# GIS TECHNIQUES FOR DIGITAL SURFACE MODELS OUTLIER DETECTION

M. A. Brovelli , D. Triglione and G. Venuti

DIIAR - Politecnico di Milano - Facoltà di Como, via Valleggio 11, 22100 Como, tel. 031.332.7517, fax. 031.332.7519, *e-mail* maria@ipmtf4.topo.polimi.it

(**)

## 1. Summary

The digital surface models are usually generated from digitized contour lines with some kind of interpolation algorithm (weighted moving average, bicubic splines, kriging,…) or (more recently) directly from automated processing of stereo aerial or satellite data by means of digital image correlation techniques. The problem we want to face in the present work refers to the set of preprocessing procedures needed to validate data, of course taking into account the producing device (namely the accuracy of the observations). This means to establish how the data correctly describe the physical world they refer to. Moreover one has to remark that the observations are to be modeled as stochastic variables to take into account the uncertainty in the knowledge of the actual world. The validation can be obtained by comparing the observed value with the one we expect assuming a certain scene model, considering as outlier (Hawkins D. M., 1980): "an observation which deviates so much from the other observations as to arouse suspicious that it was generated by different mechanism".
In this work we present, as first preprocessing and validation step, some outlier detection techniques applied to the DSMs (Digital Surface Models). The methods and the tools proposed can be applied not only to gridded geophysical data (gridded bore-hole depths, seismic velocities, amplitudes and phases, magnetic data, gravity data) but to all the data representing surface models described by grid stored information.
We decided to implement our approaches to blunders detection by adding new tools in GRASS[1].
The gross errors methods we analyse are characterized by a common localization procedure: we examine the entire dataset by considering iteratively only a small subset at a time; for each step we take into account the data belonging to a moving square window with a (2k+1) size (assuming 1 the side of the pixel in the grid) and we handle separately the attribute associated to the central pixel (corresponding to central point $P_0 = P(x_0,y_0)$) and those associated to the neighbouring points. Our basic hypothesis is that the values in the moving window (the mask) are observations affected by white noise. The methods differ for the technique used to determine the interpolating surface, estimated from the $N_k = (2k+1)^2-1 = 4k(k+1)$ surrounding points. The choice determines the estimated residual at the point $P_0$ and consequently the capability of detection of a possible outlier.
Whatever is the h(x, y) surface we examine, the observation equations are

$$h_{oss,i} = h(x_i,y_i) + v_i \,, \qquad \text{for } i = 1,2,\ldots,N_k$$

[1] The implemented commands are available at the http://geomatica.ing.unico.it/software web page.

with $\mathbf{v} = \mathrm{N}\left[\mathbf{0}, \sigma_0^2 \mathbf{I}\right]$.
The considered approaches (Brovelli et al., 1999) are: polynomial interpolation, robust estimation by using the median, collocation (or kriging) method. For each of them we built up a proper test in order to decide whether the point $P_0$ is a blunder or not. In this paper we refer to the first case, the polynomial interpolation.
The GRASS user can fix the degree of the polynomial interpolation surface or apply the "optimized polynomial interpolation" command which, taking into account the feature of the grid observations layer, chooses, for a fixed $\alpha$ significance level, both the optimal k dimension of the mask and the number NumPar of parameters of the polynomial surface. The choice of a couple (k, NumPar) is defined according to a test on the model applied to sub-samples of the DSM by maximizing a suitable criterion.
This latter strategy has been successfully tested both on synthetic examples as well as on actual cases (the Italian and Swiss national Digital Terrain Models (DTMs), gridded laser-scanning data).
In the following, after introducing some basic statistics needed for the complete understanding of the proposed method, we will focus on the two new GRASS commands for the outliers detection and removing. The first three paragraphs (§ 2, 3, 4) are devoted to recall the main formulas of the least squares estimates procedure, the basics of the statistic inference and its application to the least squares estimates. Three paragraphs follow in which the algorithms for the outliers detection in case of polynomial models (§ 5 and 6) and for an optimal automatic procedure (§ 7) are described. Finally the last two paragraphs (§ 8 and 9) describe the implemented GRASS commands.

## 2. The least squares estimate

If **h** is a vector of *n* normally distributed observables

$$\mathbf{h} = \begin{vmatrix} h_1 \\ \vdots \\ h_i \\ \vdots \\ h_n \end{vmatrix}, \quad h_i = \mathrm{N}\left[\bar{h}_i, \sigma_i^2\right], \quad f(h_i) = \frac{1}{\sqrt{2\pi}\cdot\sigma_i} e^{-\frac{1}{2\sigma_i^2}(h_i - \bar{h}_i)^2} \tag{1}$$

his distribution is given by

$$\mathbf{h} = \mathrm{N}\left[\bar{\mathbf{h}}, \mathbf{C}_{hh}\right], \quad L(\mathbf{h}) = \frac{1}{(2\pi)^{n/2}\cdot\sqrt{\det \mathbf{C}_{hh}}} e^{-\frac{1}{2}\left[(\mathbf{h}-\bar{\mathbf{h}})^T \mathbf{C}_{hh}^{-1}(\mathbf{h}-\bar{\mathbf{h}})\right]} \tag{2}$$

The **deterministic model** assumes that the vector of the averages belongs to a linear variety (variety of the allowable values) V with *m* degrees of freedom

$$\bar{\mathbf{h}} = \mathbf{A}\,\mathbf{p} + \mathbf{a} \tag{3}$$

where **A** is a (*n x m*) matrix whose *m* column vectors are linearly independent; **p** is a column vector of *m* (<*n*) scalars, the parameters.

The **stochastic model** shows us that the observations vector is obtained by the sum of the mean values and by the measurement errors (*noise*)

$$\mathbf{h} = \bar{\mathbf{h}} + \boldsymbol{\nu} \tag{4}$$

The errors are assumed to be with null mean and with covariance matrix (the same as $\mathbf{C}_{hh}$, due to the well-known covariance propagation law) proportional to a matrix **Q**

$$E[\boldsymbol{\nu}] = \mathbf{0}, \quad \mathbf{C}_{\nu\nu} = \mathbf{C}_{hh} = \sigma_0^2 \, \mathbf{Q} \tag{5}$$

By looking for the maximum for the *likelihood* function $L(\mathbf{h}|\mathbf{p},\sigma_0)$ we get that the l. s. estimate is given by

$$\hat{\mathbf{p}} = \mathbf{N}^{-1}\mathbf{A}^T\mathbf{Q}^{-1}(\mathbf{h}_{oss}-\mathbf{a}) \tag{6}$$

in which the normal matrix **N**

$$\mathbf{N} = \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A} \quad \textbf{(dim N = m x m)} \tag{7}$$

and the observations vector $\mathbf{h}_{oss}$ appears.

