Accuracy assessment of digital elevation models by means of robust statistical methods

Höhle, Joachim; Höhle, Michael

Published in:
ISPRS Journal of Photogrammetry and Remote Sensing

DOI (link to publication from Publisher):
10.1016/j.isprsjprs.2009.02.003

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

ISPRS Journal of Photogrammetry and Remote Sensing 64 (2009) 398–406

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Accuracy assessment of digital elevation models by means of robust statistical methods

Joachim Höhle[a,*], Michael Höhle[b]

[a] Department of Development and Planning, Aalborg University, Denmark

[b] Department of Statistics, Ludwig-Maximilians-Universität München, Germany

ARTICLE INFO

Article history:
Received 4 March 2008
Received in revised form
30 January 2009
Accepted 3 February 2009
Available online 19 March 2009

Keywords:
DEM/DTM
Laser scanning
Photogrammetry
Accuracy
Specifications

ABSTRACT

Measures for the accuracy assessment of Digital Elevation Models (DEMs) are discussed and characteristics of DEMs derived from laser scanning and automated photogrammetry are presented. Such DEMs are very dense and relatively accurate in open terrain. Built-up and wooded areas, however, need automated filtering and classification in order to generate terrain (bare earth) data when Digital Terrain Models (DTMs) have to be produced. Automated processing of the raw data is not always successful. Systematic errors and many outliers at both methods (laser scanning and digital photogrammetry) may therefore be present in the data sets. We discuss requirements for the reference data with respect to accuracy and propose robust statistical methods as accuracy measures. Their use is illustrated by application at four practical examples. It is concluded that measures such as median, normalized median absolute deviation, and sample quantiles should be used in the accuracy assessment of such DEMs. Furthermore, the question is discussed how large a sample size is needed in order to obtain sufficiently precise estimates of the new accuracy measures and relevant formulae are presented.

1. Introduction

Digital Elevation Models are today produced by digital photogrammetry or by laser scanning. Both methods are very efficient and accurate; the density of the elevations is very high. However, blunders may occur at both methods. From the raw data a Digital Terrain Model (DTM) and a Digital Surface Model (DSM) are generated by means of filtering (for classifying into ground and off terrain points) and interpolation (for filling gaps). Errors may also occur during such a post-processing. The quality control should detect errors and outliers in order to eliminate them. As a final step it has to be checked, whether the edited DTM and DSM achieve the accuracy of the specification. For this purpose, accurate reference values are required, and accuracy measures like the Root Mean Square Error (RMSE), mean error and the standard deviation have to be derived. The amount of data is huge, but the accuracy assessment has to be made with few check points only as it is very labour intensive to obtain them. However, the sample size should be large enough to guarantee reliable accuracy measures, which are valid for the whole DTM or DSM. Usually, the specification of accuracy measures is based on the assumption that the errors follow a Gaussian distribution and that no outliers exist. But all too often this is not the case, because objects above the terrain like vegetation, buildings and unwanted objects (cars, people, animals, etc.) are present, and the filtering program may not label all ground elevations correctly. Also system errors will occur: Photogrammetry needs structure and texture in the images and not all of the image parts fulfil this requirement. Laser light is not always reflected directly by the points to be measured and the position and altitude of the sensor may be in error. Positional errors will cause vertical errors at terrain with steep slopes and buildings. Altogether, editing of the data has to detect and correct such errors, but even with the most careful editing errors will remain. The number or percentage of outliers should be documented, for example in metadata, so that one can judge whether the derived DTM is usable for the intended application ("fit for purpose").

The derivation of accuracy measures has to adapt to the fact that outliers may exist and that the distribution of the errors might not be normal. There is thus a need for accuracy measures, which are reliable without being influenced by outliers or a skew distribution of the errors.

These facts are well known and mentioned in recently published textbooks and manuals, for example in Li et al. (2005) and Maune (2007). Recent publications, which deal in detail with

* Corresponding address: 11 Fibigerstraede, DK-9220, Aalborg, Denmark. Tel.: +45 9940 8361; fax: +45 9815 6541.

E-mail addresses: jh@land.aau.dk (J. Höhle), Michael.Hoehle@stat.uni-muenchen.de (M. Höhle).

0924-2716/$ – see front matter © 2009 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.isprsjprs.2009.02.003

accuracy assessment, are for example Carlisle (2005), Höhle and Potuckova (2006), Fisher and Tate (2007), Aguilar et al. (2007a) and Zandbergen (2008).

Our approach continues along these lines as we focus on vertical accuracy assessment in the light of outliers and non-normal distributions. It is the **objective** of this article to advocate robust statistical methods for the derivation of vertical accuracy measures for digital elevation models. Robust approaches handling outliers and detaching accuracy measures from the assumption of an underlying normal distribution have increasingly been suggested in the literature (e.g. Atkinson et al., 2005 and Aguilar et al., 2007b). It is a topic of discussion how national and international standards for DEMs should cope with these matters. So far, only the Lidar committee of the American Society for Photogrammetry and Remote Sensing deals with it and requires non-parametric accuracy measures for non-open terrain (ASPRS Lidar Committee, 2004). With this article focusing on robust estimation of variance and the estimation of sample quantiles as measures for vertical accuracy, we want to contribute to this discussion of DEMs produced either by airborne laser scanning or automated digital photogrammetry. In our approach we interpret accuracy measures directly as quantities of the error distribution — alternatives are more indirect measures such as e.g. the coefficient of variation of the sample variance (Aguilar et al., 2007a).

When validating DEMs, accurate reference data have to be available in a sufficiently large number. The question how many check points are needed can be treated within the statistical context of sample size computation. We show how required sample sizes can be calculated for the suggested quantile approach to accuracy.

