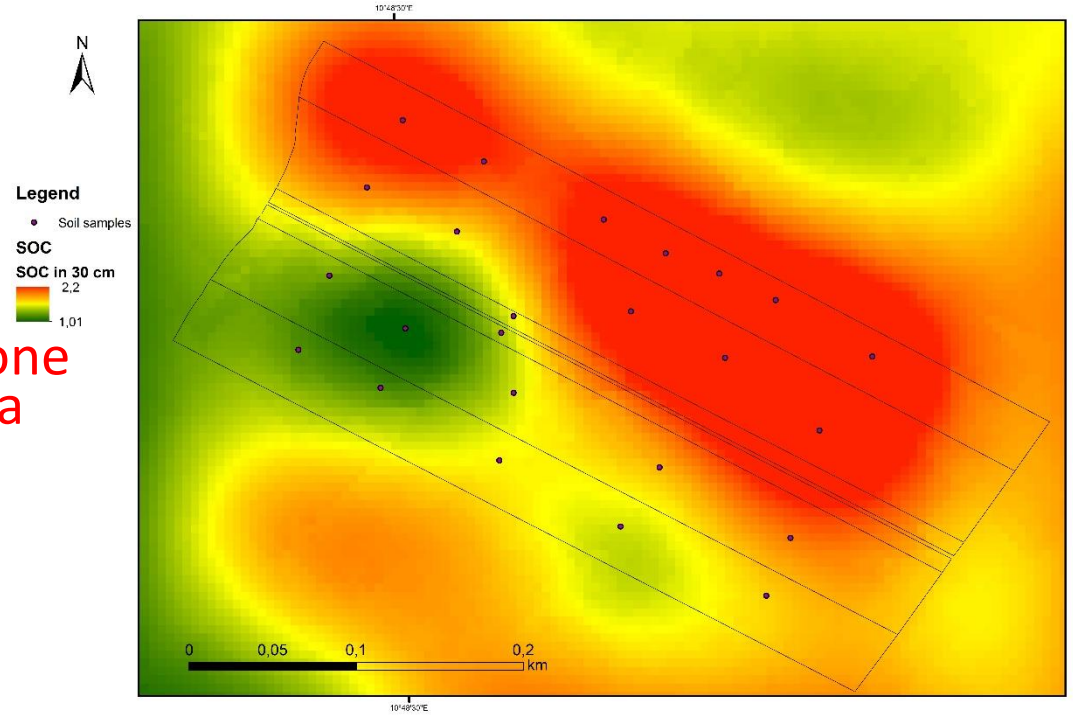
Si possono fornire all'agricoltore valori con riferimento spaziale che può usare nel suo spandiconcime automatizzato?

Rilevamento delle variabili spaziali

Il metodo tradizionale per la misura delle proprietà del suolo richiede il prelievo di campioni di suolo in un numero elevato di punti. È un approccio invasivo, e richiede lunghi tempi, sia per il prelievo dei campioni che per le analisi di laboratorio da effettuare sui campioni di suolo.

Schema di campionamento

Tecniche di interpolazione geostatistica



Mappa del carbonio organico 0-30 cm

Errore nella stima

Geostatistica ci consente stimare senza distorsioni e con il minimo errore. Ci consente di gestire proprietà che variano in modi tutt'altro che sistematici e su tutte le scale spaziali.

Gli agricoltori attenti e i gestori del territorio si affretteranno sulla parola "errore". "Le tue stime sono soggette a errori", diranno "in altre parole, sono più o meno sbagliate?

La Geostatistica ha di nuovo la risposta. Non può mai fornire informazioni complete, ovviamente, ma, dati i dati, può consentire di stimare le probabilità che i valori reali superino le soglie specificate. Ciò significa che puoi valutare i rischi dell'agricoltore di perdere la resa non facendo nulla in cui i valori reali sono inferiori alla soglia o sprecando denaro fertilizzando dove la superano.

Geostatistica

- La Geostatistica fornisce gli strumenti per quantificare la variabilità spaziale delle di una variabile, tenendo conto dei dati autocorrelazione spaziale. Consente la produzione continua mappe, a partire da dati sparsi [1] .

- Il Variogramma è il principale strumento della Geostatistica e consiste in un modello di dipendenza spaziale, che descrive la varianza dei dati tra due posizioni e la loro distanza di separazione [2].

- Usando il Variogramma e diverse tecniche di interpolazione, (ad esempio il Kriging), la variabile può essere stimata in pozioni in cui non è stata campionata [3].

Geostatistica , un po di storia

Mercer and Hall's (1911)-> plot size

Fisher (1919)-> design of experiments

- Fisher's aim was to be able to estimate the responses of crop yields to different agronomic treatments and varieties.
- Fisher have designed his experiments in such a way as to remove the effects of short-range variation by using large plots and of long-range variation by blocking

Webster and Butler (1976) -> applied this approach in soil science.

Kolmogorov (1941) -> recognized the spatial autocorrelation for which he developed the 'structure function' (now the variogram).

Kridge-> 1951

Matheron (1963) -> expanded Krige's empirical ideas, in particular the concept that neighbouring samples could be used to improve prediction, and put them into the theoretical framework of regionalized variable theory that underpins geostatistics

Geostatistica

- La superficie interpolata è concettualizzata come una delle superfici possibili che potrebbero essere osservate, tutte coerenti con i dati di input.
- Teoria delle Variabili Regionalizzate di Matheron (1965) i valori delle Variabili Regionalizzate tendono ad essere correlati (luoghi più vicini sono più simili ad altri maggiormente distanti) a certe scale. La teoria di Matheron quantifica questa correlazione.
- Una variabile regionalizzata Z(x) è una variabile con valore fortemente dipendente dalla posizione spaziale.

$$Zx = \alpha + Rx$$

α = componente casuale

R(x) = componente regionalizzata

La condizione è che R(x) sia preponderante rispetto alla componente casuale.

Kriging

Il nome viene da D. B. Krige, un ingegnere minerario sudafricano che ha definito il metodo (1951) insieme a H. S. Sichel come strumento di indagine dei giacimenti minerari. 10 anni più tardi, prendendo spunto dal loro lavoro, G. Matheron formalizzò la Teoria delle Variabili regionalizzate

Danie Krige learning the trade, 1939.

Photo credit: R.C.A. Minnitt[I, *]; W. Assibey-Bonsu[II]
http://www.scielo.org.za/scielo.php

Geostatistica, flusso di lavoro

Campionamento (progettazione, acquisizione dati)

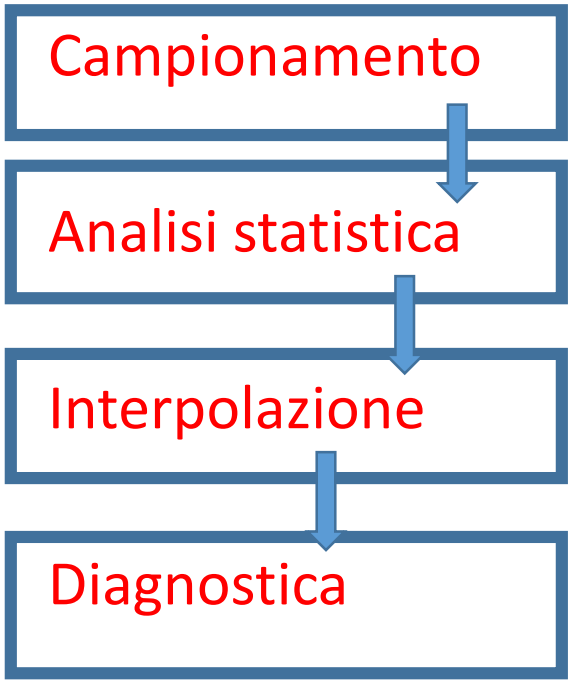
Analisi statistica (explanatory data analysis): Analizza le proprietà spaziali e statistiche dei dati esame distribuzione, identificazione ed eliminazione trend, identificazione anisotropie, ecc.

Interpolazione, calcolo superficie statistica di predizione e Carta degli errori. La predizione dei valori incogniti avviene attraverso il modello del variogramma, la configurazione spaziale dei dati e il valore dei punti misurati. Stima del variogramma sperimentale e scelta del modello da utilizzare.

Diagnostica, comprendere “quanto bene” il modello predice i valori incogniti. Analisi di errore: cross validation, validation points, ecc.

Requisiti del dataset

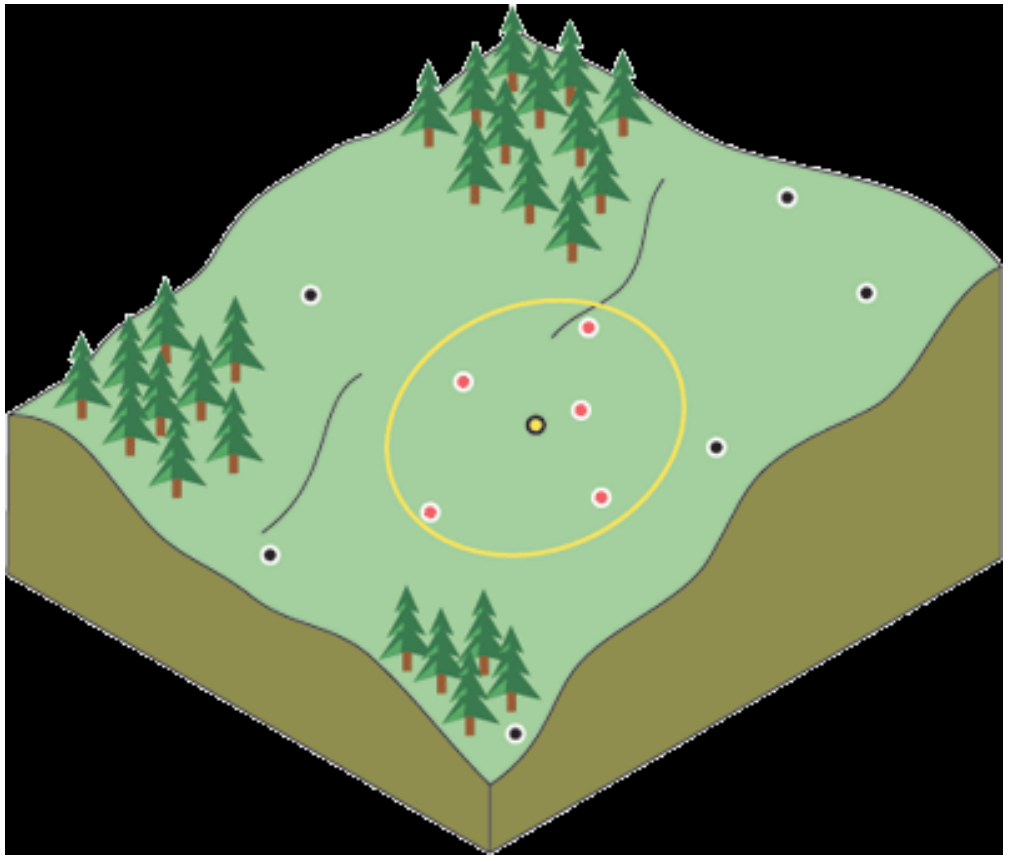
- Sufficientemente ampio (numero minimo di punti per analisi geostatistica: 30-50 osservazioni per metodi quali IDW w poligoni di thiessen, minimo 100 osservazioni per Kriging).
- Campionamento imparziale (es, nessuna preferenza verso misure nei luoghi più accessibili), Rappresentativo, Indipendente.
- Acquisito con una significativa precisione (uso di GPS differenziali con precisione <1m di errore).
- Uniformità delle misurazioni (es. stessa stagione, stessi strumenti, stesse condizioni, ecc.)
- Tenere conto delle anisotropie.
- Evitare cluster di campioni.



