Locating some types of random errors in Digital Terrain Models[1]

Carlos López
Environmental and Natural Resources Information Systems
Royal Institute of Technology
Stockholm 100 44, Sweden[2]

Abstract:
The increasing use of Geographic Information System applications has generated a strong interest in the assessment of data quality. As an example of quantitative raster data, we analyzed errors in Digital Terrain Models (DTM). Errors might be classified as systematic (strongly dependent on the production methodology) and random. The present work attempts to locate some types of randomly distributed, weakly spatially correlated errors by applying a new methodology based on Principal Components Analysis. The Principal Components approach presented is very different from the typical scheme used in image processing. A prototype implementation has been conducted using MATLAB, and the overall procedure has been numerically tested using a Monte Carlo approach. A DTM of Stockholm, with integer-valued heights varying from 0 to 59 m has been used as a testbed. The model was contaminated by adding randomly located errors, distributed uniformly within -4m. and +4m. The procedure has been applied using both spike shaped (isolated errors) and pyramid-like errors. The preliminary results show that for the former, roughly half of the errors have been located with a type I error probability of 4.6% on average checking 1 per cent of the dataset. The associated type II error of the larger errors (of exactly +4m. or -4m.) drops from an initial value of 1.21% down to 0.63%. By checking another 1 per cent of the dataset such error drops to 0.34% implying that about 71% of the ±4m errors have been located; type I error was below 11.27%. The results for pyramid-like errors are slightly worse, with a type I error of 25.80% on average for the first 1 per cent effort, and a type II error drop from an initial value of 0.81% down to 0.65%.
The procedure can be applied both for error detection during the DTM generation and by end users, and it might be of use for other quantitative raster data examples.

I Introduction:

Data quality has become an important aspect of Geographic Information Systems (GIS) applications. John (1993) stated that "...very wrong answers can be derived using perfectly logical GIS analysis techniques, if the user is not aware of the particular peculiarities of their data..."
Although this statement holds for any kind of data, we will concentrate here on the case of Digital Terrain Models (DTM). We will not consider errors in the intermediate steps in the process of DTM generation, but we will concentrate on the errors in the final product..
Östman (1987) pointed out the fact that there exists no unique criteria or single measure for the "quality" of a DTM. He suggested that at least, one should consider accuracy in height,

---

[1] Published in International Journal of Geographical Information Science, 11, **7**, 677-698, 1997

[2] Permanent address: Centro de Cálculo, Facultad de Ingeniería, Universidad de la República, CC 30, Montevideo, Uruguay

slope and also curvature. In his paper, the performance of an "on line" editor is described. It attempts to find gross-errors while the DTM is being created. This editor was intended to correct mainly those errors that affect curvature or slope, so no substantial ability to improve the height accuracy is reported. He pointed out that gross errors typically account for less than 0.5% of the whole dataset.

Day *et al.* (1988), tested three methods for the generation of DTM based on SPOT data. The three results were compared with a very carefully, manually digitized 30 m. grid DTM, in terms of height differences. Even though the goal of the work was to compare the operational behavior of the algorithms, their paper does not propose any solution for the locations of the errors. The distribution function of the absolute size of such errors is also presented for each method. Similar results arises from the work of Theodossiou *et al.* (1990).

Most of the literature concentrates on the location of gross errors in early stages of the production process. Bethel *et al.* (1984) proposed the method of maximum chi-squared ratio for on line quality control. A dense regular grid of observation is assumed. The method is restricted to the detection of gross errors in the image matching process, and is based upon the hypothesis that a least-square adjustment with bicubed spline function may fit locally a DTM. The occurrence of big residuals suggest the existence of a blunder. He tested the methodology using spike-like blunders of no more than 10 feet (about 3 m).

Even though the goal of this work is not to analyze the different ways a DTM can be produced, Ackermann, (1995) points out that the trend in DTM production is towards a move from interpolation to aproximation, because the new generation equipment is able to produce many height values, but possibly with less accuracy than traditional equipment. The surface is aproximated using many points, instead of being interpolated from few carefully obtained values.

In a review of general statistical methods, Barnett *et al.* (1984) classify the current methodologies for error detection in two classes, provided the distribution of the variable is known a priori, or not at all. The first class includes methods that typically require also the estimation of the parameters of the distribution. They are unlikely to work properly here.

For DTM applications, only those methods which do not require any particular distribution (also called Informal Methods) are suitable. Such methods include techniques related to Cluster Analysis (see for example, Fernau *et al.*, 1990), graphical methods, Principal Component Analysis, and others. Some of them were originally developed for applications in social sciences, but are increasingly used in other fields.

Principal Component Analysis (PCA) is a widely known technique, both in digital image processing and in the treatment of time series. Its ability to extract uncorrelated patterns that enhance the interpretability of the data, and the possibility to reduce the number of patterns separating the physics from the noise is well known.

In the field of remote sensing, principal components analysis is used to reduce the number of image bands of information (Chavez et al., 1989; Eklundh et al., 1993). Essentially, the remote sensing image data is re-mapped into a new coordinate system reducing the dimensionality of the data. For example, rather than analyzing data from 7 Thematic Mapper bands, we can do a principal components analysis to reduce the number of image bands to 3 or 4 bands of information that contain most of the variance of the data. Normally, we disregard the information in principal components beyond the third band, as this information relates to noise in the data set. This direct approach is not suitable for DTM analysis since at most a single model is usually available.