The maximum likelihood estimator is unbiased, i.e.

$$E[\hat{\mathbf{p}}] = \mathbf{p} \tag{8}$$

And the same happens to the maximum likelihood estimator for the variance $\sigma_0^2$: by indicating the residuals with:

$$\mathbf{U} = \mathbf{h}_{oss} - \hat{\mathbf{h}} = \mathbf{h}_{oss} - \mathbf{A}\hat{\mathbf{p}} - \mathbf{a} \tag{9}$$

we get

$$\hat{\sigma}_{0\,MV}^2 = \frac{1}{n}(\mathbf{U}^T\mathbf{Q}^{-1}\mathbf{U}), \qquad E[\hat{\sigma}_{0\,MV}^2] = \frac{n-m}{n}\sigma_0^2 \tag{10}$$

The unbiased estimator (which is not a maximum likelihood estimator) for $\sigma_0^2$ is

$$\hat{\sigma}_0^2 = \frac{1}{n-m}(\mathbf{U}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{U}) \tag{11}$$

Finally the **h** estimate is

$$\hat{\mathbf{h}} = \mathbf{A}\,\hat{\mathbf{p}} + \mathbf{a} \tag{12}$$

and the covariance matrices are:

$$\mathbf{C}_{\mathrm{vv}} = \sigma_0^2\left(\mathbf{Q} - \mathbf{A}\mathbf{N}^{-1}\mathbf{A}^{\mathrm{T}}\right), \quad \mathbf{C}_{\hat{\mathrm{p}}\hat{\mathrm{p}}} = \sigma_0^2\,\mathbf{N}^{-1} \tag{13}$$

## 3. The statistical inference

The statistical inference lets us to determine whether the hypothesis $H_0$ made on a random variable **h** can be confirmed on the base of a data sample taken out from the variable itself.

To this aim we use a statistics S(**h**), which has a known distribution *F* if the hypothesis $H_0$ is true:

$$\mathrm{S}(\mathbf{h}) \sim F \tag{14}$$

S(**h**) is built up in such a way to grow when we go away from the hypothesis $H_0$; then S(**h**) ≥ *c* determines in the sampling space $\Re^n$ a zone where $H_0$ seems to be not true or better, not very plausible.

Obviously in such a way we risk to refuse $H_0$ even if it is true: the risk is given by

$$\mathrm{P}\{\mathrm{S} \geq c \,|\mathrm{H}_0\} = \alpha \tag{15}$$

where α is named the test significance level.

## 4. The least squares inference

Assume that

$$\mathbf{h} \sim \mathrm{N}[\,\bar{\mathbf{h}}, \mathbf{C}_{\mathrm{hh}}\,] = \mathrm{N}[\,\mathbf{A}\,\mathbf{p} + \mathbf{a}, \sigma_0^2\,\mathbf{Q}\,]. \tag{16}$$

The vector of least squares estimated parameters **p** is therefore distributed as follows

By means of the l. s. theory, we can compute

$$\hat{\mathbf{p}} \sim \mathrm{N}[\,\bar{\mathbf{p}}, \sigma_0^2\,\mathbf{N}^{-1}\,] \tag{17}$$

and then

$$\hat{\mathbf{h}} = \mathbf{A}\,\hat{\mathbf{p}} + \mathbf{a}, \qquad \hat{\mathbf{h}} \sim \mathrm{N}[\,\bar{\mathbf{h}}, \sigma_0^2 \mathbf{A}\mathbf{N}^{-1}\mathbf{A}^{\mathrm{T}}\,] \tag{18}$$

For the residuals we get the expression:

$$\mathbf{U} = \mathbf{h}_{\mathrm{oss}} - \mathbf{A}\,\hat{\mathbf{p}} - \mathbf{a} = \mathbf{h}_{\mathrm{oss}} - \hat{\mathbf{h}}\,,\ \mathbf{U} \sim \mathrm{N}[\,\mathbf{0}, \sigma_0^2\,(\mathbf{Q} - \mathbf{A}\mathbf{N}^{-1}\mathbf{A}^{\mathrm{T}})\,] \tag{19}$$

As $(\hat{\mathbf{p}}, \hat{\mathbf{h}})$ and $\mathbf{U}$ are stochastically independent, the same holds for their quadratic forms

$$(\,(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}\,\mathbf{N}\,(\hat{\mathbf{p}} - \mathbf{p}) = \sigma_0^2\,\chi^2_m \tag{20}$$

$$\mathbf{U}^{\mathrm{T}}\,\mathbf{Q}^{-1}\,\mathbf{U} = \sigma_0^2\,\chi^2_{n-m} \tag{21}$$

in such a way that

$$\hat{\sigma}_0^2 = \frac{1}{n-m}\,\mathbf{U}^{\mathrm{T}}\,\mathbf{Q}^{-1}\,\mathbf{U} \sim \frac{\sigma_0^2}{n-m}\,\chi^2_{n-m} \tag{22}$$

where $\sigma_0^2$ is independent from $\hat{\mathbf{p}}$.

To verify the hypothesis $H_0$ ($\mathbf{p} = \bar{\mathbf{p}}$) on the parameters, where $\mathbf{p}$ has $m$ dimensions, if $\sigma_0^2$ is known, the statistics is

$$(\hat{\mathbf{p}} - \bar{\mathbf{p}})^{\mathrm{T}}\,\mathbf{C}_{\mathrm{pp}}^{-1}\,(\hat{\mathbf{p}} - \bar{\mathbf{p}}) = \sigma_0^{-2}\,(\hat{\mathbf{p}} - \bar{\mathbf{p}})^{\mathrm{T}}\,\mathbf{N}\,(\hat{\mathbf{p}} - \bar{\mathbf{p}}) = \chi^2_m\,. \tag{23}$$

We refuse $H_0$ if $\chi^2_{emp} > \chi^2_\alpha$.

On the contrary, if $\sigma_0^2$ is unknown, the couple of stochastically independent quadratic forms can be explited:

$$\frac{1}{m}\,(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}\,\mathbf{N}\,(\hat{\mathbf{p}} - \mathbf{p}) = \sigma_0^2\,\frac{\chi^2_m}{m} \tag{24}$$

$$\frac{1}{n-m}\,\mathbf{U}^{\mathrm{T}}\,\mathbf{Q}^{-1}\,\mathbf{U} = \hat{\sigma}_0^2 = \frac{\sigma_0^2}{n-m}\,\chi^2_{n-m} \tag{25}$$

Their ratio gives the following statistics:

$$\frac{1}{m}\,\frac{1}{\hat{\sigma}_0^2}\,(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}\,\mathbf{N}\,(\hat{\mathbf{p}} - \mathbf{p}) = F_{m,n-m}. \tag{26}$$

In this case we refuse the hypothesis $H_0$ if $F_{emp} > F_\alpha$.

