GEODATI E CLUSTER ANALYSIS, DALLE IMMAGINI IN BIANCO E NERO ALLE MAPPE 3D

Federica Migliaccio, Vincenza Tornatore, Guido Minini
Politecnico di Milano – DICA – Piazza L. da Vinci, 32 – 20133 Milano
Tel. 02-2399-6507 – Fax 02-2399-6530 – e-mail federica.migliaccio@polimi.it
Tel. 02-2399-6502 – Fax 02-2399-6530 – e-mail vincenza.tornatore@polimi.it
Tel. 02-2399-6543 – Fax 02-2399-6530 – e-mail guido.minini@polimi.it

Abstract – Il termine "Geomatica" vent'anni fa non era comunemente utilizzato, o almeno non in maniera così estesa come ora. Non si parlava di "geodati", ma eventualmente di dati "*spatially distributed*", l'analisi spaziale era "analisi dati" e certamente non avremmo definito "*spatial analysis tools*" i programmi in Fortran 77 o 90 che scrivevamo ad hoc per implementare gli algoritmi di analisi dati, in base alle esigenze delle specifiche ricerche. Certamente l'aspetto della distribuzione spaziale delle informazioni era rilevante, ma non eravamo abituati a visualizzare i dati su cui lavoravamo con la stessa facilità, velocità e versatilità di rappresentazione alle quali gli strumenti GIS e i loro "*tool*" ci hanno abituato.
Per toccare con mano, come se fosse una fotografia in "*time-lapse*", l'evoluzione delle tecnologie nelle nostre discipline, abbiamo svolto un esercizio recuperando i file originali delle anomalie di gravità e delle quote ortometriche utilizzate per alcuni test di un software di cluster analysis realizzati durante il triennio 1995/98, e descritti nella Tesi di Dottorato (Tornatore, 1998). Gli stessi dati, importati in ambiente GIS dopo quasi vent'anni, ci hanno permesso di calcolare indici statistici con la ben nota immediatezza e di realizzare facilmente modelli digitali del terreno e delle anomalie di gravità per la zona studiata. In conclusione, una riflessione su quello che nelle informazioni spaziali "vediamo" più di allora (e su come lo "vediamo") grazie ai GIS.

CLUSTER ANALYSIS – GLI ANNI NOVANTA

La *cluster analysis* ha come obbiettivo quello di distinguere in un set di dati i gruppi omogenei al loro interno.
Nella seconda metà degli anni Novanta era stata sviluppata al Politecnico di Milano una Tesi di Dottorato che proponeva una nuova strategia di analisi delle immagini (Tornatore, 1998). Durante il processamento dei dati era necessario eseguire un passo preliminare di suddivisione delle immagini in aree omogenee, o *cluster*: l'algoritmo proposto era basato su un approccio di tipo statistico-probabilistico.
Da questo punto di vista la *cluster analysis* può essere considerata come la stima di una distribuzione data dalla "mixture" di due o più distribuzioni, una volta definito il numero di gruppi omogenei nei quali possono essere suddivisi i dati. In questo modo il problema si configura come la stima dei parametri che definiscono ciascuna distribuzione associata a ciascun gruppo omogeneo. Il metodo di stima sviluppato utilizzava il principio del minimo χ^2, semplificando quello della massima verosimiglianza (che avrebbe potuto dar luogo a soluzioni di notevole complessità). Le elaborazioni erano eseguite mediante un software dedicato scritto in linguaggio Fortran 90.
Dopo aver verificato il metodo sia su dati simulati che su dati reali di immagini di vario tipo (Migliaccio et al., 1998 e 1999), esso era stato applicato anche a dati spaziali di natura diversa. Fra l'altro, era stato anche utilizzato un set di dati costituito da 6677 coppie di valori di altezze ortometriche (H) e anomalie di gravità (Δg) nell'area geografica delimitata da $44° \le \varphi \le 45°$ e $10° \le \lambda \le 12°$. In Figura 1 si vede il data set completo

rappresentato sul piano (H, Δg) e il data set ridotto (3140 coppie di valori), dopo che erano state rimosse le coppie di dati che non mostravano correlazione tra quote e anomalie di gravità. Per questo set di dati "ridotti" erano stati individuati tre *cluster*, nella Figura 1 (a destra) i dati dei tre *cluster* sono distinguibili e separati fra loro da due rette.

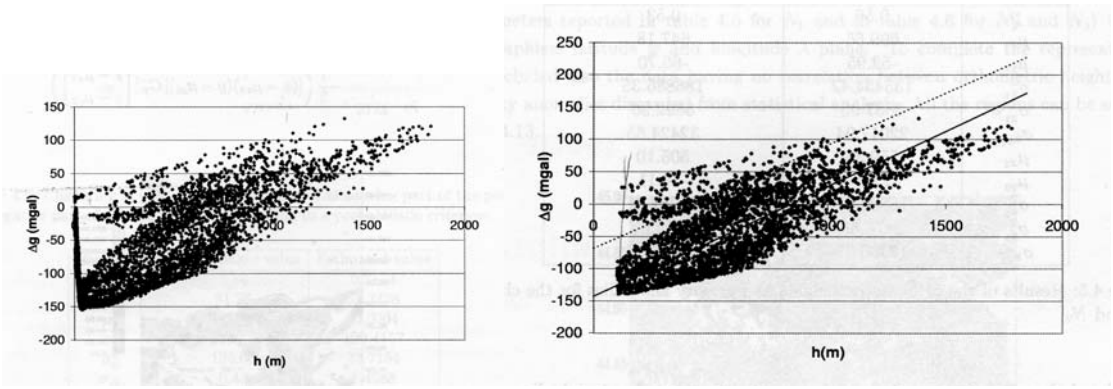

Figura 1 – Rappresentazione dei cluster nel piano (H, Δg): a sinistra, i dati completi;
a destra, dopo la rimozione dei punti che non mostravano correlazione fra quote e anomalie di gravità,
si notano i tre cluster separati fra loro da due rette