Summing up, most of the literature uses tailored procedures to locate gross errors. They concentrate mostly on the DTM production stage, disregarding the problem faced by the

end user. This paper presents a new methodology for locating not merely gross errors but also some subtle ones, which can be applied both by the producer and the end user. The methodology were tested in a real DTM using numerically simulated errors, and the results are presented.

The paper is organized in eight sections. In section II a description and quick introduction to the PCA technique is sketched. Section III introduces the methodology for elongated DTM. In section IV the proposed procedure is described in terms of as a step-by-step recipe. Section V describes the Monte Carlo experiments designed to test the methodology. Section VI show the results in a particular DTM test case, using different error shapes. Finally, section VII contains a discussion and sectionVIII is devoted to conclusions where the results and proposed future work are discussed. Acknowledgements and References are included under headings IX and X.

II Principal Component Analysis in brief

The theory of PCA can be found in many textbooks, for example, Lebart *et al.* (1977). To make clear both the notation and the terminology, a brief sketch of the major concepts and results will be presented.

Given a table of n events of w variables, they can be represented as n points in the R^W space. They are supposed to be homogeneous , i.e., share the same measuring units. Each k-point (event) corresponds to a point in R^W, and each event is composed of w scalar observations. The case of w=3 is illustrated in the figure 1, where each point M_k represents an event. The PCA attempts to find the direction $\mathbf{e}_1$ of the vector in R^W space which minimizes the sum of distances M_kH_k squared, taken over all k (see fig. 1). The origin O is the centroid of the set of points. For the sake of clarity in the figure, points with negative coordinates are not shown.

The projection OH_k, which is also the scalar product of vector $\mathbf{M_k}$-$\mathbf{O}$ with the unitary vector $\mathbf{e}_1$, is called here the score (following Richman, 1986). Thus $\mathbf{M_k}$-$\mathbf{H_k}$ is orthogonal to $\mathbf{e}_1$. There is one score value associated with vector $\mathbf{e}_1$ for each point in R^W. Let us also assume that $\mathbf{e}_1$ is unique.

If all the values M_kH_k are zero, we have reduced the problem of original dimension w, to a one-dimensional one. All the variability in the observations is explained by a single vector $\mathbf{e}_1$. If this is not the case, we may try to repeat the procedure with the remaining variability M_kH_k, which belongs to a (w-1) subspace of R^W orthogonal to $\mathbf{e}_1$. The original measurements $\mathbf{M_k}$ - $\mathbf{O}$ can be replaced with the difference $\mathbf{OM_k}$ - $\mathbf{OH_k}$, which is equal to $\mathbf{M_k}$ - $\mathbf{H_k}$.

There should be a vector $\mathbf{e}_2$ (automatically orthogonal to $\mathbf{e}_1$) which minimizes the distance in the R^W space. The process continues the same way, being each new vector $\mathbf{e}_p$ orthogonal to all the previous ones, and there are w such vectors. Those vectors are called principal components (PC).

Each event $\mathbf{M_k}$ - $\mathbf{O}$ can be expressed as a linear combination of the PC

$$\mathbf{M}_k - \mathbf{O} = a_1(k) * \mathbf{e}_1 + a_2(k) * \mathbf{e}_2 + a_3(k) * \mathbf{e}_3 + \ldots + a_w(k) * \mathbf{e}_w \qquad (1)$$

It can be shown that the scores $a_i(k)$ associated with vector $\mathbf{e}_i$ are uncorrelated with those of vector $\mathbf{e}_j$. The vectors $\mathbf{e}_i$ are the eigenvectors of the covariance matrix of the data, and its components are named *loadings* in the literature. The sum of the corresponding eigenvalues equals the sum of the squares of the distances M_kH_k (Lebart *et al.*, 1987).

PCA analysis renders a sequence of principal components, which explains most (or all, for p=w) of the variance of the data. That implies that the error in approximating the data with a linear combination of their first p vectors is minimal for a given p<w; (p=1 in fig. 1).

Typical results show that p<<w for a good approximation in a wide range of applications, including this one. Since the w PC's form a basis in R^W space, they can replicate exactly any of the n points in the set, using the scores as weights.

III Locating errors in a single strip

We will concentrate our efforts in locating errors in an elongated DTM defined over a regular grid of size w*n, w <<n. We will use the term strip to denote such DTM shape, and it is assumed that w corresponds to rows, and n to columns. Such DTM can be regarded as being composed of n cross sections (named profiles) each composed of sets of w points. The height can be referred to as h(i,j), being i bounded by w, and j by n.

In typical situations, strong correlation can be expected between "close" data points, both within each profile and between adjacent ones. On the other hand, isolated random errors are assumed to be weakly correlated with its neighbors. So a procedure may be designed to give a criteria for selecting a candidate h(i,j) as being an error, based upon some index or statistic that highlights a low correlation situation. Systematic errors might show a completely different behavior strongly related with the producing method, and they will not be considered here.

Some authors (Hawkins, 1974; López *et al.*, 1994; López *et al.*, 1993) attempted to locate errors in tabular datasets using PCA, but their results and methods could not directly be applied since DTM data cannot be automatically regarded as a time series nor a multiple replication of an experiment. Some different approach should be suggested.