If only r components of the **p** vector are to be verified, we can introduce the vector $\boldsymbol{\zeta}$ (of $r$ dimensions), extracted from **p** by means of the matrix **P**:

$$\boldsymbol{\zeta} = \mathbf{P}\,\mathbf{p}, \tag{27}$$

we have

$$\hat{\boldsymbol{\zeta}} = \mathbf{P}\,\hat{\mathbf{p}}, \qquad \mathbf{C}_{\hat{\zeta}\hat{\zeta}} = \mathbf{P}\mathbf{C}_{\hat{p}\hat{p}}\mathbf{P}^{T} = \sigma_0^2\,\mathbf{P}\,\mathbf{N}^{-1}\,\mathbf{P}^{T} \tag{28}$$

$$\frac{1}{\sigma_0^2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})^{T}(\mathbf{P}\,\mathbf{N}^{-1}\,\mathbf{P}^{T})^{-1}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) = \chi_r^2 \tag{29}$$

The hypothesis $H_0$ ($\boldsymbol{\zeta} = \overline{\boldsymbol{\zeta}}$) is verified by means of the following statistics

$$\frac{1}{r}\frac{1}{\hat{\sigma}_0^2}(\hat{\boldsymbol{\zeta}} - \overline{\boldsymbol{\zeta}})^{T}(\mathbf{P}\,\mathbf{N}^{-1}\,\mathbf{P}^{T})^{-1}(\hat{\boldsymbol{\zeta}} - \overline{\boldsymbol{\zeta}}) = F_{r,n-m} \tag{30}$$

and we refuse $H_0$ if $F_{emp} > F_{\alpha}$.

## 5. The outlier rejection in a DSM: the polynomial models

Gridded data sets are widely used in geoinformatics: either because directly produced (raw data) on a grid (e.g. a digital image, monochromatic or multi-band, or a digital terrain model derived by digital photogrammetry) or because reduced to a grid (preprocessed), to be easily managed, by means of some kind of interpolation procedures (e.g. a DTM derived from laser scanning or a mean gravity anomalies field, etc.).

In any case it is good practice to validate them in order to verify that no blunders are present into the data set or just to verify that if there are values outlying the mean statistical behaviour, this is not due to an error but because physical reality is such. This validation can be purely internal, when we have one field of values only, or external, when we have more fields of gridded data for the same phenomenon in the same area. In this second case the residuals after reducing the two datasets to a common grid, for example by using the possible higher resolution one used to predict values on the coarser, are computed. In this way we are left with a population of residuals which stem from the errors in the original data sets as well as the prediction errors; anyway we are reconducted to a unique field, where suspected high values point to either directly to an outlier in the data or to a discrepancy between the two data sets.

Furthermore there are several types of statistical anomalies we might be willing to detect in our data set; here we will look for (more or less) isolated outliers that make us think of point wise gross errors, or, in any case, point wise anomalies.

A typical approach to the problem is to compare each value at the knot of the grid with values in a suitable neighbourhood (window), of size $N_k$ depending on the mean roughness of the field (Figures 1 and 2).

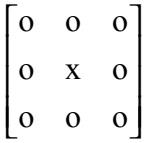


Figure 1 - Mask 3x3 (9 elements, $N_k = 8$)

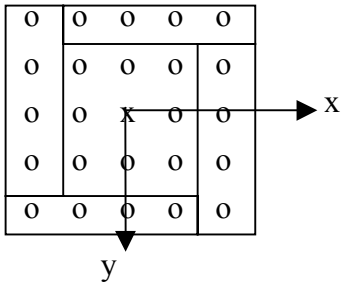


Figure 2 - Mask 5x5 (25 elements, $N_k = 24$)

Milder fields (e.g. Bouguer gravity anomalies) can exploit larger windows, while more rugged phenomena (e.g. DTM in an alpine region) might require smaller sizes down to the minimal window of 3x3 knots. More precisely the neighbourhood points are used to determine an interpolation surface whose value in the middle of the window is compared with the observed one. In our preprocessing method an automatic procedure to optimise polynomial degree and window size is set up and the use of as rigorous as possible criteria for the outlier identification is proposed.
In a cartesian coordinate system centered in the middle of the window (like in figure 2), the polynomial surfaces wetake into account have the following equations:

*1)Mean Surface*

$$h_{ms}(x,y) = a_0 \tag{31}$$

parameter $a_0$.

***2) Linear Surface***

$$h_{lin}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y \tag{32}$$

m = 3 parameters: $a_0$, $a_1$, $a_2$

***3) Bilinear Surface***

$$h_{bil}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot xy \tag{33}$$

m = 4 parameters: $a_0$, $a_1$, $a_2$, $a_3$

*4) Quadratic Surface*

$$h_{quad}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot xy + a_4 \cdot x^2 + a_5 \cdot y^2 \quad (34)$$

m = 6 parameters: $a_0$, $a_1$, $a_2$, $a_3$, $a_4$, $a_5$

*5) Biquadratic Surface*

$$h_{biq}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot xy + a_4 \cdot x^2 + a_5 \cdot y^2 + a_6 \cdot x^2 y + a_7 \cdot xy^2 + a_8 \cdot x^2 y^2 \quad (35)$$

m = 9 parameters: $a_0$, ..., $a_8$

*6) Bicubic Surface*

$$h_{bic}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot xy + a_4 \cdot x^2 + a_5 \cdot y^2 + a_6 \cdot x^2 y + a_7 \cdot xy^2 + a_8 \cdot x^2 y^2 + \\ + a_9 \cdot x^3 + a_{10} \cdot y^3 + a_{11} \cdot xy^3 + a_{12} \cdot x^3 y + a_{13} \cdot x^2 y^3 + a_{14} \cdot x^3 y^2 + a_{15} \cdot x^3 y^3 \quad (36)$$

m = 16 parameters: $a_0$, ..., $a_{15}$

## 6. The outlier rejection in a DSM: the test on the polynomial models

For all the surfaces considered, thanks to the coordinate system we choose, the value to be compared with the candidate outlier is $\hat{a}_0$. The natural statistics for the outlier detection is $h_{obs} - \hat{a}_0$, whose distribution is to be determined in the different cases. Here, to be short, we report in detail the expressions related to the low degree surfaces till the bilinear one. Note that for the higher degree surfaces the statistics becomes more complicated due to the non diagonal form of the normal matrix (Tcholesky decomposition is needed).

