A Theoretical Approach to Modeling the Accuracy Assessment of Digital Elevation Models

Fernando J. Aguilar, Francisco Agüera, and Manuel A. Aguilar.

Abstract

In this paper, a theoretical analysis is presented of the degree of correctness to which the accuracy figures of a grid Digital Elevation Model (DEM) have been estimated, measured as Root Mean Square Error (RMSE) depending on the number of checkpoints used in the accuracy assessment process. The latter concept is sometimes referred to as the Reliability of the DEM accuracy tests.

Two theoretical models have been developed for estimating the reliability of the DEM accuracy figures using the number of checkpoints and parameters related to the statistical distribution of residuals (mean, variance, skewness, and standardized kurtosis). A general case was considered in which residuals might be weakly correlated (local spatial autocorrelation) with non-zero mean and non-normal distribution. Thus, we avoided the "strong assumption" of distribution normality accepted in some of the previous works and in the majority of the current standards of positional accuracy control methods.

Sampled data were collected using digital photogrammetric methods applied to large scale stereo imagery (1:5 000). In this way, seven morphologies were sampled with a 2 m by 2 m sampling interval, ranging from flat (3 percent average slope) to the highly rugged terrain of marble quarries (82 percent average slope).

Two local schemes of interpolation have been employed, using Multiquadric Radial Basis Functions (MRBF) and Inverse Distance Weighted (IDW) interpolators, to generate interpolated surfaces from high-resolution grid DEMs. The theoretical results obtained were experimentally validated using the Monte Carlo simulation method.

The proposed models provided a good fit for the raw simulated data for the seven morphologies and the two schemes of interpolation tested ($r^2 > 0.96$ as mean value). The proposed theoretical models performed very well for modeling the non-gaussian distribution of the errors at the checkpoints, a property which is very common in geographically distributed data.

Introduction

Digital Elevation Models (DEMs) have become an important tool in many remote sensing applications like SAR simulators, classification of ground cover types, orthorectification of satellite and airborne images, and Geographic Information Systems (GIS). Consequently, DEMs provide GIS with a vertical dimension which allows them, for example, to capture 3D information from an area in a 2D environment. It is precisely in this field where a very beneficial integration has been taking place in the last few years between imagery analysis and GIS. In fact, DEMs, together with orthoimages, are becoming the main cover of GIS, contributing a rapid and cost effective methodology for updating spatial information, which allows for briefer conventional cycles for cartographic updating (Baltsavias and Hahn, 1999).

However, in spite of the usefulness and relatively high cost of this type of digital products, they are usually presented without an associated estimate of their reliability. In this context, reliability might be defined as the degree of correctness to which the DEM accuracy figures have been estimated (Li, 1991). Curiously, while nobody would purchase a television set without an instruction booklet and a warranty against potential defects, it is still common to acquire expensive digital geographic data without any kind of quality documentation. Thus, the responsible DEM user must be able to answer these three questions:

1. What precisely is the application for the DEM?
2. What type of DEM will best meet these needs?
3. How do I know that I am getting what I ordered (Daniel and Tennant, 2001)?

This work focuses precisely on the answer to, at least, part of the last question, because if GIS users are not aware of its DEM accuracy, perfectly logical GIS analysis techniques can lead to incorrect results. In other words, the data may not be fit-for-use in a certain context (Fisher, 1998). For this reason, measuring the positional error of geo-spatial data is becoming one of the major research issues in the area of quality assessment of spatial data (Shi and Bedard, 2004)

At present, the great majority of vertical accuracy standards is based on computing the vertical accuracy of a finite sample data set (checkpoints) from the differences between data set coordinate values and coordinate values from an independent source of higher accuracy for identical checkpoints (Maune *et al.*, 2001a). In some cases, the vertical Root Mean Square Error (RMSE) is converted to vertical accuracy at an established confidence level, normally 95 percent, assuming a normal distribution of the residuals. However, vertical

Department of Agricultural Engineering. Almería University, Ctra. de Sacramento s/n, La Cañada de San Urbano. Escuela Politécnica Superior. 04120 Almería, Spain (faguilar@ual.es).

Photogrammetric Engineering & Remote Sensing
Vol. 73, No. 12, December 2007, pp. 1367–1379.

0099-1112/07/7312–1367/$3.00/0

errors in a DEM do not often follow a normal distribution (López, 1997a; Maune *et al.*, 2001b).

Likewise, it would be very interesting to know the sample size (number of checkpoints) needed to compute the DEM RMSE with a certain reliability, because checkpoints should be roughly three times more accurate than the expected accuracy to be verified (FGCC, 1984) and so, a large number of checkpoints may be costly to produce (Li, 1991).

Therefore, in 1998 the Federal Geographic Data Committee (FGDC, 1998) published "The National Standards for Spatial Data Accuracy (NSSDA)" where a statistical and testing methodology was implemented for estimating the positional accuracy of points on a map and in the digital geospatial data produced, revised, or disseminated by or for the Federal Government. However, the NSSDA are perhaps not too specific with regard to the number and distribution of checkpoints sample. For example, it recommends the distribution of a minimum of 20 checkpoints to reflect the geographical area of interest and the distribution of error in the data set. Obviously, since outliers should always be eliminated, a higher number should be taken. Even in the NSSDA Appendix 3-C, Section 3, it states that, due to the diversity of user requirements for digital geospatial data and maps, it is not realistic to include statements that specify the spatial distribution of checkpoints. Thus, data and/or map producers must determine checkpoint locations and number. Some agencies have adapted the NSSDA standards, which is the case of the U.S. Federal Emergency Management Agency, which specifies that a minimum of 20 checkpoints will be used in each of three or more land-cover categories representative of the area being mapped, i.e., a minimum of 60 checkpoints.

From this brief introduction, we can deduce that there are a number of issues needing further investigation for modeling uncertainties in spatial data and analysis. These include theoretical studies, method development, and application issues (Shi *et al.*, 2002). Thus, the main objectives of this work are the following:

1. To develop theoretical models for estimating the degree of correctness, so-called, Reliability, to which a grid DEM accuracy figures, measured as RMSE, have been estimated depending on the number of checkpoints used in the accuracy assessment process. Notice that we avoided the "strong assumption" of normal distribution of errors assumed in some of the previous works (e.g., Li, 1991) and in the majority of the current standards of positional accuracy control methods (e.g., NSSDA).
2. To validate the developed models using Monte Carlo simulation.
3. To analyze the practical application of the proposed models when they are applied to a finite sample of checkpoints.

This paper has been structured in the following four main sections:

1. Theoretical approach: where a detailed theoretical development of the two proposed models to reliability calculation is outlined.
2. Experimental validation: where the data sets and experimental design used to validate the theoretical models developed in the last section are described. At the end of this section, the numerical procedure of Monte Carlo method employed to simulate observed data is also explained.
3. Results and Discussion: where the results corresponding to the residuals populations generated are shown and discussed. Furthermore, it is presented an analysis to study the degree of agreement of observed and predicted reliabilities for the two models developed in the first section. A sensibility analysis is also approached to find out the effect of the uncertainty of standardized kurtosis estimation over the calculated reliability.
4. Conclusions: where the main findings of the present work are briefly expounded.

Theoretical Approach

Suppose that the difference in height between terrain surface and interpolated DEM surface is a random variable X. This being the case, a sample of size n height differences may be used to estimate the mathematical expectation (mean value) and standard deviation (dispersion) of the said height differences on the whole surface (Li, 1988). A general case was considered in which height differences may be weakly correlated (local spatial autocorrelation) with non-zero mean and non-normal distribution. The population mean and variance of random variable X will be denoted by μ and σ^2, respectively, where μ represents the systematic error or bias for the interpolated DEM, while σ characterizes the on-systematic random component of that error.

This sample $X = \{x_1, x_2, \ldots, x_n\}$ can be considered as the residuals calculated at the n check points used for computing the DEM accuracy. Note that the theoretical analysis in this study is based on the assumption that the checkpoints are error free. Reliability, meaning the degree of correctness to which the DEM accuracy figures have been estimated, could be obtained as the coefficient of variation of the resulting sample variance (Li, 1991). In this way, reliability should be considered as a quantitative and relative value for measuring the error of the DEM accuracy test. The differences between the two theoretical models proposed having a bearing on how the said coefficient of variation is determined.