The paper is organized as follows: Section 2 discusses accuracy requirements for DEMs, Sections 3–5 deal with vertical errors and provide ordinary and robust accuracy measures. Section 6 discusses how these can be used to asses fulfilment of a specification using statistical tests, and Section 7 illustrates the methods using four examples of DTM accuracy. A discussion of the results completes the paper.

2. Requirements regarding the reference data

Accuracy assessment of the DEM is carried out by means of reference data called checkpoints. Because their position does not coincide with the posts of the DEM, an interpolation is necessary. For a DEM with a grid structure a bilinear interpolation is normally used. The accuracy of the checkpoints should be at least three times more accurate than the DEM elevations being evaluated (Maune, 2007, pp. 407). By using the formula for error propagation the influence on the DEM accuracy can be estimated:

$$\sigma^2_{\mathrm{DEM-REF}} = \sigma^2_{\mathrm{DEM}} + \sigma^2_{\mathrm{REF}} \leq \sigma^2_{\mathrm{DEM}} + \left(\frac{1}{3}\cdot\sigma_{\mathrm{DEM}}\right)^2 = \frac{10}{9}\cdot\sigma^2_{\mathrm{DEM}}. \tag{1}$$

Hence, $\sigma_{\mathrm{DEM-REF}} \leq 1.05\cdot\sigma_{\mathrm{DEM}}$. The derived DEM accuracy is then incorrect by 5%, which is acceptable. For example, if the accuracy of a DEM is specified with $\sigma = 10$ cm then the checkpoints should have an accuracy of $\sigma \leq 3.3$ cm. The DEM accuracy would then amount to 10.5 cm only.

An important issue is the spatial distribution of the checkpoints: they should be distributed randomly. If checkpoints are positioned along break lines, at steep slopes, at positions of sudden slope change, close to buildings, etc., large errors may be found. On the other hand, the number of checkpoints (aka. sample size) should be sufficiently large in order to obtain reliable accuracy measures. In Section 6 we return to the issue of how to compute the sample size in order to prove the compliance with accuracy specifications.

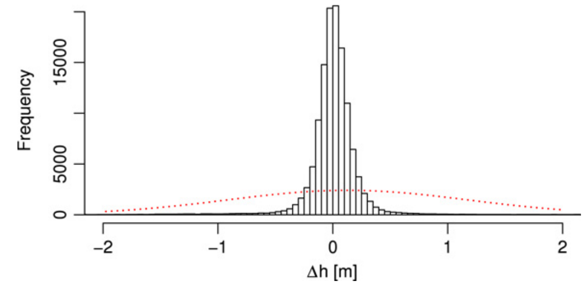

Fig. 1. Histogram of the errors Δh in metres. Superimposed on the histogram are the expected counts from a normal distribution with mean and variance estimated from the DEM data using non-robust estimators. For a better visualisation the histogram is truncated at −2 m and 2 m. The mismatch between data and estimated normal curve is due to heavy tails.

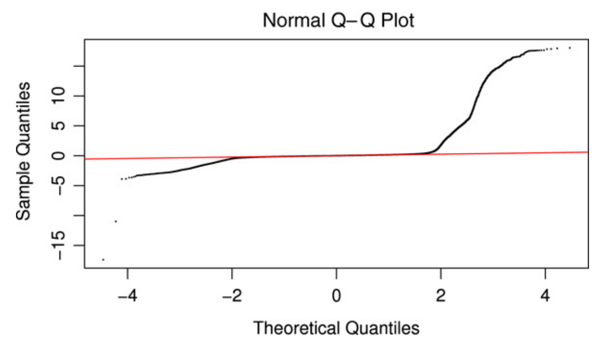

Fig. 2. Normal Q–Q-plot for the distribution of Δh.

3. The distribution of errors

The distribution of errors can be visualized by a **histogram** of the sampled errors, where the number of errors (frequency) within certain predefined intervals is plotted. Such a histogram gives a first impression of the normality of the error distribution. Fig. 1 depicts a histogram of the error distribution for photogrammetric measurements compared with checkpoints. In order to compare with normality, the figure contains a superimposed curve for a normal distribution (Gaussian bell curve) obtained by ordinary estimation of mean error and variance. Because outliers are present in the data, the estimated curve does not match the data very well. Reasons could be that the errors are not originating from a normal distribution, e.g. because a skew distribution exists which is not symmetric around its mean or because the distribution is more peaked around its mean than the normal distribution while having heavier tails. The latter effect is measured by the kurtosis of the distribution, which in this situation is bigger than zero.

A better diagnostic plot for checking a deviation from the normal distribution is the so-called quantile–quantile **(Q–Q) plot**. The quantiles of the empirical distribution function are plotted against the theoretical quantiles of the normal distribution. If the actual distribution is normal, the Q–Q plot should yield a straight line. Fig. 2 shows the Q–Q plot for the distribution of Δh in the same example as Fig. 1. A strong deviation from a straight line is obvious, which indicates that the distribution of the Δh is not normal.

It is also possible to use statistical tests to investigate whether data originate from a normal distribution, but these tests are often rather sensitive in case of large data sets or outliers. We, therefore, prefer the visual methods presented above. More details about the mentioned statistical tests for normality can be taken from e.g. D'Agostino et al. (1990), who also recommend visual methods as a component of good data analysis for investigating normality.

Table 1
Accuracy measures for DEMs presenting normal distribution of errors.