A procedure with two steps is proposed. First, locate the profiles that are likely to hold an error, and secondly (within each of them) pick the location of the error itself.

a.1.- locate the columns (profiles) likely to have candidates

In order to highlight an unusual profile some properties of PCA will be exploited. They will be illustrated using the concepts already presented. What follows is based on the ideas presented in López *et al.*, 1994.

First question is how PCA is connected with errors. Fig. 1a shows the typical distribution of the score 1, for a single strip of the DTM test problem (which will be described later), and fig. 1b, for the second score. The distribution is build from n values of the scores derived for a DTM of size w*n (25*150 in this case). For this particular strip, the first 10 PC explain more than 98% of the total variance.

As it can be seen, the distribution does not show any special shape. The mean should be zero in all cases (figs. 1a to 1d). We want to emphasize that the first score distribution is not at all symmetric, which is more likely in the second one (fig. 1b).

The progressive evolution towards a symmetric distribution is clear in the fig. 1c and 1d, which illustrates the 20th. and 25th. scores distribution. Another property that should be noted is the decay in the "width" of the distribution as an index function. There is no profile for the fig. 1a and fig. 1b which first or second score has absolute value greater than 40 m. For the fig. 1c, the same property holds, with limits 0.8 m. for the 20th. score, and 0.9 m. for the 25th. one (fig. 1d). This decay is also related with the eigenvalues associated with each PC. Almost the same behavior can be observed for other strips. Notice that the scores share the units with the DTM data (m), but *they have sign*.

In all these figures there are two small arrows pointing to an "*" and to an "O". The first one points to the values of the scores for the profile presented in fig. 2a, and the other does the same for a slightly modified profile (fig. 2b) which has an isolated "error" pointed by

the arrow which is not evident by looking at the profile alone, and will be left unnoticed in a 3-D like representation.
However, notice that for both the 20th. and 25th. scores (fig. 1c and 1d) the "O" value associated with the modified profile (fig. 2b) is now a marginal value, clearly separated from the rest of the samples.
This result holds in general for other strips. From the figures it can be seen that there are a few events (profiles, identified by the column) that renders marginally distributed values for *some* of the scores. What is the meaning of such events? Those events that behave in such a way are unusual events, and may contain the errors, as we have shown in the example. So the method will use some threshold interval for the j-score, and check all the events looking for the ones that lie outside the bounds (see Davies *et al.*, 1993). This is the first key idea.
The procedure (as sketched) disregard information coming from adjacent profiles. In other words, the profiles can be mixed, and the results will be the same. This information can in turn be used for handling some kinds of systematic errors, which is beyond our scope, or to extend the present method for non-isolated errors, which have not been considered at this stage.
Not all of the feasible set of scores will be used. Typically, the first ones are more related to physics (we meant by physics the underlying properties of the DTM) than to noise. That's why they are robust to single outliers (see the "O" in fig. 1a and 1b). So the first scores will not be checked. The appropriate limit for physics/noise qualification is subjected to direct experimentation, as will be presented later.
A related idea was proposed by Hawkins, 1974 who devised a similar approach. Instead of checking each score against a given (different) threshold, he computes the statistic T_2 and analyzes its distribution. T_2 is defined for profile (or sample) k as:

$$T_2(k) = \sum_{j=p}^{j=w} \frac{1}{\lambda_j} * a_j(k)^2 \qquad (2)$$

being λ_i the eigenvalues of the covariance matrix. In his work he says that the distribution of T_2 is a compound gamma function, assuming that the scores are normally distributed. Such property looks somewhat restrictive, but it is not strictly required. Any column profile is flagged if the value of T_2 is over a prescribed value.
Summing up, the approaches of both López and Hawkins take advantage of the fact that the non-systematic noise is likely to be important only in the weakest PC. The assumption that in this way it is possible to identify the columns containing the errors is important. What is remaining is how to identify the row for each candidate in any given column.

a.2 within each column (profile), find the rows that identify the candidates
$T_2(k)$ is supposed to be "small" in most cases, except under the effect of the outliers, when at least one of the scores will be bigger than usual. Since each score is a linear combination of the height values of the column (due to the scalar product), the most-likely error row is chosen as the one whose variation mostly affects the value of $T_2(k)$.
Therefore, a simple sensitivity analysis is carried out for each flagged "k" profile (column). The row that is responsible for the biggest change in the $T_2(k)$ function, will be classified as an "a" candidate to be an error. Also "b" and "c" candidates are selected, in decreasing order of priorities. This is the other key idea.

There exists however some reasons for using a different statistics. We may create a positive function $T_2^*(k)$ (slightly different from the one of Hawkins, 1974) which is a weighted average of the squares of some of the scores a_j, for a given column "k".

$$T_2^*(k) = \sum_{j=p}^{j=w} W_j * a_j(k)^2 \quad (3)$$

Note that as before the summation starts on the p-th score. The weights W_j can be chosen to scale the scores so that all terms in the summation have the same order of magnitude. The scores can be related with the aforementioned eigenvalues, or fixed in another way. We have used for each j-score, a weight W_j that makes $|a_j(k) * W_j| \le 1$ in 95% of the events. For example, if a_j is normally distributed with standard deviation σ_j, W_j will be exactly σ_j*1.95996.. (solution of $\frac{1}{2} erf(\frac{W_j}{\sigma_j \sqrt{2}}) = 0.95$). Hereinafter, 100-95%=5% will be called the penalty margin.