*1)Mean Surface*

The deterministic and stochastic general models are

$$\bar{\mathbf{h}} = \mathbf{A}\,\mathbf{p} \quad (37)$$

$$\mathbf{C}_{hh} = \sigma_0^2\,\mathbf{Q} \quad (38)$$

By applying the l. s. theory with $\mathbf{Q} = \mathbf{I}$, $m = 1$, $n = N_k$. we get

$$\mathbf{A} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \text{ with } \dim(\mathbf{A}) = N_k \ ^{(*)} \tag{39}$$

$$N = \mathbf{A}^T \mathbf{A} = N_k \tag{40}$$

is a scalar and then

$$N^{-1} = 1/N_k \tag{41}$$

$$\hat{a}_0 = N^{-1} \mathbf{A}^T \mathbf{h}_{oss} = \frac{1}{N_k} \sum_{i=1}^{N_k} h_{oss,i} \tag{42}$$

The estimates get:

$$\hat{h}(x,y) = \hat{a}_0 = \frac{1}{N_k} \sum_{i=1}^{N_k} h_{oss,i} \sim \mathbb{N}[\, a_0,\ \sigma_0^2 / N_k \,] \tag{43}$$

$$\hat{\sigma}_0^2 = \frac{1}{N_k - 1} \mathbf{U}^T \mathbf{U} = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (h_{oss,i} - \hat{a}_0)^2 \sim \frac{\sigma_0^2}{N_k - 1} \chi^2_{(N_k - 1)} \tag{44}$$

If the model is unbiased,

$$\Delta h = h_{oss}(0,0) - \hat{h}(0,0) = h_{oss}(0,0) - \hat{a}_0 \sim \mathbb{N}[0,\ \sigma_0^2 (1 + 1/N_k)] \tag{45}$$

and the two following distributions are independent

$$Z = \sqrt{\frac{N_k}{N_k + 1}} \frac{\Delta h}{\sigma_0} \tag{46}$$

$$\chi^2_{N_k - 1} = (N_k - 1) \frac{\hat{\sigma}_0^2}{\sigma_0^2} \tag{47}$$

Now if we take into account the ratio between the two previous expressions we get the statistics:

$$S = \sqrt{\frac{N_k}{N_k + 1}} \frac{\Delta h}{\hat{\sigma}_0} = \sqrt{\frac{N_k}{N_k + 1}} \frac{\Delta h}{\sigma_0} \frac{\sigma_0}{\hat{\sigma}_0} \sim \frac{Z}{\sqrt{\frac{\chi^2_{N_k - 1}}{N_k - 1}}} = t_{(N_k - 1)} \tag{48}$$

suitable for testing the hypothesis $H_o$: $\Delta h = 0$ (the central value is not outlier).

**2***) Linear Surface*

---

(*) $N_k = (2k+1)^2 - 1 = 4k \cdot (k+1)$ , where $(2k+1)$ is the window size.

$h_{lin}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y$

In that case $\mathbf{Q} = \mathbf{I}$, $m = 3$, $n = N_k$.

Therefore

$$\mathbf{h} = \mathbf{A}\,\mathbf{p}, \text{ con } \mathbf{p} = [\, a_0 \;\; a_1 \;\; a_2 \,]^T \tag{49}$$

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots \\ 1 & x_i & y_i \\ \vdots & \vdots & \vdots \\ 1 & x_{N_k} & y_{N_k} \end{bmatrix}, \text{ with } \dim(\mathbf{A}) = (N_k,3) \tag{50}$$

$$\mathbf{N} = \mathbf{A}^T \mathbf{A} = \text{diag} \left\{ N_k, \sum_{i=1}^{N_k} x_i^2, \sum_{i=1}^{N_k} y_i^2 \right\} \tag{51}$$

$$\mathbf{N}^{-1} = \text{diag} \left\{ 1/N_k,\ 1/\sum_{i=1}^{N_k} x_i^2,\ 1/\sum_{i=1}^{N_k} y_i^2 \right\} \tag{52}$$

$$\hat{\mathbf{p}} = \mathbf{N}^{-1}\,\mathbf{A}^T\,\mathbf{h}_{oss} \tag{53}$$

$$\hat{\mathbf{h}} = \mathbf{A}\,\hat{\mathbf{p}} \tag{54}$$

The estimate in the center of the window is $\hat{a}_0$.

Then,

$$\Delta h = h_{oss}(0,0) - \hat{h}\,(0,0) = h_{oss}(0,0) - \hat{a}_0 \;\; \sim \;\; \mathrm{N}[0,\ \sigma_0^2\,(1 + 1/N_k)] \tag{55}$$

and

$$\hat{\sigma}_0^2 = \frac{\mathbf{U}^T\mathbf{U}}{N_k - 3} = \frac{1}{N_k - 3} \sum_{i=1}^{N_k} \left( h_{oss,i} - \hat{h}_i \right)^2 \;\; \sim \;\; \frac{\sigma_0^2}{N_k - 3}\, \chi^2_{(N_k\,-\,3)} \tag{56}$$

The two following distributions are independent

$$Z = \sqrt{\frac{N_k}{N_k + 1}}\, \frac{\Delta h}{\sigma_0} \tag{57}$$

$$\chi^2_{N_k - 3} = (N_k - 3)\ \frac{\hat{\sigma}_0^2}{\sigma_0^2} \tag{58}$$

Now if we take into account the ratio between the two previous expressions we get the statistics:

$$S = \sqrt{\frac{N_k}{N_k + 1}}\frac{\Delta h}{\hat{\sigma}_0} = \sqrt{\frac{N_k}{N_k + 1}}\frac{\Delta h}{\sigma_0}\ \frac{\sigma_0}{\hat{\sigma}_0}\ \sim\ \frac{Z}{\sqrt{\frac{\chi^2_{N_k - 3}}{N_k - 3}}} = t_{(N_k - 3)} \tag{59}$$

suitable for testing the hypothesis $H_o$: $\Delta h=0$ (the central value is not outlier).

*3) Bilinear Surface*

$$h_{bil}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot xy$$

In this case $\mathbf{Q} = \mathbf{I}$, $m = 4$, $n = N_k$.

