A COMPARISON BETWEEN TWO METHODS FOR OUTLIERS DETECTION IN DTM DATA

R. Barzaghi, A. Borghi
DIIAR – Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano (Italy)
email: riccardo@ipmtf4.topo.polimi.it, ale@ipmtf4.topo.polimi.it

M.A. Brovelli, G. Sona
DIIAR - Politecnico di Milano, Campus Como, Via Valleggio 12, 22100 Como (Italy)
email: maria@ipmtf4.topo.polimi.it, giovanna@ipmtf4.topo.polimi.it

KEY WORDS : DTM, outliers,test

ABSTRACT

The paper illustrates two methods for outliers rejection that have been compared and tested on the Italian DTM data base. The first method performs a least squares interpolation, using polynomial surfaces, to estimate the height value at the center of a moving window on the DTM grid. The existing height value and the estimated one are compared through properly designed statistical tests to check for a possible outlier. The second method compute at first a mean value of the height gradient in the same moving window centered on the test point, without including the point itself, and a second mean value of height gradient on a smaller window including the test point. These two values are then compared and a simple Chebyshev relationship is used for detecting the outlier presence. By applying these different procedures on the Italian DTM data, two outlier sets have been defined and then compared. The two methods proved to be nearly equivalent since the common detected outliers are the 60% of the outlier set derived with the first procedure, and the 77% of the set defined with the second one. Moreover, an external check has been devised, using a different DTM, independent from the previous one: the Italian DTM supplied by IGM, on a 100m x 100m regular UTM grid. The outliers detected by the described methods have been confirmed through this comparison in the satisfactory proportion of 84% and 77% respectively.

1. INTRODUCTION

Outliers detection is one of the most important issue in data analysis. Data should be carefully checked for outliers before using them in any kind of computation. In this paper two methods are proposed to identify outliers in Digital Terrain Models (DTM). This is particularly relevant since DTMs are frequently used in Surveying, Photogrammetry, Geodesy and Hydrology and can induce distortions in many estimated quantities. As far as Geodesy is concerned, outliers can cause serious errors in geoid estimates since the Residual Terrain Correction can be largely affected by DTM outliers. DTMs outliers lead also to wrong estimates of terrain volumes in Surveying problems and to unrealistic estimates of water flows in Hydrology. Hence, a key point when using DTM data is to remove in a reliable and efficient way the existing outliers. The two methods that are described in the following aim at identifying outliers on the basis of their statistical properties. They have been tested on a real case using two DTMs covering the Italian area.

2. TWO METHODS FOR OUTLIERS REJECTION IN A DTM DATABASE

The problem of outliers detection in digital surface models has been studied by assuming that (Hawkins D. M., 1980): “the outlier is an observation which deviates so much from other observations as to arouse suspicious that it was generated by different mechanism".
Two different methods have been devised to detect outliers in a DTM data base.
The first approach to the problem (Method A) is to compare each value at the knot of the grid that we have with values in a suitable neighbourhood, that is, a window of size depending on the mean roughness of the digital terrain model. In such a way we examine the entire dataset by considering only a small subset at a time.
The basic hypothesis is that the values in the moving window are observations affected by normal distributed white noise. An interpolating surface (a-priori model) is computed from the points surrounding the center of the moving window (suspected blunder). The choice of the model determines the residual between the observation and the surface at the window central point P_0 and therefore the capability to detect the possible outlier.

In the test performed in the next paragraph, we used as interpolating surface the polynomial one: in the following this approach is presented, while more details about other surfaces (median, kriging,...) are available in (Brovelli et al, 1999).
If we take into account a window of size s (i.e. s x s) which is an odd integer, we assume that the digital terrain model h can be described by the model

$$h_o(x_i,y_j) = P(x_i,y_j) + \nu_{ij} \quad \text{for} \quad 0 \le i,j \le \frac{1}{2}(s-1) \quad (1)$$

where ν_{ij} is a gaussian white noise; $x_i = i$, $y_j = j$ are integer coordinates; $P(x_i,y_j)$ is a polynomial surface with d degree of freedom.
The polynomial P depends on a vector $\underline{a}$ of parameters that are estimated via l. s. from (1). With the help of $\hat{\underline{a}}$ we can compute $\hat{P}(x,y)$ at any value (x,y); in particular we define (the coordinates system is centered at the window):

$$\hat{h}_o = \hat{P}_o(0,0) \quad (2)$$

and propagate to $\hat{h}_o$ the variance due to the estimation error, i.e.

$$\sigma^2(\hat{h}_o) = \sigma_o^2 C_o . \quad (3)$$

We then compute the difference

$$\Delta h = h_o(0,0) - \hat{h}_o(0,0)$$

and we notice that if the hypothesis

$$H_o: E\{h_o\} = P(0,0) \quad (4)$$

is satisfied, i.e. h_o <u>is not</u> an outlier, then we have:

$$\Delta h = \boldsymbol{N}[0, (1 + C_o)\sigma_o^2] \quad (5)$$

so that

$$\frac{\Delta h}{\hat{\sigma}_o\sqrt{1+C_o}} = t_r \quad (6)$$

with: $\hat{\sigma}_o^2$ = l.s. estimate of σ_o^2 in (1); t_r = Student's t at r degrees of freedom, $r = s^2$-1-d = redundancy of (1).
Therefore the hypothesis H_o can then be tested by the usual design

$$\frac{|\Delta h|}{\hat{\sigma}_o\sqrt{1+C_o}} < t_r(\alpha/2) \Rightarrow \text{accept } H_o . \quad (7)$$