Such definition is more robust against the existence of outliers than the use of the eigenvalues. If the a_j are normally distributed, the statistic $T_2^*(k)$ is equal to the one of Hawkins except for a scale factor.

It should be pointed that $T_2^*(k)$ is not intended to be a weighted average of all the scores. Only those associated with the non-physical scores are included, according to the above mentioned limit p.

IV The proposed technique

Once we could locate errors in a strip, we could tackle the problem for a complete DTM. Given a regular gridded DTM model, with m rows and n columns, it can be viewed as formed by strips of width w rows and n columns. The union of all the strips is the whole DTM. In fig. 3 the whole matrix and a single strip is shown, while in fig. 4 the corresponding part of the DTM is presented.

For each row-wise strip, a set of "a", "b" and "c" error candidates can been selected, and its union will be called "row-wise candidates". This may be a complete solution for the problem. But why row-wise strips? There are no reason to discard analyzing the problem using column-wise strips. So doing the same for all column-wise strips will also cover the entire DTM and give three new global set of candidates. The union of the three sets can be called "column-wise candidates".

If the intersection of both sets ("row-wise candidates" and "column-wise candidates") is not empty, we have located those points in the DTM that behave atypically in both directions. These will form the *candidate set*, named also *guessed errors*.

The procedure involves five steps, and it can be sketched as:

Some remarks follows. We have used in the pseudo code a single value of strip width w both for rows and columns. This is not required but simplifies somewhat the tuning process,

```
Given a DTM as a matrix of size m*n
subdivide the DTM in row-wise and column-wise strips of width w

repeat until criteria are satisfied:
        a) find row-wise candidate set:
                a.1.- locate the columns likely to have candidates
                a.2.- within each column, find the rows that identify
                        the candidates
        b) find column-wise candidate set:
                b.1.-locate the rows likely to have candidates
                b.2.- within each row, find the columns that identify
                        the candidates
        c) intersect both sets
        d) evaluate criteria
        e) correct all errors
end
```

as will be presented later.

The method is in fact iterative, since all the distribution functions (covariance matrix, etc.) are modified as soon as an error is removed. In each iteration, a candidate set is obtained with the presented methodology. After the candidates are corrected (if they are true errors) or they have been verified, they will be excluded from the feasible set of "a", "b" or "c" candidates, and a new iteration takes place. The process is supposed to stop when some criterion is fulfilled; for example, it stops if the type I error is too big (see below). Each iteration will be named "step" in the following discussion.

V The experiment:

To test the methodology we have used a Monte Carlo simulation procedure. Using a real DTM as a test problem we first selected at random about 5% of the points within the dataset. Then all of them were modified by adding an error, which were also randomly selected from a given feasible set, and finally the methodology is applied in order to locate the errors and its results were recorded for later statistical processing.

The DTM used as a test problem covers an area of 7.5x5 km in Stockholm with 150x100 points with a 50 m grid spacing and 1 m height resolution. The area consists mainly of hilly terrain, with height values ranging from 0 to 59 m. Fig. 5 shows a mesh view of the DTM, while in fig. 6 its height distribution is presented. The 0 m areas can be seen to the left of the fig. 5 . The DTM has a mean height value of 20.83 m and its standard deviation is 9.47 m. For this type of DTM, errors within [-4 m,+4 m] are typical. Since the data is rounded to the nearest meter, there will be little chance to pick errors of one meter. Any "feasible" error should also be an integer number.

There is scarce guidance in the literature about the spatial distribution of real errors, and they are strongly related with the generation procedure. Since we have a single replication of the DTM, we were not able to test the procedure with real errors. We follow some authors (Bethel *et al*., 1984) in modeling errors as additive and isolated, being the added height chosen from a given set. As a feasible set we have used as a first example the values [-4,-3,-2,-1,+1,+2,+3,+4] meters, with equal probability, which is considered as a difficult

case. This is expected to model spatially uncorrelated errors, and we named it as *spike-like* errors (see fig. 7, left).

The selection of 5% as a typical value looks somewhat high when compared to the one reported by Östman, 1987. He found a typical value of 0.5% for the number of *gross* error occurrences, but here the worst errors are of absolute size 4m and they account for only 1/4 of the total.

As another alternative for an error shape model, we also tried a more structured one, which resembles a pyramid; once a point is selected, it is modified by adding a 2Δ meter error, and the eight points surrounding it adds only Δ (see fig. 7, right) We have selected Δ uniformly from the set [-2,-1,+1,+2]. We named this model *pyramid-like*, and it is expected to model some degree of spatial correlation in errors.

Once the model has been contaminated with errors, the described procedure is applied. Since one possible application for the method is to help locate errors while the model is being created, various measurements of success have been provided. The most important is the probability of type I error (Ounping, 1988), which measures the probability of classifying a correct value as wrong. It is estimated with the quotient between the number of good points classified as errors, in relation to the number of candidates suggested. This statistic can be calculated for each step (even in real applications), and the user can take the appropriate decision to continue the process or to stop.