Utilizzando il software di *cluster analysis* sviluppato in Fortan 90 erano stati stimati i parametri delle tre distribuzioni normali associate ai singoli *cluster* e individuate tutte le coppie di dati appartenenti a ciascun gruppo omogeneo. Associando simboli diversi a ciascun gruppo, questi erano stati rappresentati nel piano (φ, λ) (si veda la Figura 2) insieme ai dati che erano stati eliminati (quelli con nessuna correlazione tra quote e anomalie di gravità). A suo tempo questa rappresentazione era stata confrontata con la carta della densità crostale media nella stessa area (Vecchia, 1952), ed era stato confermato che i tre cluster individuati corrispondevano a regioni di diversa densità media. Ciò aveva portato a concludere che questa strategia di analisi dei dati poteva essere applicato anche ad altre zone per distinguere regioni di densità diversa.

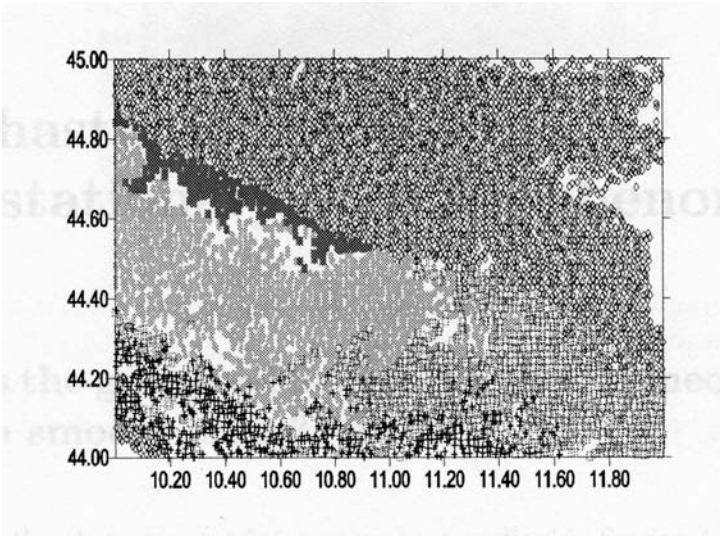

Figura 2 – Rappresentazione dei cluster nel piano (φ, λ)

CLUSTER ANALYSIS – VENT'ANNI DOPO

Utilizzare in ambiente GIS i dati analizzati così tanti anni prima mediante il software sviluppato appositamente in linguaggio Fortran 90 ha rappresentato un esercizio interessante e ha fornito lo spunto per una serie di riflessioni, che svolgeremo poi a conclusione di questo lavoro.

Attualmente quando si lavora con dati georeferenziati (termine non di uso comune a metà degli anni Novanta) l'idea è di visualizzarli in ambiente GIS e di sfruttare le funzionalità di questi tipi di software per svolgere i diversi tipi di analisi ai quali si è interessati. Nello specifico caso qui illustrato si è sfruttato l'applicativo ESRI ArcGIS 10.2, ampiamente diffuso sia in ambienti professionali che di ricerca. L'importazione dei dati in ambiente GIS permette immediatamente di inquadrare sulla cartografia la porzione di territorio di interesse grazie alla definizione di una "bounding box" (corrispondente alle coordinate geografiche riproiettate in UTM WGS84) e alla sua sovrapposizione su una base cartografica. Nel caso in esame sono stati utilizzati i dati OpenStreetMap, direttamente caricati in ambiente ArcMap, e si è ottenuta la rappresentazione mostrata in Figura 3.