Model 1

In Model 1, the general case, reliability would be calculated according to Equation 1:

$$R = \frac{Sd(Sd_x)}{Sd_x} \tag{1}$$

where R is reliability and $Sd(Sd_x)$ the standard deviation of the standard deviation computed for sample X (residuals at the n checkpoints). But, which is the statistical distribution of the sample variance Sd^2? First, the mathematical expression of the sample variance will be drawn as:

$$Sd_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}; \quad \text{where } \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \text{ and} \tag{2}$$

$\overline{x}$ being an unbiased estimator of the population mean because $E(\overline{x}) = \mu$ (operator E denotes the mathematical expectation). Note that the expected value of Sd_x^2 for a sample size n is then given by Equation 3. That is to say, the sample variance, expressed as shown in Equation 2, is not an unbiased estimator of population variance. This subject will be further reconsidered.

$$\begin{aligned} E(Sd_x^2) &= E\left(\frac{\sum_{i=1}^{n}(x_{i-}\ \mu)^2}{n} - (\overline{x} - \mu)^2\right) \\ &= \frac{\sigma^2 n}{n} - E(\overline{x} - \mu)^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \end{aligned} \tag{3}$$

Similarly, the expected variance of the sample variance is given by the following expression:

$$Var(Sd_x^2) = \frac{1}{n}\left(\frac{(n-1)^2}{n^2}\mu_4 - \frac{(n-3)(n-1)}{n^2}\sigma^4\right); \quad \text{where } \mu_4 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n}, \text{ and} \tag{4}$$

μ_4 being the fourth central moment of sample X. The algebra from which Equation 4 is derived is rather tedious and has not been included in this paper.

In any case, our objective is to deduce the expression $Var(Sd_x)$. For this, a change of variable is carried out as follows: $Sd_x = W$ and $Sd_x^2 = Y \Rightarrow Y = W^2$, i.e., we have obtained a general expression $Y = G(w_1, w_2, \ldots, w_m)$ that can be approximated by a linear function using the Taylor series method (see Equation 5). The first order of the Taylor series expansion of G may be considered locally as a good approximation and simplifies the definition of error propagation because we can write the following equation (e.g., Burrough and McDonnell, 1998):

$$Var(Y) \approx \sum_{i=1}^{m} \sum_{j=1}^{m} \rho_{ij} Sd_{wi} Sd_{wj} \frac{\partial Y}{\partial w_i} \frac{\partial Y}{\partial w_j} \tag{5}$$

where ρ_{ij} is the correlation coefficient between w_i and w_j. Applying the latter equation to our case ($m = 1$) the following expression can be deduced:

$$Var(Y) = Sd_w^2 \left(\frac{\partial Y}{\partial W}\right)^2 = Sd_w^2 4w^2 \Rightarrow Var\left[Sd_x^2\right] = 4Var\left[Sd_x\right] Sd_x^2. \tag{6}$$

That is to say,

$$4Var(Sd_x) Sd_x^2 = \frac{1}{n}\left(\frac{(n-1)^2}{n^2}\mu_4 - \frac{(n-3)(n-1)}{n^2}\sigma^4\right). \tag{7}$$

Operating in Equation 7 and supposing that $Sd_x^2 \approx \sigma^2$, we can write:

$$\frac{Sd^2[Sd_x]}{Sd_x^2} = \frac{1}{4n}\left(\frac{(n-1)^2}{n^2}\frac{\mu_4}{\sigma^4} - \frac{(n-3)(n-1)}{n^2}\right). \tag{8}$$

But the expression $\mu_4/\sigma^4 - 3$ is known as the standardized kurtosis denoted by γ_2. The standardized kurtosis measures the relative peak (positive value) or flatness (negative value) of a given distribution compared to a normal distribution which presents a γ_2 value of zero. Substituting γ_2 and operating in Equation 8, it yields:

$$100\frac{Sd[Sd_x]}{Sd_x} = R(\%) = \frac{100}{2\sqrt{n}}\sqrt{\left(\frac{(n-1)^2}{n^2}(\gamma_2 + 3) - \frac{(n-3)(n-1)}{n^2}\right)} \tag{9}$$

with R(%) being reliability measured as a percentage. Thus, reliability measures the error, specifically the variation coefficient of this error, purely due to sampling the continuous terrain surface with a finite number of checkpoints for computing the global interpolated DEM surface accuracy (Aguilar *et al.*, 2006). Equation 9 will be considered as "Model 1."

As we noticed before, the sample variance shown in Equation 3 is a biased estimator of population variance. Thus, if sample variance was written with an unbiased estimator, it could be expressed as follows:

$$\hat{S}d_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \Rightarrow E(\hat{S}d_x^2) = \sigma^2. \tag{10}$$

Operating in the same way as above, we can derive the following expression for reliability:

$$R(\%) = \frac{100}{2\sqrt{n}}\sqrt{\left(\gamma_2 + 3 - \frac{n-3}{n-1}\right)}. \tag{11}$$

Note that if the residual population distribution trend is normal, i.e., $\gamma_2 = 0$, then the last equation can be simplified:

$$R(\%) = \frac{100}{\sqrt{2(n-1)}}. \tag{12}$$

Equation 12 reproduces the model developed by Li (1991), where a normal distribution of the residuals population is assumed.

Model 2

Because the most widely used global accuracy measure for evaluating the performance of DEMs is the Root Mean Square Error (RMSE) (Li, 1988; Wood, 1996), and bearing in mind the practicality of the proposal methodology for computing reliability, a model based on the variation coefficient of the resulting RMSE has been proposed:

$$R = \frac{Sd(RMSE_x)}{RMSE_x} \tag{13}$$

where R is reliability and $Sd(RMSE_x)$ the standard deviation of the RMSE computed from sample X (vertical residuals at the n checkpoints) following the next expression:

$$RMSE_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \tag{14}$$

with x_i being the vertical residual at the i^{th} checkpoint and n the sample size. Notice that $RMSE_x$ is not referred to x-coordinates, but it is related to the vertical error or z-coordinates.

Again, the algebra to solve Equation 13 is rather tedious and has not been included in this paper because it would occupy a lot of space. In any case, the following expression could be obtained as follows:

$$R(\%) = \frac{100\sigma^2}{2\sqrt{n}(\sigma^2 + \mu^2)}\sqrt{\gamma_2 + 3 + \frac{4\mu\gamma_1}{\sigma} + \frac{4\mu^2}{\sigma^2} - 1} \tag{15}$$

with γ_2 being the standardized kurtosis and γ_1 the skewness parameter obtained from the expression μ_3/σ^3, where μ_3 is the third central moment of sample X. Skewness characterizes the degree of asymmetry of the residuals distribution around its mean.

However, the last model may be quite complex to apply because four residuals distribution parameters need to be estimated. It can be simplified if the systematic errors are considered null, i.e., the mathematical expectation of X is zero ($E(X) = \mu = 0$). Now Equation 15 can be rewritten as:

$$R(\%) = \frac{100}{2\sqrt{n}}\sqrt{\gamma_2 + 2}. \tag{16}$$

The last expression will be considered as "Model 2".

Experimental Validation of the Proposed Models

Study Sites and Original Datasets

The two study areas are located in Almería, Southeastern Spain. The first one is situated in what is known as "Comarca del Mármol," specifically in the municipal area of Macael. It is a zone of marble quarries with a high level of extraction activity, which has formed a terraced and artificial relief, with a predominance of steep slopes and even vertical walls. The second study area is situated in "Comarca del Campo de Níjar," bordering on the "Cabo de Gata" Nature

TABLE 1. GENERAL CHARACTERISTICS OF THE TOPOGRAPHIC SURFACES STUDIED. ALL SURFACES ARE GRID DEMS COMPOUNDED BY 10,000 POINTS (100 BY 100 POINTS WITH A 2 METER BY 2 METER SPACING. THE STANDARD DEVIATION OF UNITARY VECTORS PERPENDICULAR TO THE TOPOGRAPHIC SURFACE IS DENOTED BY SDUV

	NÍJAR					MACAEL	
Terrain descriptive statistics	Flat	Rolling1	Rolling2	Slightly mountainous	Mountainous	Steep-rugged hillside	Highly rugged
Average elevation (m)	166.54	176.94	195.42	178.96	215.16	762.29	922.5
*Z*max-*Z*min (m)	6.70	17.25	23.08	34.43	45.17	201.32	116.48
Z coefficient of variation (%)	0.97	2.28	2.20	3.79	3.98	6.52	4.28
Average Slope (%)	3.30	9.27	10.01	19.42	31.18	82.14	65.12
Slope coefficient of variation (%)	48.02	45.66	69.59	58.62	38.27	30.91	77.01
SDUV (m)	0.03	0.09	0.10	0.19	0.31	0.35	0.64

Reserve. This is an area with a smooth relief sculpted by natural agents.

