Effects of Check Points on the Reliability of DTM Accuracy Estimates Obtained from Experimental Tests

*Zhilin Li**
Department of Geography and Topographic Science, University of Glasgow, United Kingdom

ABSTRACT: This paper is an attempt to discuss the effects of the characteristics (accuracy, number, and distribution) of the set of check points used for experimental tests of DTM accuracy on the resulting accuracy figures. In this paper, first of all, the concept of reliability in the context of DTM accuracy tests is introduced and alternative measures for this are sketched. Then, the effects of the characteristics of a set of check points on the DTM accuracy estimates are investigated both through a theoretical analysis and by experimental tests.

INTRODUCTION

IN THE EXPERIMENTAL TESTS ON THE ACCURACY of digital terrain models (DTM), a set of check points is used as the "ground truth." Then the points interpolated from the constructed DTM surface are checked against the corresponding check points. After that, the difference of the two heights (DH) at each DTM point is obtained. These differences are used to compute statistical values such as the mean and standard deviation which are used as a measure of DTM accuracy. In these circumstances, DH is considered as a random variable.

In the case of the experimental tests on DTM accuracy, it is clear that the final DTM accuracy figures estimated from the test results – in this case, the mean and standard deviation values – are definitely affected by the characteristics of the set of check points. In other words, it can be said that the characteristics of the set of check points which were used as the ground truth in the experimental tests have effects on the reliability of the final DTM accuracy figures obtained from these tests.

It is obvious that the reliability of the accuracy figures, estimated from an experimental test, is also a problem which is of considerable importance in DTM accuracy tests, becuase only if the accuracy figures are reliable to a certain level, can one use the accuracy estimates to evaluate the "goodness" of the digital terrain model which has just been tested. Therefore, this study is an attempt to obtain an insight into the effects of the set of check points used in the experimental tests on the reliability of the DTM accuracy figures estimated from the test results.

In this paper, the concept of reliability in the context of DTM accuracy tests will be introduced and alternative measures for this will be sketched. Then, the effects of the characteristics of a set of check points on the DTM accuracy estimates will be investigated both through a theoretical analysis and by experimental tests.

RELIABILITY IN THE CONTEXT OF DTM ACCURACY TESTS

Reliability is a concept which is widely used in engineering (including photogrammetric engineering) and industry. It seems pertinent to have a look at how this concept is defined and used in these areas before it can be adopted into the methodology and context of experimental tests on DTM accuracy.

*Presently with the North East Regional Research Laboratory, Centre for Urban and Regional Development Studies, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, United Kingdom.

THE CONCEPT OF RELIABILITY IN ENGINEERING AND INDUSTRY

Due to the many differing operational requirements and varying environments existing in engineering and industry, the concept of reliability may mean quite different things to different people. Nevertheless, a generally acceptable definition given by the B.S.I. (British Standards Institution) is as follows (Dummer and Winton, 1986):

> "Reliability is the characteristic of an item expressed by the probability that it will perform a required function under stated conditions for a stated period of time."

For example, suppose that the life of the bulbs made by a lamp manufacturer is declared to be 1,000 hrs (which is the stated period of time required by the above definition), then the reliability is 98 percent if one tested 100 bulbs of this make and found that 2 of them had shorter lives than declared. This might belong to one of the simplest examples. In practice, the reliability of an engineering system or structure is much more complicated. However, the detailed discussion of this matter lies outside the interest of this study. What is intended here is to adopt the concept of reliability into the context of DTM accuracy estimates.

RELIABILITY IN THE CONTEXT OF DTM ACCURACY TESTS

Obviously, in the context of experimental tests on the accuracy of a digital terrain model, there is nothing which is concerned with "a required function under stated conditions for a stated period of time." Instead, what is of concern in this context is "with what probability are the estimated accuracy figures (i.e., the mean and standard deviation values) likely to be correct" or "to what degree of correctness will the accuracy results have been estimated." In any case, it is an obvious fact that the DTM accuracy results obtained from experimental tests are not absolutely certain and one can accept these results only to a certain confidence level. Therefore, the concept of reliability can be adopted into this context because reliability is concerned only with uncertainty.

In some sense, the concept of reliability in this context might be defined as the degree of correctness to which the DTM accuracy figures have been estimated.

Of course, the reliability of DTM accuracy estimates in the context of experimental tests may be affected by several factors such as the capabilities of the person who has undertaken the work; the program by which the accuracy figures have been calculated and recorded; and the characteristics of the set of check points which have been used as ground truth against which the DTM points have been checked. However, in this

PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING,
Vol. 57, No. 10, October 1991, pp. 1333–1340.

0099-1112/91/5710–1333$03.00/0

study, it is assumed that other factors are absolutely reliable; therefore, only the effect of certain characteristics of the set of check points on the reliability of DTM accuracy estimates will be considered.

A set of check points can be characterized by three main parameters: (1) the sample size (i.e., the number of points in the data set); (2) their accuracy; and (3) the distribution of the data points. The main discussion in this paper will be about how each of these three main parameters of a set of check points affect the reliability of DTM accuracy estimates in the context of experimental tests.

Alternative Measures of Reliability

As one can imagine, a measure is required for the reliability of DTM accuracy estimates. There may be two types of measure available. One is qualitative (or descriptive) and the other type is quantitative (or numerical). For the former, words such as absolutely reliable, most reliable, very reliable, quite reliable, fairly reliable, not so reliable, not reliable, unreliable, most unreliable, absolutely unreliable, etc., can be used. However, in the scientific community, such a statement is not acceptable because the definition of such a term is usually too loose.