Caso studio

- La Geostatistica utilizza variabili misurate in località/posizioni specifiche ma possono essere teoricamente misurate ovunque (e.g. stazioni meteo, proprietà del suolo, rese di una coltura),
- A causa delle limitazioni di tempo e monetarie le variabili vengono misurate solo in alcuni luoghi. Comunque noi vorremmo ottenere una stima di esse per tutti I posti di interesse.
- Per esempio, non abbiamo una stazione meteorological attiva alla Facoltà di Agraria ma vorremmo sapere come stimare la tamperatura basandoci sulle misure ottenute dale stazioni vicine.
- Introdurremo i concetti attraverso l'esempio di mappatura della temperatura.

- Posizioni sconosciute vengono stimate basandosi sulla somma pesata delle osservazioni vicine. Essenzialmente, il cerchio giallo viene stimato usando le osservazioni dei punti rossi attorno ad esso.
- Il modo con il quale I peso sono calcolati determina il tipo di interpolazione. Un metodo popolare è l'inverso pesato della distanza o meglio Inverse Distance Weighted IDW.
- Com'è suggerito dal nome, i pesi sono calcolati come l'inverso della distanza tra il giallo e ogni rosso



Inverso delle distanze IDW

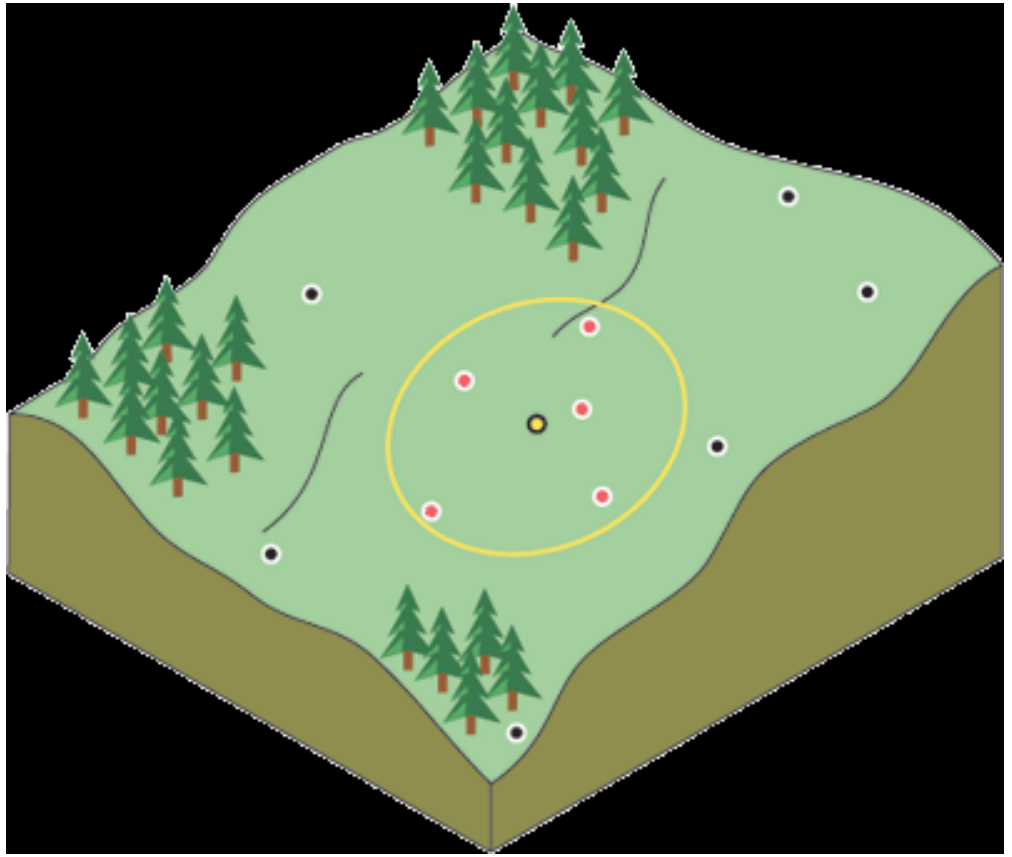
I pesi λi sono inversamente proporzionali alla distanza dell' i-esimo punto del vicinaggio rispetto al **cerchio giallo.**

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

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

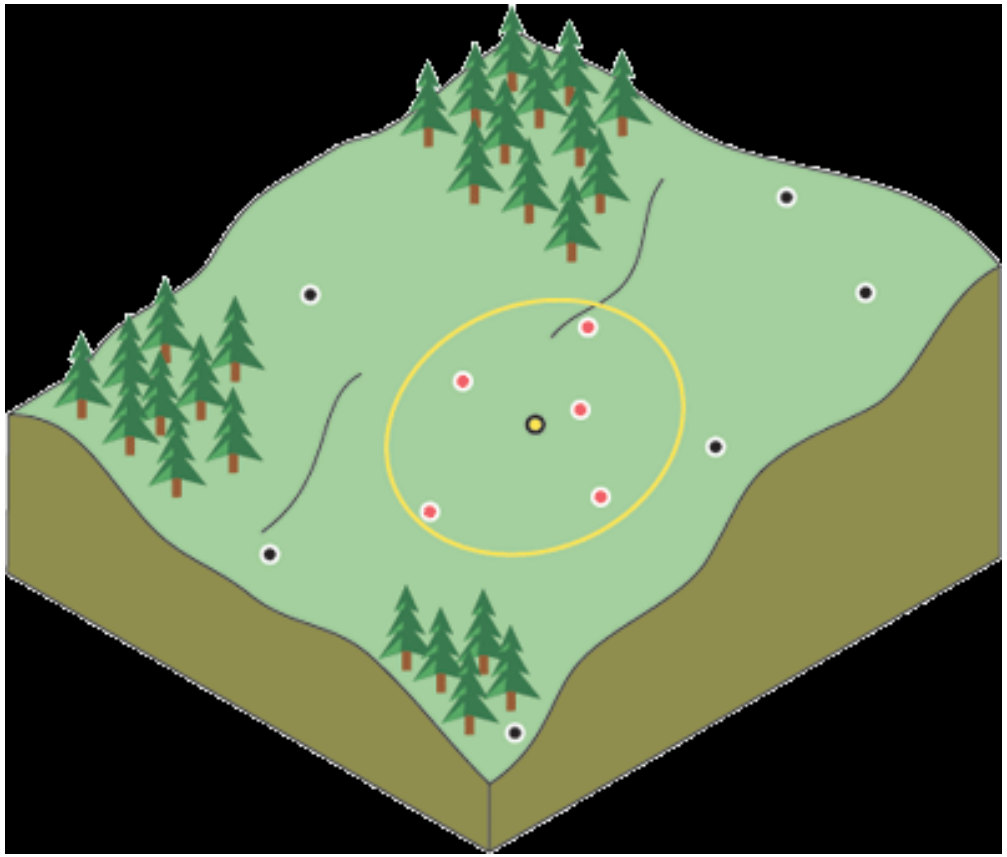
$\varphi(d_i) = d \quad \Rightarrow$ **Inverso della distanza**

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



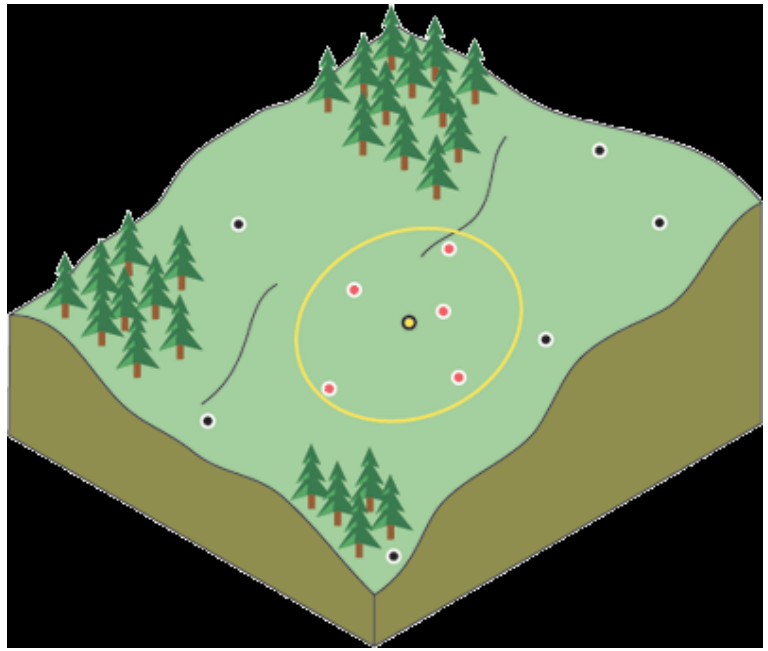
Kriging

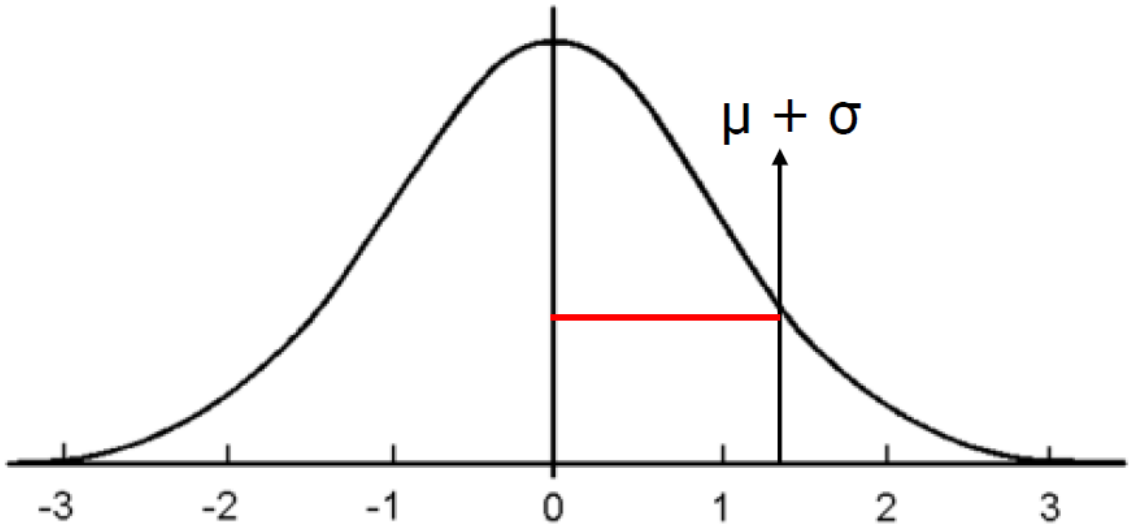
- É un metodo di intepolazione statistico.
- Il Kriging assume la condizione di stazionarietà. Questo significa che il Kriging assume che in un'area piccola attorno al punto giallo la media della variabile (temperature ad esempio) sia costante.
- I valori in ogni punto rosso cambieranno, perchè la variabile avrà una cera varianza locale. Comunque, la media generale, assumendo che si possa campionare ovunque, rimarrà costante e uguale alla media artimetica dei punti rossi