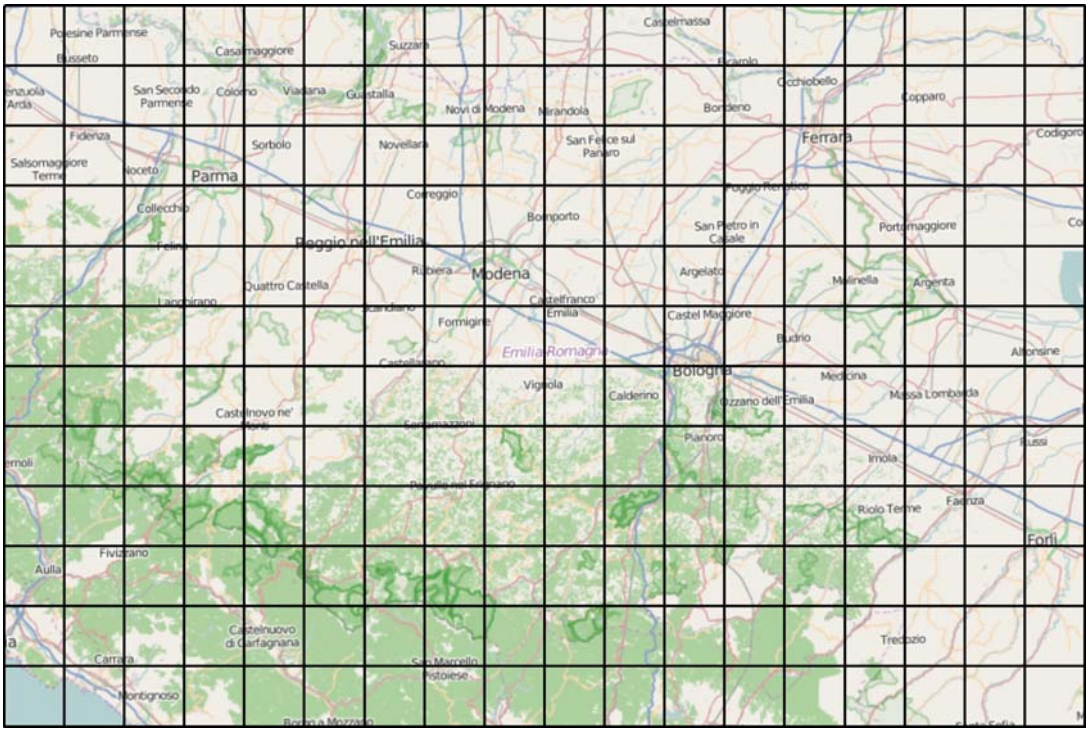

Figura 3 – Inquadramento della zona di interesse sulla base di una cartografia OpenStreetMap: la "bounding box" è suddivisa in una griglia di 10 km x 10 km

Passando alla parte analitica del lavoro, replicare la *cluster analysis* svolta in (Tornatore, 1998) e (Migliaccio et al., 1998 e 1999) non è stato possibile in quanto non è stato individuato fra gli strumenti offerti da ArcGIS uno che implementasse un algoritmo equivalente. Si è però utilizzato il *tool* "*Clusters and Outliers Analysis*" presente in ArcMap, che calcola il "*Local Moran Index*" (Anselin, 1995) che permette di individuare similarità di comportamento (raggruppamenti di dati con valori tendenzialmente "alti" o "bassi") fra dati spazialmente prossimi fra loro. Dato un set di dati questo indice evidenzia sia le zone in cui valori alti o valori bassi sono raggruppati in *cluster* sia zone nelle quali sono presenti valori che risultano molto differenti da quelli circostanti (*outlier*). Per i *dataset* delle quote e delle anomalie di gravità i risultati sono riportati nelle Figure 4 e 5. La planimetria è rappresentata da coordinate kilometriche (Est, Nord).

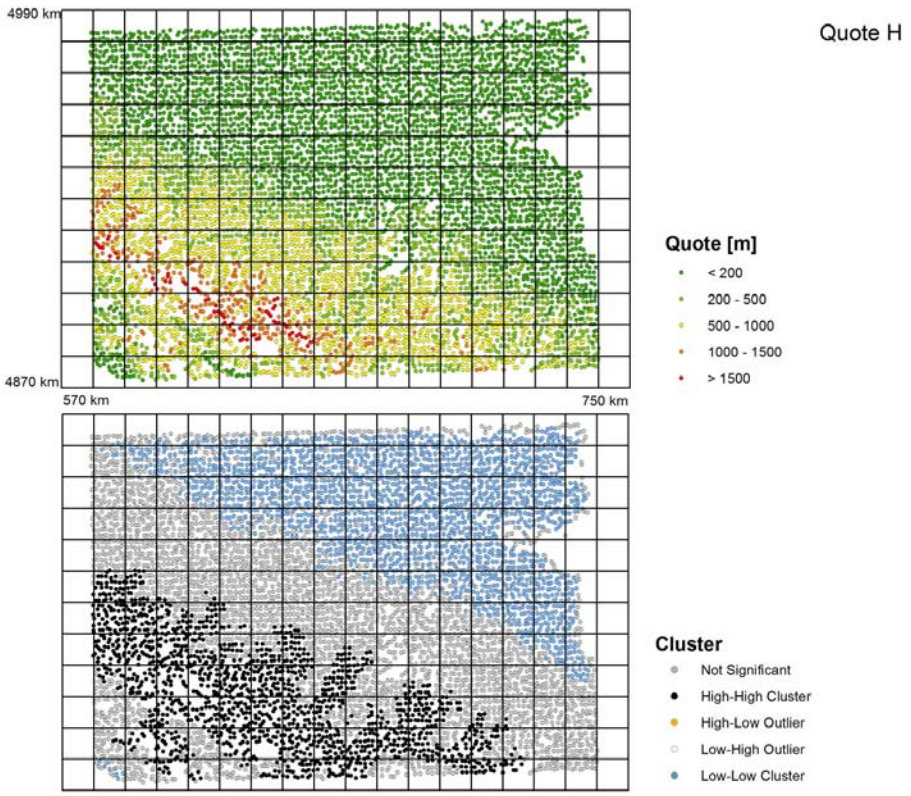

Figura 4 – Punti di misura e valori delle quote (in alto)
e risultato del *tool* "Clusters and Outliers Analysis" di ESRI ArcGIS 10.2 (in basso)