Therefore

$$\mathbf{h} = \mathbf{A}\ \mathbf{p},\ \text{con}\ \mathbf{p} = [\ a_0\ \ a_1\ \ a_2\ \ a_3\ ]^T \tag{60}$$

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 & y_1 & x_1 y_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_i & y_i & x_i y_i \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N_k} & y_{N_k} & x_{N_k} y_{N_k} \end{bmatrix},\ \text{with}\ \dim(\mathbf{A}) = (N_k, 4) \tag{61}$$

$$\mathbf{N} = \mathbf{A}^T \mathbf{A} = \text{diag}\left\{ N_k, \sum_{i=1}^{N_k} x_i^2, \sum_{i=1}^{N_k} y_i^2, \sum_{i=1}^{N_k} x_i^2 y_i^2 \right\} \tag{62}$$

$$\mathbf{N}^{-1} = \text{diag}\left\{ 1/N_k,\ 1/\sum_{i=1}^{N_k} x_i^2,\ 1/\sum_{i=1}^{N_k} y_i^2,\ 1/\sum_{i=1}^{N_k} x_i^2 y_i^2 \right\} \tag{63}$$

$$\hat{\mathbf{p}} = \mathbf{N}^{-1}\ \mathbf{A}^T\ \mathbf{h}_{oss} \tag{64}$$

$$\hat{\mathbf{h}} = \mathbf{A}\,\hat{\mathbf{p}} \tag{65}$$

The estimate in the window center is even equal to $\hat{a}_0$.

Then,

$$\Delta h = h_{oss}(0,0) - \hat{h}\,(0,0) = h_{oss}(0,0) - \hat{a}_0 \;\sim\; \mathrm{N}[0,\ \sigma_0^2\,(1 + 1/N_k)] \tag{66}$$

and

$$\hat{\sigma}_0^2 = \frac{\mathbf{U}^T\mathbf{U}}{N_k - 4} = \frac{1}{N_k - 4}\sum_{i=1}^{N_k}\left(h_{oss,i} - \hat{h}_i\right)^2 \;\sim\; \frac{\sigma_0^2}{N_k - 4}\,\chi^2_{(N_k\,-\,4)} \tag{67}$$

Analogously as the previous cases we get the statistics:

$$S = \sqrt{\frac{N_k}{N_k + 1}}\,\frac{\Delta h}{\hat{\sigma}_0} = \sqrt{\frac{N_k}{N_k + 1}}\,\frac{\Delta h}{\sigma_0}\,\frac{\sigma_0}{\hat{\sigma}_0} \;\sim\; \frac{Z}{\sqrt{\dfrac{\chi^2_{N_k\,-\,4}}{N_k\,-\,4}}} = t_{(N_k\,-\,4)} \tag{68}$$

suitable for testing the hypothesis $H_o$: $\Delta h=0$ (the central value is not outlier).

## 7. The optimal outliers detection

In the following, the criteria for the optimal choose of the window size and the polynomial degree for the local interpolation procedure are described. For that we consider three different window sizes, i.e. s=3,5,7 and polynomial surfaces up to degree 3. We suggest to perform the following analysis on a sub-sample of the whole dataset.

In addition in an effort of keeping not too high the ratio ρ between m (number of estimated parameters) and $N_k$ (information available) only the following alternatives have been considered:

| Model | Polynomial Surface | m | 2k+1 | ρ |
|---|---|---|---|---|
| a | SM (simple mean) | 1 | 3 | 0.125 |
| b | BL (bilinear) | 4 | 3 | 0.500 |
| c | BL (bilinear) | 4 | 5 | 0.167 |
| d | BQ (biquadratic) | 9 | 5 | 0.375 |
| e | BQ (biquadratic) | 9 | 7 | 0.188 |
| f | BC (bicubic) | 16 | 7 | 0383 |

Table 1 - Alternative to choose the optimal interpolator

The idea is to perform a test on the parameters of the different surfaces as described in § 4.

The tests we take into account are:

1) “a” (simple mean (ms), 3x3) versus “b” (bilinear (bil), 3x3)

2) “c” (bilinear, 5x5) versus “d” (biquadratic (biq), 5x5)

3) “e” (biquadratic, 7x7) versus “f” (bicubic (bic), 7x7)

The comparison between the surfaces returns “a” (i.e. we choose the simple mean surface) if the parameters of the bilinear surface except $a_0$ are significantly equal to zero. In the other cases we consider also the window size, as it is summarized in the following table:

```
if (Result(a,b) == a)
        Cod_opt = a;
else
  if (Result(c,d) == c)
        Cod_opt = b;
else
     if (Result(e,f) == e)
           Cod_opt = d;
else
           Cod_opt = f;
```

(69)

*1) "a" (simple mean, 3x3) versus "b" (bilinear, 3x3)*

$$h_{ms}(x,y) = a_0 \tag{70}$$

$$m_{ms} = 1 \tag{71}$$

$$p_{ms} = a_0 \tag{72}$$

$$\hat{p}_{ms} \sim \mathrm{N}[\, p_{ms} \,,\, (1/N_k)\sigma_0^2 \,] \tag{73}$$

$$h_{bil}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot xy \tag{74}$$

$$m_{bil} = 4 \tag{75}$$

$$\mathbf{p}_{bil} = [a_0 \;\; a_1 \;\; a_2 \;\; a_3]^{\mathrm{T}} \tag{76}$$

$$\hat{\mathbf{p}}_{bil} \sim \mathrm{N}[\, \mathbf{p}_{bil} \,,\, \sigma_0^2 \, \mathbf{N}_{bil}^{-1} \,] \tag{77}$$

Bilinear surface coincides with a simple mean with $a_1$, $a_2$, $a_3$ set to zero.

$$\boldsymbol{\zeta} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \mathbf{R}\, \mathbf{p}_{bil} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} \tag{78}$$

$$\hat{\boldsymbol{\zeta}} = \mathbf{R}\, \hat{\mathbf{p}}_{bil} \,, \quad \mathbf{C}_{\hat{\zeta}\hat{\zeta}} = \mathbf{R}\, \mathbf{C}_{\hat{p}\hat{p},bil}\, \mathbf{R}^{\mathrm{T}} = \sigma_0^2\, \mathbf{R}\, \mathbf{N}_{bil}^{-1}\, \mathbf{R}^{\mathrm{T}} \tag{79}$$

