Individuazione e rimozione di errori grossolani nei Modelli Digitali di Superfici

Presso il Laboratorio di Geomatica del Politecnico di Milano - Facoltà di Ingegneria di Como - si stanno implementando, all'interno di alcuni Sistemi Informativi Geografici, dei moduli specifici per la modellizzazione digitale delle superfici (DSM): il modello digitale del terreno (DTM), rientrando completamente nella classe di superfici appena definita, è solo un esempio al quale possono essere applicati gli algoritmi e le procedure sviluppate.
I metodi e i conseguenti comandi predisposti hanno valenza del tutto generale e possono essere applicati a un qualsiasi modello numerizzato che descriva un campo di osservazioni. Da questo punto di vista la produzione e/o l'analisi di un DTM rappresenta solo un sottocaso di produzione e/o analisi di una qualsiasi carta tematica, a partire da osservazioni puntuali del fenomeno fisico considerato.
Il primo passo da prendere in esame, nell'analisi di un qualunque DSM (ovviamente la considerazione è generalizzabile ad un qualsiasi insieme di osservazioni), è la validazione dei dati disponibili, ossia un insieme di procedure di *preprocessing* che separi dall'insieme delle osservazioni i dati validi dagli eventuali outlier. Gli outlier, rifacendoci alla definizione intuitiva di Hawkins (Hawkins, 1980) sono osservazioni che si discostano così tanto dalle altre da far sorgere il sospetto che siano influenzate e/o generate da meccanismi differenti rispetto a quelli che determinano la restante popolazione statistica osservata.
L'esame approfondito dei dati anomali è imposto dalla consapevolezza che un solo outlier può rovinare completamente un insieme di misure, compromettendo, quindi, i risultati dell'analisi del fenomeno e soprattutto la precisione dei parametri che lo descrivono.
La nozione di dato "valido" si riferisce, ovviamente, sia al dispositivo che lo ha generato, sia alla capacità del dato di descrivere la realtà fisica a cui si rapporta. Un valore non valido può essere sia causato da un mal funzionamento dello strumento utilizzato per la misura che da una perturbazione del processo di generazione del dato (ad esempio cattivo campionamento, errore di archiviazione, ecc.).
La difficoltà sta nel fatto che, non conoscendo il dato "reale", ma solo sue osservazioni (affette quindi da errore), per capire se è affidabile o meno, l'unico criterio è quello di confrontare il dato in esame con quello che ci si aspetta di osservare a partire da un certo modello che ci siamo costruiti della realtà. Ad esempio nel caso del DTM si ipotizza che:

- il DTM sia una superficie regolare che non presenta rilevanti discontinuità;
- le altezze dei punti siano correlate a quelle dei punti vicini, ma siano indipendenti da quelle di punti lontani.