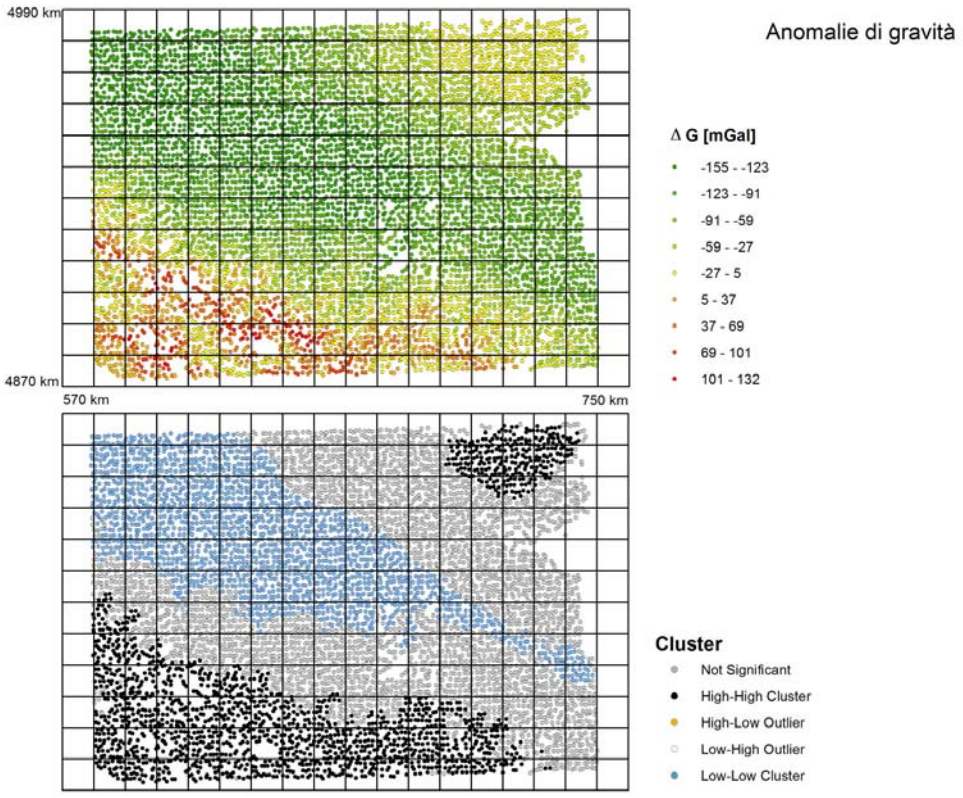

Figura 5 – Punti di misura e valori delle anomalie di gravità (in alto)
e risultato del *tool* "*Clusters and Outliers Analysis*" di ESRI ArcMap10.2 (in basso)

Poter lavorare con i dati georeferenziati in un ambiente che fornisce rapidamente visualizzazioni spaziali ha comunque consentito di ottenere quasi immediatamente (grazie a un'operazione di selezione sui dati) una rappresentazione interessante. Essa è mostrata nella Figura 6, dove sono evidenziati in colore più scuro i punti di misura corrispondenti a dati di quota e gravità che nelle analisi precedenti erano stati individuati come non correlati fra loro (e che per questo motivo erano stati esclusi dalla "vecchia" *cluster analysis*). Confrontando tali punti con la carta in Figura 3, si vede che corrispondono alla porzione di territorio pianeggiante nella quale quindi le quote si mantengono su valori pressoché uniformi, mentre le anomalie di gravità presentano significative variazioni (si vedano anche le Figure 4 e 5, dove è possibile riscontrare questo diverso comportamento dei valori delle quote e delle anomalie di gravità). Questo tipo di rappresentazione non era stata realizzata in precedenza.

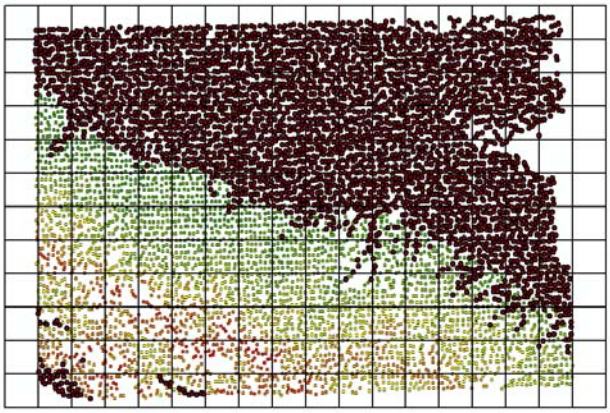

Figura 6 – Punti di misura corrispondenti a dati di quota e gravità non correlati fra loro (punti evidenziati con colore più scuro)

La possibilità di ottenere rappresentazioni cartografiche in maniera immediata come prodotto accessorio dell'esecuzione di calcoli automatici è naturalmente il grande punto di forza dei software GIS.

A questo riguardo, sono stati prodotti i modelli della superficie del terreno (DTM) e delle anomalie di gravità, interpolando rispettivamente i dati delle quote e quelli di anomalia di gravità applicando due diversi metodi, implementati nei *tool* "*IDW*" e "*Kriging*" di ArcMap. L'algoritmo IDW produce una superficie raster basata su un modello deterministico nel quale i dati sono pesati con l'inverso del valore della distanza (Inverse Distance Weighted"). Il kriging è una procedura di stima di una superficie interpolante basata su un modello stocastico dei dati e richiede di studiare preliminarmente il comportamento spaziale del fenomeno da rappresentare; tale studio è possibile grazie alla definizione della cosiddetta "semi-varianza" e alla costruzione del corrispondente "semi-variogramma". ArcMap fornisce gli strumenti per costruire un semi-variogramma, tramite la componente "*Geostatistical Wizard*" dell'estensione "*Geostatistical Analyst*". Le Figure 7 e 9 mostrano uno dei passi della costruzione dei variogrammi rispettivamente per i dati delle quote e per quelli delle anomalie di gravità. Come si vede, nella finestra di dialogo sono riportati (oltre al variogramma) diversi tipi di dati e un grafico particolare che aiuta a visualizzare eventuali comportamenti anisotropi nella distribuzione dei valori studiati.