For the quantitative (or numerical) measures, there are three alternatives as follows:

- One possible measure is to use the absolute values of the accuracy of each of the obtained accuracy estimates, e.g., the value of standard deviation of the obtained standard deviation estimate. Suppose that the estimated standard deviation values for the DH accuracy is SD(DH), then such an *absolute* value may denoted as SD(SD(DH)).
- Another possible measure is to use a relative value, similar to the term "per mil of flying height" which is commonly used to state the accuracy of photogrammetrically measured data. Thus, in this case, a percentage value may be quite adequate and thoroughly acceptable. For example, the percentage value of the ratio SD(SD(DH))/SD(DH) might well be adequate.
- The third possible way is to use the concept of "membership" as used in the context of fuzzy sets. Percentage values between 0 percent and 100 percent can be used and these values represent the degree of reliability to which an accuracy estimate belongs. In this case, it is not necessary that the percentage value be obtained from the ratio SD(SD(DH))/SD(DH). Instead, the SD(SD(DH)) value is converted into a figure expressing the degree of reliability by a pre-defined function.

There is no fundamental difference between the second and the third approaches. The second value will become the same as the third if the former is stretched into the range of 0 percent to 100 percent.

After these introductory discussions and definitions, it is time to look into the matter of the effect of check points on the DTM accuracy estimates.

EFFECT OF SAMPLE SIZE (NUMBER) ON THE RELIABILITY OF THE DTM ACCURACY ESTIMATES

It seems obvious that the inclusion of more check points in the data set will lead to a more reliable result. So researchers try to use large sample sizes in order to ensure that the obtained accuracy values will be reliable. For example, in the ISPRS DTM test which was conducted by Commission III's Working Group No. 3 (Torlegård *et al.*, 1986), more than 1,800 check points were used in each test area. However, a large number of check points may sometimes be costly to produce and, in some cases, even impossible to provide in the context of DTM accuracy testing. Therefore, an important question which arises is whether such a large number of check points is necessary. If not, then the obvious follow-up question is "what is the minimum number of check points required for a given degree of reliability for the accuracy estimates?" That is to say, the important matter in this case is to determine the required minimum sample size (number) for the given degree of reliability required for the accuracy estimates (i.e., the estimated mean and standard deviation values).

Ley (1986) tried to provide a solution to this problem based on his own experience and pointed out that "a sample size of 150 points will guarantee that the subsequent accuracy statement possesses a standard deviation of 10 percent (of the obtained value of standard deviation estimate). This number (150 points) is over 10 times smaller than that used in the ISPRS test. However, he didn't provide any information about how this figure was obtained nor the context in which it occurred. Therefore, a theoretical deduction may be both revealing and important.

In an attempt to answer the questions raised above, this section starts with a theoretical analysis; then the theoretical results will be validated with experimental data. The theoretical analysis in this study is based on the assumption that the check points are free of error.

Effect of Sample Size on the Accuracy of the Estimated Mean Value

From statistical theory, it can be found that the sample size, required with a given degree of accuracy requirement for the accuracy figures to be estimated, depends on the variation associated with the random variable, i.e., DH in the case of the DTM accuracy tests. The smaller the variation, the smaller the sample size that is needed to achieve a given degree of accuracy required for the accuracy estimates. For an extreme example, suppose the standard deviation (SD) of the height difference (DHs) was equal to zero, then one check point would be enough no matter how large the test area or the size of the data set. The required minimum sample size also depends on the given degree of the accuracy requirement itself. A general discussion about the relationship between the sample size, the value of SD, and the given degree of the accuracy requirement is given in the following paragraphs.

Let M be the mean of a random sample of size n from a particular distribution, and u be the true value of the random variable. Then the ratio as follows:

$$Y = \frac{M - u}{\mathrm{SD}/\sqrt{n}} \tag{1}$$

is the standardized variable and has approximately the normal distribution $N(0,1)$, even though the underlying distribution is not normal, as long as n (the sample size) is large enough (Hogg and Tanis, 1977).

Suppose the SD of a distribution is known but the value of u (the true value of the random variable) is unknown. Then, for the probability r and for a sufficiently large value of n, a value Z can be found from the statistical table for $N(0,1)$ distribution such that the probability that Y will be within the range from $-Z$ to Z is approximately equal to r; or mathematically

$$P(-Z \le Y \le Z) \approx r. \tag{2}$$

The closeness of the approximate probability r to the exact probability depends upon both the underlying distribution and the sample size. When the underlying distribution is unimodal (with only one mode) and continuous, the approximation is usually quite good for even a small value of n (e.g., $n=5$). If the underlying distribution is "less normal" (i.e., badly skewed or discrete), a large sample size is required to keep a reasonably accurate approximation. However, 20 or 30 is the number which is quite adequate for n in all cases (Hogg and Tanis, 1977).

Substituting Equation (1) into Equation (2) and rearranging it, the following expression can be obtained:

$$P(M - Z \cdot \mathrm{SD}/\sqrt{n} \le u \le M + Z \cdot \mathrm{SD}/\sqrt{n}) \approx r \tag{3}$$

For a given constant S, the percentage of the probability, $(100r)$ percent, of the random interval $M \pm S$ including u is called the confidence interval, where S is the specified degree of accuracy for the mean estimate, M in this case. In general, if the required confidence interval $(100r)$ percent $= 100(1-\alpha)$ percent, then the sample size n can be expressed as the following according to Equation 3:

$$n = \frac{Z_r^2 \cdot \mathrm{SD}^2}{S^2} \tag{4}$$

where SD is the standard deviation of the random variable, S is the given degree of accuracy for the mean estimate, and Z_r is the limit value within which the values of the random variable Y will fall with a probability of r. Its value can be found in the statistical table for the $N(0,1)$ distribution. The mathematical expression is the following:

$$\Phi(Z) = 1 - \alpha/2 \tag{5}$$

and the commonly used values are as follows:

$$Z_{(r=0.95)} = 1.960;\ Z_{(r=0.98)} = 2.326;\ Z_{(r=0.99)} = 2.576$$

and n is the required minimum sample size for the check points with a given confidence level which is expressed by Z_r.

In the case of a DTM accuracy test, the SD in Equation 4 is the expected standard deviation of the DTM, and an approximate estimate is required before starting to measure the check points. The discussion of how to estimate such a rough value lies outside this study but has been given elsewhere (e.g., Li, 1990). The value of r is commonly selected as 95 percent, 98 percent, or even 99 percent.