La seconda ipotesi ci permette di costruire test statistici per l'individuazione degli outlier che si basino su "tecniche di localizzazione", cioè sul confronto, nello stesso punto, di un osservazione con il valore predetto, partendo solo dalle osservazioni circostanti. La prima ipotesi ci permette di scegliere, come modelli che interpolino i dati, delle funzioni "semplici": ad esempio polinomi di grado basso o, in un test basato sul *kriging*, funzioni di covarianza o variogrammi dei dati di tipo sferico.
Il problema specifico, legato ai DTM, è la numerosità dei dati (dal milione al centinaio di milioni), che rende necessario studiare delle procedure automatizzate e veloci per poter processare in tempi rapidi le osservazioni, individuando i dati sospetti, che saranno in generale un sottoinsieme (esiguo) del campione totale.
Una volta individuati gli outlier si dovrà procedere ad un analisi particolareggiata che verifichi se l'anomalia del dato rientra nella normale variabilità (eventualmente come caso estremo) o se si tratta effettivamene di un errore.
L'ultima considerazione, relativamente all'insieme delle procedure messe a punto presso il Laboratorio di Geomatica, riguarda il problema relativo a test da eseguire su stimatori robusti delle osservazioni. Per chiarire di cosa si tratta partiamo dalla constatazione che, in generale, quando si esegue una qualsiasi interpolazione dei dati per predire il valore in un punto differente dai punti di osservazione, si utilizza il metodo statistico dei minimi quadrati. Questo metodo risente però fortemente della presenta di outlier; si dice quindi che il metodo è poco robusto, cioè le stime che deduciamo sono distorte se nell'insieme delle osservazioni sono presenti degli errori grossolani.
Un semplice esempio è costituito dal caso della media. Basta un errore grossolano nell'insieme dei dati per avere una media anche molto distorta; infatti, se, ad esempio, osservando una certa grandezza si ottiene 99 volte 0 e una sola volta 100.000, la media è pari a 1.000! Viceversa un semplice stimatore robusto è la mediana, cioè il valore argomentale di una variabile casuale monodimensionale per il quale i valori minori e quelli maggiori hanno pari probabilità. Una stima della mediana si ottiene ordinando i valori di un campione estratto dalla variabile casuale e scegliendo quello centrale, nel caso di campione dispari, oppure un qualunque valore compreso tra i due valori centrali per un campione pari. Rifacendoci all'esempio precedente, la mediana è in questo caso uguale a 0 e quindi non distorta.
Il test per l'identificazione di outlier mediante stimatori robusti si basa proprio sulla mediana; data la complessità del test, per una descrizione approfondita si rimanda a Brovelli et all, 1999.
Le procedure descritte sono state introdotte come nuovi comandi nel GIS GRASS. I modelli digitali sinora sottoposti a validazione sono: il modello digitale italiano, con risoluzione 7.5" in latitudine per 10" in longitudine, parte del modello digitale IGMI, con risoluzione 100 m per 100 m, il modello digitale svizzero, con risoluzione 25 m per 25 m (queste ultime elaborazioni sono state eseguite presso l'Istituto di Geodesia e Fotogrammetria dell'ETH di Zurigo).
Il primo DTM considerato ha una storia più articolata. Si tratta, infatti, di un prodotto che, inizialmente in uso presso il Servizio Geologico Nazionale, ha conosciuto una lunga fase di integrazioni: prima che il D.I.I.A.R. (Dipartimento di Idraulica, Ingegneria Ambientale e Rilevamento del Politecnico di Milano) e il Gruppo di Geomatica della Facoltà di Ingegneria di Como del Politecnico di Milano lo rendessero disponibile riveduto e corretto, era privo delle misure batimetriche e di alcuni dati, sia interni alla penisola sia ai confini con Svizze-

ra, Austria, Francia e Slovenia. Un ulteriore contributo alla eterogeneità delle misure è dato dal fatto che alcuni valori di quota dei laghi sono riferiti al pelo dell'acqua, mentre altri alla quota condensata (la massa equivalente); infine, alcuni punti sono tratti da un DTM mondiale di produzione americana (GTOPO30).

Il materiale usato per costruire il DTM originale comprendeva tavolette 1:25.000 dell'Istituto Geografico Militare Italiano e fogli 1:100.000 dell'Istituto Idrografico della Marina. Da queste carte sono stati digitalizzati i punti quotati e le curve di livello, con algoritmi di interpolazione; da tali dati è stato poi ricavato il grigliato di altezze del DTM. I valori sono quote ortometriche medie, ciascuna delle quali è relativa ad un'area di 7.5" in latitudine per 10" in longitudine (approssimativamente 230 m per 230 m).

L'estensione del DTM in coordinate (φ,λ) è

$$34.05103414° < \varphi < 48.99895843°$$
$$3.998168142° < \lambda < 21.00094186°$$

e corrisponde ad una matrice di 7176 righe e 6122 colonne (oltre 4 milioni di dati).

Nelle figure seguenti (Fig. 1, 2 e 3) si riportano rispettivamente la carta del DTM, la relativa carta delle pendenze e la carta delle esposizioni (cioè la direzione di massima pendenza).