I modelli delle superfici generati per i due tipi di dati sono poi mostrati nelle Figure 8 e 10 (sia per la procedura IDW che per il kriging). Come si vede, si tratta di classiche rappresentazioni "bidimensionali" di superfici, visualizzate grazie alle curve di livello o isoipse.

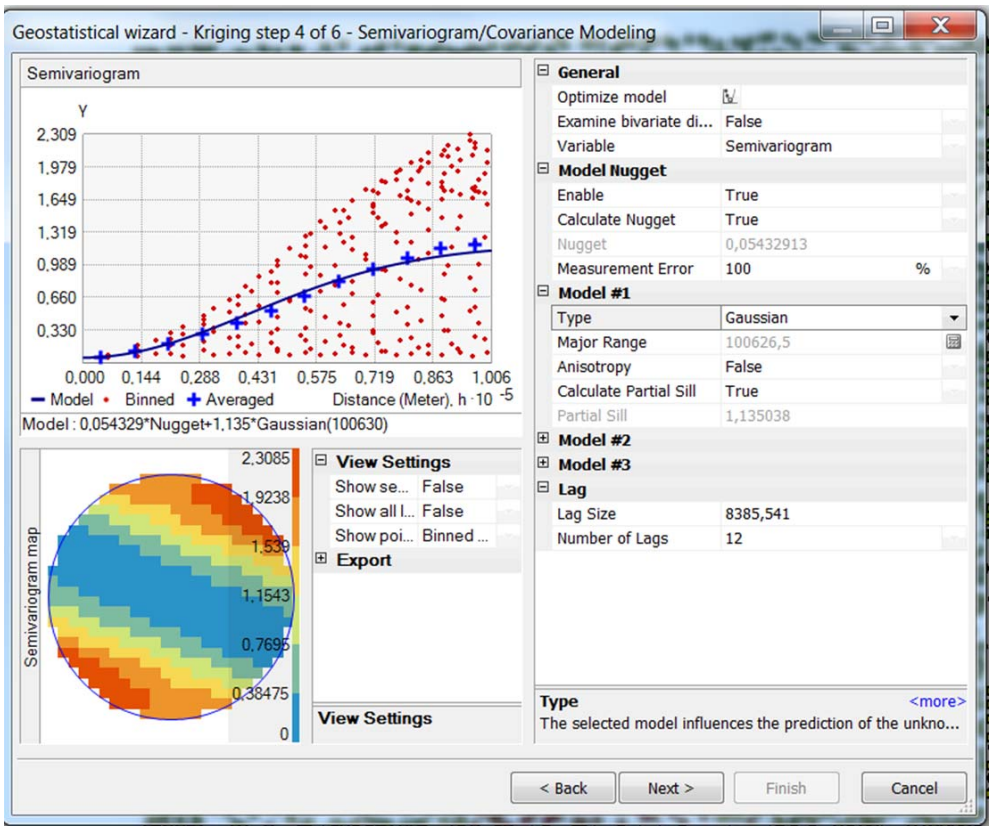

Figura 7 – Costruzione del semi-variogramma per i dati delle quote

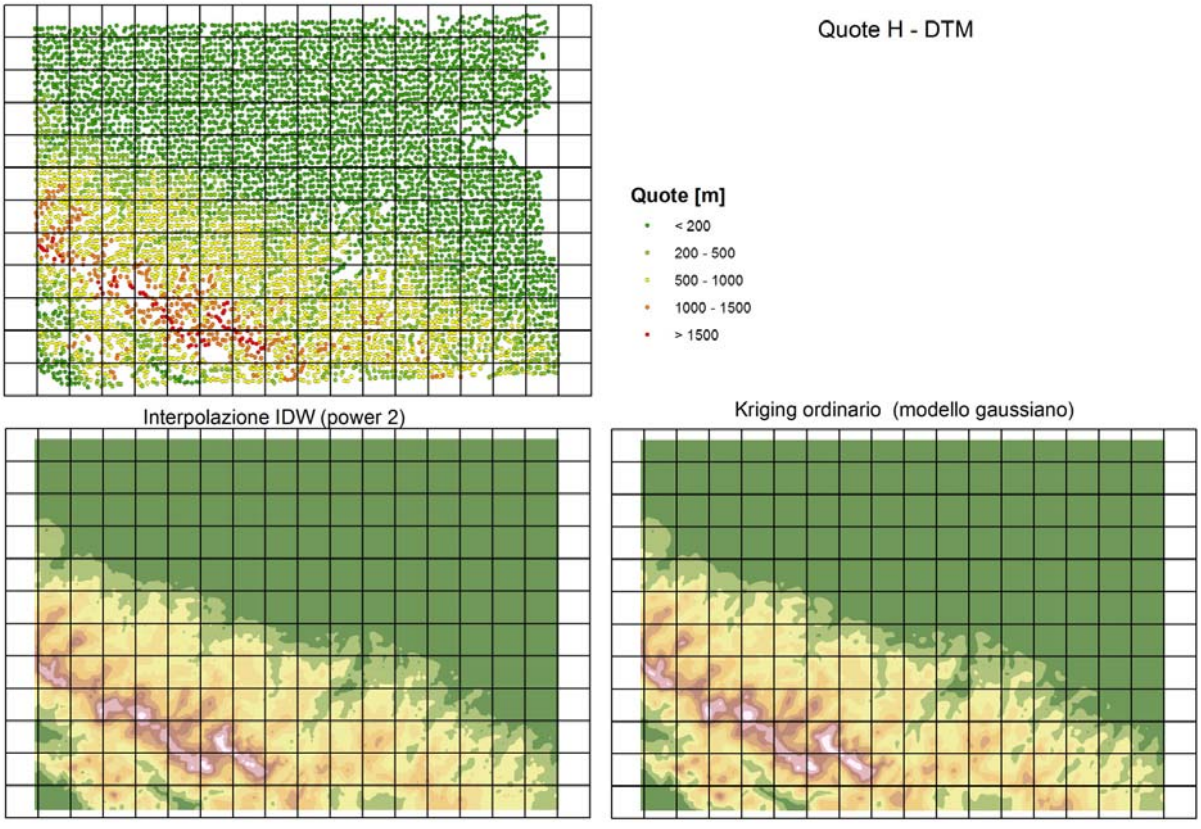

Figura 8 – Punti di misura e valori delle quote (in alto)
e DTM generati con interpolazione (media pesata) e kriging ordinario (in basso)