It seems that the ratio S/SD is a value which can be used as the reliability of the estimated mean value. If it is denoted as $R(M)$, then Equation 4 can be rewritten as follows:

$$n = \frac{Z_r^2}{R^2(M)}. \tag{6}$$

The diagrammetric presentation of Equation 6 is given in Figure 1(a). Equation 6 can also be rewritten as follows:

$$R(M) = \frac{Z_r}{\sqrt{n}} \tag{7}$$

Effect of Sample Size on the Reliability of Estimated SD Value

Next, the influence of sample size on the reliability of the SD estimate should be considered. It can be shown that the variance of the standard deviation estimated from a sample can be approximately expressed as follows (Burington and May, 1970, p.194):

$$\mathrm{VAR}(\mathrm{SD}(\mathrm{DH})) = \mathrm{VAR}\ (\mathrm{DH}) / 2(n-1) \tag{8}$$

In the context of a DTM set, this would mean that the estimated standard deviation of the DTM errors possesses a standard deviation of $\sqrt{1/(2n-2)}$ times itself if the check points are free of error, or with a variance smaller than the critical value which will be discussed later.

This can be expressed as a percentage of the estimated variance. It can be rewritten as follows:

$$R(\mathrm{SD}) = \frac{1}{\sqrt{2(n-1)}} \times 100\% \tag{9}$$

where R(SD) is used as the reliability of the SD estimate. For example, a sample size of 150, which was given by Ley (1986) as an example, will provide a standard deviation of 6 percent times itself for the standard deviation estimate. This value is very close to that presented by Ley (1986). To give another example, a sample size of 1,800 will produce a standard deviation of 2 percent times itself for the standard deviation estimate.

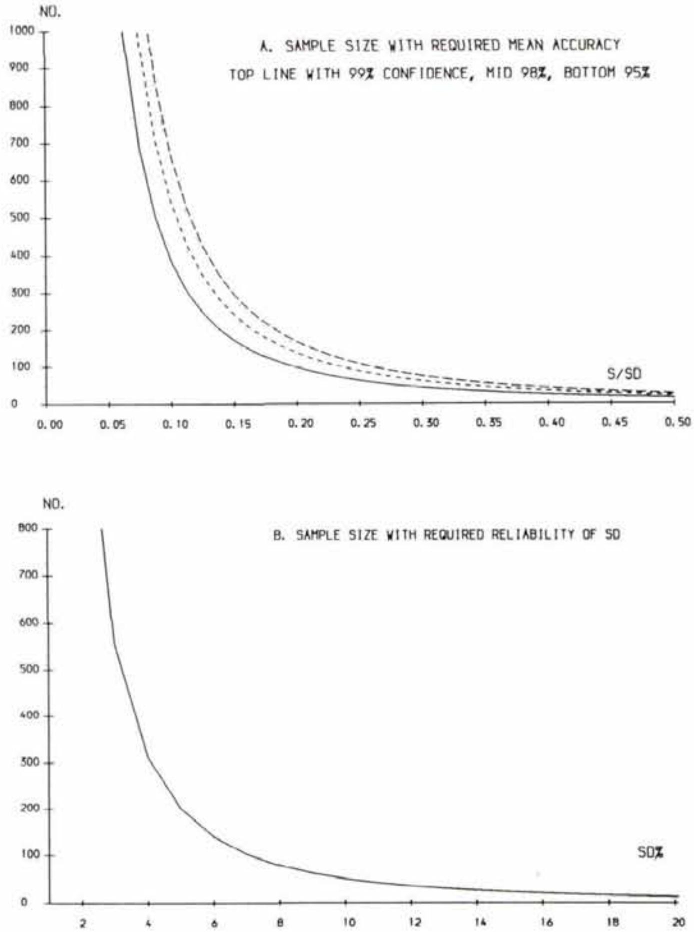

FIG. 1. Sample size (number of check points) with required accuracy.

Accordingly, if the reliability requirement for the standard deviation estimate of the DTM is given beforehand, then the required minimum sample size can also be computed from the following:

$$n = \frac{1}{2 \cdot R^2\ (\mathrm{SD})} + 1 \tag{10}$$

In practice, a relatively large sample size is usually used for experimental tests on DTM accuracy; therefore, for convenience, Equation 10 can be approximated as folows:

$$n = \frac{1}{2 \cdot R^2(\mathrm{SD})} \tag{11}$$

where R(SD) is a percentage value. For example, if $R(\mathrm{SD}) = 10$ percent is the reliability required, then from Equation 11, it can be computed that the required number for this example is 50. The graphical presentation of Equation 11 is shown in Figure 1b.

Experimental Validation

The discussion given in the previous two sections is purely theoretical. One very important question arising from this discussion is whether these criteria can really be applied in practice. To answer this question, some experimental tests are necessary.

This has been done using the data sets which had been generated and used in the ISPRS test (Torlegård *et al.*, 1986). Detailed information about these data sets is omitted here, but it has given in the author's thesis (Li, 1990) and in a previous paper (Li, in press). Two areas, i.e., Uppland and Sohnstetten, were selected for this experiment because, from the tests which have been described in the previous paper (Li, in press), it was found that the occurrence frequencies of large residual errors was very low. Thus, the data sets for these two areas were assumed to be very reliable.

The check points were originally arranged in a grid form. For Uppland and Sohnstetten areas, the grid sizes are 69 by 36 = 2,484 points and 20 by 104 = 2,080 points, respectively. However, not every point was measured because a certain number fell in a woodland area or on some other unsuitable features. In fact, only 2,314 grid nodes were measured for Uppland and 1,892 for the Sohnstetten area. From these points, several subsets were selected. These data sets were selected simply by choosing every nth point from the data file. The test results are shown in Tables 1 and 2, where the symbol "±" before SD and RMSE values is simply omitted.