Fig. 1 - Il DTM italiano con risoluzione 7.5" in latitudine per 10" in longitudine

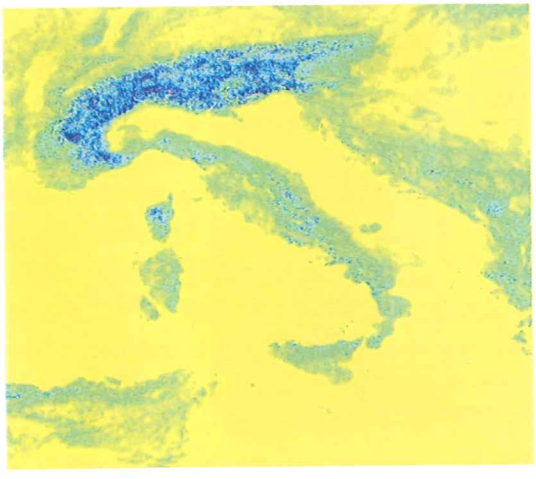

Fig. 2 - La carta delle pendenze relativa al DTM precedente

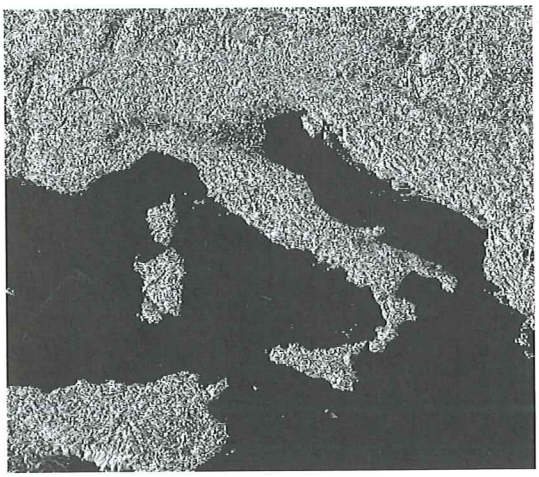

Fig. 3 - La carta delle esposizioni relativa al DTM precedente

Applicando la procedura per l'individuazione di outlier, si sono evidenziati 385 valori anomali (Fig. 4), molti dei quali sono risultati poi errori grossolani. Tali valori sono stati sostituiti con quelli predetti via interpolazione polinomiale.

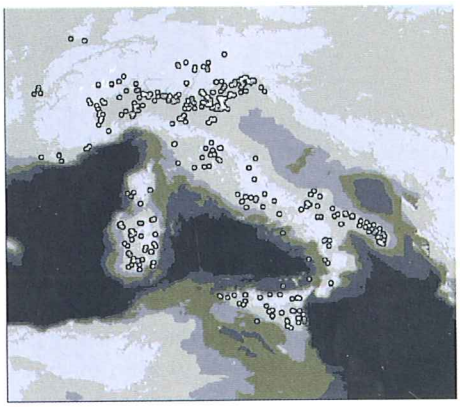

Fig. 4 - Validazione sul DTM precedente mediante interpolazione bilineare, finestra 5x5, e livello di significatività del test a = 0.0001%

Un'analisi particolareggiata del DTM ha inoltre evidenziato che le procedure di interpolazione, adottate per produrre la griglia di quote, sono state probabilmente applicate tavoletta per tavoletta: infatti in alcune zone è evidente, dall'analisi della carta delle esposizioni, il bordo delle tavolette stesse (si vedano le Fig. 5, 6, 7).