Also, the relative importance of the errors is important, and not merely how many they are. Since in this test we know in advance how the original errors are distributed, the distribution of the identified errors can also be calculated. That will not be possible in the real operation, but it will render useful information at this stage. We will show results for errors of any size, but most figures will concentrate on errors of absolute size 4 m. The reason will be clear later. The number of errors remaining in the dataset (classified as good by previous steps) in relation with the initial population evolves monotonically as the process continues but might become stationary (no more errors are found irrespective of the number of iterations).

The method has some free parameters, and some previous estimations should be done in order to fix its value. They include

a) which scores will be considered as associated with noise.

b) what threshold interval will be used, to separate the marginal from the typical values in each distribution.

c) the penalty margin value

The strip width has been kept fixed, that means w has been held constant for all strips (both row-wise and column-wise). Instead of selecting a different number of scores for each strip, a single value has been chosen.

Further theoretical guidance for the choice of w for a given m and n has not been investigated in full. The number w can be assumed as a common divisor of m and n, and the minimum ratio n/w can be derived following Hawkins (1974). He gives some theoretical results, assuming that all the scores are normally distributed. Figure 8 gives guideline values depending on the confidence value α, and they are derived after Hawkins, 1974, 1994. We used n/w=150/25, 150/15 and 150/10 at best in our examples, which are denoted in the figure as "+", "o" and "*" respectively. They imply somewhat low values for the confidence α. Under some assumptions (Hawkins, 1974) these results may indicate that our results using w=10 should be more reliable than those of w=25.

However, from our results will be clear that given a DTM an appropriate value for w can be estimated by means of a similar experiment that the one in this paper. Further guidance will be given later.

Different and separate series of experiments were carried out, using values for w of 10, 15 and 25 elements and also applying *spike-like* and *pyramid-like* shaped errors.
From some runs that were done previously, it became clear that once w was fixed, the most important parameter is the number of uncontrolled scores, being the limit between the physics and the noise. This limit may be found in each particular DTM by means of a simulation like the one described in IV, or by applying some rule of thumb.
The other parameters were initially fixed at 2.5% for the threshold level, and at 5% for the penalty margin. Increasing the first one renders more candidates to the "a", "b" and "c" sets by means of selecting more columns (profiles). In order to avoid an empty candidate set, we increased the threshold level if too few candidates are suggested. The threshold interval is derived from the cumulative sampled histogram, using appropriate bounds. The penalty margin, provided it is not too small, is somewhat insensitive to the gross-errors, and so are the weights W_j.
The goal is to minimize both the probability of type I and II error. However, not both of the objectives are equivalent, and in some situations one may be more important than the other, depending upon the user. This will be discussed later.

VI The results from the Monte Carlo simulation

Results for spike-like errors:
We will denote as *candidates* or *guessed errors* the group of coordinates (i,j) suggested by any single step of the procedure. The *true errors* are those elements in the previous set that also belong to the known errors set.
Fig. 9a shows the average Type I error evolution up to 5 per cent depuration effort (see below). The y-axis shows the evolution of the type I error calculated as the number of points missclasiffied as errors compared with the number of candidates, averaged after 50 replications of the random error set. The x-axis shows the effort, defined as the fraction of the dataset already revised. An effort of 100 per cent implies that all possible points have been checked, since the effort per step depends on the number of uncontrolled scores and the threshold level. We interpolated our results to prescribed effort values using splines. Each polyline corresponds to different strip width and also number of uncontrolled scores. For w=10, 15 and 25 we left uncontrolled 6 scores.
From this result it is clear that in the first 1 per cent effort the measured Type I error is low, being below 5 per cent for the lower values of w. The dashed horizontal line corresponds to the limit of 95%. Obtaining an error rate over that level is worse than to pick the points at random, since that level is the noise initially seeded into the DTM. The limit was shown as being constant, despite the fact that such probability grows slightly as soon as an increasing number of errors has been found. For 2 per cent effort and over somewhat poorer results may be achieved, but certainly better than chance.
But good results in the type I error is not the whole picture. From a DTM producer's point of view, what is more important is to minimize errors still in the DTM, so the type II error is more representative. It measures the probability of classifying as good a wrong value. The type I error in fig. 9a only *count* the success and the failures, but do not reflect the relative importance of the errors. The absolute value of the error is not considered.
Table 1 shows the evolution for a single replication of the experiment of the distribution of the errors. The original distribution of the error size is shown in bold. For example there were originally 116 points with error +3. The rows below show the remaining errors after finishing each step of the cleaning process. For example, after step 4 only 49 of the 116 original errors of size +3 remain in the dataset.

As expected the errors of size +1 and -1 were very difficult to locate, since the original DTM has a resolution of that size. We omitted the corresponding effort involved.