Root mean square error	$\hat{\text{RMSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Delta h_i^2}$
Mean error	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\Delta h_i$
Standard deviation	$\hat{\sigma} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(\Delta h_i - \hat{\mu})^2}$

4. Accuracy measures for the normal distribution

If a normal distribution can be assumed and no outliers are present in the data set, the following accuracy measures for DEMs can be applied (cf. Table 1).

In the table Δh_i denotes the difference from reference data for a point i, and n is the number of tested points in the sample (sample size). When an underlying normal distribution of the errors can be assumed, it is well known from the theory of errors that 68.3% of the data will fall within the interval $\mu \pm \sigma$, where μ is the systematic error (aka. bias) and σ is the standard deviation, see e.g. Mikhail and Ackermann (1976). If the accuracy measure should be based on a 95% confidence level (aka. a 95% tolerance interval), the interval is $\mu \pm 1.96 \cdot \sigma$ instead. However, for **DEMs derived by laser scanning or digital photogrammetry** a normal distribution of the errors is seldom due to e.g. filtering and interpolation errors in non-open terrain. In this work we describe and compare several approaches how to deal with this situation.

One approach to deal with outliers is to remove them by applying a threshold. For example, the threshold can be selected from an initial calculation of the accuracy measures. In the DEM tests described in Höhle and Potuckova (2006), the threshold for eliminating outliers was selected as three times the Root Mean Square Error (RMSE), i.e. an error was classified as outlier if $|\Delta h_i| > 3 \cdot \text{RMSE}$. Another approach is to use $3 \cdot \sigma$ as the threshold for the outlier detection (aka. 3 sigma rule), where σ is the specified vertical accuracy or a preliminary value for the standard deviation which is derived from the original data set (Daniel and Tennant, 2001). But not all of the outliers can be detected in this way, and the DEM accuracy measures (mean error and standard deviation) will therefore be wrong or inaccurate. Furthermore, it can be shown that even the best methods based on outlier removal do not achieve the performance of robust methods, because the latter are able to apply a more smooth transition between accepting and rejecting an observation (Atkinson et al., 2005). Robust methods for the derivation of accuracy measures should therefore be applied for the assessment of DEM accuracy.

5. Robust accuracy measures suited for non-normal error distributions

If the histogram of the errors reveals skewness, kurtosis or an excessive amount of outliers, another approach for deriving accuracy measures has to be taken. Such an approach has to be resistant to outliers, and the probability assumptions to be made should not assume normality of the error distribution. Our suggestion in this case is to use the sample quantiles of the error distribution.

The **quantile** of a distribution is defined by the inverse of its cumulative distribution function (CDF), F, i.e.

$$Q(p) = F^{-1}(p)$$

for $0 < p < 1$. For example, the 50% quantile, $Q(0.5)$, is the median of the distribution. An alternative definition using the minimum is necessary in cases where F is a step function and thus no unique definition of the inverse exists:

$$Q(p) = \min\{x : F(x) \geq p\}.$$

Sample quantiles are non-parametric estimators of the distributional counterparts based on a sample of independent observations $\{x_1, \ldots, x_n\}$ from the distribution. We use the so-called order statistic of the sample $\{x_{(1)}, \ldots, x_{(n)}\}$, where $x_{(1)}$ denotes the minimum and $x_{(n)}$ the maximum of $\{x_1, \ldots, x_n\}$. Thus, for a sample $\{0.1, -0.3, -0.5, 0.4, 0.1\}$ of size $n = 5$, the order statistic is $\{-0.5, -0.3, 0.1, 0.1, 0.4\}$.

A simple definition of the sample quantile is now $\hat{Q}(p) = x_{(j)}$ where $j = \lceil p \cdot n \rceil$ and $\lceil p \cdot n \rceil$ denotes rounding up to the smallest integer not less than $p \cdot n$.

If an interest exists in the 10%, 20%, 50% and 90% quantile, the following values for j are obtained:

$$j = \lceil 0.1 \cdot 5 \rceil = 1, \quad j = \lceil 0.2 \cdot 5 \rceil = 1, \quad j = \lceil 0.5 \cdot 5 \rceil = 3,$$
$$j = \lceil 0.9 \cdot 5 \rceil = 5.$$

The corresponding sample quantiles of the distribution are then:

$$\hat{Q}(0.1) = x_{(1)} = -0.5, \quad \hat{Q}(0.2) = x_{(1)} = -0.5,$$
$$\hat{Q}(0.5) = x_{(3)} = 0.1, \quad \text{and} \quad \hat{Q}(0.9) = x_{(5)} = 0.4.$$

In other words, the 50% quantile or median of this sample is 0.1 and the 90% quantile equals 0.4.

An often desired property of $\hat{Q}(p)$ is that it should be a continuous function of p. To obtain this and other desirable properties, one can extend the above definition by using a linear interpolation between the two relevant successive data points (Hyndman and Fan, 1996). The calculation for the practical examples of Section 6 will be carried out by the software "R"— a free software environment for statistical computing and graphics (R Development Core Team, 2008).

With respect to the application of accuracy measures of DEMs we use the distribution of Δh **and** $|\Delta h|$. One robust quality measure is the median $\hat{Q}_{\Delta h}(0.5)$, also denoted $m_{\Delta h}$, which is a robust estimator for a systematic shift of the DEM. It is less sensitive to outliers in the data than the mean error and provides a better distributional summary for skew distributions.

A robust and distribution free description of the measurement accuracy is obtained by reporting sample quantiles of the distribution of the **absolute** differences, i.e. of $|\Delta h|$. Absolute errors are used, because we are interested in the magnitude of the errors and not their sign. Furthermore, absolute errors allow us to make probability statements without having to assume a symmetric distribution.