For the development of this study topographic surfaces were selected covering an area of 198 meters by 198 meters (approximately 3.92 hectares), two situated in Macael and five in Níjar; the morphological characteristics can be examined in Table 1. It is interesting to observe the great variability of the morphologies utilized, both in terms of their roughness and of their average slope. As a roughness descriptor, we used the Standard Deviation of Unitary Vectors (SDUV) perpendicular to the topographic surface calculated as described by Aguilar *et al.* (2006).

The DEM of each topographic surface was obtained automatically by stereo image matching. Later on, a revision and a manual edition of the grid DEM (areas with poor digital correlation because the lack of texture, moving or deleting badly posed points, adding mass points, etc.) were carried out to improve its adjustment to the real terrain surface. The photogrammetric flight presented an approximate scale of 1:5 000 and was taken with a Zeiss RMK TOP 15 metric camera using a wide-angle lens with a focal length of 153.33 mm. The negatives were digitized with a Vexcel 5000 photogrammetric scanner with a 20 μm geometric resolution and a radiometric resolution of 24-bits (8-bits per RGB channel). In the case of the Níjar study area, the DEM was constructed using the module *Automatic Terrain Extraction* of the digital photogrammetric system Leica Geosystems SOCET SET® NT 4.3.1. For the study area of Macael, the DEM was constructed using the modules *ImageStation Automatic Elevations* and *ImageStation DTM Collection* of the digital photogrammetric system Z/I Imaging ImageStation® SSK. In both cases, we obtained a final DEM in grid format with a spacing of 2 m by 2 m, orthometric elevations, map projection UTM Zone 30 North, and European Datum 1950.

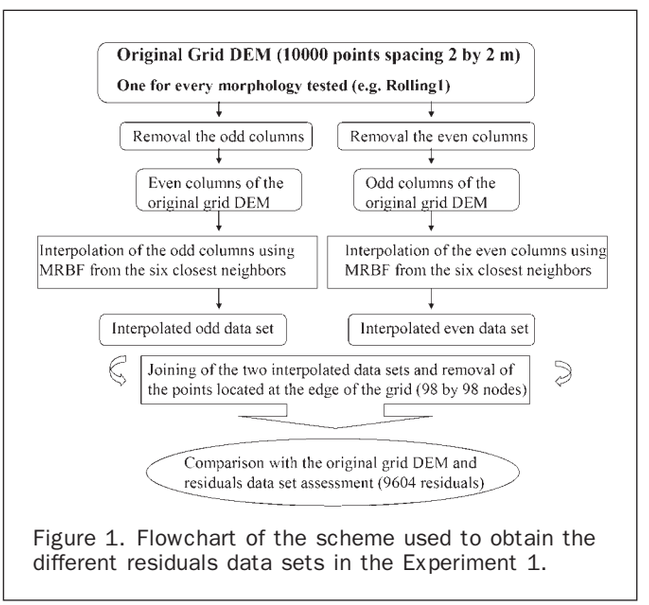

Figure 1. Flowchart of the scheme used to obtain the different residuals data sets in the Experiment 1.

Experimental Design

The generation of the different residual data sets from the interpolated grid DEM and from every morphology tested were obtained by means of two different schedules, called Experiment 1 and Experiment 2, respectively. In both of them, residuals populations were generated uniformly distributed to reflect the geographical area of interest with a grid spacing of 2 m by 2 m.

Experiment 1

In this first experiment the residuals population for every morphology was obtained by means of the algorithm shown in Figure 1. The interpolation schedule employed, six closest neighbors from adjacent columns, guarantees that the separation distance between sample points and interpolated points takes values from 2 m to 2.82 m (see Figure 2). This decrease in interpolation distance, together with the introduction of a systematic interpolation schedule based on mobile square windows, sought to generate residuals population with a low absolute value of errors and local spatial autocorrelation (weak autocorrelation) following Tobler's Law. Since dispersion value around the mean is low, it would be very probable to find a high presence of outliers, defining outliers as errors over three times the size of standard deviation.

In as much as the method of interpolation, Multiquadric Radial Basis Function, has been applied due to its suitability for interpolating from high resolution DEMs, it performs very well with local support and will produce residuals with a low absolute value (Aguilar *et al.*, 2005), which is the main objective of experiment one. The selected value of the smoothing factor was zero because it usually yields better results (Aguilar *et al.*, 2005).

Experiment 2

In the second experiment the residuals population was obtained using the schedule shown in Figure 3. The seed data set for every morphology, compound by 128 ground points including their *X, Y,* and *Z* coordinates, was extracted from each original grid DEM by stratified random sampling (four by

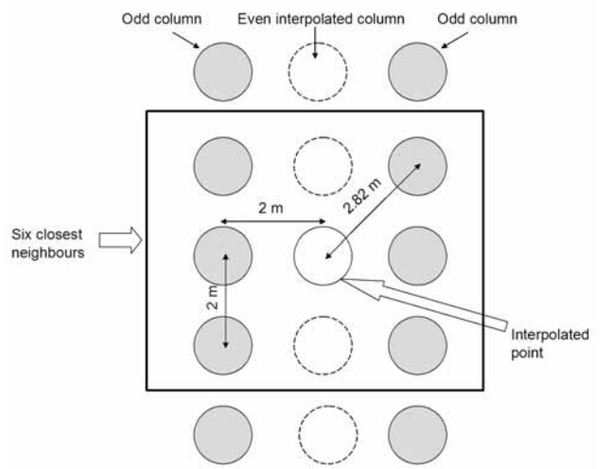

Figure 2. Scheme of the interpolation used in Experiment 1. Even interpolated columns from odd original grid DEM columns.

four sampling quadrants), which guarantees a homogenous distribution of the seed data over the whole working area. Based on the initial data sets obtained, a grid DEM with 2 m by 2 m spacing was generated for every morphology using the Inverse Distance Weighted (IDW) interpolation method from the local support of the six closest neighbors. The IDW is a well-known interpolation method where weighting is assigned to data using the inverse separation distance to a power as a weighting function (Aguilar *et al.*, 2005). To be specific, a weighting power parameter of 2 was used in our case.

Note that in Experiment 2, the separation distance between the interpolated points and the set of local support points will be greater and more variable than in Experiment 1, and so the interpolation errors will be greater. That is to say,

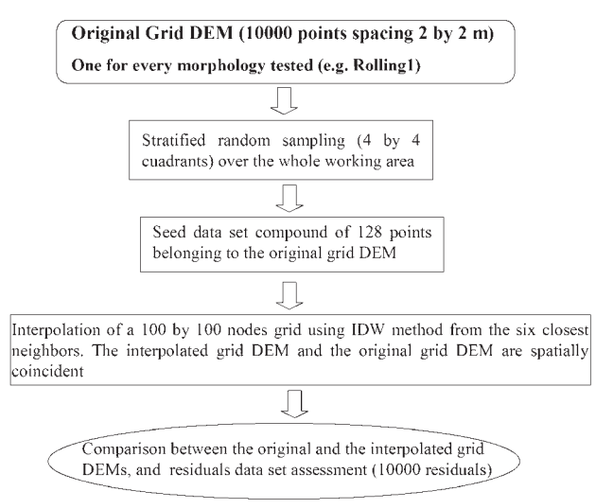

Figure 3. Flowchart of the scheme used to obtain the different residuals data sets in Experiment 2.

the probability of outliers occurring will also be lower. In this sense, the interpolation schedule employed in Experiment 2 can be considered less sophisticated and artificial than the one used in Experiment 1.

Since the interpolated and the original grid DEMs are spatially coincident and equally spaced, it is easy to compute the height differences between both DEMs at each node to obtain a set of 10,000 residuals which will be considered as the residuals population.

For both experiments it is also important to find out whether the errors at the checkpoints may be spatially autocorrelated, because the two proposed models have not taken into account the effect of spatial autocorrelation between residuals. Thus, we explored this property for every data set using the grid correlogram, which indicates how well grid values correlate across the grid. The correlograms were calculated using the methodology implemented in the software SURFER® 8 (Golden Software, Inc., 2002). As an example, in Figure 4 we can see how the profile of the grid correlogram computed for Rolling 2 terrain, belonging to Experiment 1, shows a positive spatial correlation (coefficient of correlation $\rho > 0$) between residuals located at close points (separation distance < 10 m), but a weak or non-existent correlation ($\rho \approx 0$) between residuals located at a distance greater than 10 m. Notice the sharp hill located at the center of Figure 4 (high ρ values for low separation distances) and the wide sill (low ρ values for separation distances > 10 m) which surround the said hill. Therefore, the residuals showed a local spatial autocorrelation. Roughly the same pattern was observed in the rest of the data sets for both experiments.