Size of errors	-4	-3	-2	-1	1	2	3	4	Total
Original # errors	**93**	**85**	**89**	**82**	**83**	**98**	**116**	**93**	**739**
# after step 1	36	53	78	82	81	83	72	43	528
# after step 2	21	26	66	82	79	74	56	29	433
# after step 3	20	19	57	81	79	69	52	25	402
# after step 4	18	17	50	80	78	68	49	24	384
# after step 5	18	17	48	80	78	67	47	24	379

Table 1 Error´s size distribution for a single experiment, when 5 terms are left uncontrolled

We limit our Type II error calculations to errors of absolute size exactly 4m, because as table 1 shows, the other cases are less prone to be located and its type II error will be unaffected. In fig. 9b the evolution of the Type II error is presented. Notice that the best results are for w=25, but for w=15 very similar results are achieved. The initial Type II errors is 1.21% in all cases, so it can be reduced to 0.64 with only 1 per cent effort.
The same behaviour were noticed for other combinations of w and number of uncontrolled scores; better results are obtained for the Type I error for lower w values, while the opposite happends for the Type II error.
In fig. 10 the type I error up to the 1 per cent effort is represented as a function of the number of scores not controlled. The continuos line were obtained by spline interpolation. Notice that, irrespective of w, the relative minimum lies between 5 to 10, and the absolute minimum correspond to the option w=15, nearly twice the optimum number of uncontrolled terms.
In fig. 11 the Type II error up to the 1 per cent effort is analyzed in relation with the number of uncontrolled scores. The results for w=10 are worse than the other options, but it is not so evident the differences between w=15 and w=25. So a bigger w is preferred, while the optimum choice for the scores to be left uncontrolled is less crucial. On the other hand, from fig. 10, the w=10 option is worse as far as the type I error is concerned, while w=10 results in similar figures.
This result supports some conclusions:

a) the optimum choice for the number of controlled scores is more or less independent of the goal (minimize the type I or II error), while the situation for w is slightly different. In the former, a smaller w is preferred while in the later a bigger one is better.

b) there seems to be a compromise w value for a given DTM, which can be estimated from simulations before truly apply the methodology, or in another way. In this case it should be near w=15.

c) the results for type II error are based only on errors of size 4m. Slightly different figures can result from considering other categories.

As mentioned before, a crude estimation for the optimum w can be done without performing a simulation. From the outlined results, the best w is nearly twice the optimum number of uncontrolled scores. That value should be in turn near the limit between the physically meaningfull and the noisy eigenvectors, and to define the former there is no unique rule in the literature. We applied the one suggested by Hawkins, 1974 which in our case proves to

give similar values for different w. See the original reference for a justification. The procedure might be the following:

1) choose some suitable initial value, based on fig. 8.

2) for each strip, calculate the covariance matrix, and its eigenvectors. We will assume that they are sorted in ascending order, depending on the corresponding eigenvalue. Find the p that makes

$$\min_{j=1..p}\left(\max_{i=1..w}\left(abs\left(e_{ij}^{(p)}\right)\right)\right) > 0.25 / \sqrt{w}$$

Since this value depends on each strip, we select the median of the results for all strips. In our DTM, for w=10, 15, 20, 25 and 30, we obtain p=3, 4, 4, 3 and 4. Estimate the optimum number of uncontrolled scores as twice that number and the proper w as four times. In our case, it will suggest 6 to 8 uncontrolled scores, and for w something between 12 and 16.

3) choose the definitive w as close as possible to the value obtained in the previous step, while being a divisor for both m and n.

This will give a rough estimate, while a simulation like the one described in this work will give more reliable results. Simple or flat terrain will be well represented with less terms, so w will be smaller, while for complex ones it should be greater.

Results for pyramid-like errors

In previous results, all the calculations have been performed assuming that typical errors are completely isolated ones, uncorrelated in space. This is not true in practice even though the shape of the errors has not been analyzed in the literature. Bethel *et al.*, 1984 used spike like errors only, and we present here the results of a different model: the pyramid-like error.

Since pyramid errors are spatially autocorrelated, the chance of locating them is lower. That´s due to the underlying design of the methodology, strongly oriented towards independent errors in space.

There are also some other details regarding the "accounting" procedure. We will consider as a candidate not only any point which is both a row-wise and column-wise candidate, but also its immediate neighbors. So for every candidate, nine points are checked. However, due to computational simplicity no effort has been made to take into account the overlap of candidates for the same step (i.e. if both a point and its neighbor are selected, there are points that count twice). So the results are somewhat pessimistic in terms of the type I error.

The behavior observed in figs. 12a and b are very similar to the ones observed for the spike-like error shape model, althoug the numbers are more pessimistic. The first 1 per cent effort renders an acceptable Type I error, but the second and others are somewhat higher. The analysis of the Type II error should take into account that the initial value in this case is 0.81%, while in fig. 9b it was 1.21. This imply that the procedure reduces at most the Type II error in 79% with only 1 per cent effort.

For comparison purposes, fig. 13 shows the type I errors for the first step in terms of the uncontrolled scores, for different values of w, and for 50 replications of the experiment.

The type II values in fig. 14 shows that (as happened before with isolated errors) the weak dependence upon the number of uncontrolled scores; the associated error is fairly "high" (0.64% per cent up to the 1 per cent effort, and 0.42 up to the 5 per cent effort, starting with an initial value of 0.81%,) when compared with the other error shape model. This imply that 79% of the gross errors remain in the dataset with 1 per cent effort, and 51% cannot be located even with a 5 per cent effort. So again, as expected this results are worse than those for isolated errors, but still might be of use.