Kriging

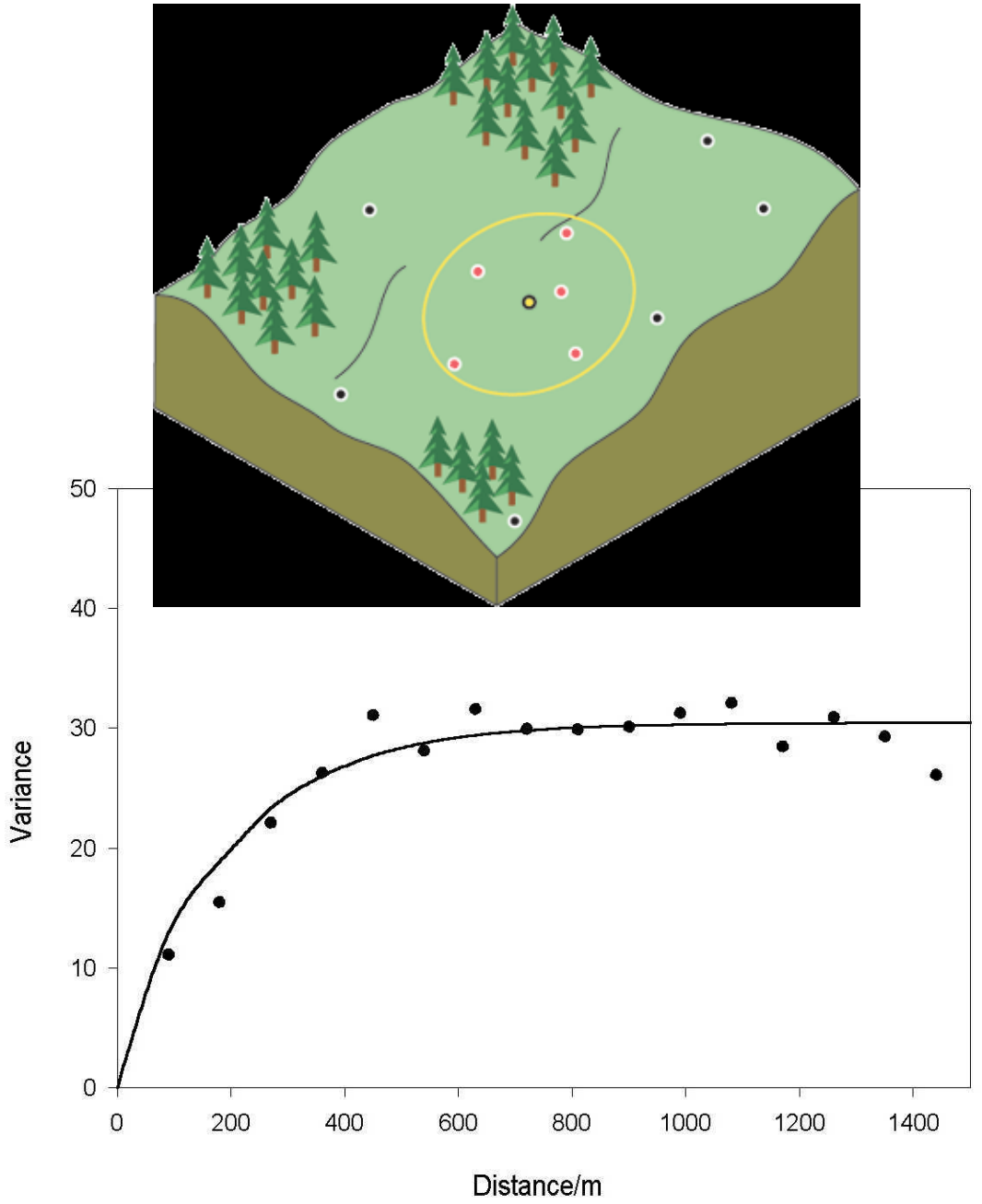
Assumendo che la nostra variabile si possa approssimare ad una distribuzione normale, se conosciamo la media possiamo calcolare il valore sconosciuto alla posizione n (pallino giallo) aggiungendo o sottrendo la varianza





Kriging

- La varianza locale, la differenza tra il punto giallo e i vicini rossi dipenderà allora dalla distanza tra loro.
- Il Kriging modella questa differenza creando il variogramma, che consiste in una funzione che ci dice *quant'è la varianza del pallino giallo in relazione alla sua distanza dai vicini rossi*.
- La semivarianza è la metà della somma delle varianze (quadrato della deviazione standard) tra ogni valore Z e ognuno dei punti alla stessa distanza.
- Una misura della interdipendenza dei valori di Z basata su quanto vicini essi sono (una misura del grado di dipendenza spaziale tra i campioni)



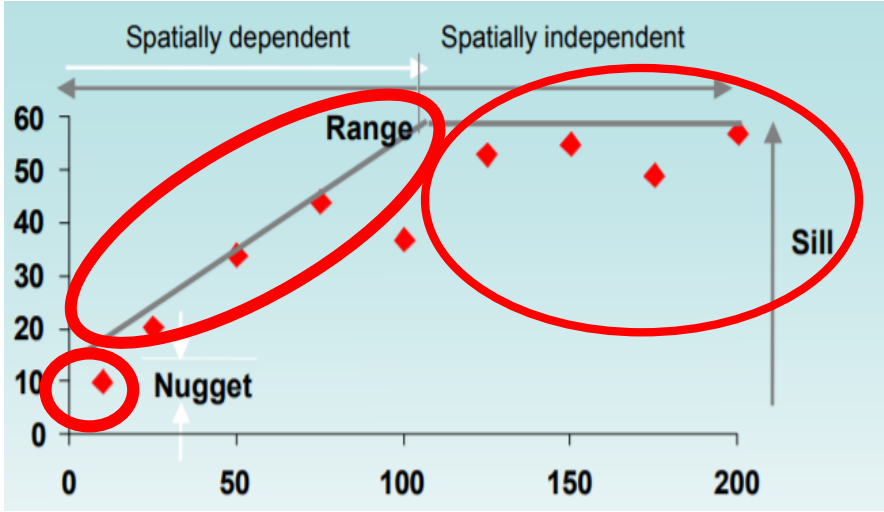
Variogramma

Il variogramma stima il grado medio varianza tra valori in funzione della loro distanza nello spazio geografico.

Il nugget rappresenta le variazioni casuali, della variabilità su piccola scala (inferiore cioè alla scala del campionamento), la variazione che ci sarebbe tra due punti più vicini rispetto alla distanza minima dei punti campionati.

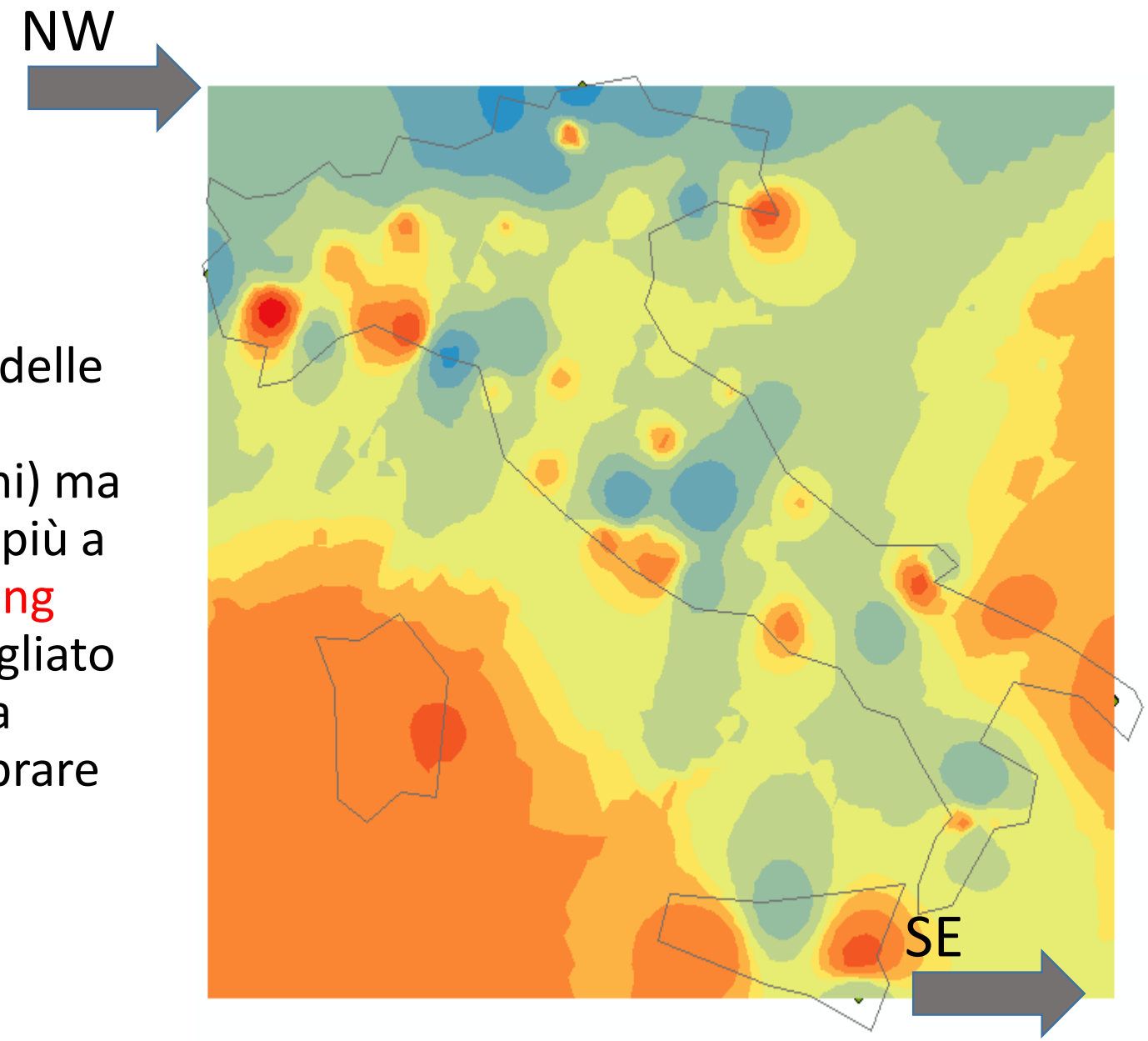
Il range corrisponde alla distanza in cui il semivariogramma (in un modello stazionario) raggiunge un valore costante detto sill. Intuitivamente è la massima distanza per cui due punti possono essere considerati spazialmente correlati.