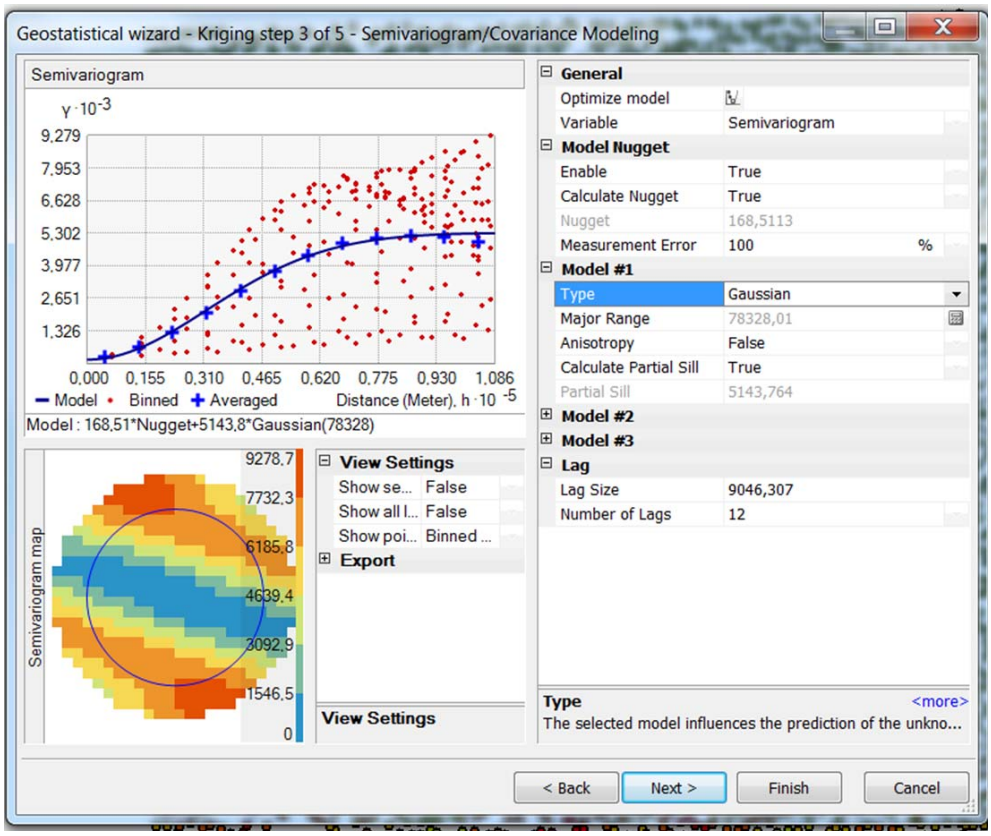

Figura 9 – Costruzione del semi-variogramma per i dati delle anomalie di gravità

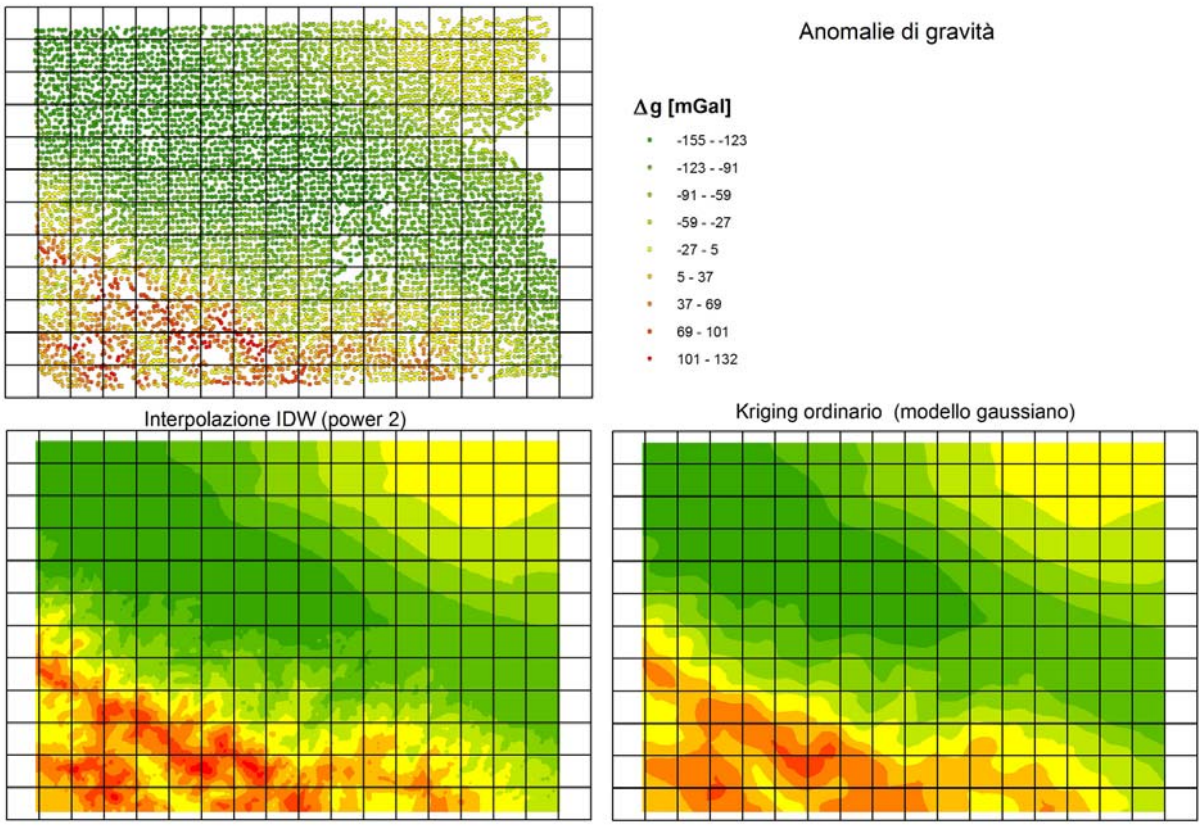

Figura 10 – Punti di misura e valori delle anomalie di gravità (in alto)
e modelli generati con interpolazione (media pesata) e kriging ordinario (in basso)

Naturalmente dal punto di vista analitico le elaborazioni sono complete, ma dal punto di vista della rappresentazione dei risultati è possibile fare un passo in più nella direzione della visualizzazione "tridimensionale" grazie al software ArcScene. Infatti, importando in ArcScene le superfici ottenute interpolando i dati, si possono realizzare delle viste 3D: in questo modo si riesce anche a sovrapporre con una certa facilità superfici ottenute tramite algoritmi diversi e ad eseguire dei confronti che, per quanto qualitativi, risultano essere molto efficaci.

Le Figure 11 e 12 mostrano tali confronti per i DTM e per le anomalie di gravità rispettivamente. E' immediato notare che nel caso dei DTM i due metodi utilizzati generano superfici praticamente indistinguibili (almeno visivamente). Nel caso delle anomalie di gravità, invece, i due metodi di interpolazione producono due superfici con aspetto nettamente diverso: l'algoritmo di kriging fornisce infatti una superficie molto più "liscia", dove i valori alle alte frequenze sono stati di fatto tagliati.

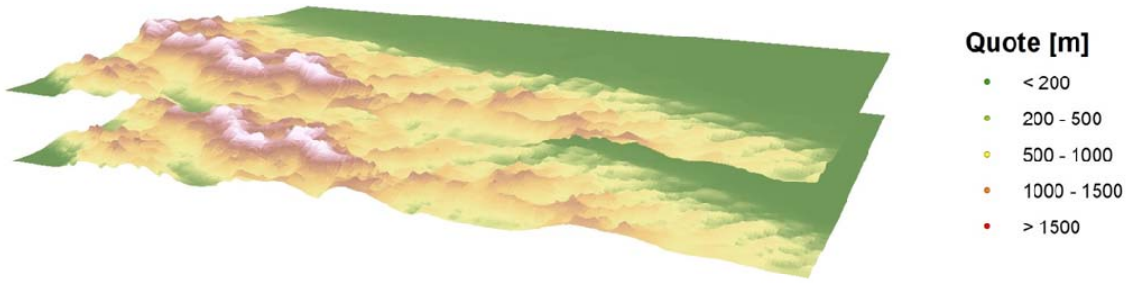

Figura 11 – Visualizzazione in ArcScene dei due DTM,
generati con l'algoritmo IDW (sotto) e con il kriging (sopra)

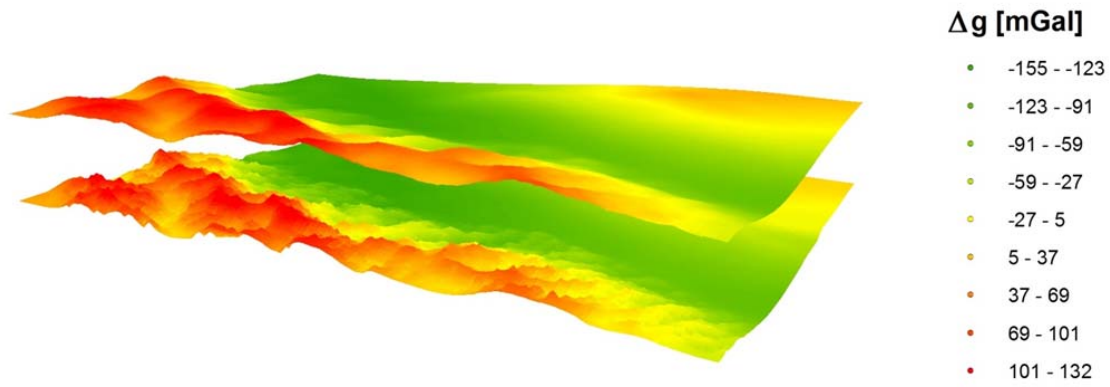

Figura 12 – Visualizzazione in ArcScene dei due modelli delle anomalie di gravità,
generati con l'algoritmo IDW (sotto) e con il kriging (sopra)

E' evidente che a questo punto sarebbe necessaria un'analisi dei risultati dei due metodi di interpolazione e in particolare dei residui e delle loro caratteristiche per comprendere il diverso comportamento dei dati, ma questo esula dallo scopo del presente lavoro. E' qui di per sé interessante notare l'efficacia grafica dell'ambiente di lavoro GIS, peraltro già abbondantemente sottolineata sia dagli sviluppatori del software sia dalla letteratura.