Table 1 shows the variation in the parameters defining the accuracy of the DTM from composite data sets for Uppland with the number of check points used. The SD is estimated as 0.59m as determined from the entire sample. According to Equation 4, if the estimated mean should lie within a range of ±0.05m from the true value with 95 percent confidence, then 535 check points are required for the purpose. However, with the same confidence level and value for SD, 134 and 273 points will give estimated means within a range of ±0.10m and ±0.07m from the true value, respectively. From the same table, it can also be seen that the results obtained using more than 578 check points are very consistent not only for the mean values (varying within a range of ±0.016m) but also for the SD and RMSE. Below this number, the mean, the SD, and the RMSE all show bigger variations. When the number of check points lies within the range between 257 and 578, the mean varies within the range of ±0.065m. When fewer than 115 check points were used, the figures of these accuracy parameters become very unstable. The results in this table show more or less similar trends to those expressed by Equations 4 and 10.

Table 2 shows the variation in DTM accuracy with the number of check points for the Sohnstetten area. The SD value for the Sohnstetten data set using all check points is ±0.401m. Also according to Equation 4, with 95 percent confidence, 683, 245, 125, and 62 check points will give the estimated means within the ranges of ±0.03m, ±0.05m, ±0.07m, and ±0.10m from the true value, respectively. From Table 2, it can be found that the mean varies from 0.153 to 0.173 in a range of 0.02m. At this stage, the SD and RMSE values are very stable.

When the number of check points falls within the range 379 to 237, the mean varies over a greater range of 0.035 (0.154 to 0.189). Also, the SD and RMSE values vary over a greater range. When the number of check point lies within the range of 211 to 119, the means vary with a range of 0.063m (0.144 to 0.207m). When the number of check points lies within the range between 106 to 64, the mean varies from 0.085m to 0.236m. It is 0.068m lower and 0.083m higher than the value of 0.0153m which is that obtained using all the check points (i.e., 100 percent) in this test. Accordingly, the RMSE and SD values also vary with a greater range when fewer check points are used. This test again shows that Equations 4 and 10 are appropriate.

More intuitively, these data values are presented graphically in Figure 2 and Figure 3. The continuous lines represent the variation ranges which are predicted from purely theoretical considerations. The lines in Figure 2 are produced according to Equations 3 and 4, where a 95 percent confidence level is selected; SD = 0.590m and 0.401m; and 0.220m and −0.155m are used as the "true" values of the means for the Uppland and Sohnstetten areas, respectively. Here it needs to be pointed out that the term S used in Equation 4 is the given degree of absolute accuracy but not the precision. The latter is well-known to topographic scientists as follows:

$$\mathrm{VAR}(M) = \mathrm{VAR}(\mathrm{DH}) / n \tag{12}$$

The lines in Figure 3 are produced according to Equation 10, using 0.575 and 0.395 as the SD values for Uppland and Sohnstetten areas, respectively (because it can be seen from the tables that these two values seem more alike than 0.590 and 0.401).

TABLE 1. VARIATION OF DTM ACCURACY WITH NUMBER OF CHECK POINTS FOR THE UPPLAND THE AREA

Parameters for check points			Parameters for DTM accuracy		
Fraction	Percent	Number	RMSE (m)	SD (m)	Mean (m)
1/1	100.0	2,314	0.636	0.590	0.238
5/6	83.3	1,928	0.618	0.575	0.227
3/4	75.0	1,735	0.614	0.573	0.222
7/12	58.3	1,349	0.618	0.574	0.229
1/2	50.0	1,157	0.615	0.574	0.222
1/3	33.3	771	0.622	0.576	0.235
1/4	25.0	578	0.612	0.571	0.220
1/5	20.0	462	0.592	0.566	0.175
1/6	16.7	385	0.619	0.579	0.218
1/7	14.3	330	0.597	0.547	0.240
1/8	12.5	289	0.587	0.545	0.218
1/9	11.1	257	0.623	0.568	0.256
1/10	10.0	231	0.606	0.586	0.155
1/20	5.0	115	0.566	0.545	0.166
1/30	3.3	77	0.574	0.571	0.058
1/40	2.5	58	0.570	0.548	0.157
1/50	2.0	47	0.759	0.743	0.154
1/60	1.7	39	0.580	0.580	0.016
1/70	1.4	34	0.437	0.346	0.268
1/80	1.25	29	0.554	0.546	0.095
1/90	1.1	26	0.630	0.613	0.148

TABLE 2. VARIATION OF DTM ACCURACY WITH NUMBER OF CHECK POINTS FOR THE SOHNSTETTEN AREA

Parameters for check points			Parameters for DTM accuracy		
Fraction	Percent	Number	RMSE (m)	SD (m)	Mean (m)
1/1	100.0	1,892	0.429	0.401	−0.153
5/6	83.3	1,580	0.422	0.391	−0.159
3/4	75.0	1,419	0.429	0.398	−0.160
7/12	58.3	1,104	0.430	0.393	−0.173
1/2	50.0	946	0.423	0.395	−0.152
1/3	33.3	631	0.421	0.385	−0.170
1/4	25.0	473	0.442	0.405	−0.178
1/5	20.0	379	0.443	0.405	−0.159
1/6	16.7	316	0.409	0.377	−0.159
1/7	14.3	271	0.437	0.401	−0.172
1/8	12.5	237	0.439	0.396	−0.189
1/9	11.1	211	0.415	0.389	−0.144
1/10	10.0	190	0.417	0.372	−0.189
1/12	8.3	158	0.406	0.363	−0.181
1/14	7.1	136	0.449	0.401	−0.201
1/16	6.25	119	0.460	0.410	−0.207
1/18	5.56	106	0.398	0.379	−0.122
1/20	5.0	95	0.477	0.414	−0.236
1/25	4.0	76	0.371	0.361	−0.085
1/30	3.3	64	0.349	0.291	−0.193
1/40	2.5	48	0.518	0.420	−0.302
1/50	2.0	38	0.346	0.315	−0.143
1/60	1.7	32	0.358	0.298	−0.197
1/70	1.4	28	0.492	0.410	−0.272