Il sill rappresenta la varianza massima tra i punti misurati, che per modelli stazionari corrisponde alla varianza di campionamento.



Kriging

- Una nota un po spiacevole dell'analisi geostatistica in GIS e che secondo le impostazioni di base, vengono create delle mappe che non seguono il limite geografico della variabile (I suoi confini) ma il suo spazio definito dalle coordinate più a nord-ovest e sud-est in gergo "bounding box" del dataset, che deve essere ritagliato sul campo di esistenza del dato o della regione geofrica di interesse per sembrare più professionale.



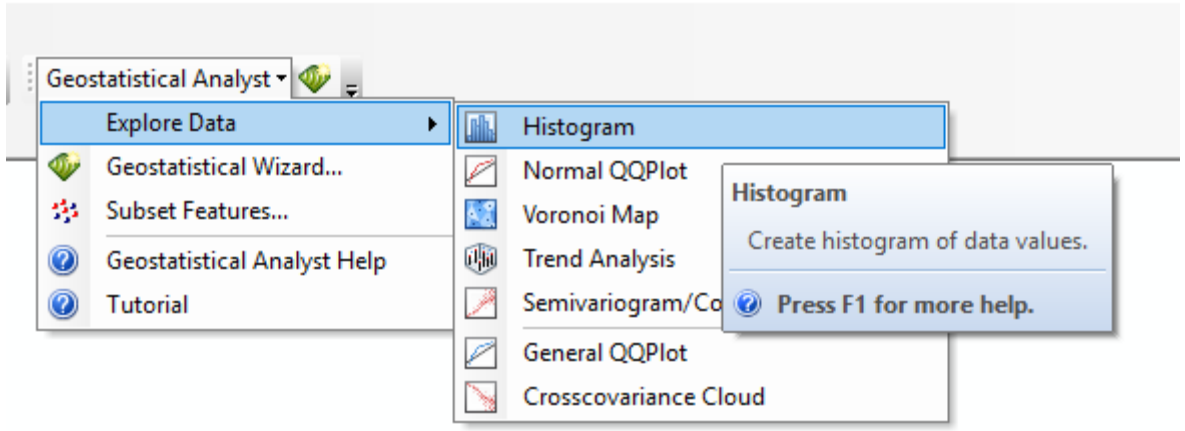
Kriging

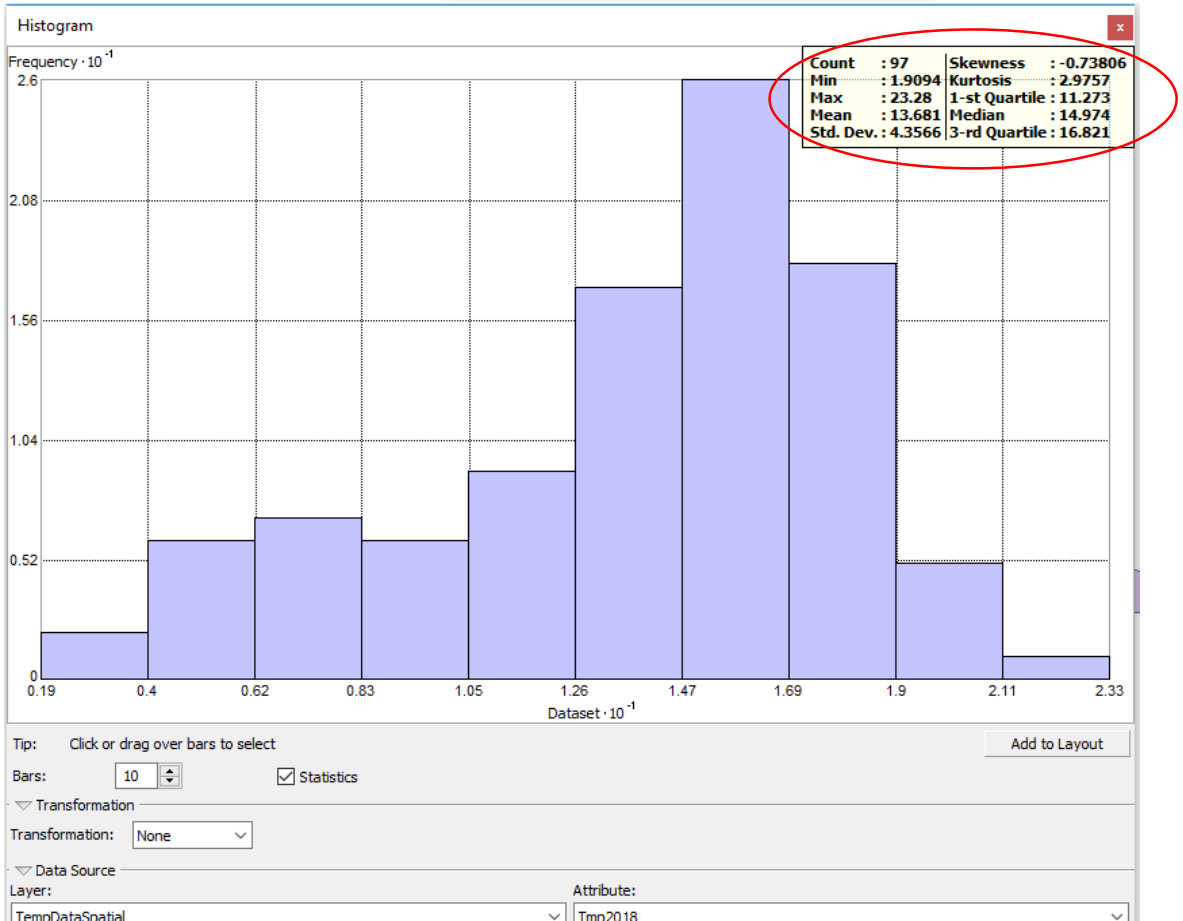
Kriging richiede qualche altra specifica, ci sono altre assunzioni che devono essere soddisfatte oltre alla stazionarietà per applicare in maniera esaustiva il metodo.

- **Normalità**
- **Scelta del modello del variogramma**

Kriging

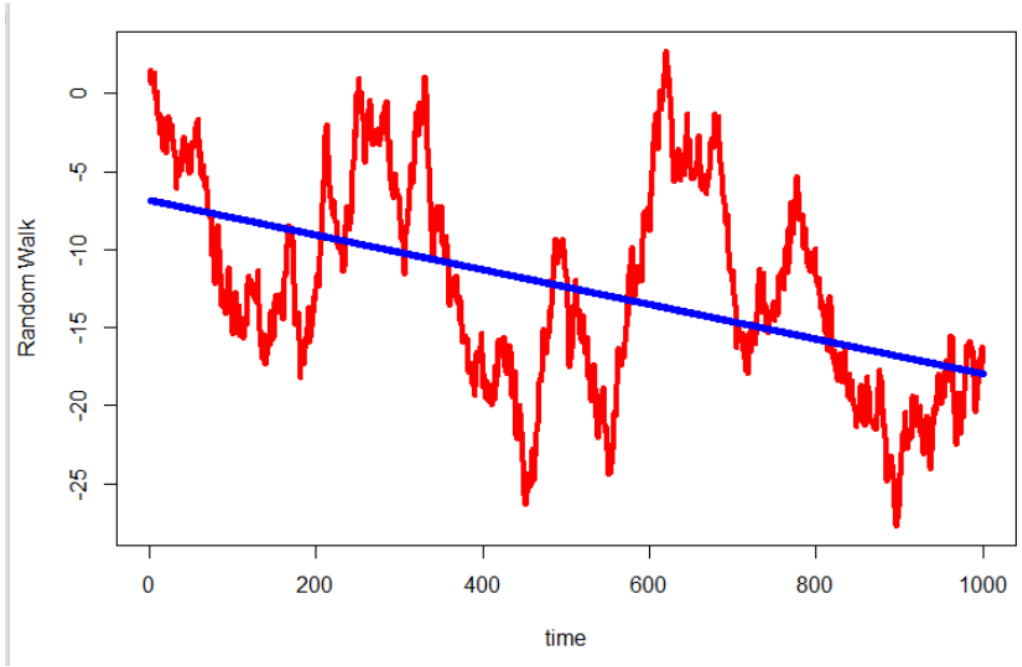
- Una delle assunzioni del kriging è infatti la Normalità, ovvero che la variabile segua una distribuzione normale. Questa distribuzione è la sola che ci consente, conoscendo la media di stimare i valori sconosciuti partendo dalla varianza.
- Dobbiamo quindi controllare se nel nostro caso abbiamo dei dati normalmente distribuiti altrimenti dobbiamo applicare delle trasformazioni.
- Secono fonti in letteratura (2), la via più semplice di verificare il requisito di normalità, è osservare il valore di skewness (il quale è 0 per una distribuzione perfettamente normale). Se la skewness è maggiore o inferior di ±0.5, allora sarà necessario tenere in considerazione la trasformazione dei dati, in questo caso è stata effettuata una trasformazione Box-Cox con lambda equal 1.3 1.4