CONCLUSIONI

Negli ultimi vent'anni l'evoluzione delle tecnologie a supporto delle discipline cartografiche ha subito il ben noto sviluppo, accompagnato da una estesa diffusione degli strumenti GIS anche fra utenti non professionisti del settore, grazie alla loro facilità d'uso e alle interfacce intuitive. Gli aspetti di visualizzazione dei dati e di produzione di cartografia sono spesso quelli prevalentemente sfruttati, e sono quelli che abbiamo esplorato in ambiente ArcGIS su di un *dataset* di valori di quote e anomalie di gravità già utilizzati in passato per alcuni test di un software di *cluster analysis* sviluppato in Fortran 90.

Naturalmente non vi è nulla di originale (e nemmeno di insolito) nell'utilizzare l'ambiente GIS per lavorare con dati a referenza spaziale, tuttavia riteniamo che si sia trattato di un esercizio interessante perché ci ha fatto fare per così dire un balzo di venti anni nel giro di qualche ora (il tempo dedicato alle elaborazioni dei "vecchi" dati). L'efficacia del software GIS nella produzione di visualizzazioni bi e tridimensionali ci ha consentito di ottenere rappresentazioni quali quelle della Figura 3 e della Figura 6 che in passato non avevamo realizzato (essenzialmente per motivi di tempi di esecuzione, perché avrebbero richiesto uno specifico lavoro di disegno cartografico da parte di personale esperto).

Vale la pena di non perdere di vista il fatto che la produzione automatica di cartografia fornita dall'ambiente GIS richiede di lavorare sulla base di dati (sia pure con diversi livelli di immediatezza e affrontando eventuali problemi di notevole complessità, a seconda dei casi), attivando algoritmi di algebra relazionale che, per quanto "trasparenti" all'utente, sono il motore di gran parte delle funzionalità di un software GIS. Gli altri algoritmi "trasparenti" all'utente sono quelli di geometria computazionale, che svolgono il lavoro grafico alla base di qualunque operazione sulle finestre di visualizzazione dei dati. La notevole sinergia fra questi tipi di algoritmi (e molti altri) costituisce il punto di forza dei software GIS, e il motivo della loro grande diffusione e del loro successo.

La rapidità di esecuzione delle funzionalità di visualizzazione non deve tuttavia mettere in secondo piano il lavoro di analisi dei dati che ne è il fondamento, e che si suppone sia guidato dall'utente, non dall'applicativo GIS. In diversi casi peraltro si verifica che l'utente si debba "adattare" alle funzionalità e alle peculiarità dello specifico ambiente GIS (e non viceversa), soprattutto quando si utilizza software proprietario a codice non aperto. Ciò accade ogni volta che non sono disponibili *tool* specifici per le operazioni desiderate o quando le sequenze di elaborazione dei dati variano a seconda che le stesse funzionalità vengano attivate a partire da diverse barre dei menù (tutte condizioni che si sono verificate nel corso del nostro esercizio).

Tutto questo ci ricorda che vent'anni fa non eravamo "utenti" di applicativi software: in tutti i casi sviluppavamo noi stessi i software di supporto al nostro lavoro, che erano quindi adeguati alle esigenze delle specifiche ricerche (con lo svantaggio talvolta di essere troppo "adattati" e poco generali). Questo non ci vuole portare a concludere che era migliore il modo di lavorare di allora. La conclusione è che avendo a disposizione molti più strumenti, e tecnologicamente più avanzati, è opportuno usarli (e insegnare ad usarli) con consapevolezza e adeguata conoscenza degli algoritmi presenti negli applicativi.

BIBLIOGRAFIA

Anselin L., "Local Indicators of Spatial Association - LISA", Geographical Analysis, Vol. 27, No.2, April 1995.

Burrough, P. A. Principles of Geographical Information Systems for Land Resources Assessment. New York: Oxford University Press. 1986.

Migliaccio F., F. Sansò, V. Tornatore, “Clusters and probabilistic models for a refined estimation theory”. Bollettino di Geodesia e Scienze Affini, Anno LVII, N. 3, 1998.
Migliaccio F., F. Sansò, V. Tornatore, “Clusters and probabilistic models: the bidimensional case”. Bollettino di Geodesia e Scienze Affini, Anno LVIII, N. 2, 1999.
Oliver, M. A. "Kriging: A Method of Interpolation for Geographical Information Systems." International Journal of Geographic Information Systems 4: 313–332. 1990.
Philip, G. M., and D. F. Watson. "A Precise Method for Determining Contoured Surfaces." Australian Petroleum Exploration Association Journal 22: 205–212. 1982.
Tornatore V., “Analysis and denoising of images: a new strategy”. Ph.D. Thesis, Ph.D. in Geodetic and Surveying Sciences, XI Course, Politecnico di Milano, 1998.
Vecchia O., “Carta della densità sino al livello del mare nell'Italia settentrionale”, Bollettino di Geodesia e Scienze Affini, Anno XI, No. 1, pp. 337-345, 1952.
Watson, D. F., and G. M. Philip. "A Refinement of Inverse Distance Weighted Interpolation." Geoprocessing 2:315–327, 1985.