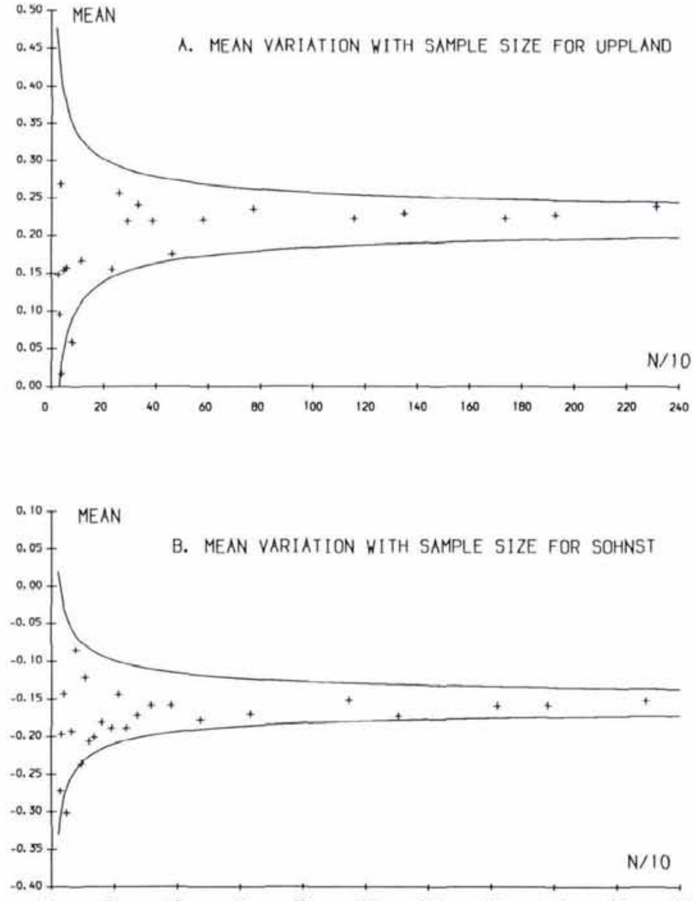

FIG. 2. Variation of mean with sample size (number). "+" represents the experimental results. Solid lines represents "mean of the tested means" plus (higher line) and minus (lower line) the achievable accuracy of the mean with the given sample size.

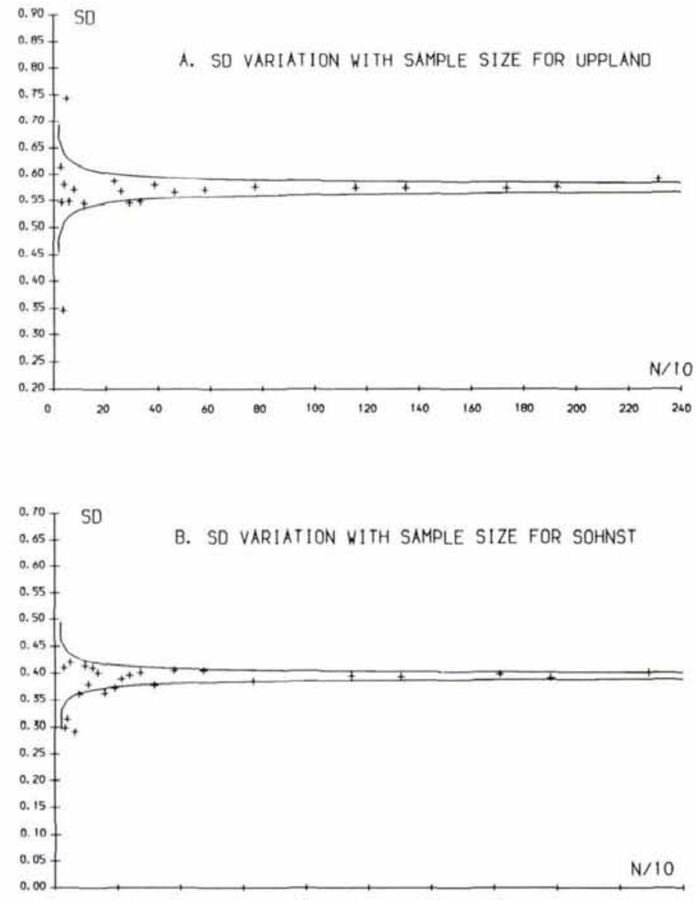

FIG. 3. Variation of SD with sampling size (number). "+" represents the experimental results. Solid lines represents mean of the tested SDs plus (higher line) and minus (lower line) the SD of the SDs.

Symbols "+" represents the points obtained from experimental tests which have been listed in Table 1 and Table 2. These diagrams appear to prove the validity of the theoretical discussions of the previous two sections, at least in the context of the ISPRS test data.

EFFECTS OF ERRORS IN THE CHECK POINTS ON THE RELIABILITY OF THE DTM ACCURACY ESTIMATES

Next, the matter of the errors in the check points which were used as ground truth in the experimental tests and their effects on the reliability of the DTM accuracy estimates requires discussion. An attempt is made to establish the relationship between the accuracy of the check points and the reliability of the resulting DTM accuracy estimates so that the maximum tolerable accuracy for the check points can be determined for a given degree of reliability for the final DTM accuracy figures to be estimated.

ACCURACY REQUIREMENT FOR THE CHECK POINTS

In the context of topographic science, in most cases, the accuracy of the check points is specified in terms of root-mean-square error (RMSE). In the present discussion, this RMSE value is assumed to be the same as the standard deviation (SD). Therefore, in this context, the important thing is to find the relationship between the SD value of the check points and the given degree of reliability required for the final standard deviation estimate which depends on the sample size of the check points.