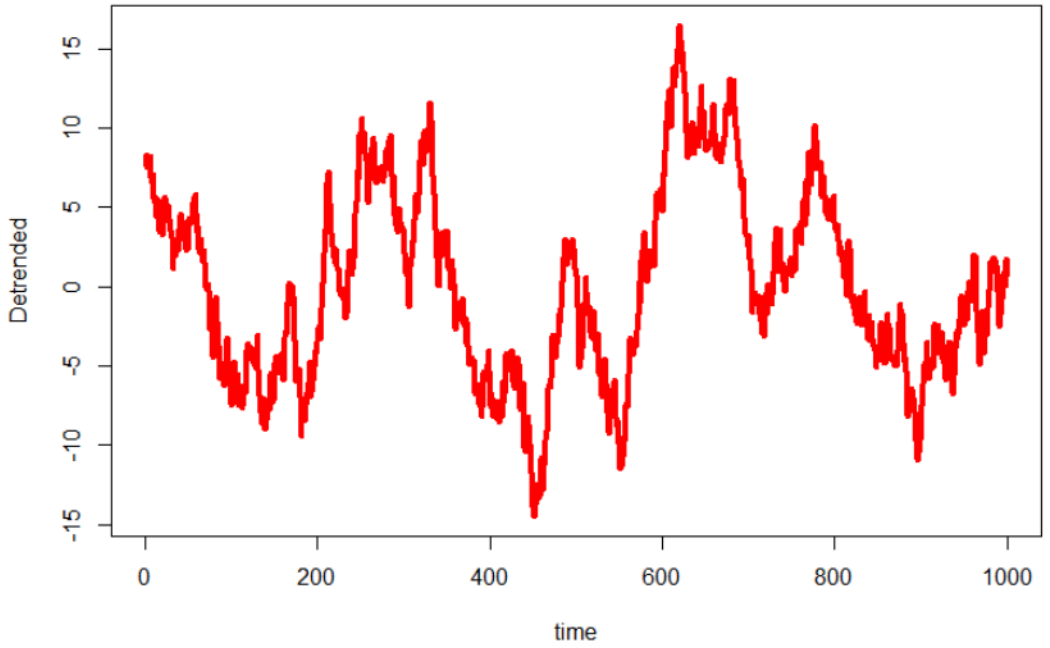




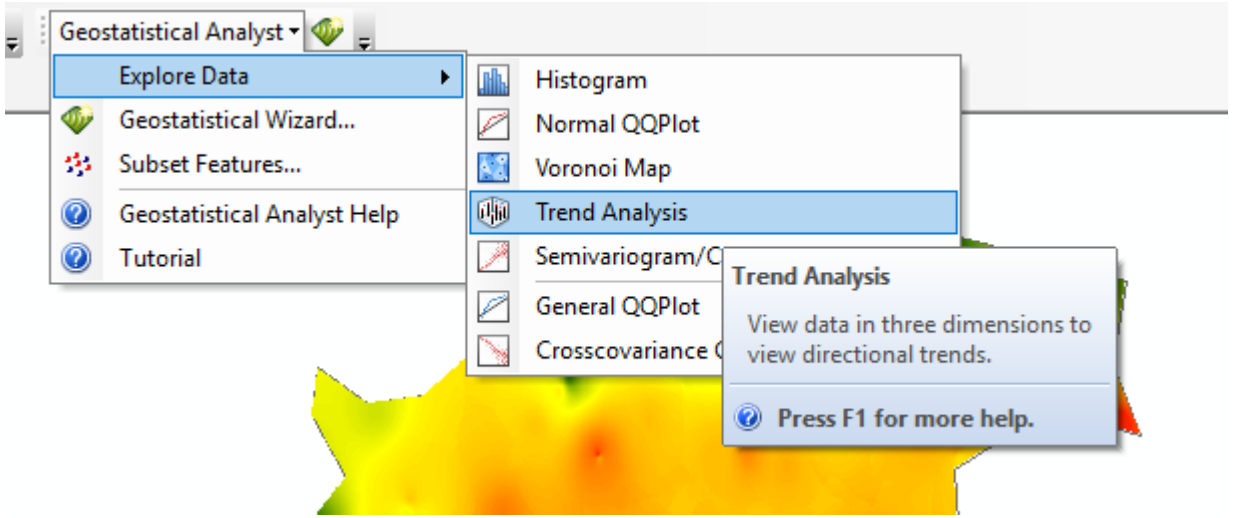
Trend

Un’altra assunzione è la Stazionarietà, ovvero che la media sia costante. Alle volte però la variabile ha una variazione lineare nello spazio, questo fenomeno è noto come Trend, ed è necessario rimuoverlo.

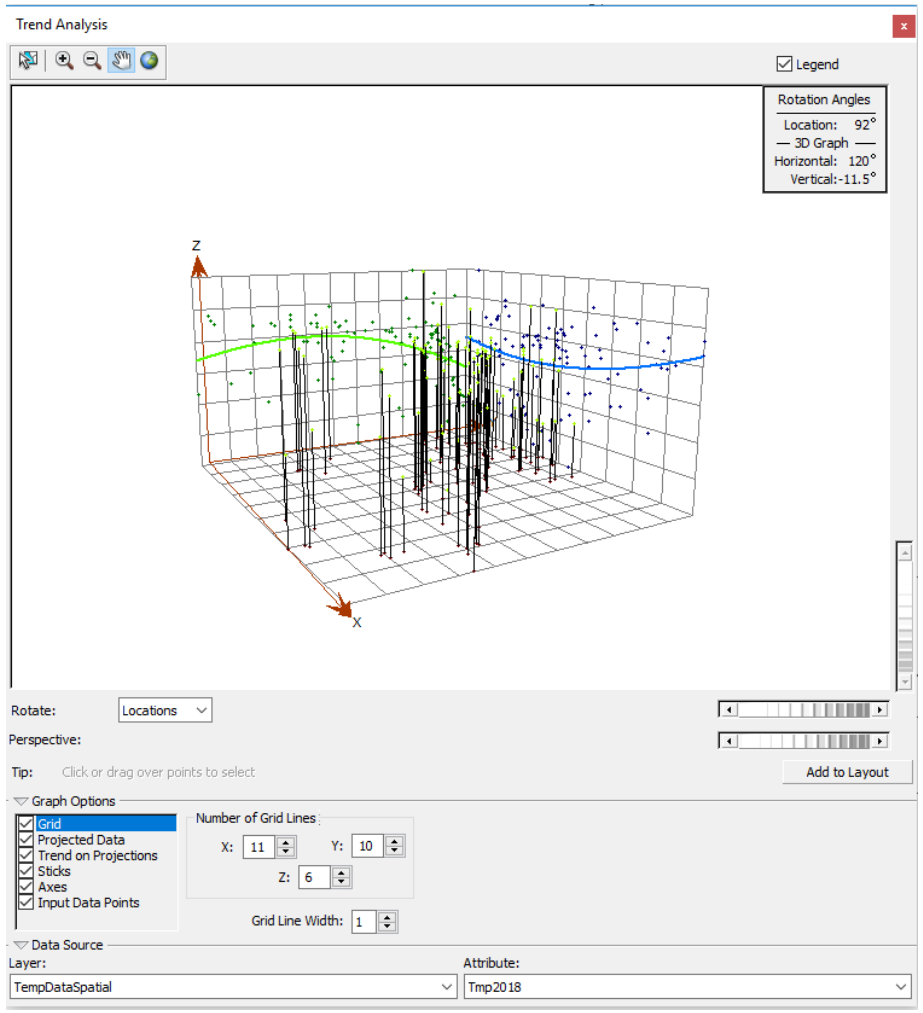




Analisi del Trend

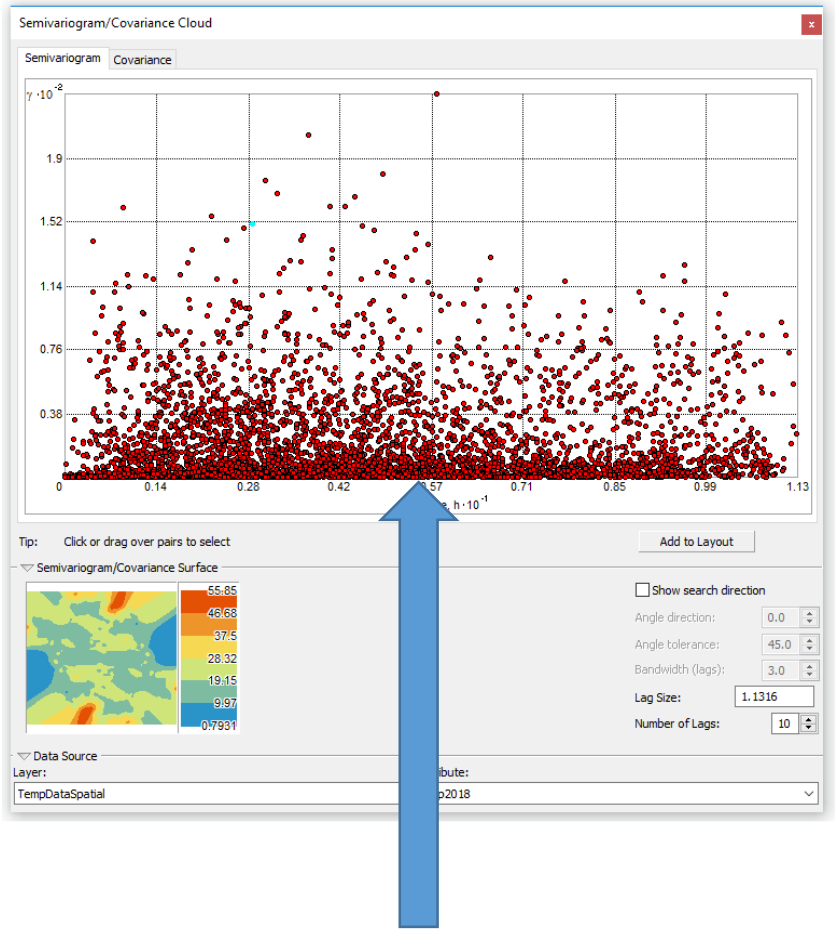


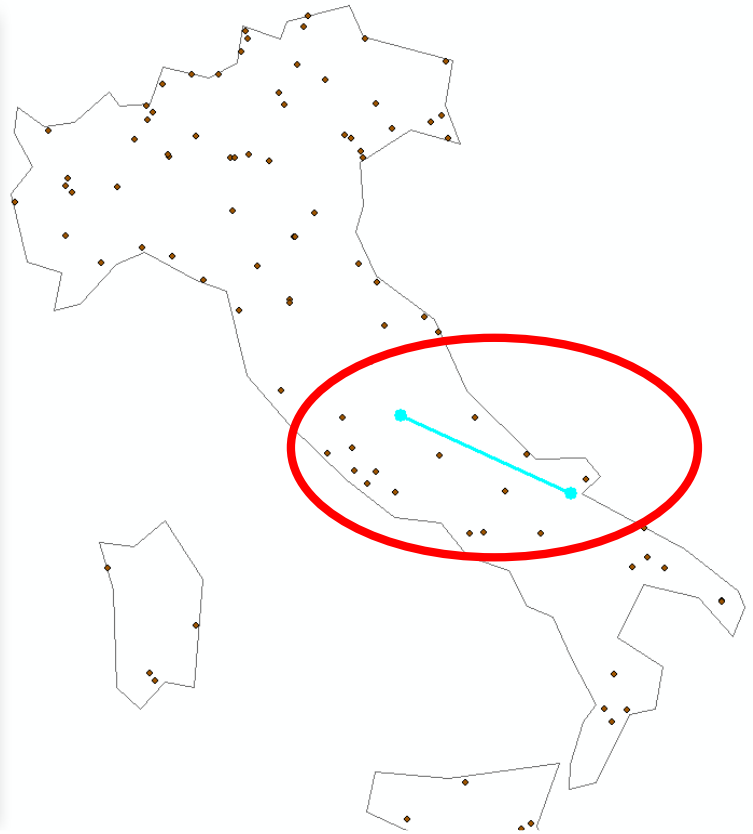
- La tool (Trend Analysis) consente di ruotare le posizioni per controllare se c'è un trend.
- Le line blu e verde suggeriscono la presenza di un trend quadratico, ed è quindi necessario rimuoverlo.