The second method that we propose to test for outliers (Method B) is based on the same window analysis approach applied to height gradients.
A moving 5 x 5 grid knots window is considered around any DTM point (Fig. 1). For each point contained in the window, except for the central investigated point P, the absolute values of the height gradient are computed with respect to the closest surrounding points

$$\delta H = \left|\frac{\Delta H}{\Delta d}\right| \quad (8)$$

ΔH = height increment
Δd = distance between the points

Combining all the gradients computed in the 5 x 5 window, the mean $\overline{\delta H}$ and the standard deviation $\sigma_{\delta H}$ are evaluated.
Then, a 3 x 3 grid knots window centered in P (Fig. 1) is taken into account and the height gradients with respect to P of the eight surrounding points in the window are computed. The average of these values value $\overline{\delta H}_P$ is compared with the value $\overline{\delta H}$ using a 2σ condition:

$$\left|\overline{\delta H}_P - \overline{\delta H}\right| < 2\sigma_{\delta H} \Rightarrow H_P \text{ is not an outlier} \quad (9)$$

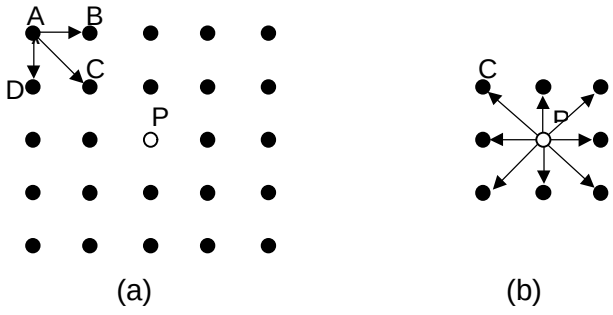

Fig. 1: Masks around the checking point P: (a) 5x5 grid knots; (b) 3x3 grid knots.

So, basically, both methods try to detect an outlier by considering the neighbouring points and their characteristics for defining the mean properties of the DTM surface in the selected window, in order to check for a possible anomalous value in its center. In the following paragraph, both methods are applied to the Italian DTM (Carrozzo et al., 1982) to verify if

outliers are still present (a check for outliers was performed also by the authors of DTM using a different approach).

3. THE OUTLIERS REJECTION PROCEDURES APPLIED TO THE ITALIAN DTM

The two methods that we previously described have been applied to the Italian DTM, which covers entirely Italy in the areas above sea level. It is a mean height DTM distributed on a regular geographical grid, having latitude and longitude steps of 7.5" and 10" respectively.
To apply the first approach, several tests have been performed to properly define the window size and the significance level of the test. In the end, we found that the "best" interpolating surface is a bilinear surface based on a 5x5 window and that the significance level α has to be set equal to 10^{-6} (Brovelli et al, 1999). By using these parameters finally we get 330 suspected outliers distributed as shown in Fig. 2

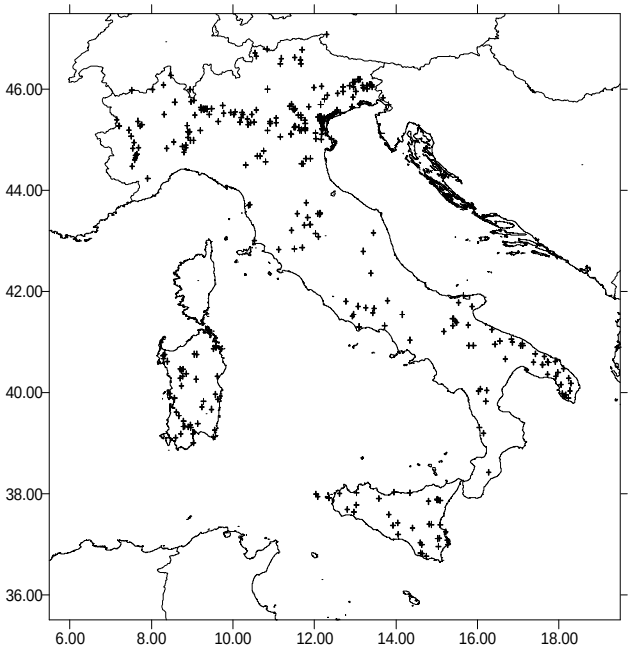

Fig. 2 - Method A: distribution of the 330 suspected outliers.

The second method was then applied , following the scheme described in the previous paragraph.
In Fig.3 are represented the 426 suspected outliers found in this way.