Let DH_2 be the error involved in the check points and DH_1 be the true height difference. Then the overall error (DH) is as follows:

$$DH = DH_1 + DH_2. \tag{13}$$

By applying the error propagation law to Equation 13, the following expression can be obtained:

$$VAR(DH) = VAR(DH_1) + VAR(DH_2). \tag{14}$$

The variance of DH could be the sum of a few random variables. In this study, it is split into two, i.e., $VAR(DH_1)$ and $VAR(DH_2)$. The value of VAR(DH) itself is not of interest, but the value of $VAR(DH_1)$ is. An attempt may be made to estimate the latter through the former because only the former can be known. The attempt which is made here is to find a critical value for $VAR(DH_2)$ so that the value of VAR(DH) is still acceptable as being representative of $VAR(DH_1)$.

Also, as expressed as Equation 8 in the section on sample size, the standard deviation $SD(DH_1)$ estimated from a sample of size n has a variance approximately as follows:

$$VAR(\,SD(DH_1)\,) = VAR(DH_1)\,/\,(2n-2). \tag{15}$$

Therefore, the acceptable range for SD(DH) to deviate from $SD(DH_1)$ can be expressed as follows:

$$SD(DH_1) - SD(DH_1)/\sqrt{2n-2} < SD(DH)$$
$$< SD(DH_1) + SD(SD(DH_1)/\sqrt{2n-2}. \tag{16}$$

It is much more convenient to use a single vlaue, so the square root of these two terms is used as the representative value because they are independent. Then the following equation can be obtained.

$$\begin{aligned} \mathrm{VAR(DH)} &= \mathrm{VAR(DH_1)} + \mathrm{VAR(DH_1)} / (2n-2) \\ &= (2n-1)\cdot \mathrm{VAR(DH_1)} / (2n-2). \end{aligned} \quad (17)$$

Combining Equations 17 and 14 with a simplification, the following expression can be derived:

$$\mathrm{VAR(DH_2)} = \mathrm{VAR(DH)} / (2n-1). \quad (18)$$

It is more convenient to express this criterion in terms of the standard deviation. So Equation 18 can be converted to the following form:

$$\mathrm{SD(DH_2)} = \mathrm{SD(DH)} / \sqrt{(2n-1)}. \quad (19)$$

Let $K = \mathrm{SD(DH_2)/SD(DH)}$, then Equation 19 can be rewritten as follows:

$$K = 1 / \sqrt{(2n-1)} \quad (20)$$

where K is a function of the sample size, n. A graphic presentation is shown in Figure 4. For a given sample size which is determined by the reliability requirement discussed in the previous sections, the critical value for the required accuracy of the check points can be determined by Equation 19. In which case, $\mathrm{SD(DH_2)}$ may be given a special annotation, thus denoted as $\mathrm{SD}(t)$ in this context.

Obviously, the value of K decreases with an increase in n. This means that, with the increase in n, the variance of the estimated $\mathrm{SD(DH_1)}$ value becomes smaller. And the smaller the variance of $\mathrm{SD(DH_1)}$, the greater the influence of the check points with the same accuracy on the reliability of the estimated $\mathrm{SD(DH_1)}$ which is approximated by SD(DH).

As discussed before, if the accuracy of the check points is higher than $\mathrm{SD}(t)$, then SD(DH) can be used to approximate $\mathrm{SD(DH_1)}$ and reliability of SD(DH) can still be approximated by Equation 10.

On the other hand, if the standard deviation of the check points is larger than the value of $\mathrm{SD}(t)$, then the estimated value of SD(DH) is not as reliable as it should be in theory with the same sample size. Alternatively, it can be said that the value of SD(DH) possesses a larger variance than the theoretical value for that sample size. Therefore, the value of SD(DH) is not reliable enough to be used to represent $\mathrm{SD(DH_1)}$.

Accuracy of Check Points and the Reliability of the Standard Deviation Estimate

From the discussions conducted in the previous section, it can be concluded that, if the standard deviation of the check points is smaller than the critical value set out in this section, then their effect is negligible. However, if the check points have a standard deviation larger than the critical value, then the accuracy of check points itself affects the reliability of the estimated accuracy figures. In this section these effects are discussed.

Substituting Equation 20 into Equation 9, the following relationship can be obtained:

$$\mathrm{R(SD)} = \frac{K}{\sqrt{(1-K^2)}} \times 100\% \quad (21)$$

This formula expresses the relationship between K (the ratio of the standard deviation of the check points to the standard deviation of the final DTM) and the reliability of the standard deviation estimate. For example, if $K=0.09$, then $R=9.0$ percent. The graphical representation of Equation 21 is shown in Figure 5.

The reliability of the estimated standard deviation figure R(SD) derived from both Equation 10 and Equation 21 should be very similar if the accuracy of the check points is higher than the critical value. However, when the accuracy of the check points is lower than the criterion which has been set, then the value of reliability computed from Equation 21 will be much lower than that from Equation 10. Equation 21 also shows that, if the SD of the check points is 70.7 percent of the DTM SD, then the SD of the estimated standard deviation, SD(SD(DH)), will be equal to the SD(DH) itself. This confirms what has been stated before — namely, that the accuracy of the check points affects the reliability of the standard deviation estimate if it is lower than the critical value set by the formula given in Equation 19.