Numerical Approximation Using the Monte Carlo Method

The Monte Carlo simulation method has been applied to validate the proposed models because it is easily implemented and generally applicable for simulating error propagation without using analytical equations (Heuvelink *et al.*, 1989). The aim of this method is to calculate the reliability $R = Sd(RMSEx)/RMSEx$ repeatedly, with size n input samples X that are randomly selected from each original residual data set.

It must be remarked that the application of model one assumes the use of the standard deviation as a measure of uncertainty, meanwhile model two employs a more common measure as RMSE. In fact, RMSE is likely the most widely used global accuracy measure for evaluating the performance of DEMs. Furthermore, let us remember that standard deviation and RMSE present similar value when residuals tend to an expected value near to zero ($E(X) = 0$). Only under that hypothesis it is possible to apply the model one using the

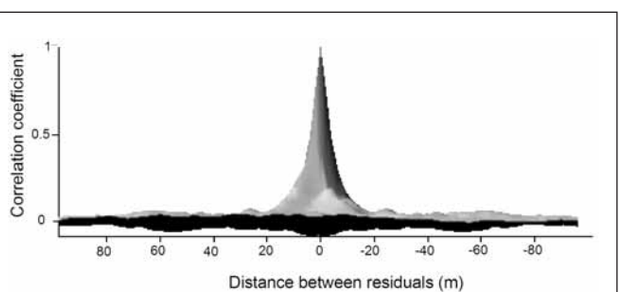

Figure 4. Profile of the grid correlogram computed from the residuals population data of Rolling 2 terrain for Experiment 1.

RMSE as an uncertainty measurement. Taking into account this subject and bearing in mind the practicality of the proposed methodology, the simulated reliability was computed as the coefficient of variation of the RMSE instead of the standard deviation.

In our case, a stratified random sampling in 4 by 4 quadrants was carried out to guarantee the homogeneous distribution of sampling points within the scope of the original grid DEM (Burrough and McDonnell, 1998). The sample size of residuals, i.e., the number of checkpoints n, took the following values: 16, 32, 64, 128, 192, 288, 384, 576, 960, and 1,440. If the RMSE is computed for every random size n sample X, and the process is run M times, a random sample of size M from the distribution of $RMSEx$ can be obtained in the same way as the mean and standard deviation can be estimated from the sample, and so the reliability corresponding to samples of checkpoints n could be simulated. Because the accuracy of the Monte Carlo method is inversely related to the square root of the M number of runs, a reasonably high number of M runs = 1,000 was used in this work for the robustness of the process.

Results and Discussion

Characteristics of the Residuals Populations Generated

Table 2 shows some statistics of the residuals population for the seven morphologies studied in the case of Experiment 1. The values of skewness, in general, may be considered as close to a normal distribution (an absolute value < 0.5 could be accepted (Daniel and Tennant, 2001). At the same time, the mean values are very close to zero, which indicates a low presence of systematic errors. However, it must be stressed that the high values of standardized kurtosis obtained (> 0 in all cases), i.e., the residuals distribution, are leptokurtic for all morphologies. Leptokurtosis indicates there are more occurrences far away from the mean than predicted by a normal distribution, probably because of the presence of outliers or gross errors. In other words, the probability distribution function peaks at the center and it has longer tails.

Table 3 shows the statistics of the residuals population for every morphology tested in the case of Experiment 2. A quick analysis of these figures allows us to observe higher values in the standard deviation (σ) than the ones registered from Experiment 1. That is to say, the residuals or errors are greater and present more variability. Likewise, these standard deviations are highly correlated with the roughness of every morphology, lower for flat terrain and greater for highly rugged terrain. Meanwhile, the skewness continues to be similar to the one observed in Experiment 1, the standardized kurtosis is much lower and, therefore, the residuals populations show a moderate leptokurtosis. Again the presence of mean values lower than the standard deviations indicates the absence of systematic errors.

The outliers can corrupt the true statistical distribution of the errors. From a statistical point of view, they cannot be considered as belonging to the same population as the other observations (López, 1997a). Therefore, outliers should be removed to normalize, in a certain way, residuals distribution. Taking into account that outliers are generally defined as errors over three times the size of standard deviation σ, the "3-sigma" rule ($|\text{error-}\mu| > 3\sigma$) was applied to the different data sets to remove outliers and obtain "corrected" data sets. This technique is generally applied for removing outliers in the DEM quality assessment of lidar data (Daniel and Tennant, 2001).

In Tables 4 and 5 are shown the statistics of the residuals population from Experiments 1 and 2, respectively, when the 3-sigma rule was applied. According to Torlegard *et al.* (1986), and analyzing errors in real DEMs, outliers typically account for less than 3 percent of the data set, which is fairly coherent with the residuals removed applying the 3-sigma rule for our case. Notice that, after 3-sigma rule application, a remarkable decrease in kurtosis values must be highlighted, which is a sign indicating a tendency toward normalization of the residuals distribution. Anyway, it is important to point out that the residuals data sets continued being non-normal when they were checked using the Smirnov-Kolmogorov test (Royston, 1982) with a confidence level of 95 percent. It was especially true in the data from Experiment 1.

TABLE 2. STATISTICS OF THE RESIDUALS POPULATION FOR EVERY MORPHOLOGY TESTED. RAW DATA FROM EXPERIMENT 1 (INTERPOLATOR: MULTIQUADRIC RADIAL BASIS FUNCTION)

Morphology	μ (cm)	σ (cm)	Skewness (γ_1)	Standardized Kurtosis (γ_2)
Mountainous	0.02	11.16	0.39	12.18
Rolling 1	0.01	2.72	0.84	13.20
Flat	0.01	2.08	0.64	23.99
Steep rugged hillside	0.48	41.01	0.60	21.55
Highly rugged	−0.87	135.02	0.12	31.95
Slightly mountainous	0.12	6.36	1.12	21.12
Rolling 2	−0.08	1.84	−0.37	29.66

TABLE 3. STATISTICS OF THE RESIDUALS POPULATION FOR EVERY MORPHOLOGY TESTED. RAW DATA FROM EXPERIMENT 2 (INTERPOLATOR: INVERSE DISTANCE WEIGHTED)

Morphology	μ (cm)	σ (cm)	Skewness (γ_1)	Standardized Kurtosis (γ_2)
Mountainous	−2.15	166.84	−0.49	1.67
Rolling 1	1.59	52.08	0.54	2.50
Flat	−1.65	16.75	−0.69	3
Steep rugged hillside	−68.44	448.56	−0.64	1.81
Highly rugged	99.45	884.57	1.08	4.73
Slightly mountainous	−18.57	122.32	−0.49	2.90
Rolling 2	6.24	53.06	−0.44	4.93

TABLE 4. STATISTICS OF THE RESIDUALS POPULATION FOR EVERY MORPHOLOGY TESTED. "3-SIGMA" RULE CORRECTED DATA FROM EXPERIMENT 1 (INTERPOLATOR: MULTIQUADRIC RADIAL BASIS FUNCTION)

Morphology	μ (cm)	σ (cm)	Skewness (γ_1)	Standardized Kurtosis (γ_2)	Residuals removed (%)
Mountainous	−0.12	8.20	−0.04	3.07	2.32
Rolling 1	−0.13	1.88	0.41	3.79	2.73
Flat	−0.01	1.35	−0.04	4.15	2.16
Steep rugged hillside	0.22	30.34	0.04	3.16	1.90
Highly rugged	−1.51	87.57	0.02	6.05	2.25
Slightly mountainous	0.03	4.42	0.10	4.39	2.49
Rolling 2	−0.03	1.11	−0.25	5.11	2.15

TABLE 5. STATISTICS OF THE RESIDUALS POPULATION FOR EVERY MORPHOLOGY TESTED. "3-SIGMA" RULE CORRECTED DATA FROM EXPERIMENT 2 (INTERPOLATOR: INVERSE DISTANCE WEIGHTED)

Morphology	μ (cm)	σ (cm)	Skewness (γ_1)	Standardized Kurtosis (γ_2)	Residuals removed (%)
Mountainous	4.22	154.19	−0.14	0.80	1.30
Rolling 1	0.10	45.86	0.22	0.82	1.84
Flat	−0.88	14.75	−0.16	1.05	1.72
Steep rugged hillside	−51.66	414.75	−0.28	0.53	1.21
Highly rugged	43.72	732.46	0.38	2.45	2.37
Slightly mountainous	−13.61	107.76	−0.09	1.58	1.90
Rolling 2	6.64	45.60	0.10	1.31	1.70

Validation of the Proposed Models

To understand how the models predict the observed data obtained in the Monte Carlo simulation, the regression coefficients (r^2) of the different fits are shown in Tables 6 and 7 for Experiments 1 and 2, respectively. In the case of the raw residuals data sets for Experiment 1, Model 1 presented the best fit, closely followed by Model 2 (Table 6, raw residuals). The same occurs with the raw residuals from Experiment 2, where, once more, Model 1 appears to be slightly better than Model 2 (Table 7). We must remember that the expression used for Model 2 is really a simplification of the general Model 2 given by Equation 15, where the absence of systematic errors is supposed. When a great part of the outliers are removed by means of the 3-sigma rule, the skewness and standardized kurtosis diminish, and Model 2 becomes as good as Model 1 with r^2 values > 0.98 for the two experiments. In Figures 5 and 6 (a and b), we can see the plots of the observed data versus the predicted data for Model 1 for the two sets of experimental data, highlighting how the fit results slightly improve when the 3-sigma rule is applied, especially in the case of the higher values of standardized kurtosis originated in Experiment 1.