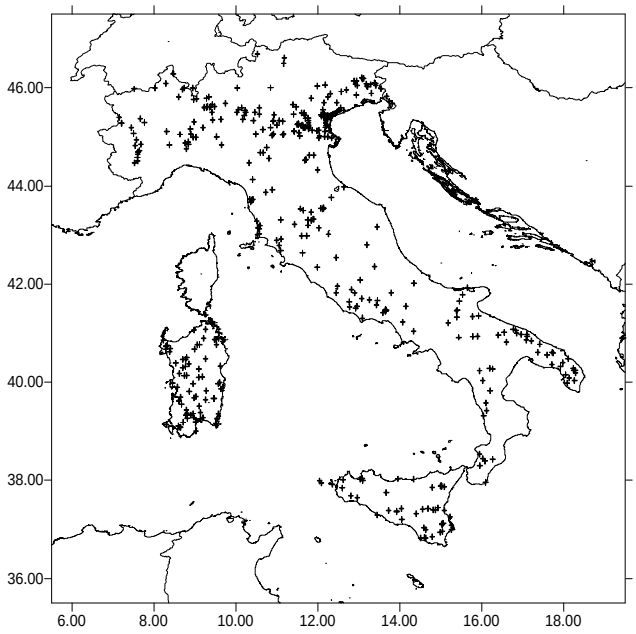

Fig. 3 - Method B: distribution of the 426 suspected outliers.

As one can see, a common behaviour is clearly visible, at least on such a small scale representation. An important check has been done on these results to confirm them (or not) through a comparison with an independent DTM data base which was supplied by the IGM. This is a 100m $\times$ 100m DTM, which covers the whole Italy, which was set up at the IGM and made available to IGeS for scientific purposes. DTM values are known on a regular UTM grid.
The principle we adopted for confirming an outlier, that was detected by applying one of the two methods, is the following.
If the height of a point P is supposed to be an outlier, the heights of all the points contained in the 3 x 3 mask centered on P are compared with the corresponding values extracted from the IGM DTM. If the maximum absolute difference value corresponds to the point P, the outlier presence is confirmed. Using this approach, we obtained the statistics listed in Table 1.

	Method A	Method B
Detected outliers	330	426
Confirmed outliers	276	327

Table 1. Detected and confirmed outliers

Furthermore, the common outliers to Method A and B are 255: 214 of them are confirmed through the comparison with the IGM DTM (see Fig. 4).

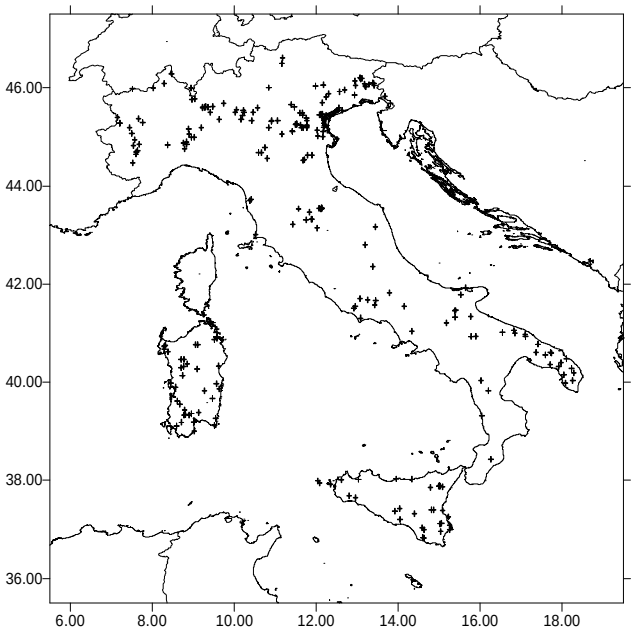

Fig. 4 – The common 255 suspected outliers

4. CONCLUSIONS

The two methods that we propose work properly and lead to satisfactory results. Method A seems to be more conservative and gives better results with respect to Method B in the external check with the IGM DTM. Using such an approach, a higher rate of confirmed outliers is obtained since in this case we get the 84% of confirmed values to be compared with the 77% of Method B. In this comparison a good percentage (84%) is also reached in the common confirmed outlier data set. Probably a proper tuning of Method B should be carried out, based on some assumption on the statistical properties of the height gradients.

Acknowledgements

The authors wish to thank IGM that supplied its 100m $\times$ 100m DTM used in the comparisons.

References

Hawkins D. M., 1980, Identification of outliers – Monographs on applied probability and statistics, Chapman and Hall, London.

Brovelli M.A., F. Sansò and Triglione D., 1999, Different Approaches For Outliers Detection In Digital Terrain Models And Gridded Surfaces Within The GRASS Geographic Information System Environment, The International Archives of Photogrammetry and Remote Sensing, Volume XXXII, Part 4W12, pp. 1-8.

Carrozzo, M.T., Chiarenti, A., Giadam, M., Luzio, D., Margiotta, C., Miglietta, D., Pedone, M., Quarta, T., Zuanni, F. (1982) - *Data bases of mean height values and gravity values* - Proceeding of the 2nd International Symposium on the Geoid in Europe and Mediterranean Area.