Therefore,

$$\frac{1}{\sigma_0^2}\,(\hat{\boldsymbol{\zeta}}-\boldsymbol{\zeta})^{\mathrm{T}}\,(\mathbf{R}\,\mathbf{N}_{\mathrm{bil}}^{-1}\,\mathbf{R}^{\mathrm{T}})^{-1}\,(\hat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}) = \chi_r^2 = \chi_3^2 \tag{80}$$

$$\hat{\sigma}_0^2 = \frac{1}{\mathrm{N}_k - m_{bil}}\,\mathbf{U}_{\mathrm{bil}}^{\mathrm{T}}\;\mathbf{U}_{\mathrm{bil}} = \frac{\sigma_0^2}{\mathrm{N}_k - m_{bil}}\,\chi^2_{N_k - m_{bil}} \tag{81}$$

that is

$$\hat{\sigma}_0^2 = \frac{1}{\mathrm{N}_k - 4}\,|\mathbf{U}_{\mathrm{bil}}|^2 = \frac{\sigma_0^2}{\mathrm{N}_k - 4}\,\chi^2_{N_k-4} \tag{82}$$

As (80) ad (82) are independent, the hypothesis $H_0$: $\boldsymbol{\zeta} = \overline{\boldsymbol{\zeta}} = \mathbf{0}$ is true when

$$\frac{1}{3}\,\frac{1}{\hat{\sigma}_0^2}\;\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\,(\mathbf{R}\,\mathbf{N}_{\mathrm{bil}}^{-1}\,\mathbf{R}^{\mathrm{T}})^{-1}\;\hat{\boldsymbol{\zeta}} = F_{3,N_k-4} \tag{83}$$

or

$$\frac{\mathrm{N}_k - 4}{3}\,\frac{1}{|\,\mathbf{U}_{\mathrm{bil}}\,|^2}\;\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\,(\mathbf{R}\,\mathbf{N}_{\mathrm{bil}}^{-1}\,\mathbf{R}^{\mathrm{T}})^{-1}\;\hat{\boldsymbol{\zeta}} = F_{3,N_k-4} \tag{84}$$

As we have $k = 1$, or $N_k = 8$, then $N_k$-4 = 4. And also the numerator can be decomposed by $|\mathbf{U}_{\mathrm{ms}}|^2 - |\mathbf{U}_{\mathrm{bil}}|^2$ ; the previous relation can be written as

$$\frac{4}{3}\,\frac{|\,\mathbf{U}_{\mathrm{ms}}\,|^2 - |\,\mathbf{U}_{\mathrm{bil}}\,|^2}{|\,\mathbf{U}_{\mathrm{bil}}\,|^2} = F_{3,4} \tag{85}$$

If $F_{emp} \le F_\alpha$ ($\alpha$ is the significance level), we accept $H_0$, i.e. we choose the simple mean.

*2) "c" (bilinear, 5x5) versus "d" (biquadratic, 5x5)*

$$\mathrm{h}_{\mathrm{bil}}(\mathrm{x},\mathrm{y}) = a_0 + a_1\cdot\mathrm{x} + a_2\cdot\mathrm{y} + a_3\cdot\mathrm{xy} \tag{86}$$

$$m_{\mathrm{bil}} = 4 \tag{87}$$

$$\mathbf{p}_{\mathrm{bil}} = [a_0\;\; a_1\;\; a_2\;\; a_3]^{\mathrm{T}} \tag{88}$$

$$\hat{\mathbf{p}}_{\mathrm{bil}} \sim \mathbb{N}[\,\mathbf{p}_{\mathrm{bil}}\,,\;\sigma_0^2\,\mathbf{N}_{\mathrm{bil}}^{-1}\,] \tag{89}$$

$$\mathrm{h}_{\mathrm{biq}}(\mathrm{x},\mathrm{y}) = a_0 + a_1\cdot\mathrm{x} + a_2\cdot\mathrm{y} + a_3\cdot\mathrm{xy} + a_4\cdot\mathrm{x}^2 + a_5\cdot\mathrm{y}^2 + a_6\cdot\mathrm{x}^2\mathrm{y} + a_7\cdot\mathrm{xy}^2 + a_8\cdot\mathrm{x}^2\mathrm{y}^2 \tag{90}$$

$$m_{\mathrm{biq}} = 9 \tag{91}$$

$$\mathbf{p}_{\mathrm{biq}} = [a_0\;\; a_1\;\; \ldots\; a_7\;\; a_8\,]^{\mathrm{T}} \tag{92}$$

$$\hat{\mathbf{p}}_{\mathrm{biq}} \sim \mathbb{N}[\,\mathbf{p}_{\mathrm{biq}}\,,\;\sigma_0^2\,\mathbf{N}_{\mathrm{biq}}^{-1}\,] \tag{93}$$

The biquadratic surface coincides with the bilinear if the coefficients $a_4$, $a_5$, $a_6$, $a_7$, $a_8$. are null. As previously,

$$\hat{\sigma}_0^2 = \frac{1}{\mathrm{N}_k - m_{biq}} \mathbf{U}_{\mathrm{bil}}^{\mathrm{T}} \mathbf{U}_{\mathrm{bil}} = \frac{\sigma_0^2}{\mathrm{N}_k - m_{biq}} \chi^2_{N_k - m_{biq}} \tag{94}$$

or

$$\hat{\sigma}_0^2 = \frac{1}{\mathrm{N}_k - 9} |\mathbf{U}_{\mathrm{biq}}|^2 = \frac{\sigma_0^2}{\mathrm{N}_k - 9} \chi^2_{N_k - 9} \tag{95}$$

and then

$$\frac{15}{5} \frac{|\mathbf{U}_{\mathrm{bil}}|^2 - |\mathbf{U}_{\mathrm{biq}}|^2}{|\mathbf{U}_{\mathrm{biq}}|^2} = F_{5,15} \tag{96}$$

If $F_{emp} \leq F_\alpha$) we choose the bilinear surface.

*3) "e" (biquadratic, 7x7) versus "f" (bicubic, 7x7)*

$$\mathrm{h}_{\mathrm{biq}}(\mathrm{x},\mathrm{y}) = a_0 + a_1 \cdot \mathrm{x} + a_2 \cdot \mathrm{y} + a_3 \cdot \mathrm{xy} + a_4 \cdot \mathrm{x}^2 + a_5 \cdot \mathrm{y}^2 + a_6 \cdot \mathrm{x}^2\mathrm{y} + a_7 \cdot \mathrm{xy}^2 + a_8 \cdot \mathrm{x}^2\mathrm{y}^2 \tag{97}$$

$$m_{\mathrm{biq}} = 9 \tag{98}$$

$$\mathbf{p}_{\mathrm{biq}} = [a_0 \;\; a_1 \;\; \ldots \; a_7 \;\; a_8]^{\mathrm{T}} \tag{99}$$

$$\hat{\mathbf{p}}_{\mathrm{biq}} \sim \mathbb{N}[\, \mathbf{p}_{\mathrm{biq}} \,,\; \sigma_0^2 \mathbf{N}_{\mathrm{biq}}^{-1}] \tag{100}$$