On the other hand, Li's model presented low r^2 values when applied to the raw data derived from Experiment 1, although it performed somewhat better in the case of corrected data (Table 6). This evidence confirms its inefficacy to model non-gaussian residuals population with a high presence of outliers, or what is the same, of high standardized kurtosis. In fact, when the input data are closer to a normal distribution, as in Experiment 2 (Table 7), it can be observed how the results offered by Li's model improve notably, even though they are very close to the ones offered by the models proposed in this article for the 3-sigma corrected data. Nonetheless, as we can see in Figures 5 and 6 (c and d), the errors registered when using Li's model are systematic. Thus, Li's model underestimated the observed data in all the morphologies because it is based on the restrictive hypothesis of residuals normality. The quantitative differences between Model 1 or Model 2 and Li's model are significant both in raw residuals data set and in corrected data set. As could be expected, the results offered by Li's model tend to improve when the residuals distribution is closer to a normal distribution. Shortly, Li's model shows notable difficulties to model the reliability of the accuracy assessment from non-gaussian error distributions.

The three models tested have not taken into account the effect of spatial autocorrelation between residuals, although this DEM error property has been reported by several authors (Wood, 1996; Fisher, 1998; Weng, 2002). If the error at the checkpoints is in fact not independent, we could state that the actual sample size is lower than n. Therefore, the use of a data set of n spatially autocorrelated residuals would lead to a variability of RMSE which is lower than it should be. Thus, the models proposed in this paper perform quite well with weakly correlated errors (local spatial autocorrelation shown in Figure 4), but they could have difficulties to handle the problem of heavily correlated errors in space. Similar problems have been reported by López (1997a) about certain error detection procedures which obviate the spatial autocorrelation phenomenon (López, 1997b; Felicísimo, 1994). This problem should be approached in another study in the future.

Finally, Figure 7 shows a 3D graphical representation of Model 1. A high leptokurtic residuals population forces us to use a great number of checkpoints to obtain a low value theoretical error in the assessment of DEM accuracy. Conversely, normal residuals distributions allow us to obtain low values of reliability with a reasonably low number of checkpoints. It must be pointed out that surveyed checkpoints can be expensive to obtain. So we need to limit the number of checkpoints because the DEM accuracy assessment should not cost more than the DEM data acquisition.

That way, in Figure 7 we can observe an interesting property. The slope of the curve for a constant value of γ_2 significantly diminishes for a number of around 150 checkpoints. That is to say, in terms of efficiency it would be suitable to measure the DEM error in at least 150 checkpoints distributed over the whole working area. Any effort focused on using more than around 150 checkpoints will mean little improvement in the accuracy figures of the DEM assessment

TABLE 6. REGRESSION COEFFICIENTS R^2(%) FOR THE FITTING OF THE DIFFERENT MODELS TESTED TO THE SIMULATED DATA FROM THE MONTE CARLO METHOD. DATA FROM EXPERIMENT 1 (INTERPOLATOR: MULTIQUADRIC RADIAL BASIS FUNCTION)

	Raw residuals population			"3-sigma" rule corrected residuals		
Morphology	Model 1	Model 2	Li`s model	Model 1	Model 2	Li`s model
Mountainous	99.83	99.46	64.52	99.02	99.73	86.01
Rolling 1	99.75	98.56	64.58	99.67	99.86	83.08
Flat	99.01	96.83	53.92	99.76	99.43	85.89
Steep rugged hillside	97.13	93.61	58.68	99.35	99.86	86.37
Highly rugged	94.17	89.31	52.32	98.66	99.11	75.01
Slightly mountainous	95.39	91.24	60.19	99.17	99.31	82.15
Rolling 2	95.96	91.66	52.34	99.33	99.30	79.77
Average value	97.32	94.38	58.07	99.28	99.51	82.61

TABLE 7. REGRESSION COEFFICIENTS R^2(%) FOR THE FITTING OF THE DIFFERENT MODELS TESTED TO THE SIMULATED DATA FROM MONTE CARLO METHOD. DATA FROM EXPERIMENT 2 (INTERPOLATOR: INVERSE DISTANCE WEIGHTED)

	Raw residuals population			"3-sigma" rule corrected residuals		
Morphology	Model 1	Model 2	Li`s model	Model 1	Model 2	Li`s model
Mountainous	98.86	97.69	96.70	97.99	96.64	99.78
Rolling 1	99.43	98.56	92.65	99.48	99.08	98.95
Flat	99.80	99.35	89.27	99.68	99.64	97.43
Steep rugged hillside	99.79	99.90	94.07	99.55	99.89	98.82
Highly rugged	98.11	96.35	85.07	98.11	96.78	93.51
Slightly mountainous	97.95	96.27	92.07	99.15	98.34	96.38
Rolling 2	94.39	90.53	88.45	99.36	99.05	96.75
Average value	98.33	96.95	91.18	99.04	98.49	97.37

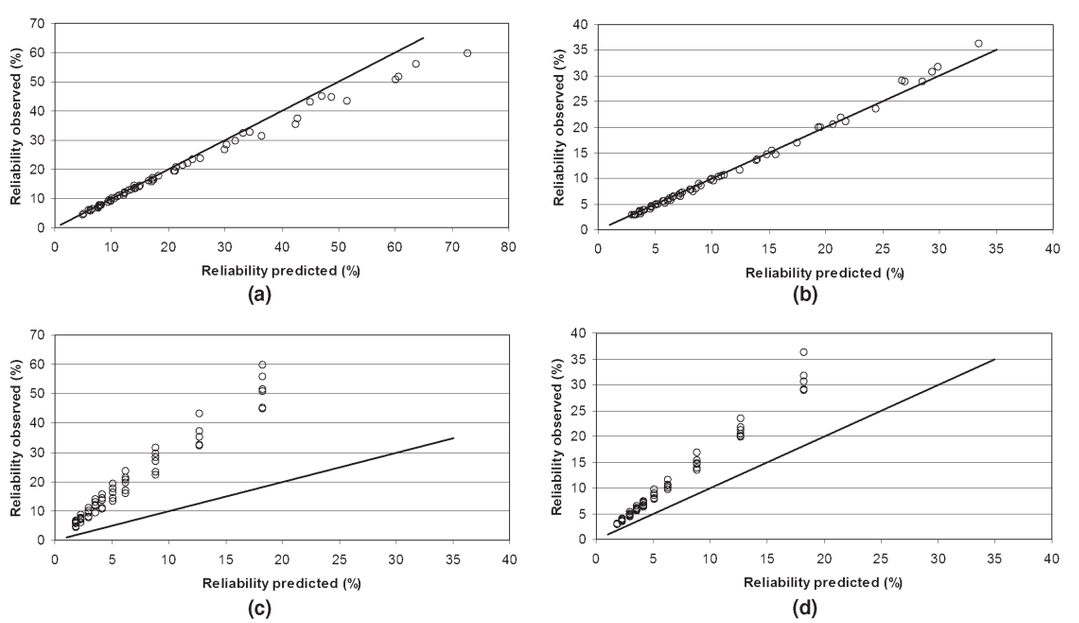

Figure 5. Plots of the Reliability observed and predicted by the model one (a and b) and Li's model (c and d) for the case of Experiment 1. (a) and (c) are models fitted using the raw residuals population. (b) and (d) are models fitted afterward "3-sigma" rule residuals correction. Diagonal line would mean a full agreement between observed and predicted reliability (r^2 = 100%).