Variogramma

- É possibile capire meglio il Variogramma guardando la nuvola di punti che si genera plottando le semivarianze di tutte le coppie di osservazioni (explore data, semivariogram).
- Ogni punto rappresenta un paio di dati osservati, sull'asse X c'è la distanza (metri), mentre sull'asse Y c'è la varianza.

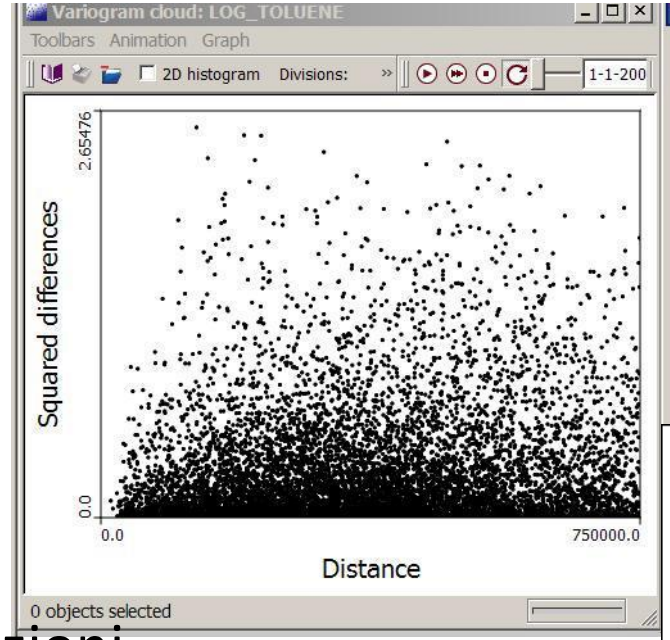


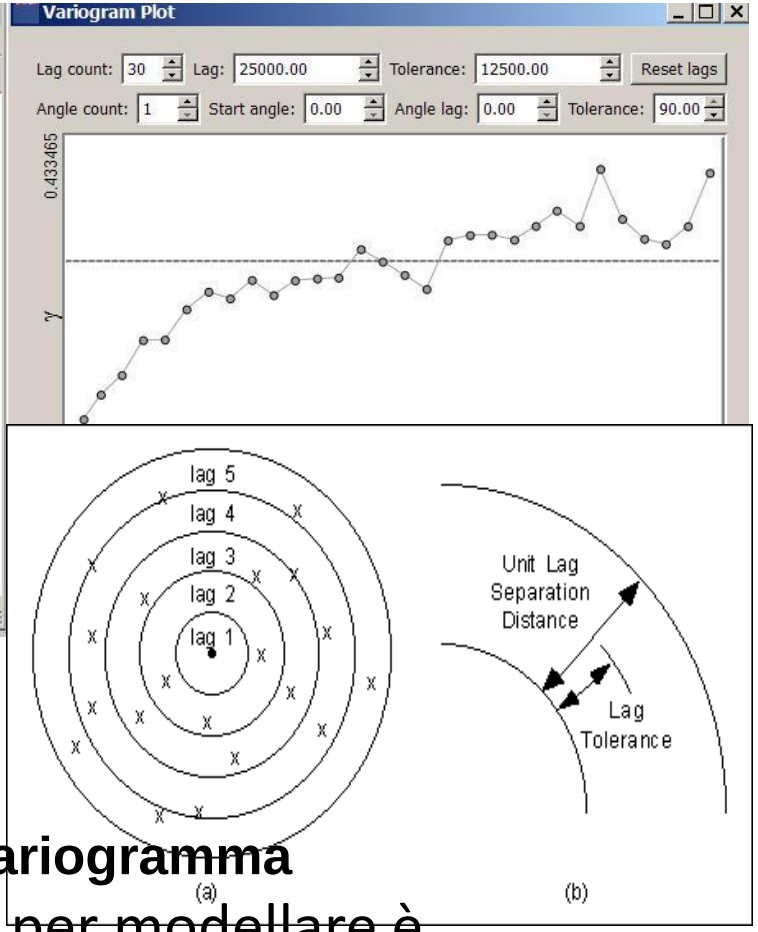


La nuvola si ottiene graficizzando le semivarianze di ogni singola coppia con i rispettivi lag (distanze)

Variogramma,

modellizzare la distribuzione spaziale





Il variogramma è semplicemente una rappresentazione sintetica (media) della nuvola di punti delle semivarianze vs distanze.

La nube di punti contiene tutte le relazioni spaziali nei dati per tutte le possibili distanze tra i campioni, ma non è una funzione continua. E' difficie interpretarla e comprendere l'esistenza di correlazioni spaziali. Si rende necessario prendere in considerazione un numero ristretto di lag (distanze). In questo modo sono più riconoscibili gli outlier ed è possibile modellizzare la distribuzione spaziale

Modello del variogramma

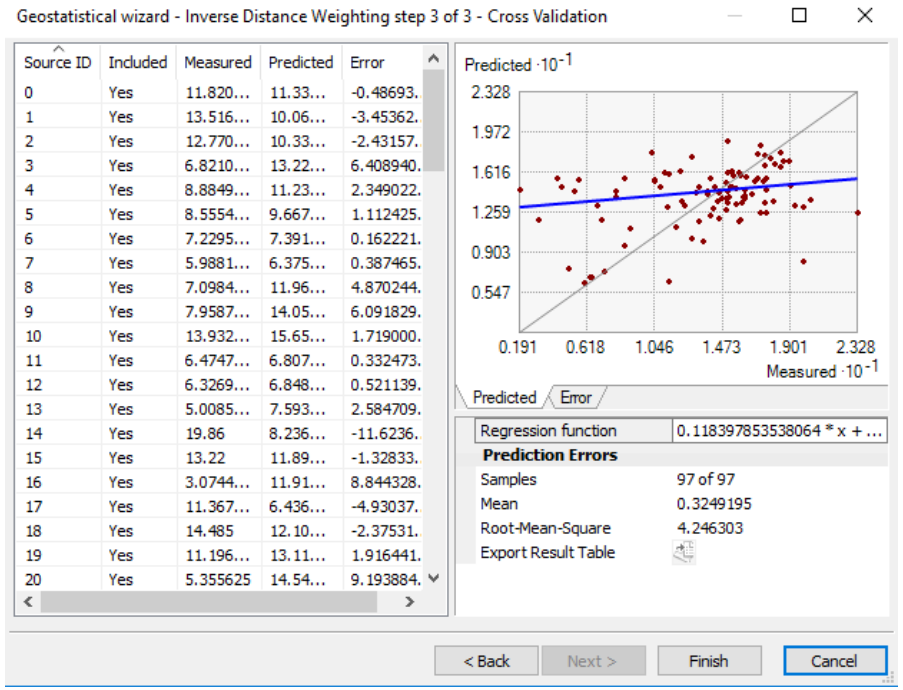
Una funzione per modellare è necessaria per ottenere un valore di varianza per ogni possibile distanza tra i dati alle posizioni note e le posizioni che devono essere predette.

Validazione incrociata o cross-validation

- Per ogni punto che avremo stimato con il kriging, conosciamo il suo valore mosurato e il suo valore predetto. I Residui possono essere positive (sovrastima) oppure negative (sottostima).

- La media dei residui può essere positva o negative indica se il modello sovrastima o sottostima.

- Il root-mean-square error (RMSE) è un indice che usa I quadrati dei residui per calcolare la differenza in vlore assolutotra dati predetti e osservati



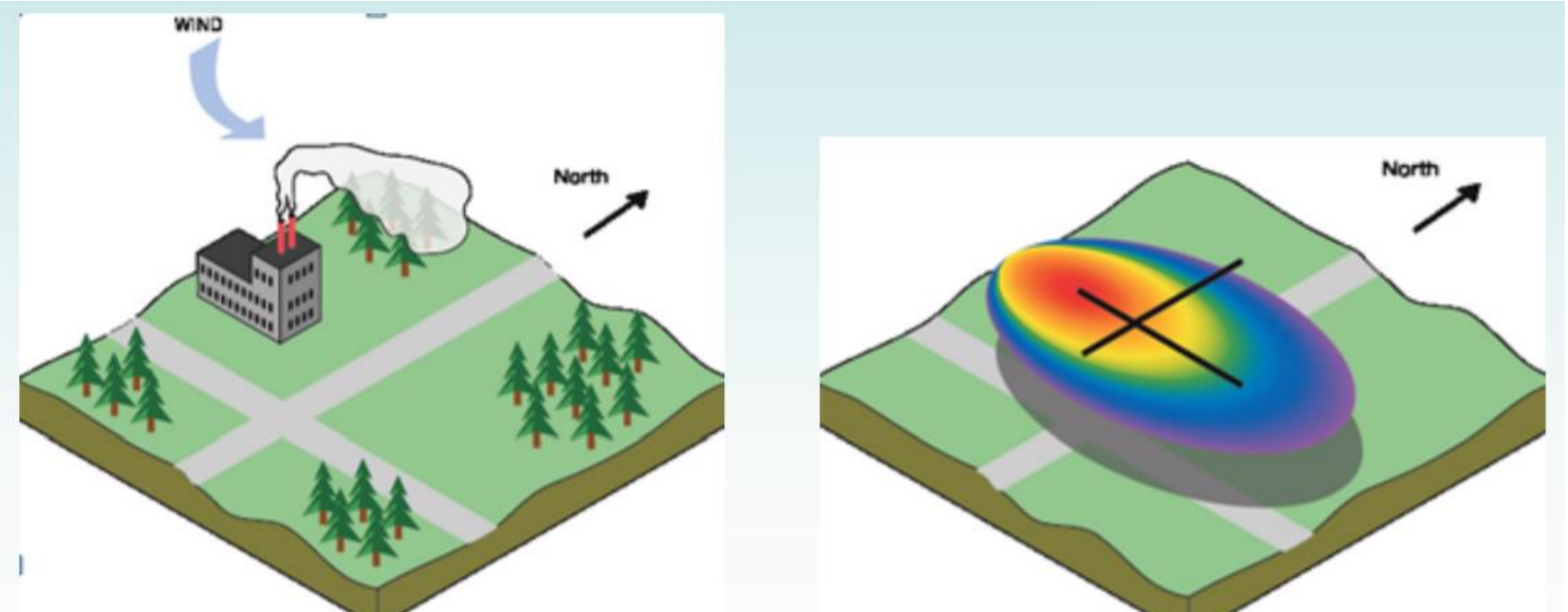
Valutazione della predizione

Una funzione integrate nella tool di ArcGIS ci fornisce il valore di incertezza su ogni punto stimato ”local uncertainty”. Questa è la chiave per valutare l’accuratezza delle mappe in regioni dove la copertura di osservazioni è scarsa