EFFECT OF THE DISTRIBUTION OF CHECK POINTS ON THE RELIABILITY OF THE ACCURACY ESTIMATE

Another important concern with the check points used for the DTM accuracy test is their distribution. The distribution of the check points can be characterized by their locations and patterns. In the ISPRS test, the check points are in a grid pattern. The question must be raised as to whether such a pattern is suitable. If not, then it poses the question as to what kind of distribution is desirable. Ley (1986) has made some efforts to answer this question. He stated that "an accuracy assessment of a DTM should be based on a sample of heights taken from the entire model." He also points out that such "a sample of points should include both the recorded (measured) and inter-

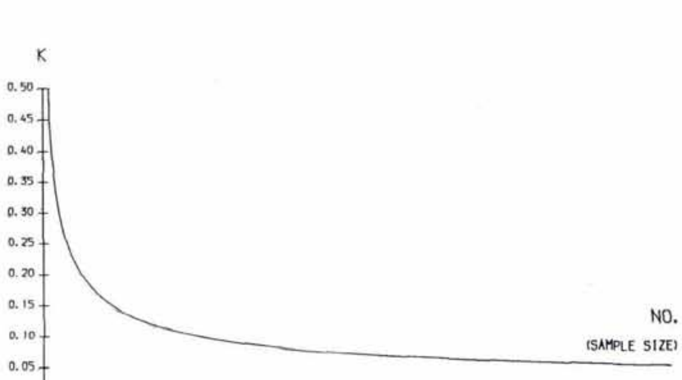

FIG. 4. Required accuracy of check points (in terms of the ratio of the SD of check points to that of the DTM) with sample size.

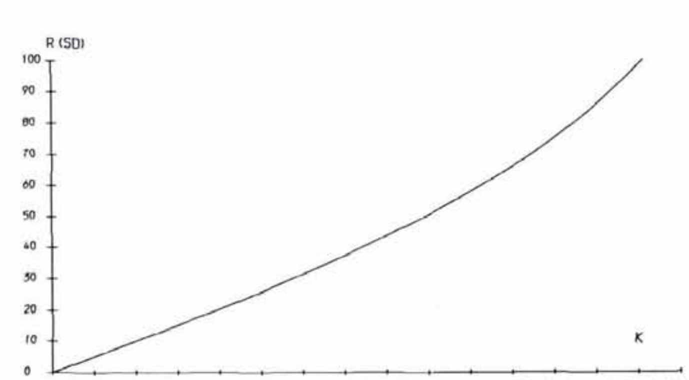

FIG. 5. The reliability of the estimated SD of the DTM with the accuracy of the check points (in terms of the ratio of the SD of the check points to that of the DTM).

polated heights." However, the answer to the question as to their distribution is still not complete.

Therefore, it is of interest to know how this factor affects the reliability of the accuracy figures for the DTM to be estimated from test results. If this were known, then the desirable distribution of check points could be determined. In this section, an attempt will be made to discuss this particular matter from the viewpoint of statistical theory. An experimental test has also been carried out to see if such a theoretical analysis is applicable to DTM practice.

Theoretical Discussion

A serious shortcoming of using check points located in a grid pattern is that they then represent a systematic sample. In this case, if the first point is sampled, then the positions (locations) of all other points are definitely determined. Such a sample is evenly distributed whereas the procedure which has been discussed for use in a DTM accuracy assessment is based on random sampling. From this point of view, a gridded data pattern is not so appropriate. Thus, from the purely theoretical standpoint, in order to make such a statistical procedure applicable, random sampling is desirable.

The use of a grid pattern for check points may be the result of the thought that the DH values in some parts of the area being tested may be greater than those in other parts and that the sample is representative only if the points are so distributed. Such a line of thinking would ignore the prerequisite for such a statistical test; namely, that the sample should come from the same distribution, because of the fact that the DH values in some parts of the test area are greater than those in other parts and are not from the sample space or population. If the stated prerequisite should be applied, then the large values of DH should also be randomly distributed. Therefore, the use of a gridded pattern of check points is not always sound. The advantages of using it are (1) its convenience and efficiency in terms of implementing a sampling and measuring strategy in a stereo-plotting machine, and (2) its convenience in terms of the resulting data structure which can be implemented in the computer used for the processing of the data.

In this case, the concept of random sampling is very clear. It means that there is no intention to select a point in specific position so that any point, including the recorded points, has an equal chance of being measured at every time of sampling.

Finally, it should be noted that the remarks made in this section are based solely on a purely theoretical analysis and may not be so suitable in practice because the terrain surface is certainly not the result of a purely stochastic process. Therefore, some experimental tests will be conducted to see how far the statistical theory varies from DTM practice.

Experimental Test

The two ISPRS test areas, Uppland and Sohnstetten, have again been used for this purpose. The aim of the test is to find how DTM accuracy estimates vary with different distributions of check points.

The first step in this test is to select randomly some sets of check points with a certain size (number) from the original data sets (1,892 points for Sohnstetten area and 2,314 for Uppland). In this test, for each area, 15 sets of check points have been used, each with a sample size of 500 points. The randomness of the selection was achieved by using a set of random numbers from a uniform distribution which was generated by computer using an NAG (Numerical Algorithm Group) routine. (More discussion about the limitations of this test will be given in the next section). In generating the random numbers, the range is determined by the total number of points in the original data set. For example, for the Uppland area, the random numbers lie within 1 to 2,314. After this, those check points with the same numbering as the generated random numbers are taken from the data set and form the sample.

The test results are listed in Table 3. Some standard statistical parameters compiled from these results are given in Tables 4 and 5. In the computation of the percentage values, the arithmetic means are assumed to be the best estimates of these values. The expected tolerable values are computed according to the theoretical formulae set out in the previous sections.

Discussion of the Test Results

From these results, it can be seen that the standard deviation of the standard deviation estimate for the Uppland data set behaves very well, but that for the Sohnstetten data set is much larger than expected. Using another measure – the mean, all the values derived from both the Uppland data set and the Sohnstetten data set fall within the range expected.