$$\begin{aligned} \mathrm{h}_{\mathrm{bic}}(\mathrm{x},\mathrm{y}) = {} & a_0 + a_1 \cdot \mathrm{x} + a_2 \cdot \mathrm{y} + a_3 \cdot \mathrm{xy} + a_4 \cdot \mathrm{x}^2 + a_5 \cdot \mathrm{y}^2 + a_6 \cdot \mathrm{x}^2\mathrm{y} + a_7 \cdot \mathrm{xy}^2 + a_8 \cdot \mathrm{x}^2\mathrm{y}^2 \cdot + \\ & + a_9 \cdot \mathrm{x}^3 \cdot + a_{10} \cdot \mathrm{y}^3 \cdot + a_{11} \cdot \mathrm{xy}^3 \cdot + a_{12} \cdot \mathrm{x}^3\mathrm{y} \cdot + a_{13} \cdot \mathrm{x}^2\mathrm{y}^3 \cdot + a_{14} \cdot \mathrm{x}^3\mathrm{y}^2 \cdot + a_{15} \cdot \mathrm{x}^3\mathrm{y}^3 \end{aligned} \tag{101}$$

$$m_{\mathrm{bic}} = 16 \tag{102}$$

$$\mathbf{p}_{\mathrm{bic}} = [a_0 \;\; a_1 \;\; \ldots \; a_{14} \;\; a_{15}]^{\mathrm{T}} \tag{103}$$

$$\hat{\mathbf{p}}_{\mathrm{bic}} \sim \mathbb{N}[\, \mathbf{p}_{\mathrm{bic}} \,,\; \sigma_0^2 \mathbf{N}_{\mathrm{biq}}^{-1}] \tag{104}$$

The bicubic coincides with the biquadratic if we take as null the $r = m_{\mathrm{bic}} - m_{\mathrm{biq}} = 16 - 9 = 7$ coefficients: $a_9$, $a_{10}$, $a_{11}$, $a_{12}$, $a_{13}$, $a_{14}$, $a_{15}$.

We have

$$\hat{\sigma}_0^2 = \frac{1}{\mathrm{N}_k - m_{bic}} \mathbf{U}_{\mathrm{bic}}^{\mathrm{T}} \mathbf{U}_{\mathrm{bic}} = \frac{\sigma_0^2}{\mathrm{N}_k - m_{bic}} \chi^2_{N_k - m_{bic}} \tag{105}$$

or

$$\hat{\sigma}_0^2 = \frac{1}{\mathrm{N}_k - 16} |\mathbf{U}_{\mathrm{bic}}|^2 = \frac{\sigma_0^2}{\mathrm{N}_k - 16} \chi^2_{N_k - 16} \tag{106}$$

The statistics to be used is

$$\frac{32}{7} \frac{|\mathbf{U}_{\mathrm{biq}}|^2 - |\mathbf{U}_{\mathrm{bic}}|^2}{|\mathbf{U}_{\mathrm{bic}}|^2} = F_{7,32} \tag{107}$$

If $F_{emp} \leq F_\alpha$ we choose the biquadratic.

## 8. Description of the GRASS command r.oldetect

The procedure, aiming at the outliers detection, works on raster files and has the following syntax:

usage:

r.oldetect [-abq] input=name [output=name] method=name size=value nalfa=value [dwsites=name] [residuals=name] [factor=value] [binrast=name]

flags:

-a Do not align output with the input
-b Do validation on borders
-q Run quietly

parameters:

input Name of existing raster file
output Name of the output raster file

method Type of validation
options:
average
linear
bilinear
quadratic
biquadratic
bicubic

median
collocation

size Neighborhood size
options:
1,3,5,7,9,11,13,15,17,19,21,23,25

nalfa Significance test level
dwsites Name of site list file for outliers
residuals Name of residuals output raster file
factor Factor with which multiply residuals
binrast Name of binary output raster file

The input file (input), is re-sampled in the current region; the east - west and south - north limits and resolutions of this input file are assumed as the default ones, in order to avoid the re-sampling done by GRASS. If we want to resample by using the grid parameters stored in the region, the flag "-a" has to be specified.

The data processing performs the interpolation of the DTM values in the moving mask. The window size varies from 3 (3x3) to 25 (25x25); when the observation is closer to the edge than the half-size, the mask is not filled and the procedure fails.
In order to avoid this problems we make available two possibilities: the first one is to fill the mask with zero values (flag -b); the second consists in disregarding these border observations (no flag).

The flag -q avoids the display of the progressive elaboration percentage (by default it appears).

The significance test level may assume real values from 0.0 to 100.0. Obviously the use of higher values for nalfa corresponds to more conservative approach, while the use of lower values is more suitable in case of large files in order to do not have too numerous tails.

The method parameter allows the choice of the interpolating model following the procedures previously presented. As the biquadratic and bicubic surface require the determination of many coefficients, in order to have enough degrees of freedom, the software prevents the user from jointly choose the size = 3 and the method = biquadratic or bicubic.

Finally as output files, four possibilities are predisposed (at least one must be specified):

- the raster file output, obtained by substituting in the input file the values, suspected to be outliers, with the interpolated ones;
- the raster file residuals, in which at each cell the difference between observed and interpolated value are assigned. To overcome the limit of this GRASS version (limit no more present in the newest version 5) which imposes the use of integers for the category values a factor (factor) can be used to multiply the differences to obtain whatever we want precision level;
- the raster file binrast, with zero or one value according as the input cell corresponds to regular or outlier value. This output file may be very useful in case of algebraic operations by using the r.mapcalc GRASS command;
- the site-list file dwsites, containing the list of suspected outliers, called downweighted sites, stored in ASCII format.

The dwsites contains:

the file name;
a general description;
the data list: east and north coordinates and description;
possible comments.

The output GRASS files are stored (in the same current mapset) in the <cell> or <site_lists> location depending on the file typology.

Between the raster elements, the color file, in the <colr> location), plays an important role in the validation map drawing; in fact the command r.outldetect automatically creates one of the files for every input raster file, with the following rules:

- the output color file is the same as the input one (input);
- the binrast color file associates to the zero and one values respectively the black white color (black is the standard color for the graphic GRASS windows);
- the residuals color file adopts a gray scale for residual values from 0 to 255, while other gradations from white to red characterize the residuals from 255 to the greatest absolute value reached. In this way the slight differences between residuals around to zero are more visible and with red shades, we perceive, with the corresponding detail, where the interpolated value deviates more from the observation.