For example, the 95% sample quantile of $|\Delta h|$ literally means that 95% of the errors have a magnitude within the interval $[0, \hat{Q}_{|\Delta h|}(0.95)]$. The remaining 5% of the errors can be of any value making the measure robust against up to 5% blunders. Such probability statements about a certain proportion of the errors falling within a given range are usually obtained by assuming a normal distribution. For example, one assumes that the symmetric interval of $\pm 1.96 \cdot \hat{\sigma}$ (where $\hat{\sigma}$ is the estimated standard deviation) contains 95% of the errors (assuming no systematic error). Equivalently, this means that 95% of the absolute errors are within $[0, 1.96 \cdot \hat{\sigma}]$. Thus, if the distribution of Δh is really normal then $\hat{Q}_{|\Delta h|}(0.95)$ converges to the estimator of $1.96 \cdot \hat{\sigma}$.

If the problem is the heavy tails of the error distribution due to a large amount of outliers, an alternative approach to estimate the scale of the Δh distribution is to use a robust scale estimator such as the **Normalized Median Absolute Deviation (NMAD)**:

$$\text{NMAD} = 1.4826 \cdot \text{median}_j(|\Delta h_j - m_{\Delta h}|), \quad (2)$$

where Δh_j denotes the individual errors $j = 1, \ldots, n$ and $m_{\Delta h}$ is the median of the errors. The NMAD is thus proportional to the median of the absolute differences between errors and the median error. It can be considered as an estimate for the standard deviation more resilient to outliers in the dataset. In

Table 2
Proposed accuracy measures for DEMs.

Accuracy measure	Error type	Notational expression
Median (50% quantile)	Δh	$\hat{Q}_{\Delta h}(0.5) = m_{\Delta h}$
Normalized median absolute deviation	Δh	$\text{NMAD} = 1.4826 \cdot \text{median}_j(\lvert\Delta h_j - m_{\Delta h}\rvert)$
68.3% quantile	$\lvert\Delta h\rvert$	$\hat{Q}_{\lvert\Delta h\rvert}(0.683)$
95% quantile	$\lvert\Delta h\rvert$	$\hat{Q}_{\lvert\Delta h\rvert}(0.95)$

case of an underlying normal distribution this value is equivalent to the standard deviation if the number of checkpoints (i.e. n) is sufficiently large. More details about such robust estimation can be taken from Hoaglin et al. (1983).

In summary, as a robust and distribution free approach handling outliers and non-normal distribution we suggest the following accuracy measures given in Table 2.

One could furthermore assess non-normality using estimators for the skewness and kurtosis of the distribution. However, to be consistent, robust estimators such as the Bowley coefficient of skewness and a standardized kurtosis measure suggested by Moors (1988) should be used. However, we will in this text use the more intuitive visual inspection using QQ-plots and histograms.

In a statistical context all estimates of population quantities should be supplied by measures quantifying the uncertainty of the estimator due to estimation from a finite sample. One way to achieve this is to supply with each point estimator a **confidence interval** (CI) with a certain coverage probability. For example, a 95% CI $[c_1, c_2]$ for the sample median says that in 95% of the errors the interval $[c_1, c_2]$ contains the true but unknown median of the error distribution.

Using the bootstrap is one approach to assess the uncertainty of the above sample quantiles as estimators of the true quantiles of the underlying distribution (Davison and Hinkley, 1997). Here one draws a sample of size n with replacement from the available data $\{x_1, \ldots, x_n\}$ and then uses this new sample to compute the desired $\hat{Q}(p)$. This procedure is repeated a sufficiently large number of m times; in our case we choose $m = 999$, which yields 999 values $\hat{Q}^1(p), \ldots, \hat{Q}^{999}(p)$. A bootstrap based 95% confidence interval of $Q(p)$ can then be obtained as the interval from the 2.5% to the 97.5% sample quantiles of the 1000 available values $\{\hat{Q}(p), \hat{Q}^1(p), \ldots, \hat{Q}^{999}(p)\}$.

We prefer such a bootstrap approach over classical large sample arguments to construct confidence intervals (as e.g. in Desu and Raghavarao (2003)), because the bootstrap is more robust to the small number of check points used and can be extended to handle autocorrelated data.

So far all calculations in our work are based on the assumption that errors are independent and identically distributed. However, as Fig. 7 shows, there is substantial spatial autocorrelation present in the data, which will make the proposed bootstrap estimated confidence intervals too narrow. One approach to treat this problem is to modify the above bootstrap procedure to take the dependence of data into account using e.g. a block bootstrap (Lahiri, 2003). Another approach would be dividing data into a number of terrain classes and compute sample quantiles within each class. A statistical framework for this task is **quantile regression** (Koenker, 2005). In its simplest form a quantile regression model for the pth quantile of the absolute error distribution is

$$|\Delta h_i| = \beta_0 + \varepsilon_i,$$

where the ε_i are independent realisations of a random variable having a CDF F which is completely unspecified having the only requirement that $F(0) = p$. The value of β_0 depends on the quantile p, but for ease of exposition we have omitted this from the notation. It is then possible to show that with this formulation we have $\beta_0 = Q(p)$. An estimator for β_0 is obtained from the n observed absolute errors by solving the following minimization problem:

$$\hat{\beta}_0 = \arg\min_{\beta_0} \sum_{i=1}^{n} \rho_p(|\Delta h_i| - \beta_0)$$

where $\rho_p(u)$ is the so-called check function for the pth quantile defined as $\rho_p(u) = u \cdot (p - I(u < 0))$ with $I(u < 0)$ being one if $u < 0$ and zero otherwise. The above minimization problem can be solved by linear programming and is implemented in the R add-on package quantreg, see Koenker (2005). It can be shown that $\hat{\beta}_0$ is equivalent to the previous definition of the pth sample quantile, i.e. $\hat{\beta}_0 = \hat{Q}(p)$.