Summing up, as expected for the pyramid error shape, for most of the cases the 1 per cent effort still renders a low type I error, being rather insensitive to the number of uncontrolled terms (but in the range 4 to 8) and to the alternative w=10 or w=15.
In terms of the type II error, about 21 per cent of the worst errors can certainly be located with only 1 per cent effort of the procedure, and it may go up to 49% with 5 per cent effort. The best results follow the ones obtained with isolated errors, being w=25 the best option. The number of uncontrolled scores is again between 5 and 10.
The conclusion is that the method proves to be effective in identifying a significant amount (up to one third) of the big errors with limited effort. For better Type I results, a smaller w is suggested, while for Type II optimization a somewhat greater might be the option irrespective of the shape model assumed. The number of uncontrolled scores is between 5 and 10 in any case.
No special pattern of the location of the errors found were noticed during the runs. Such aspect may be investigated in the future, with a wider set of DTM.

VII Discussion:

From the results obtained, a process can be devised to pinpoint an important part of the larger random errors in a raster dataset. Further actions strongly depend on the application the user is involved with.
In a production environment, some action can be taken to check these identified isolated values. In photogrammetric measurements these checks can be done before removing the stereopair. The goal here is to remove most of the errors, i.e. diminishing the type II error, while the type I error is less crucial.
On the other hand, the end user is left alone in most cases, because he may not be able to go to the original data sources. Therefore he should be worried by the risk of modifying a value that is correct, so the type I error is more important.
The results show that it can be assumed that up to the 1 per cent effort, most candidates are errors. The associated Type I error can be less than 5%, as has been shown for isolated errors, and around 25% for pyramid-like ones for proper choice of the parameters. The Type II error is defined here only for errors of absolute size 4 m, and it can be reduced 64% (for isolated errors) and 21% (for pyramid-like errors) checking only 1 per cent of the dataset.
Every step produces a candidate set, and once this set is obtained, any standard procedure can be used to replace the outliers with suitable values. As long as the dataset is progressively being corrected the risk that a point classified as an error is correct is higher, and some caution should be taken.
The procedure may be unable to locate non random errors, i.e., if a region has been affected by an improper choice of the control points, or there are edges along the rows or columns that arise from an improper matching of a partial DTM, for example. Further studies will be necessary to clarify this aspect.
The test area is considered to be a difficult one. Rough terrain, narrow channels, steep hills, and small water areas are typical, all of them may easily mask errors. The DTM itself should not be considered as free of errors, and it has been used "as is". This fact is common to most users of this kind of data, so it is believed that such a situation will not limit the range of applications of the ideas presented.
It should be noticed that there are two integer parameters free: the strip width and the optimum number of scores that should be left uncontrolled. Here we have taken a fixed strip width w for both row-wise and column-wise strips, and we have considered a single value also for the optimum number of scores for all strips.

Loosely speaking, the optimum number of uncontrolled scores should be somewhat stable, provided that w is not too small. This assertion comes from a weather analogy. In that case, increasing w is the same as adding a new weather station to the set. The optimum value is related with the number of typical weather systems, and that is certainly independent of the number of observations being taken.

If w is less than this (unknown) optimum value, poor results will be achieved. We noticed it using w=10. On the other hand, if w is too big, the number of events will be small compared with w, and unstable results may appear. In the extreme case of w>number of events, the covariance matrix will be no longer positive definite, and zero eigenvalues will appear. This optimum may serve as an objective number which characterize terrain complexity being lower for smooth terrain. Further work with other DTM may render a closer relationship between roughness and this number, also linked with grid size. We provide some rough estimation rule, which has to be tested in other cases.

There are two other non-integer parameters: the penalty margin and the threshold level to be considered. Some calculations have been carried out and the penalty margin value has been found to have a rather low influence on the final result.

The threshold level is the key for the amount of work to be done. If its value is too low, only very few values will be chosen as candidate errors. They certainly have a good chance to be true errors (i.e., the Type I error is lower). But some others, which are also errors, may not be picked even in further steps (i.e., the Type II error will stay high). This should be the choice for an end user.

On the other hand, if its value is higher, more candidate errors will be selected, but the efficiency will be lower. That also implies that the type I error may be higher, and that may be unacceptable. In an automatic production environment, where there is a chance to check the values, this may not be a problem. In a semi-automatic one, the operator may quickly become bored of checking points that are correct, and the overall procedure may be considered as poor, even if most of the errors are in fact removed.

The end-user has limited ways to check the values. Maybe he can use a three-dimensional plot of the interpolated surface near the candidates, or a contour plot, or something else to get an idea of the surroundings. But taking into account that a) he certainly will not check that way too many candidates, and b) he also has no definitive way to find the true value, he will limit the search to a small set, where may be included the "worst" errors. So he will choose a somewhat lower threshold.

A comment about the computer time requirement: the procedure involves for each step, the computation of (m/w).(n/w) covariance matrices of size (w,w), which takes $\left[(n/w).O(n^2)+(m/w).O(m^2)\right].O(w^2)$ operations, find its eigenvectors after $\left[(n/w)+(m/w)\right].O(w^2)$ operations, and project each strip to calculate the scores (which in turn requires (m+n).w operations). Some other operations are required but depend linearly on m and n. In our example, for a DTM of size m=150, n=100, and for w=10, it requires about 3 seconds in a PC486 (and 2 in SUN Sparc 10) per step, both working with MATLAB, so the overall procedure is considered cheap in terms of computer time.

VIII Conclusions:

A new methodology to locate random errors in quantitative raster data has been presented, and tested in a grid-based DTM as an example. The methodology is iterative, and proves to be robust, rendering type I error rates near 5% for isolated errors. Roughly speaking, it also located half of the 4 m isolated errors checking only 1 per cent of the database.