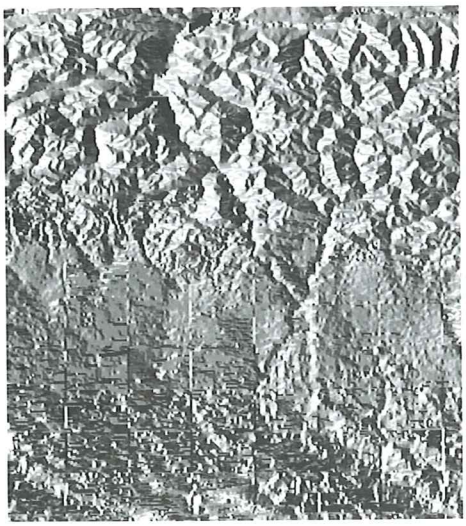

Fig. 5 - Carta delle esposizioni nella Pianura padana centrale

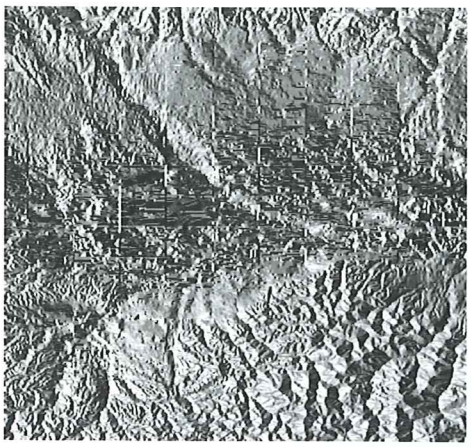

Fig. 6 - Carta delle esposizioni nella Pianura padana occidentale

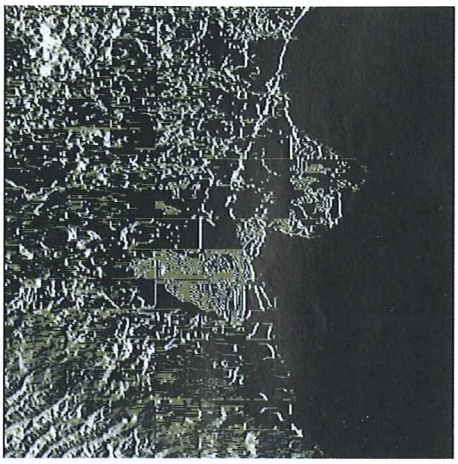

Fig.7 - Carta delle esposizioni nella zona del delta del Po.

Il secondo DTM analizzato è stato quello nazionale svizzero con risoluzione 25 metri; in questo caso si è eseguita un'analisi più elaborata in quanto gli scarti tra valori osservati e predetti mediante il modello sono piccoli: l'analisi dimostra l'esistenza di altezze da considerare solo debolmente come valori anomali. È stato quindi deciso di applicare un'analisi specifica solo su quei dati che fossero contemporaneamente probabili outlier secondo metodi diversi (interpolazione polinomiale o superficie mediana) e con diversi livelli di significatività del test (Fig. 8).

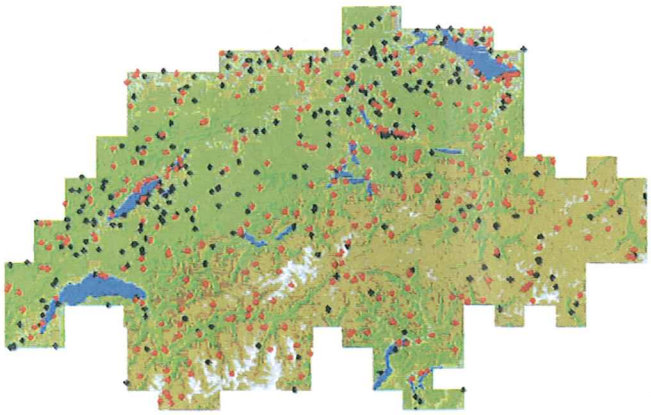

Fig. 8 - Punti con quote anomale secondo il metodo polinomiale e secondo la superficie mediana.

In questo modo ci si è ridotti ad un esiguo numero di dati da studiare dettagliatamente: il test definitivo per decidere se si trattasse o no di outlier è stato condotto analizzando la carta raster digitale svizzera (con risoluzione di 1 m) nei punti di coordinate corrispondenti ai valori anomali sospetti.
Nelle Fig. 9, 10, 11 e 12 si riportano alcuni dei dati anomali rilevati (il valore anomalo corrisponde al centro della circonferenza rossa). Come si può notare ognuno di questi casi trova una giustificazione più che plausibile. Gli unici errori rilevati nel DTM svizzero sono state due strisce di quote errate su un lago. Globalmente gli altri valori anomali rientrano nelle categorie illustrate nelle figure e non possono essere, quindi, considerate errori.