OID *	Shape *	Included	Predicted	Standard Error
1	Point	Yes	14.099718	4.15867
2	Point	Yes	14.070244	4.139364
3	Point	Yes	14.078638	4.141978
4	Point	Yes	14.087528	4.144672
5	Point	Yes	14.040341	4.127555
6	Point	Yes	14.048133	4.130099
7	Point	Yes	14.056529	4.132745
8	Point	Yes	14.065481	4.135484
9	Point	Yes	14.074941	4.138304
10	Point	Yes	14.017362	4.117932
11	Point	Yes	14.025034	4.120486
12	Point	Yes	14.033379	4.123157
13	Point	Yes	14.042348	4.125933
14	Point	Yes	14.051889	4.128804
15	Point	Yes	14.06195	4.13176
16	Point	Yes	14.072482	4.134791
17	Point	Yes	13.98679	4.10555
18	Point	Yes	13.993469	4.107958
19	Point	Yes	14.000952	4.110511
20	Point	Yes	14.009186	4.113196
21	Point	Yes	14.018118	4.116003
22	Point	Yes	14.027693	4.118919

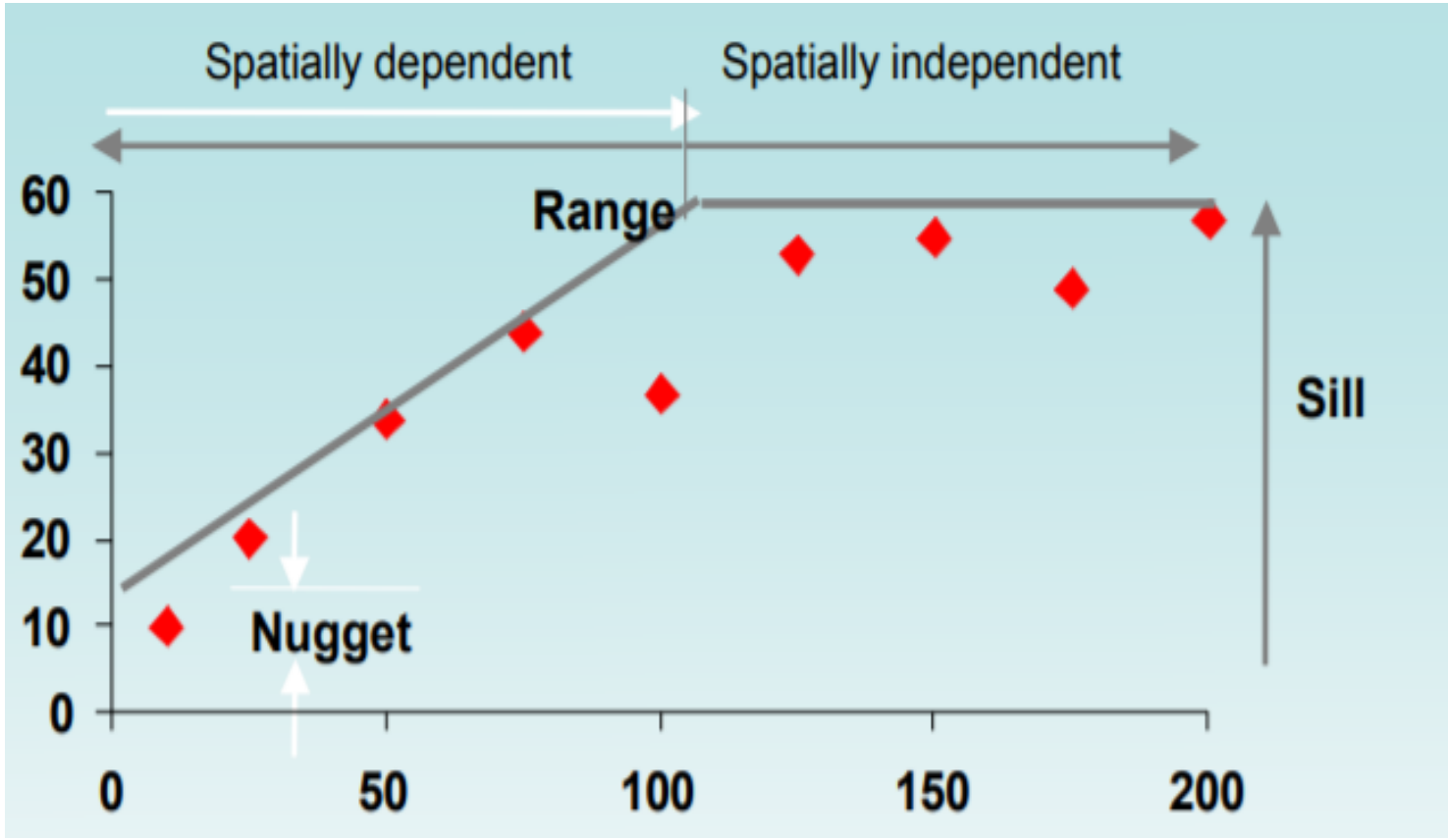
Isotropia-> la proprietà dell'indipendenza dalla direzione

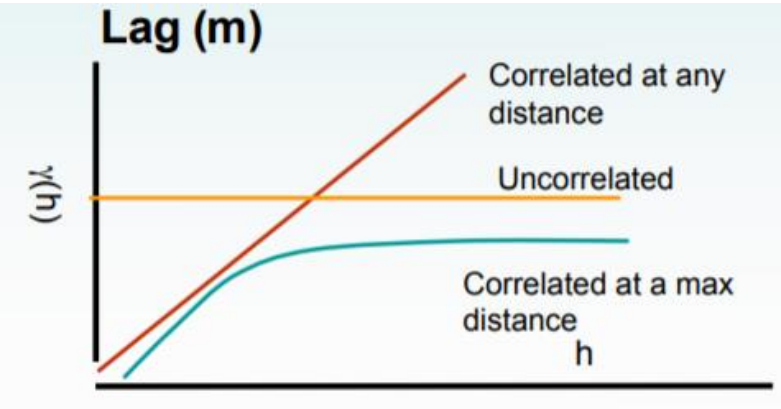
- Spesso si nota la presenza di anisotropia (l'autocorrelazione dipende dalla direzione nei dati). Esempi: inquinanti atmosferici nella direzione del vento prevalente, flussi idrici sotterranei o superficiali, ecc.
- Tenere in considerazione questa caratteristica vuol dire consentire al variogramma di cambiare forma quando incontra cambiamenti locali



Modello del variogramma

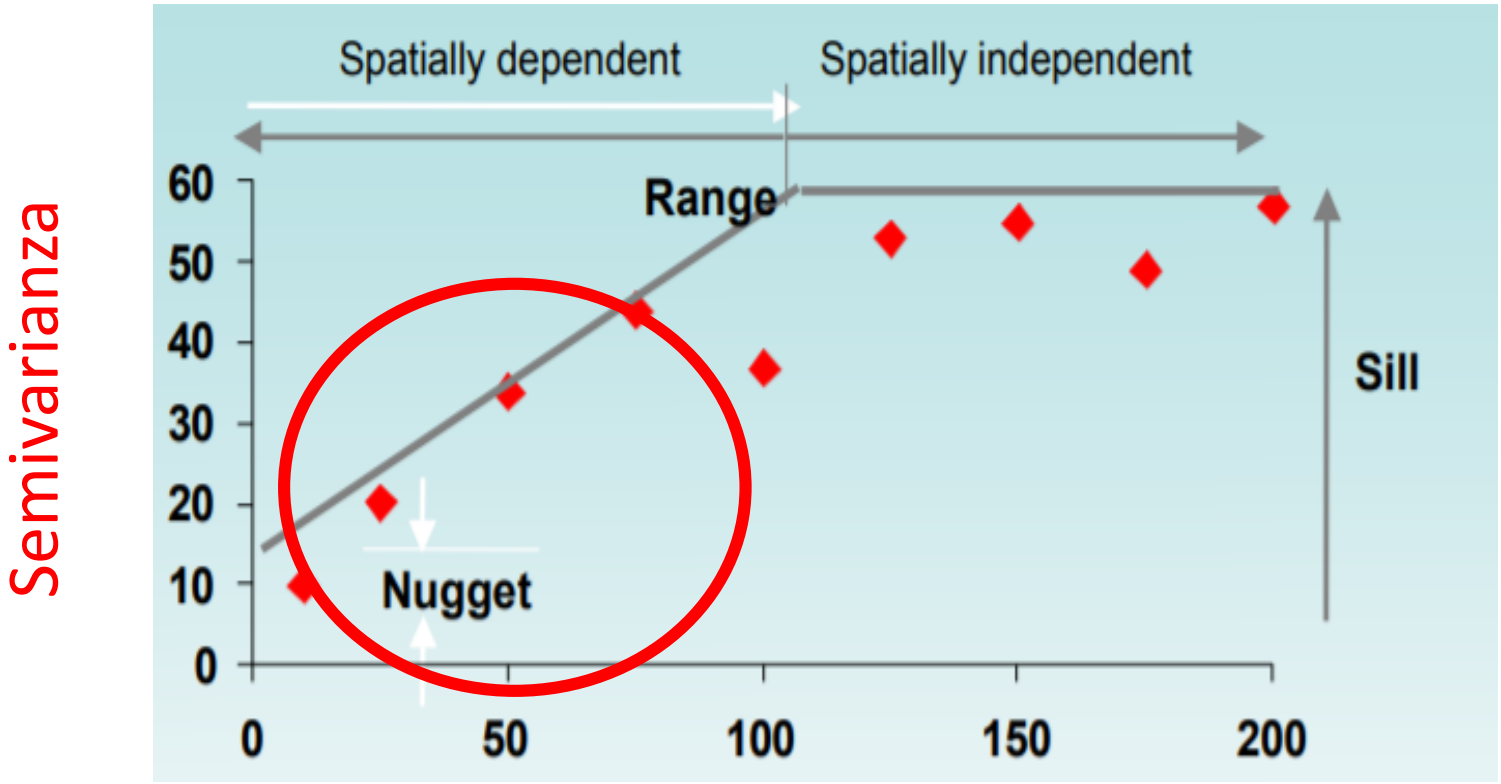
La parte più importante del variogramma è la sua forma nei pressi dell'origine poichè ai punti più vicini verrà dato un peso maggiore durante l'interpolazione