Of course, the variation in the accuracy results may also be related to the roughness and/or the steepness of the terrain surface. The fact that the results for Uppland behave better could be due to the smaller slope angles which prevail in the area. The results could also have been affected by the errors in the check points themselves. However, such an effect in this particular case is not significant here because the accuracy of

Table 3. Accuracy Results for the Randomly Selected Check Points for the Uppland and Sohnstetten Areas (No. = 500)

	Results for Uppland Area			Resultsfor Sohnst Area		
File No.	RMSE (±m)	SD (±m)	Mean (m)	RMSE (±m)	SD (±m)	Mean (m)
1	0.628	0.589	0.219	0.426	0.397	−0.155
2	0.589	0.543	0.228	0.421	0.390	−0.158
3	0.603	0.565	0.204	0.417	0.387	−0.157
4	0.648	0.585	0.278	0.412	0.382	−0.155
5	0.633	0.595	0.216	0.403	0.369	−0.160
6	0.621	0.586	0.207	0.407	0.384	−0.136
7	0.637	0.597	0.224	0.437	0.415	−0.138
8	0.601	0.565	0.205	0.448	0.425	−0.143
9	0.629	0.570	0.227	0.453	0.427	−0.152
10	0.630	0.594	0.212	0.398	0.372	−0.142
11	0.637	0.593	0.232	0.451	0.426	−0.149
12	0.623	0.570	0.252	0.431	0.389	−0.186
13	0.622	0.586	0.208	0.417	0.385	−0.162
14	0.618	0.567	0.249	0.424	0.397	−0.148
15	0.617	0.562	0.235	0.408	0.381	−0.145

Table 4. Statistical Estimates for the SD

		Uppland Area	Sohnst Area
Average Value (AV)		0.5778	0.395
SD of Distribution		0.015	0.019
SD/AV	Computed	2.64%	4.72%
	Expected	3.16%	3.16%

Table 5. Statistical Estimates for the Mean

		Uppland Area	Sohnst Area
AV	(Average Value)	0.226	−0.152
SD	Computed	0.020	0.012
	Expected	0.026	0.018
S	Max. (computed)	0.052	0.034
	99% Confidence	0.067	0.046
	98% Confidence	0.060	0.041
	95% Confidence	0.051	0.035

the check points is much higher than that of the DTMs (Li, in press). This test shows that, to a certain extent, random sampling over the entire area (without taking into account geographical location) is a method which is acceptable for the creation and acquisition of check points.

Before ending this discussion, some remarks on the randomness of the check points used in this test need to be made. In this experimental test, nominally, the check points were randomly sampled with a size of 500 points from the entire set of check points. However, in practice, truly random numbers can only be obtained by rolling dice or dealing cards or be generated by special mechanical machines. The randomness which was achieved by the NAG routine is always doubtful because the random numbers generated by computer software follow certain rules specified by algorithms (Frodesen *et al.*, 1979). Also, for this particular test, the locations of the check points have not been changed so the data set is still not a random sample of the whole area covered, but only of the original grid. Such a limitation may affect the conclusion made in this study.

DISCUSSION AND CONCLUSION

All the theoretical discussions set out in the previous sections of this paper are based purely on statistical theory – that is to say, they are valid in theory. The theory regarding sample size has been confirmed by the limited tests in this study. No tests on the theory related to the accuracy of check points have been carried out in this study because it is not easy to acquire a set of samples of check points with different accuracies. On the random sampling of check points, one test shows quite encouraging results, but the other is not so satisfactory. Therefore, more studies on the applicability of these theories to the practice of DTM accuracy tests still need to be carried out because practical situations are not as perfect as those required by statistical theories.

The standard deviation of the DTM is assumed to be known beforehand, both for the estimation of required minimum sample size and the required accuracy of check points. It needs, therefore, to be estimated at first. However, a rough estimate may serve the purpose. As stated before, the problem of how to obtain such a value lies outside of this study but has been discussed elsewhere (e.g., Li, 1990).

From the preceding discussion, some conclusions might be drawn for the check points used in the experimental tests on DTM accuracy as follows:

- The accuracy and the reliability of the final DTM accuracy estimates are affected by the sample size (the number) of the check points used in the experimental tests. In a reverse way, it can be said that the required minimum sample size is determined by the given degree of accuracy or the reliability requirement. A general guide to the required values can be derived from Equations 4 and 10, if the check points are free of errors and relief reflects a stochastic process.
- The reliability of the estimated DTM accuracy figures is also affected by the accuracy of check points. Again, the accuracy of check points required for a given degree of reliability can also be determined by Equation 19.
- The reliability of the estimated standard deviation figure was expressed in terms of percentage of the estimated value. It can be obtained through the use of Equations 10 and 21. When the variance of the check points is larger than the critical value, then Equation 21 should be used to compute the reliability factor.
- The check points could be sampled randomly from the entire testing area (and preferably as a result of a very even distribution). In this context, the use of the word "randomly" is meant to convey the concept that every point, including the recorded points, has the same chance of being selected every time sampling is carried out.
- Only if the sample size is increased and the accuracy of the check points is improved at the same time, can the reliability of the final estimates be improved. It may be very difficult to implement the second of these criteria.

The discussion carried out in this study might be also applicable to other experimental tests using check points.

ACKNOWLEDGMENTS

This paper is based on the materials included in the author's Ph.D. thesis. The author would like to thank the University of Glasgow and CVCP for their financial support during this study. Thanks are also due to Prof. G. Petrie and Prof. B. Makarivic for their comments. The kind comments and suggestions made by three referees are also very appreciated.

REFERENCES

Burington, R., and D. May, 1970. *Handbook of Probability and Statistics with Tables*. 2nd edition. McGraw-Hill Book Company. 264p.

Dummer, G., and R. Winton, 1986. *An Elementary Guide to Reliability*. 3rd edition. Pergamon Press. 47p.

Frodesen, A., O. Skjeggestad, and H. Tofte, 1979. *Probability and Statistics in Particle Physics*. Universitetsfortlaget. Oslo. 501p.

Hogg, R., and E. Tanis, 1977. *Reliability and Statistical Inference*. Macmillan Publishing Co., Inc., N.Y. 450p.

Ley, R., 1986. Accuracy assessment of digital terrain models. *Proceedings, Auto Carto London*, 1:455–464.

Li, Zhilin, 1990. *Sampling Strategy and Accuracy Assessment for Digital Terrain Modelling*. Ph.D. thesis. University of Glasgow. 299p.

———, in press. Variation of the accuracy of digital terrain models with sampling interval. *Photogrammetric Record*.

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

(Received 9 August 1990; revised and accepted 20 February 1991)