An high level vision of the software structure is shown in Figure 1. The user interaction determines the information stream (1) needed to the data processing: the input file name, the validation method, the window dimension, the significance level, the flags and the input files names. The other necessary information comes from the raster database (2). The raster database itself is the destination of the three output raster files (3). On the contrary the site list file corresponds to a different stream (4).

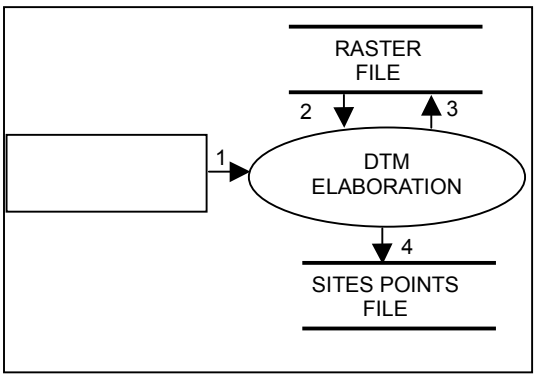


Fig. 1 –Functional structure of the r.outldetect command

Referring to the data processing let us recall only that, after the parameters parsing and the allocation of the dynamic space needed for the validation of a single value, the data into the window are iteratively read, the interpolation and the test are performed, the result is stored.

When the raster file scanning is completed, the dynamic memory is delivered and the open file are closed.

## 9. Description of the GRASS command r.outldetopt (outliers detection optimal procedure)

The procedure works on raster files and presents the following syntax:

usage:
r.oldetopt [-aq] input=name method=name [nalfa=value] wedsratio=value nsdsratio=value

flags:
-a Do not align output with the input
-q Run quietly

parameters:

input Name of existing raster file

method Type of enquiry
options: polin, colloc

nalfa Category of significance level (1->5%, 2->1%, 3->0.1%, 4->0.01%) (Only for polin)
options: 1,2,3,4

wedsratio Downsampling percentage along w-e axis
nsdsratio Downsampling percentage along n-s axis
Also in this case, as for r.outldetect, the input file is resampled into the current region by using the sampling e-w and n-s resolutions of input itself, unless the region parameters are forced by specifying the -a flag.
In this case, differently from the previous command, the user doesn't insert the window size, as into the program the alternatives are already stored. Between these, the software determines that corresponding to the maximum value of k, in such a way to identify the points that have to be left out of the research (they are those belonging to the frame with k+1 thickness). This is the reason of the absence of the -b flag.
The *method* parameter allows to choose the search type: polynomial or via collocation.
The flag -q avoids the display of the progressive elaboration percentage (by default it appears).
The significance level $\alpha$ (nalpha) refers only to the polynomial search and determines the preferableness threshold between two different polynomial surfaces with the same size mask. There are four possible values for nalpha (1, 2, 3, 4) corresponding to the probability values 5%, 1%, 0,1%, 0.01%. An error occurs if on the command line are indicated the collocation method and an alpha value.
The *wedsratio* parameter adjusts the down-sampling percentage along the w-e axis in such a way to speed up the optimal search by performing it only on a subset in the DTM. If, for instance, this value is set equal to 33, the procedure is applied only to the first of three adjacent consecutive points on the row; if it is set equal to 50 the procedure is applied every other column in each row. Analogously the *nsdsratio* down-samples the DTM but along the south-north axis.
There are no files as output and the result is simply the histogram of success of the different options and the corresponding mean value.

The structure
In the figure 2 we show the high level functional structure vision. The user interaction always determines the information stream needed to the data processing.: the input file name, the validation method, the significance level, the flags and the down-samples ratios. However in this case the raster archive only acts as data source: the input file is acquired (2) but the output is addresses to the screen (3).
Referring to the data processing, after the parameters parsing and the dynamic space allocation, the data into the moving windows are iteratively read. For each sub-sample datum the different interpolations are performed and compared. The result is stored in order to build the histogram.

When the raster file scanning is completed, the dynamic memory is delivered and the open files are closed.

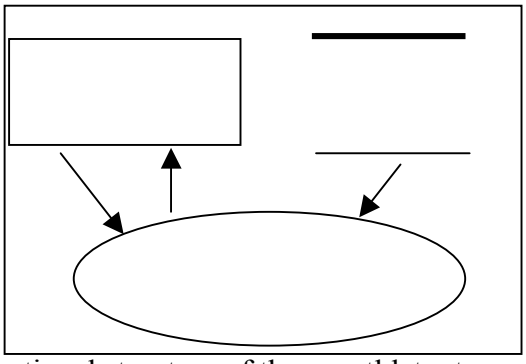

Fig. 2 –Functional structure of the r. outldetopt command

## 10. Conclusions

Methods for outliers detection in gridded surfaces are presented. Two new GRASS commands (that you can find attached to this paper) have been implemented. The paper presents the case of polynomial interpolation of the moving window data and the outlier detection by subsequent designed tests. Other methods have been also studied and implemented by the authors: documentation about these approaches can be found in the papers cited in the references. The implemented commands have been completely validate as refer to the interpolation techinques. As for the connected test procedures only average, linear and bilinear cases have been verified.

**Acknowledgements**

The researches here reported were financed by means of the "Programma MURST di ricerca scientifica di interesse nazionale COFIN 98: Riprese con laser a scansione, integrato da GPS, per la produzione di modelli numerici finalizzati alla realizzazione di cartografia 3D e ortofoto digitali", coordinated by Prof. Riccardo Galetto.

## 11. References

Benciolini B., Mussio L. and Sansò F., 1982, **An approach to gross error detection more conservative than Baarda Snooping**, Symposium of Commision III, I.S.P.

Brovelli M. A., Sansò F. 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.

Betti. B., Brovelli M. A., Venuti G., 2000, **Procedure di localizzazione per rimozione di outlier in dati laserscanning grigliati**, Atti della IV Conferenza Nazionale ASITA – Genova, 3-6 Ottobre 2000 Vo.1, pp183-188.

Brovelli M. A., 2000, **Individuazione e rimozione di errori grossolani in modelli digitali di superfici,** Geomedia 4/2000 pp. 24-27

Brovelli M. A., Reguzzoni M., Sansò F. and G. Venuti, **Outliers detection in data sets: a strategy with applications to DTM validation**, in print.

Clamons S. F: and Byars B. W., 1997, **GRASS 4.2 Programmer's Manual**, Baylor University GRASS Research Group.

Forlani G., 1990, **Metodi robusti di stima in geodesia e fotogrammetria**, in "Ricerche di Geodesia, Topografia, Fotogrammetria, 8", pp. 123-311.

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

USACERL, 1993, **GRASS 4.1 User's Reference Manual**, Champaign, Illinois, United States Army Corps of Engineers Construction Engineering Research Laboratories.