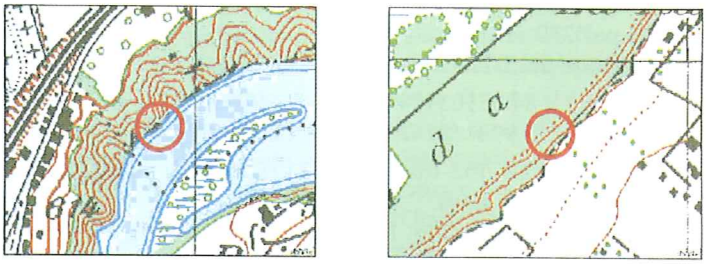

Fig. 9 - Particolari della carta digitale svizzera in punti corrispondenti a sospetti outlier nel DTM

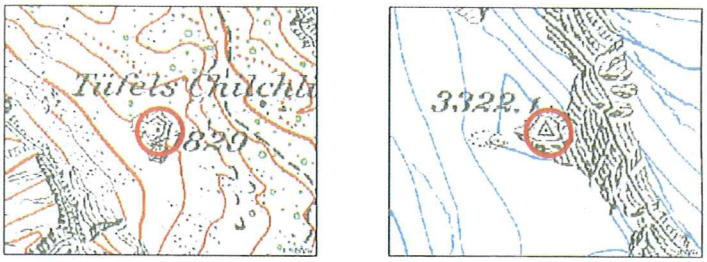

Fig.10 - Particolari della carta digitale svizzera in punti corrispondenti a sospetti outlier nel DTM

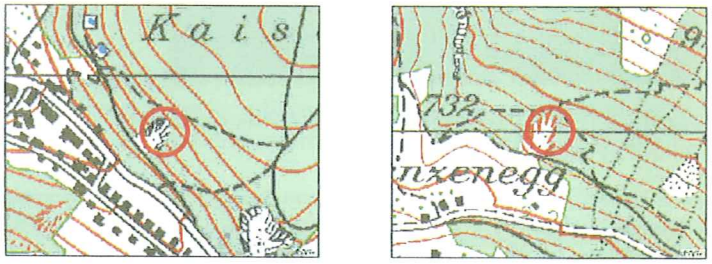

Fig. 11 - Particolari della carta digitale svizzera in punti corrispondenti a sospetti outlier nel DTM

Fig. 12 - Particolari della carta digitale svizzera in punti corrispondenti a sospetti outlier nel DTM

Infine l'ultimo DTM, analizzato solo in parte - nella regione dell'Etna -, è il modello digitale con risoluzione 100 m dell'IGMI. Nelle figure 13, 14 e 15 si riportano la carta delle altezze, delle pendenze e delle esposizioni relative alla regione considerata.

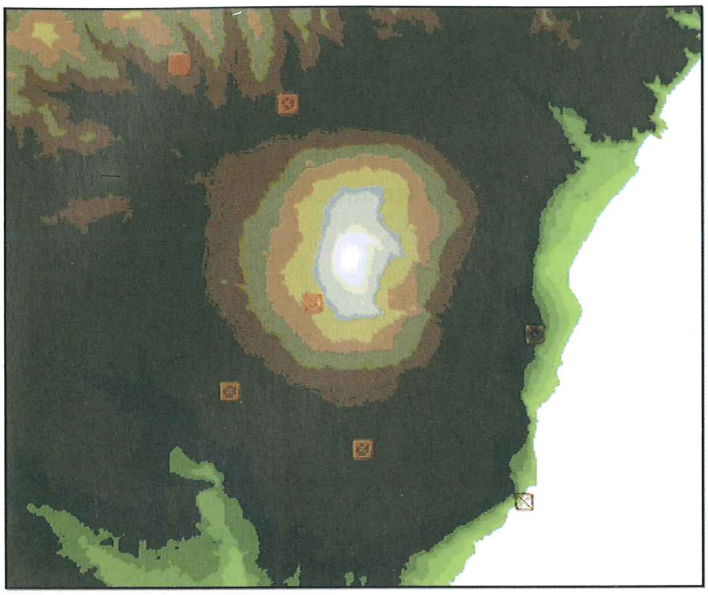

Fig. 13 - Carta delle altezze IGMI con risoluzione 100 m nella zona dell'Etna.