The process involves the decomposition of the DTM into strips, and requires a Principal Component Analysis (PCA) of each one. That is not the usual way of using the technique in image processing. Further simple calculation renders three sets of candidates to be considered. The stripping process is done both row-wise and column-wise as a cross check, and a even more reduced set of candidates is obtained.
Some experiments were performed using a DTM with heights between 0 and 59 m seeded with randomly located additive errors, with amounts up to 4 m. This value was considered to be on the order of the errors in the model. Two error shape models were considered: one completely isolated (like a spike) and the other with some arbitrary regular shape (pyramid like). Even though it has been assumed that those shapes are typical, their representativeness for real DTM errors is still to be investigated.
The method has some parameters left free to the user, and some guidance is provided. However, at least for the first candidates, the high rate of success obtained proved to be fairly insensitive to some of the parameters.
In the case of using the algorithm in a semi-automatic production environment, the method warns the operator about possible errors before the stereopair is unmounted, enabling a new measurement. In a fully digital production environment, some correlation thresholds have been usually fixed weakly to diminish computer time. This method may help in selectively strengthening the correlation thresholds in unlikely points.
In the case that there is no possibility to verify the errors, e.g. for end users, the algorithm will help to locate the most unlikely values; they may be replaced with the aid of some suitable interpolation method. If there are some independent sources (cartographic maps, etc.) they could be used for checking.

IX Acknowledgments:

This work was started during a stay at the Environmental and Natural Resources Information Systems of the Royal Institute of Technology, Stockholm, Sweden, under the supervision of Prof. Friedrich Quiel. Dr. Juan Echague made several suggestions which certainly improved this work. The author wishes to acknowledge BITS (Swedish Agency for International Technical and Economical Cooperation) and CONICYT (Consejo Nacional de Investigaciones Científicas y Técnicas, Uruguay) for partial funding. Some of the basic routines have been programmed by Elías Kaplan, and his collaboration is also acknowledged. Computer resources have also been kindly supplied by the Department of Numerical Analysis and Computer Science at the Royal Institute of Technology, and by the Centro de Cálculo of the School of Engineering, University of the Republic, Montevideo, Uruguay.

X References:

Ackermann, F. 1995. Digitale Photogrammetrie - Ein Paradigma-Sprung. Zeitschrift für Photogrammetrie und Fernerkundung, 3/95, 106-115

Barnett, V.; Lewis, T. 1984. Outliers in statistical data. John Wiley and Sons. 463 pp.

Bethel, J. S.; Mikhail, E. M. 1984. Terrain surface approximation and on-line quality assessment. International Archives of Photogrammetry and Remote Sensing. Commission III, V25, A3a, 23-32

Chavez, P.S.; Yaw Kwarteng, A. 1989. Extracting Spectral Constrast in Landsat Thematic Mapper Image Data using Selective Principal Component Analysis. Photogrammetric Engineering and Remote Sensing, V 55, N 3, March 1989, 339-348.

Davies, L.; Gather, U. 1993 "The identification of multiple outliers" Journal of the American Statistical Society. Sept. 1993. V 88, N 423, 782-801.

Day, T.; Muller, J.P. 1988. Quality assessment of Digital Elevation Models produced by automatic stereo matchers from SPOT image pairs. International Archives of Photogrammetry and Remote Sensing, V27, B3, Commission III, 148-159.

Eklundh, L; Singh, A. 1993. A comparative analysis of standardized and unstandardized Principal Components Analysis in remote sensing. Int. J. Remote Sensing, V 14, N 7, 1359-1370.

Fernau, M. E.; Samson, P. J. 1990. Use of Cluster Analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part I: Transport Patterns. Journal of Applied Meteorology, V 29, N 8, 735-750

Hawkins, D. M. 1974. The detection of errors in multivariate data, using Principal Components. Journal of the American Statistical Association, V 69, 340-344.

Hawkins, D. M. 1994. Personal communication

John, S. A. 1993. Data integration in a GIS - The question of data quality ASLIB Proceedings, V 45, N 4, 109-119.

Lebart, L.; Morineau, A.; Tabard, N. 1977. Techniques de la description statistique: Methodes et logiciels pour l'analyse des grands tableaux. Ed. Dunod, Paris, 344 pp.

López, C.; Kaplan, E. 1993. "Principal Component analysis applied for outlier location and missing value problem wish surface wind and pressure data" Internal report, 22 pp. Available from http://www.fing.edu.uy/~carlos/papers/rep93_5/viento.htm.

López, C.; González, E.; Goyret, J. 1994. Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos. Estadística V46, N 146-147, 25-54.

Östman, A. 1987. Quality control of Photogrammetrically sampled Digital Elevation Model. Photogrammetric Record, 12 (69) 333-341.

Ounping, Gong. 1988. Experiment and discussion on the method of blunder location. International Archives of Photogrammetry and Remote Sensing, V27, B3, Commission III, 841-849.

Richman, M. B. 1986 Review article: Rotation of principal components. Journal of Climatology, V 6, 293-335.

Theodossiou, E. I.; Dowman, I. J. 1990. Heighting accuracy of SPOT. Photogrammetric Engineering & Remote Sensing, V 56, 1643-1649.

Figures