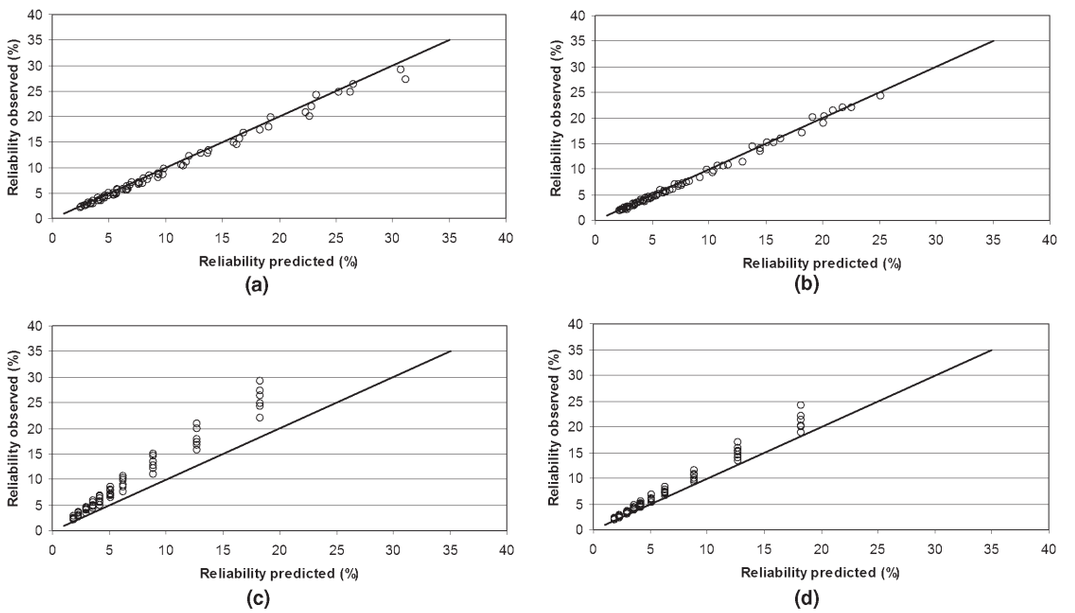

Figure 6. Plots of the Reliability observed and predicted by the model one (a and b) and Li's model (c and d) for the case of Experiment 2. (a) and (c) are models fitted using the raw residuals population. (b) and (d) are models fitted afterward "3-sigma" rule residuals correction. Diagonal line would mean a full agreement between observed and predicted reliability (r^2 = 100%).

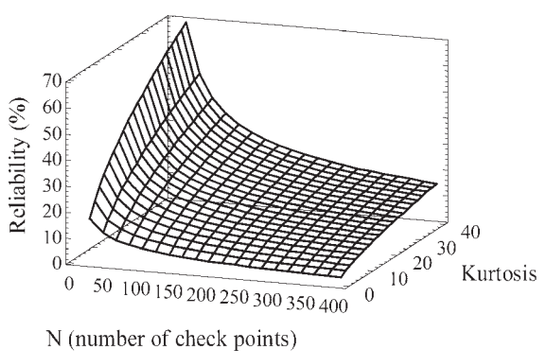

Figure 7. 3D graphical representation of the model one.

process. Recently, Ariza *et al.* (2005) found that the use of the minimum sample size (20 checkpoints) proposed by the NSSDA method leads to a variability of the horizontal accuracy results in the order of 11 percent, which means an insufficient confidence level of 89 percent with regard to the theoretical value of 95 percent. Along these lines, the same authors stated that the checkpoints sample size must be in the order of 100 to have a 95 percent confidence level in the planimetric or horizontal accuracy estimated using the standard NSSDA. It must be stressed that the synthetic residuals population used in the above mentioned article presented a perfectly normal distribution, and so, the situation could clearly be worse working with non-normal distributions.

Standardized Kurtosis Estimation

From a practical point of view, the models developed could not be used to design tests of DEM accuracy for a fixed level of reliability because, although the number of necessary checkpoints could be selected *a priori*, the standardized kurtosis of residuals population would be unknown. However, the theoretical models could be applied to estimate the reliability obtained in the measure of global DEM accuracy (RMSE) from a certain sample of n residuals $X = \{x_1, x_2, x_n\}$. In this case the population standardized kurtosis could be estimated from the following expression:

$$\gamma_2^* = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)Sd_x^4}\sum_{i=1}^{n}(x_i-\overline{x})^4\right) - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (17)$$

with γ_2^* being the estimated standardized kurtosis, n the number of checkpoints, and Sd_x the sample standard deviation. Obviously, a good estimation of γ_2 will result in a good approach to the theoretical value of the reliability of the DEM test accuracy computed from sample X. Thus, attempts to circumvent the problem of kurtosis estimation from size n samples have concentrated on using the Monte Carlo method to find out an answer to the following question: how good is Equation 17 for estimating γ_2? The procedure employed was the following. We extracted 1,000 size n samples (n = 16, 32, 64, 128, 192, 288, 384, 576, 960, 1,440, and 2,880) by means of the stratified random sampling of 4 by 4 quadrants from every morphology (original grid DEM). For every sample, γ_2^* was computed according to Equation 17, obtaining its expected value (mean value) and dispersion (standard deviation) over the 1,000 runs for each sample size. Thus, the 95 percent confidence interval upper and lower limits of

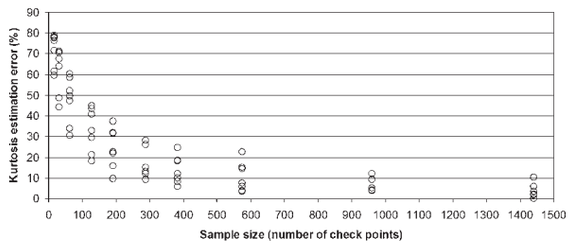

Figure 8. Evolution of the standardized kurtosis estimation error found versus the sample size used for its calculation. All data come from Experiment 1: (a) raw data, and (b) "3-sigma" rule corrected data. Note: Error = 100 $|\gamma_2 - \gamma_2^*| / \gamma_2$, being γ_2^* the estimated standardized kurtosis and γ_2 the population standardized kurtosis or true kurtosis.

γ_2^* were calculated, to obtain a measure of the uncertainty of the kurtosis estimation.

In Figure 8 we can observe the expected error of the standardized kurtosis estimation (γ_2^*) with regard to the population kurtosis (γ_2) and for all the morphologies tested, as a function of the sample size employed for its calculation. Raw data from experiment one are shown in Figure 8a, where it must be highlighted that a relatively large sample size of around 400 checkpoints is needed to reach a mean estimation error value lower than 20 percent. However, when the original data set is corrected by means of the 3-sigma rule, thus removing many outliers, the sample size needed to try low mean estimation errors diminishes notably (Figure 8b), reaching values lower than 10 percent even with less than 100 checkpoints.

Bearing in mind the findings shown in Figure 8, it seems clear that the problem is not the estimation of the expected standardized kurtosis, but the uncertainty of the estimated value. In fact, in Figure 9 the evolution of the uncertainty (measured as standard deviation) of γ_2^* is plotted versus the sample size used for its calculation. For the sake of clarity, only the flat, mountainous, and highly rugged morphologies are shown. At any rate, it can be verified how the 3-sigma corrected data set presented very low values of uncertainty, specifically lower than 1.5, when sample sizes with more than 128 checkpoints were taken (Figure 9b). Note that uncertainty is much larger when raw data are employed (Figure 9a), even if large sample sizes are utilized. This feature was more pronounced in the case of highly rugged terrain, where the outliers presented high extreme values. Thus, we strongly recommend the application of the 3-sigma rule for the correction of sample data and quality control, removing, as well as possible, the outliers or gross errors. This is necessary, especially in the case of a residuals distribution with many outliers or high kurtosis.

Naturally, there are other methods for the detection and management of outliers, which have not been checked in this work. For instance, Atkinson *et al.* (2005) proposed the use of several estimators, called robust estimators, less sensitive to extreme observations or outliers than classic estimators. The approach is based on the weighting of sample data, i.e., the data that are found further from the most probable values will have a lower weight in the calculation of the average value and standard deviation. Atkinson *et al.* (2005) proposed the Danish method as the most suitable if it is applied cutting down the weighting on those data that exceed 2.5-sigma. It should probably be considered as a good alternative when the sample size is too small so that any loss of information will be relevant.