General quantile regression for a specific quantile p proceeds by replacing the single parameter model $Q(p) = \beta_0$ with a linear predictor $Q(p) = x_i^T \beta$ as in ordinary linear regression. This allows e.g. modelling the p-quantile as a function of terrain classes as in Carlisle (2005), who used ordinary least squares regression for this task. Modelling autocorrelation in an even better way would be to model the pth quantile as a function of the (x, y) position in the plane, i.e. $Q(p) = f(x, y)$, where the function f could for example be a tensor product of univariate basis functions or a triogram function as described in Koenker and Mizera (2004). In both cases, penalization is used to ensure an appropriate smoothness of the function. Using the Koenker and Mizera (2004) approach a continuous, piecewise linear function over an adaptively selected triangulation over the (x, y) plane for e.g. the 95% quantile of the absolute error distribution can thus be obtained.

6. Statistical tests and sample size calculations

It becomes obvious from the confidence intervals of the preceding section that an important issue in the above quality control is the question of how large a **sample size** is needed in order to obtain sufficiently precise estimates of $\hat{\sigma}$ and the sample quantiles. One approach to this problem is to solve it by means of a sample size methodology for statistical tests, which will be described in the following for the case where errors are assumed to have no autocorrelation.

Considering a normal distribution of the errors, a typical specification would be to demand $\sigma < \sigma_{\text{spec}}$ with e.g. $\sigma_{\text{spec}} = 10$ cm, i.e. the true (but unknown) standard deviation of the error distribution is smaller than 10 cm. Thus, we may formulate a statistical test with null hypothesis $H_0 : \sigma^2 = \sigma^2_{\text{spec}}$ and an alternative hypothesis $H_A : \sigma^2 < \sigma^2_{\text{spec}}$. A sample of size n is now drawn and the hope is to be able to reject the null hypothesis (thus proving the desired specification). This is done (see e.g. Desu and Raghavarao (1990)) if

$$\hat{\sigma}^2 < \frac{\sigma^2_{\text{spec}} \cdot \chi^2_{\alpha,n-1}}{n-1}, \tag{3}$$

where α is the pre-specified type I error probability, i.e. the probability of erroneously rejecting H_0 when $\sigma = \sigma_{\text{spec}}$ and $\chi^2_{\alpha,n-1}$ denotes the α quantile of the χ^2 distribution with $n-1$ degrees of freedom. The parameter α is also called the level of the test — in our work we shall use $\alpha = 0.05$. In other words: To check if the DEM specification $\sigma < 10$ cm is fulfilled we specify that at the extreme setting where the specification is not fulfilled, i.e. at $\sigma = 10$ cm, we want to detect this lack of compliance based on (3) with a probability of 95%.

Let $\sigma_1 < \sigma_{\text{spec}}$ and β be two predefined constants, for example $\sigma_1 = 7.5$ cm and $\beta = 0.05$. If we require that the probability of correctly rejecting H_0 when $\sigma = \sigma_1$ is equal to $1 - \beta$, the

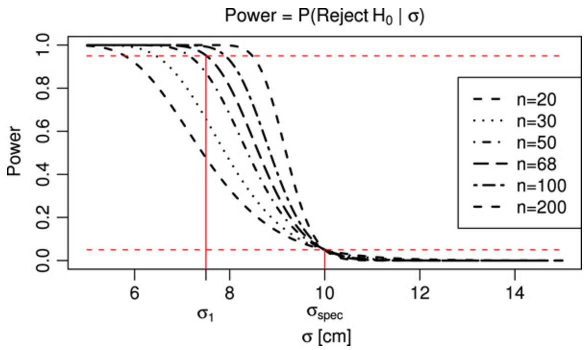

Fig. 3. Power of the discussed variance test for the normal distribution as a function of sample size n. In all cases a significance level of $\alpha = 0.05$ is used.

necessary sample size is found as the positive smallest integer n, which satisfies

$$\sigma^2_{\text{spec}}\chi^2_{\alpha,n-1} - \sigma^2_1\chi^2_{1-\beta,n-1} \geq 0.$$

Here, we say that the test has a power of $1-\beta$ at σ_1, i.e. β is the probability of erroneously keeping H_0 even though $\sigma = \sigma_1$ and thus H_1 applies because $\sigma < \sigma_{\text{spec}}$. For the above selection of values for σ_{spec}, σ_1, α and β we obtain $n = 68$. Fig. 3 shows the involved quantities and how different sample sizes n lead to different powers at σ_1. With $n = 68$ the desired power of 0.95 is achieved.

Because $(10\text{ cm})^2 \cdot \chi^2_{0.05,67}/67 = 73.38\text{ cm}^2$ we have that $\hat{\sigma}^2 < 73.38\text{ cm}^2$ in 95% of the cases when $\hat{\sigma}^2$ is computed from 68 normally distributed random variables having $\sigma = 10$ cm. Similarly, at $\sigma = 7.5$ cm, where the specification is fulfilled, we want to be sure to detect this compliance based on (3) with a probability of 95% — for σ greater than 7.5 cm this probability will be smaller as shown in Fig. 3.

For comparison, the American Society of Photogrammetry and Remote Sensing (ASPRS) recommends a minimum of 20 checkpoints in each of the major land cover categories. In the case of three landcover classes (e.g. open terrain, forested areas, and urban areas) a minimum of $n = 60$ checkpoints are required (ASPRS Lidar Committee, 2004).