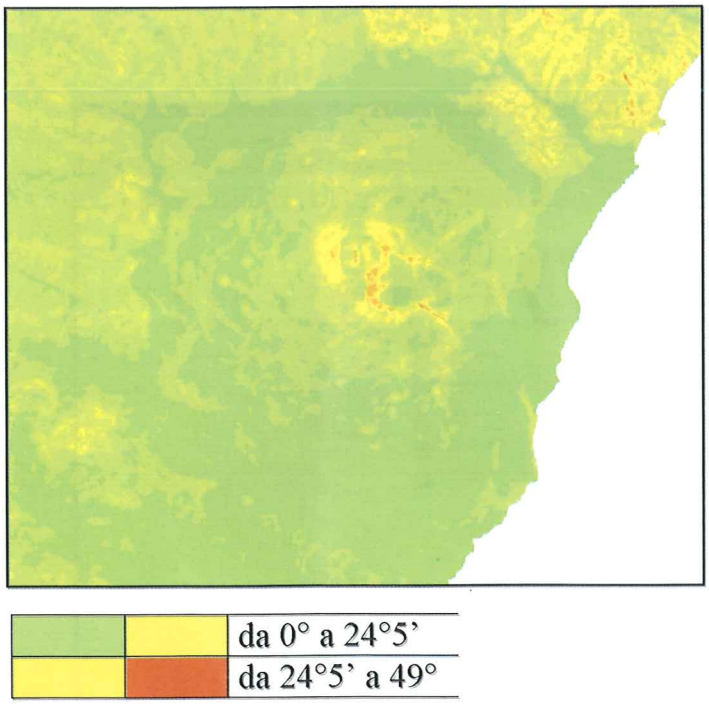

Fig. 14 - Carta delle pendenze IGMI con risoluzione100 m nella zona dell'Etna.

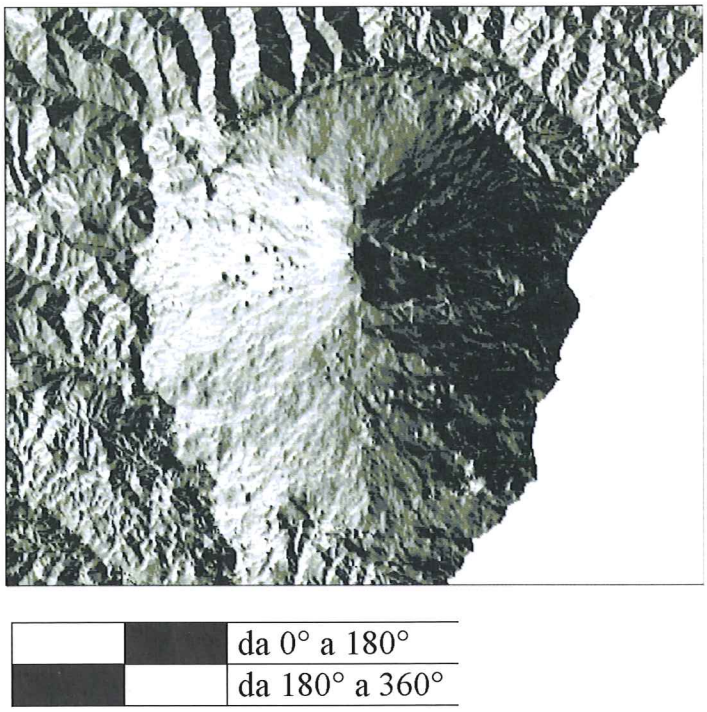

Fig. 15 - Carta delle esposizioni IGMI con risoluzione 100 m nella zona dell'Etna.

Un'analisi dettagliata dei pochi sospetti outlier (visibili in figura 13) evidenzia come il DTM non presenti, in tale regione, errori isolati grossolani.
Un'analisi sistematica di questo tipo è auspicabile che venga effettuata su tutto il territorio nazionale, possibilmente con l'ausilio di strumenti di confronto di buona qualità, così come è stato possibile nel caso del DTM svizzero.

MARIA ANTONIA BROVELLI
Politecnico di Milano
Facoltà di Ingegneria di Como

RIFERIMENTI BIBLIOGRAFICI

M. A. Brovelli, F. Sansò, D. Triglione, *Different Approaches for Outliers Detection in Digital Terrain Models and Gridded Surfaces within the GRASS Geographic Information System Environment*, Atti del Covegno DMGIS'99, Pechino, 4-6 ottobre 1999, pp.1,8.

D.M. Hawkins, *Identification of Outliers*, Chapman and Hall, 1980.