Tables 8 through 11 show a more detailed description of the effect of the 3-sigma correction on the standardized kurtosis estimation and subsequent reliability calculation. Thus, Tables 8 and 9 present the mean value and standard deviation of the estimated standardized kurtosis, from raw and corrected data pertaining to Experiments 1 and 2, respectively, for a sample size of 128 checkpoints. Once more, we can verify that data from the 3-sigma correction procedure, both in experiment one and two, perform better than in the case when raw data are used because they point

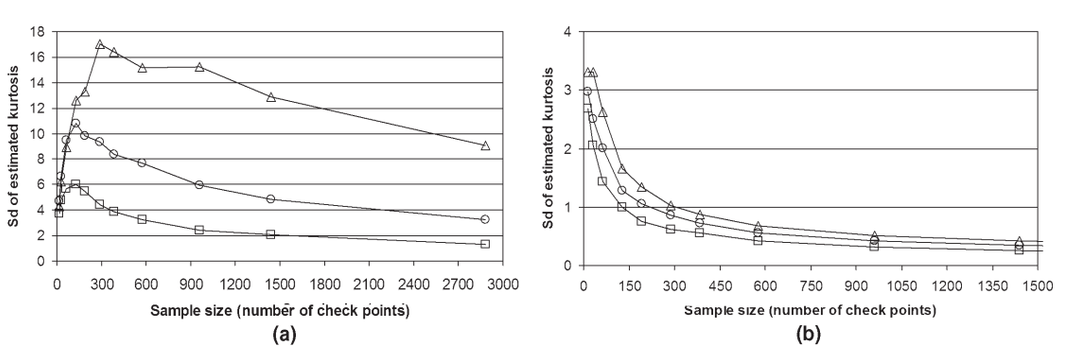

Figure 9. Evolution of the uncertainty (standard deviation: Sd) of the estimated estandardized kurtosis versus the sample size used for its calculation. Triangles, circles and squares denote highly rugged, flat, and mountainous terrain, respectively. All data comes from Experiment 1: (a) raw data, and (b) "3-sigma" rule corrected data.

TABLE 8. EXPECTED VALUE (MEAN VALUE) AND DISPERSION (STANDARD DEVIATION) OF THE ESTIMATED STANDARDIZED KURTOSIS FOR RAW AND 3-SIGMA RULE CORRECTED DATA COMPUTED FROM 1,000 SAMPLES OF 128 CHECKPOINTS; DATA FROM EXPERIMENT 1 (INTERPOLATOR: MULTIQUADRIC RADIAL BASIS FUNCTION)

	Raw residuals population			"3-sigma" rule corrected residuals		
Morphology	Mean value of estimated kurtosis γ_2*	Standard deviation of γ_2*	Population kurtosis γ_2	Mean value of estimated kurtosis γ_2*	Standard deviation of γ_2*	Population kurtosis γ_2
Flat	16.96	10.81	23.99	4.13	1.28	4.15
Mountainous	9.95	6	12.17	3.19	0.99	3.07
Rolling 1	10.42	5.62	13.20	3.73	1.18	3.79
Steep rugged hillside	12.2	11.13	21.55	3.22	0.95	3.16
Highly rugged	17.66	12.6	31.95	5.79	1.66	6.05
Slightly mountainous	12.52	9.57	21.12	4.42	1.26	4.39
Rolling 2	19.97	11.13	29.66	4.88	1.54	5.10

TABLE 9. EXPECTED VALUE (MEAN VALUE) AND DISPERSION (STANDARD DEVIATION) OF THE ESTIMATED STANDARDIZED KURTOSIS FOR RAW AND 3-SIGMA RULE CORRECTED DATA COMPUTED FROM 1,000 SAMPLES OF 128 CHECKPOINTS; DATA FROM EXPERIMENT 2 (INTERPOLATOR: INVERSE DISTANCE WEIGHTED)

	Raw residuals population			"3-sigma" rule corrected residuals		
Morphology	Mean value of estimated kurtosis γ_2*	Standard deviation of γ_2*	Population kurtosis γ_2	Mean value of estimated kurtosis γ_2*	Standard deviation of γ_2*	Population kurtosis γ_2
Flat	2.99	1.62	3	1.13	0.51	1.04
Mountainous	1.74	0.8	1.67	0.94	0.47	0.80
Rolling 1	2.58	1.1	2.50	0.95	0.53	0.82
Steep rugged hillside	1.88	1.67	1.81	0.67	0.42	0.53
Highly rugged	4.58	2.2	4.73	2.56	0.69	2.45
Slightly mountainous	2.77	1.15	2.89	1.85	0.62	1.57
Rolling 2	4.50	2.92	4.93	1.52	0.57	1.31

TABLE 10. MEAN VALUE AND UPPER AND LOWER LIMITS (95 PERCENT CONFIDENCE INTERVAL) OF THE ESTIMATED RELIABILITY (R*) USING THE PROPOSED MODEL 1 FOR A SAMPLE SIZE OF 128 CHECKPOINTS AND THE VALUES OF THE ESTIMATED STANDARDIZED KURTOSIS (MEAN AND STANDARD DEVIATION) SHOWN IN TABLE 8. THE TRUE REALIABILITY (R) HAS BEEN CALCULATED STARTING FROM THE POPULATION STANDARDIZED KURTOSIS. THE TERM "ERROR" MEANS THAT THE SQUARE ROOT SIGN IN THE MATHEMATICAL EXPRESSION OF MODEL 1 IS NEGATIVE AND SO A REAL SOLUTION DOES NOT EXIST; DATAFROM EXPERIMENT 1 (INTERPOLATOR: MULTIQUADRIC RADIAL BASIS FUNCTION)

	Raw residuals population				"3-sigma" rule corrected residuals			
	Estimated Reliability R*(%)			True	Estimated Reliability R*(%)			True
Morphology	Mean R*	Upper R*	Lower R*	Reliability R (%)	Mean R*	Upper R*	Lower R*	Reliability R (%)
Flat	19.10	27.78	Error	22.36	10.87	12.89	8.36	10.89
Mountainous	15.16	21.35	1.98	16.52	10	11.72	7.92	9.89
Rolling 1	15.46	21.23	5.22	17.10	10.51	12.44	8.12	10.56
Steep rugged hillside	16.53	26.32	Error	21.28	10.03	11.68	8.05	9.97
Highly rugged	19.45	29.20	Error	25.55	12.25	14.58	9.35	12.45
Slightly mountainous	16.71	25.30	Error	21.09	11.12	13.08	8.73	11.10
Rolling 2	20.56	29.02	1.81	24.68	11.51	13.80	8.63	11.70

out more accurate population kurtosis estimation and, at the same time, produce lower uncertainty.

Note that any uncertainty on the standardized kurtosis estimation will result in uncertainty in the reliability estimation, although somewhat mitigated, as can be verified observing Model 1 (Equation 9) or Model 2 (Equation 16).

In fact, Tables 10 and 11 show the mean value (expected value) and upper and lower limits of a 95 percent confidence interval for the reliability estimated from the proposed model one using a sample size of 128 checkpoints. The variability of estimated reliability is very pronounced, working with samples coming from clear leptokurtic data (Table 10, raw residuals from experiment one). Thus, we can obtain estimated kurtosis values that are very far removed from the actual reliability in some cases. However, confidence intervals tend to be narrower and closer to true reliability when model one is applied to 3-sigma corrected data (Tables 10 and 11) or the original leptokurtic scenario is less pronounced (Table 11, raw data from experiment two).

Finally, it can be interesting to compare the results offered by model one with the reliability values proposed by other authors. Ley (1986), based on experience, recommended the use of a sample size of 150 checkpoints in order to guarantee a reliability of around 10 percent. Li (1991)

TABLE 11. MEAN VALUE AND UPPER AND LOWER LIMITS (95 PERCENT CONFIDENCE INTERVAL) OF THE ESTIMATED RELIABILITY (R*) USING THE PROPOSED MODEL 1 FOR A SAMPLE SIZE OF 128 CHECKPOINTS AND THE VALUES OF THE ESTIMATED STANDARDIZED KURTOSIS (MEAN AND STANDARD DEVIATION) SHOWN IN TABLE 9. THE TRUE REALIABILITY (R) HAS BEEN CALCULATED STARTING FROM THE POPULATION STANDARDIZED KURTOSIS; DATA FROM EXPERIMENT 2 (INTERPOLATOR: INVERSE DISTANCE WEIGHTED)

	Raw residuals population				"3-sigma" rule corrected residuals			
	Estimated Reliability R*(%)			True	Estimated Reliability R*(%)			True
Morphology	Mean R*	Upper R*	Lower R*	Reliability R (%)	Mean R*	Upper R*	Lower R*	Reliability R (%)
Flat	9.81	12.54	5.93	9.82	7.77	8.92	6.42	7.67
Mountainous	8.49	10.11	6.48	8.42	7.53	8.63	6.25	7.36
Rolling 1	9.40	11.39	6.84	9.32	7.55	8.77	6.08	7.38
Steep rugged hillside	8.65	11.74	3.45	8.58	7.18	8.21	5.98	7
Highly rugged	11.26	14.48	6.62	11.39	9.37	10.67	7.87	9.26
Slightly mountainous	9.59	11.63	6.97	9.72	8.62	9.88	7.13	8.31
Rolling 2	11.19	15.34	3.90	11.55	8.24	9.45	6.81	8

developed a more thorough theoretical analysis obtaining previously discussed Equation 12. Applying the said equation to a sample size of 128 checkpoints, a reliability value of 6.3 percent is obtained. These values are similar to the estimation performed by the proposed model one when the residuals populations presented low kurtosis, and were, therefore, closer to a normal distribution (Tables 10 and 11, corrected data).