The above test is, however, very sensitive to deviations from the normal distribution. Here our suggestion was to use the quantiles of the absolute error distribution. Similarly, we suggest proving compliance with a specification using statistical tests for the quantiles of the error distribution. To test whether the 68.3% quantile of the absolute error distribution is below 10 cm, for each observation Δh_i a zero-one variable $Y_i = I(|\Delta h_i| < 10\text{ cm})$ is created, where $I(|\Delta h_i| < 10\text{ cm})$ is one if $|\Delta h_i| < 10$ cm and zero otherwise. Assuming that Y_i is Bernoulli distributed with parameter p we thus want to test if H_0: $p = p_0$ against the alternative H_A: $p > p_0$. In our example of the 68.3% quantile $p_0 = 0.683$, i.e. we want to investigate if more than 68.3% of the absolute errors are smaller than 10 cm. If so, the desired accuracy specification is fulfilled. The H_0 is rejected if $Y > c$, where $Y = \sum_{i=1}^{n} Y_i$ and the constant c is found as the smallest integer so that

$$1 - F(c-1; n, p_0) \quad \leq \alpha,$$

with $F(c-1; n, p_0) = \sum_{x=0}^{c-1}\binom{n}{x}p_0^x(1-p_0)^{n-x}$ being the CDF at $c-1$ of a binomially distributed variable with size n and success probability p_0.

Again the sample size methodology can be used to compute the necessary sample size for such a Bernoulli test. A mathematical derivation can be found in Desu and Raghavarao (1990). The final approximate formula for the sample size is:

$$n = \left\lceil \left(\frac{z_\alpha + z_\beta}{2\left(\arcsin\left(\sqrt{p_1}\right) - \arcsin\left(\sqrt{p_0}\right)\right)} \right)^2 \right\rceil,$$

where z_α denotes the α quantile of the standard normal distribution and p_1 denotes the CDF value F (10 cm) of the absolute error distribution at which we want to achieve the desired power $1-\beta$.

With a comparable formulation as used in the normal setting, i.e. $p_1 = 0.818$, $\alpha = 0.05$ and $\beta = 0.05$ one obtains a required sample size of $n = 110$ to prove that the 68.3% quantile of the error distribution is below 10 cm. Here, the p_1 value was computed as the CDF of $|X|$ at 10 cm when X is normal with mean zero and variance $(7.5\text{ cm})^2$. Because no distributional assumptions are made, this number is higher than in the case of a normal distribution. In this binomial setting H_0 is rejected if $Y > 84$.

7. Results of four examples of DTM accuracy

This section illustrates the robust methods from Section 5 through four practical examples. The DTMs are derived by digital photogrammetry and laser scanning. Reference data were derived by ground surveying using GPS/RTK. In addition, the DTM derived by digital photogrammetry was directly compared with the DTM derived by laser scanning. Because the sample size is very different in these three examples special attention has to be paid to the calculated confidence intervals, which illustrate the uncertainty of the obtained estimators.

7.1. Test of the photogrammetrically derived DEM by means of GPS/RTK data

In the examples the checkpoints were distributed over the whole area with a low density or a small area with a high density. The accuracy measures were determined for both cases (large sample area or small sample area).

Images of suburbs of the city of Aalborg, Denmark, were taken by the digital large-format frame camera UltraCam D from an altitude of 640 m. One stereopair was used to derive a DEM by the software package 'Image Station Automatic Elevations', v. 5.1, of Z/I Imaging. Editing and filtering of the data were not applied.

7.1.1. Small sample area

The sample area is a fraction of the photogrammetric model (about 370m^2) and the checkpoints have a relatively high density, which leads to a relatively large sample size. The test area is mainly covered by grass, trees and a few large houses. The results of the DEM evaluation assuming a normal distribution of the errors are summarized in Table 3.

It is obvious from Table 3 that the outliers have a great influence on the estimated standard deviation. The value dropped from 20 cm to 15 cm. A histogram and a Q–Q plot shown in Figs. 4 and 5 illustrate that the distribution of the errors is non-normal.

The histogram shows that the kurtosis of the error distribution is slightly positive, i.e. the distribution has a more acute peak around the mean than the normal distribution and fatter tails. The Q–Q plot deviates from the straight line and shows more extreme positive outliers than negative ones. Table 4 summarizes the results of the robust methods.

The NMAD value and the 68.3% quantile are about the same. These values are somewhat smaller than the standard deviation of Table 3 ($\hat{\sigma}^* = 15$ cm) and the 95% confidence interval is relatively narrow at this large sample size ($n = 587$). Note also, that the estimated 95% quantile is greater than two times the 68% quantile, which clearly indicates non-normality and is caused by the acute peak and fat tails of the distribution.

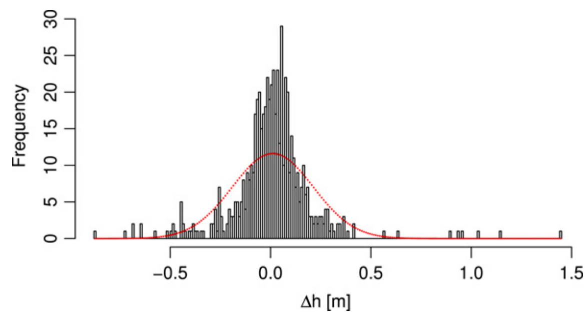

Fig. 4. Histogram of the errors Δh at a large sample size ($n = 587$). Superimposed on the histogram are the expected counts from a normal distribution with mean and standard deviation estimated from all data.