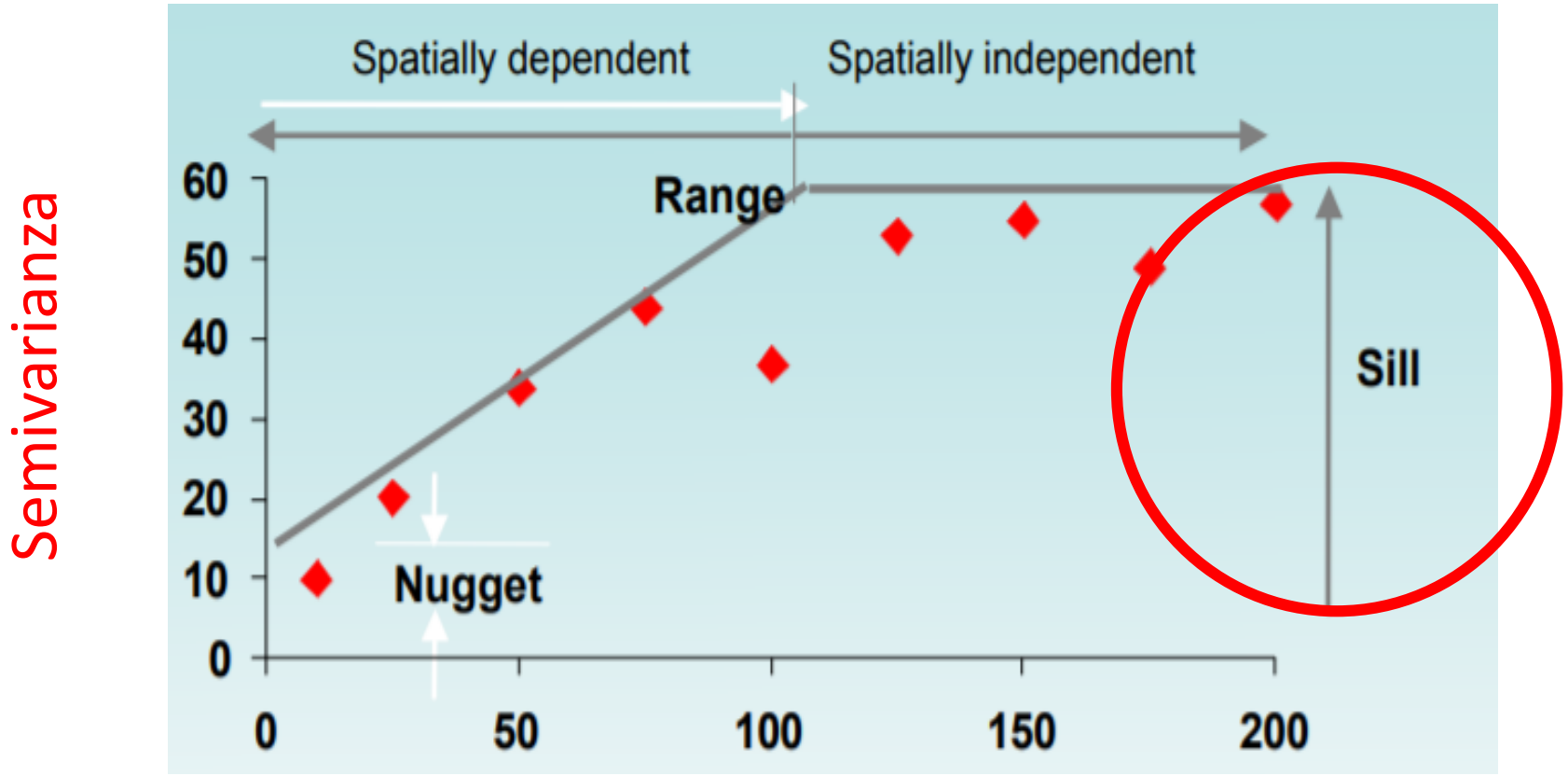
Nugget

La (semi)varianza a scala spaziale infinitesimale (variabilità di corto raggio); Rappresenta una stima del residuo, rumore spazialmente non correlato. Combina le variazioni residue degli errori di misurazione con le variazioni spaziali che si verificano su distanze più brevi del passo (lag) di campionamento.



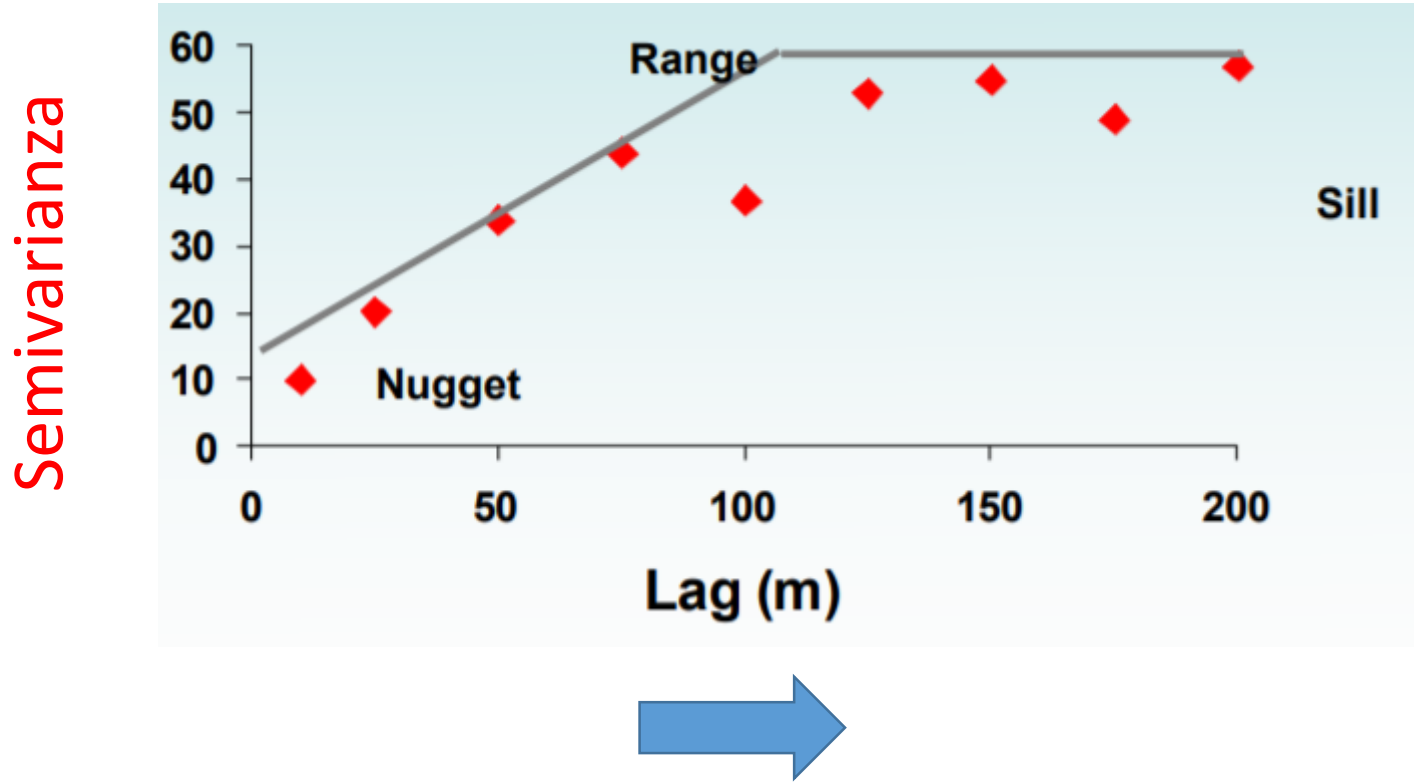
Sill

la (semi)varianza che ci si aspetta su una scala spaziale infinita (molto grande)



Lag (scelta della distanza di autocorrelazione)

La scelta del **lag** influenza fortemente il variogramma così questa distanza va definita con cura.



Lag troppo piccoli = troppe semivarianze medie (una per ogni lag) e alta probabilità di variogramma inesatto.

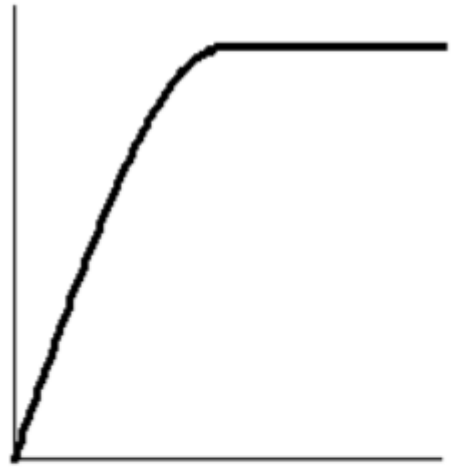
Lag troppo grandi = poche stime di varianza,perdita di dettaglio ed eccessivo smoothing del variogramma.

Scelta del modello di correlazione spaziale

- Dal punto di vista pratico la scelta del modello non presenta poi particolari difficoltà se si dispone di software geostatistici adeguati.
- Il modello da adottare viene suggerito in maniera inequivocabile dal comportamento dei dati rispetto a certe operazioni. E ciò non è casuale dato che i modelli proposti traggono spunto dalla realtà.

Scelta del modello di correlazione spaziale

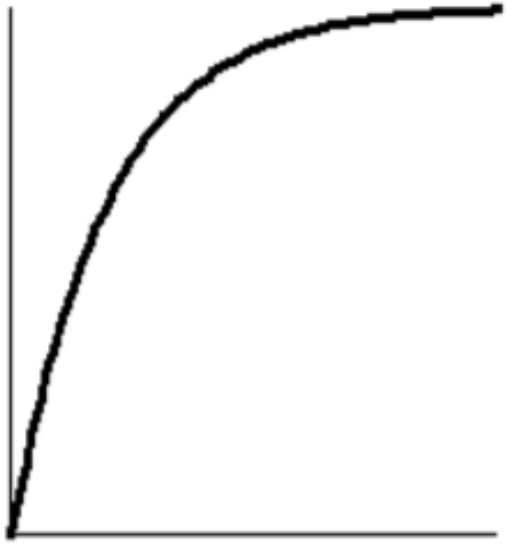
- **Sferico**: Si usa se il nugget è presente ma piccolo. E' il modello più utilizzato come default in molti software.



Spherical Model
Pannatier (1996, p. 48)

$$\gamma(h) = \begin{cases} C\left[1.5h - 0.5h^3\right] & h < 1 \\ C & h \geq 1 \end{cases}$$

- **Esponenziale**: Si usa se il nugget è rilevante e la crescita verso il sill è poco ripida.



Exponential Model
Cressie (1991, p. 61)

$$\gamma(h) = C\left[1 - e^{-h}\right]$$

Conclusioni

- La geostatistica è uno strumento utile e potente e viene utilizzata spesso nel campo della regionalizzazione di variabili continue (variabili meteorologiche, proprietà del suolo, siti contaminati).
- L'esperienza nella valutazione degli elaborati progettuali relativi a mappe di prescrizione di interesse locale e regionale 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.
- Spesso si vogliono usare modelli complessi (es. kriging) non applicabili ai dati disponibili o con limitati set di dati.
- Importante non limitarsi solo a fornire delle "mappe", ma indicare tutte le valutazioni che le hanno prodotte oltre all'incertezza dei risultati.

UNIVERSITÀ
DEGLI STUDI
DI MILANO

DiSAA
DIPARTIMENTO
di SCIENZE
AGRARIE e
AMBIENTALI

Grazie per l'attenzione

References

1- **Matheron** G. The Theory of Regionalised variables and its Applications, Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau. 1971.

2- **Webster** R, **Oliver** MA. Geostatistics for environmental scientists. Second ed. Chichester. Wiley. 2007.

3- **Chilès** JP, Delfiner P. Geostatistics: modelling spatial uncertainty. Second ed. Wiley. 2012.