Conclusions

The results obtained in this work allow us to conclude that the two theoretical models developed for estimating the reliability of the DEM accuracy figures provided a good fit for the simulated data offered by the Monte Carlo method. In fact, the two models performed well for the seven morphologies tested when raw data observed were fitted, both in Experiment 1 (r^2 = 95.85% as mean value) and in Experiment 2 (r^2 = 97.64% as mean value).

Bearing in mind the variability of the morphologies tested, from flat terrain to highly rugged terrain of marble quarries, and the two different methods employed to generate the residuals data populations, the results can be considered as highly applicable in the practice of DEM production. Thus, we must point out that the residuals data sets were not subordinated to the "strong assumption" of distribution normality commonly accepted in the majority of standards for spatial data accuracy. Nowadays, it is known that the non-gaussian distribution of errors at the checkpoints is very common in geographically distributed data.

The removal of outliers from simulated data sets applying the 3-sigma rule increased the regression coefficients r^2 to a value of over 0.98 for all the cases, i.e., not significantly better than when raw data were modeled. The key was, without doubt, the introduction of standardized kurtosis in the model, which rightly described the leptokurtosis scenario produced by the presence of anomalous values in the residuals data set.

Thus, the standardized kurtosis from a finite sample of checkpoints, estimated using the expression shown in Equation 17, could be considered as a good indicator of the sample quality or goodness. Therefore, it should be included in the calculation of the reliability of the accuracy test, as it is done precisely by the two models developed.

Finally, because of the great uncertainty observed in standardized kurtosis estimation working with high leptokurtic residuals distributions and low sample sizes, the application of the 3-sigma rule is strongly recommended for the removal of outliers only within a definite land-cover category. For example, let us suppose that land-cover such as bare earth terrain and fully forest terrain are grouped in the quality control of a lidar DEM. Because lidar pulses often do not penetrate the vegetation but map the top surfaces, it would be very important to recognize that checkpoints from forest terrain will yield larger elevation errors than those situated on bare terrain. So, it would be a big error to remove as outliers all checkpoints located at forest terrain because we are throwing out the information necessary to identify that we have a problem in such vegetation category, i.e., the DEM is not representing the bare terrain as it should. So we can conclude that an adequate and careful removal of outliers within a definite land-cover category will lead to narrower confidence intervals for the estimation of reliability.

Acknowledgments

The authors would like to thank the anonymous reviewers whose comments and suggestions much improved the submitted draft of the article.

References

Aguilar, F.J., F. Agüera, M.A. Aguilar, and F. Carvajal, 2005. Effects of terrain morphology, sampling density and interpolation methods on grid DEM accuracy, *Photogrammetric Engineering & Remote Sensing*, 71(7):805–816.

Aguilar, F.J., M.A. Aguilar, F. Agüera, and J. Sánchez, 2006. The accuracy of grid digital elevation models linearly constructed from scattered sample data, *International Journal of Geographical Information Science*, 20(2):169–192.

Ariza, F.J., and A.D.J. Atkinson, 2005. Sample size and confidence when applying the NSSDA, *Proceedings of the 21st* International Cartographic Conference, 09–16 July, A Coruña, Spain, The International Cartographic Association, unpaginated CD-ROM.

Atkinson, A.D.J., F.J. Ariza, and J.L. García-Balboa, 2005. Positional accuracy control using robust estimators, *Proceedings of the 21st* International Cartographic Conference, 09–16 July, A Coruña, Spain, The International Cartographic Association, unpaginated CD-ROM.

Baltsavias, E.P., and M. Hahn, 1999. Integration of image analysis and GIS, *International Archives of Photogrammetry and Remote Sensing*, 32(part 7–4–3W6):12–19.

Burrough, P.A., and R.A. McDonnell, 1998. *Principles of Geographical Information Systems*, Oxford University Press, New York, 333 p.

Daniel, C., and K. Tennant, 2001. DEM quality assessment, *Digital Elevation Model Technologies and Applications: The DEM Users Manual* (D.F. Maune, editor), American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, pp. 395–440.

Felicísimo, A., 1994. Parametric statistical method for error detection in digital elevation models, *ISPRS Journal of Photogrammetry and Remote Sensing*, 49(4):29–33.

FGCC, 1984. Standards and Specifications for Geodetic Control Networks, U.S. Federal Geodetic Control Committee, Rockville, Maryland URL: http://www.ngs.noaa.gov/FGCS/tech_pub/1984-stds-specs-geodetic-control-networks.pdf (last date accessed: 23 August 2007).

FGDC, 1998. Geospatial Positioning Accuracy Standards, Part 3: National Standard for Spatial Data Accuracy, U.S. Federal Geographic Data Committee, Reston, Virginia, URL: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3, (last date accessed: 23 August 2007).

Fisher, P., 1998. Improved modeling of elevation error with geostatistics, *GeoInformatica*, 2(3):215–233.

Golden Software, Inc., 2002. *Surfer 8 Users' Guide*, Golden Software, Inc., Golden, Colorado, 640 p.

Heuvelink, G.B.M., P.A. Burrough, and A. Stein, 1989. Propagation of errors in spatial modeling with GIS, *International Journal of Geographical Information Systems*, 3(4):303–322.

Ley, R.G., 1986. Accuracy assesstment of digital terrain models, *Proceedings of AutoCarto*, September, London, England, International Cartographic Association, Volume 1, pp. 455–564.

Li, Z., 1988. On the measure of digital terrain model accuracy, *The Photogrammetric Record*, 12(72):873–877.

Li, Z., 1991. Effects of check points on the reliability of DTM accuracy estimates obtained from experimental tests, *Photogrammetric Engineering & Remote Sensing*, 57(10):1333–1340.

López, C., 1997a. On the improving of elevation accuracy of Digital Elevation Models: A comparison of some error detection procedures, *Proceedings of the 6th* Scandinavian Research Conference on Geographical Information Systems, ScanGIS'97, 01–03 June 1, Stockholm, Sweden, Centre of Geoinformatics, The Royal Institute of Technology, Stockholm, Sweden, pp. 85–106.

López, C., 1997b. Locating some types of random errors in digital terrain models, *International Journal of Geographical Information Science*, 11(7):677–698.

Maune, D.F., J. Binder Maitra, and E.J. McKay, 2001a. Accuracy standards, *Digital Elevation Models and Applications: The DEM Users Manual* (D.F. Maune, editor), American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, pp. 61–82.

Maune, D.F., T.A. Blak, and E.W. Constance, 2001b. DEM user requirements, *Digital Elevation Models and Applications: The DEM Users Manual Manual* (D.F. Maune, editor), American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, pp. 441–460.

Royston, J.P., 1982, Expected normal order statistics (exact and approximate), *Applied Statistics*, 31:161–165.

Shi, W.Z., M.F. Goodchild, and P.F. Fisher, 2002. A prospective on spatial data quality, *Spatial Data Quality* (W.Z. Shi, P.F. Fisher, and M.F. Goodchild, editors), Taylor and Francis, London, pp. 304–309.

Shi, W.Z., and Y. Bedard, 2004. Advanced techniques for analysis of geo-spatial data, *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(1–2):1–5.

Torlegard, K., A. Ostman, and R. Lindgren, 1986. A comparative test of photogrammetrically sampled digital elevation model, *Photogrammetria*, 41(1):1–16.

Weng, Q., 2002. Quantifying uncertainty of digital elevation models derived from topographic maps, *Advances in Spatial Data Handling* (D. Richardson and P. van Oosterom, editors), Springer-Verlag, New York, pp. 403–418.

Wood, J.D., 1996. *The Geomorphological Characterisation of Digital Elevation Models*, Ph.D. Thesis, University of Leicester, UK, 185 p.

(Received 13 October 2005: accepted 23 March 2006; revised 05 May 2006)