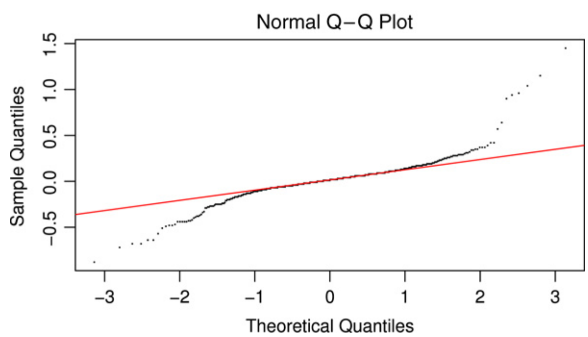

Fig. 5. Q–Q plot for the distribution of Δh at a large sample size ($n = 587$).

Table 3
Result of a DEM evaluation with the assumption of normal distribution at $n = 587$ checkpoints of which 13 are classified as outliers by the 3 · RMSE threshold.

Accuracy measures	Value (cm)
RMSE	20
Mean ($\hat{\mu}$)	1
Standard deviation ($\hat{\sigma}$)	20
Mean (after removal of outliers) ($\hat{\mu}^*$)	1
Standard deviation (after removal of outliers) ($\hat{\sigma}^*$)	15

Table 4
Accuracy measures of the robust methods for the large sample size ($n = 587$).

Accuracy measure	Error type	Value (cm)	95% confidence interval (cm)
50% quantile (median)	Δh	1	[1, 4]
NMAD	Δh	12	[10, 13]
68.3% quantile	$\|\Delta h\|$	13	[12, 16]
95% quantile	$\|\Delta h\|$	41	[36, 53]

7.1.2. Large sample area

This example is a photogrammetric model (about 0.19 km^2) and the checkpoints were distributed over the whole area of the same photogrammetric model as in Section 7.1.1. The sample size is, however, very small ($n = 19$). The used checkpoints were well defined regarding their elevation. Results for this example are given in Tables 5 and 6.

One outlier was present which leads to large values for all standard accuracy measures (RMSE, μ, σ). Moreover, the RMSE and the standard deviation have the same value although the sample mean is not zero. After removal of the outlier the results for the mean and the standard deviation are about the same as in the previous example (cf. Table 3). When using a robust method the following accuracy measures are obtained (cf. Table 6).

The accuracy measures NMAD and 68.3% quantile differ slightly, but there is again a smaller value for the 68.3% quantile, namely 8 cm instead of 14 cm when a normal distribution is assumed. Confidence intervals are relatively large at such a small sample size. Note also the extreme value of the estimated 95% quantile: the computations are robust against 5% outliers, but with one out of 19 points being an outlier this value enters the calculations.

Table 5
Result of a DEM evaluation with the assumption of normal distribution by means of $n = 19$ checkpoints of which one is classified as outlier by the 3 · RMSE threshold.

Accuracy measures	Value (cm)
RMSE	389
Mean ($\hat{\mu}$)	88
Standard deviation ($\hat{\sigma}$)	389
Mean (after removal of outliers) ($\hat{\mu}^*$)	−1
Standard deviation (after removal of outliers) ($\hat{\sigma}^*$)	14

Table 6
Accuracy measures of the robust methods for a small sample size ($n = 19$).

Robust accuracy measures	Error type	Value (cm)	95% confidence interval (cm)
Median (50% quantile)	Δh	−4	[−5, −1]
NMAD	Δh	4	[3, 11]
68.3% quantile	$\|\Delta h\|$	8	[7, 31]
95% quantile	$\|\Delta h\|$	205	[34, 1694]

Table 7
Accuracy measures of a DEM derived by laser scanning by means of $n = 41$ checkpoints of which 1 is classified as outlier by the 3 · RMSE-threshold.

Accuracy measures	Value (cm)
RMSE	33
Mean ($\hat{\mu}$)	−4
Standard deviation ($\hat{\sigma}$)	33
Mean (after removal of outliers) ($\hat{\mu}^*$)	1
Standard deviation (after removal of outliers) ($\hat{\sigma}^*$)	7

Table 8
Accuracy measures of the robust methods for the sample of size $n = 41$.

Accuracy measure	Error type	Value (cm)	95% confidence interval (cm)
50% quantile (median)	Δh	−2	[−2, 1]
NMAD	Δh	6	[6, 12]
68.3% quantile	$\|\Delta h\|$	7	[6, 10]
95% quantile	$\|\Delta h\|$	14	[13, 150]

7.2. Test of the DTM derived by laser scanning

In this section the test area is identical with the one described in Section 7.1.2. Elevations of the checkpoints were derived by ground surveying and of the DTM derived by laser scanning. The checkpoints are randomly distributed. For reasons of space we do not report histograms and Q–Q plots for this example, but refer to Tables 7 and 8 for the results.

The standard deviation after removal of one outlier is much lower as with the outlier included.

The NMAD value and the 68.3% quantile are nearly the same (6 and 7 cm, respectively). The achieved standard deviation (after removal of outliers) is $15/7 = 2.1$ times better than at the DEM derived by digital photogrammetry.

The same improvement can be found at the NMAD value and the 68.3% quantile. The condition of a three times higher accuracy is not completely fulfilled but nevertheless we will use the DTM of the whole model area as reference data for the DEM derived by photogrammetry (cf. next section). The derived accuracy measures are then **relative** errors.

7.3. Test of photogrammetric data by means of laser scanned data

The availability of an accurate and very dense DTM, which was derived from airborne laser scanning (ALS) and automatic labelling of ground points, gives the possibility of checking the

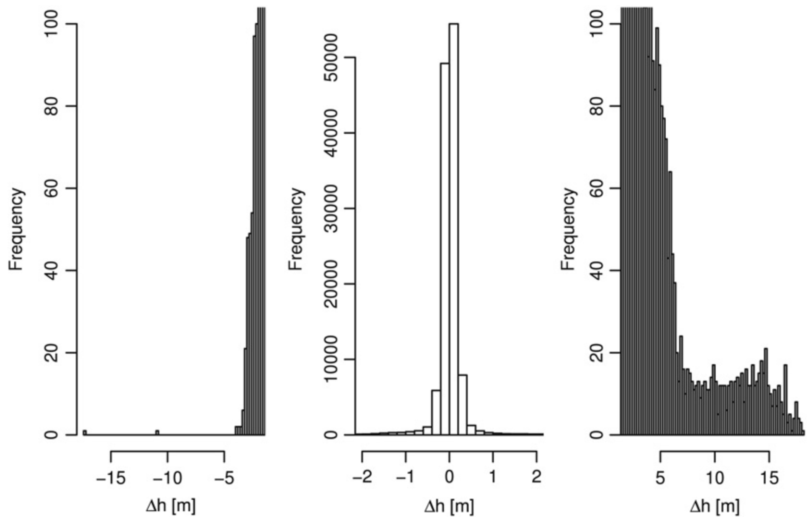

Fig. 6. Histogram of the error distribution. For better visualisation the histogram for the 200 classes is given in three separate plots: −18 m to −2 m and +2 m to +18 m with different scaling of the *y*-axis.

Table 9
Accuracy measures of a DEM derived by digital photogrammetry and checked by a dense point cloud derived from laser scanning (126559 checkpoints of which 2052 are classified as outliers by the 3 · RMSE threshold).

Accuracy measures	Value (cm)
RMSE	106
Mean ($\hat{\mu}$)	13
Standard deviation ($\hat{\sigma}$)	105
Mean (after removal of outliers) ($\hat{\mu}^*$)	2
Standard deviation (after removal of outliers) ($\hat{\sigma}^*$)	34

Table 10
Accuracy measures of the robust methods for the sample of size $n = 126\,559$.

Accuracy measure	Error type	Value (cm)	95% confidence interval (cm)
50% quantile (median)	Δh	2	[1, 2]
NMAD	Δh	12	[12, 13]
68.3% quantile	$\|\Delta h\|$	13	[13, 14]
95% quantile	$\|\Delta h\|$	68	[62, 88]

elevations of automated photogrammetry thoroughly. The applied filter is the two step approach described by Axelsson (2000), where a temporary TIN model is created first and then densified with new points by checking distance and angle parameters. Both DEMs were determined with a high density (grid spacing = 3 m for photogrammetry, grid spacing ≈ 2 m for ALS); the sample size is therefore very high. No editing occurred at the photogrammetrically derived DEM and new elevation values were calculated by bilinear interpolation at the position of a point (footprint) from laser scanning. The differences between the two elevations were evaluated with the accuracy measures assuming a normal distribution (cf. Table 9).

1.6% of the differences are outliers. After removal of the outliers the standard deviation is considerably reduced (from 105 cm to 34 cm) and the accuracy at 95% confidence level amounts to 67 cm. Again, the histogram shown in Fig. 6 reveals that other approaches have to be taken in order to obtain reliable accuracy measures for the DEM. The histograms depicted in Fig. 6 show heavy positive tails of the distribution, which are caused by a large amount of outliers at the 2–18 m range.

The median of the differences is 2 cm which is a value for the systematic shift between the two DEMs. The robust estimator (NMAD) of the standard deviation is 12 cm. Quantiles of the distribution of absolute differences ($|\Delta h|$) are: 13 cm (68.3%) and 68 cm (95%) as given in Table 10. These values are not influenced by outliers or non-normality of the error distribution.

The NMAD value and the 68.3% quantile agree well with each other. In comparison with the quality measures when a normal distribution is assumed, considerable differences can be observed. The sample mean (13 cm) and median (2 cm) differ by 11 cm while the standard deviation (105 cm) and the 68.3% quantile (13 cm) differ by 92 cm. Even the standard deviation after removal of outliers (34 cm) differs substantially from the 68.3% quantile (13 cm), which illustrates the problems of a 3·RMSE based removal of blunders in case the underlying distribution differs substantially from a normal distribution.

A spatial visualisation of the deviations may give hints where outliers have occurred.

The plot in Fig. 7 shows the spatial distribution of the differences between the DTM derived by laser scanning including filtering and the DTM/DSM derived by photogrammetry. Elevations for laser points were derived through bilinear interpolation with the dense net of photogrammetrically derived points. Blunders are concentrated especially at the top right corner of the figure. The large amount of blunders in the data is due to the fact that **no** editing of the photogrammetrically derived DEM occurred.

Finally, Fig. 8 shows a contour plot of the 95% quantile of the absolute error distribution as a function of location using the Koenker and Mizera (2004) approach, i.e. a quantile regression using $Q(0.95) = f(x, y)$. For computational reasons the plot is only based on a subset of 12,000 points. The plot provides an overview for direct accuracy checking using quantiles, which better takes autocorrelation into account.

7.4. Summarizing results with the proposed accuracy measures

Several examples of DEM quality control were presented with different numbers of outliers and checkpoints. The presented histograms revealed skewness and kurtosis, which should be taken into account when deriving accuracy measures. In order to derive reliable values for the systematic error and the standard deviation three different approaches have been used: Estimation using all data, blunders removed using a RMSE based threshold, and estimation of location and scale using a robust method. Fig. 9 depicts the differences between the different approaches for the example of Section 7.3.

From the graph it is obvious that the robust approach fits the histogram best. The removal of outliers by a threshold ($T \geq 3 \cdot \text{RMSE}$) does not eliminate all of the outliers. Therefore, the use of

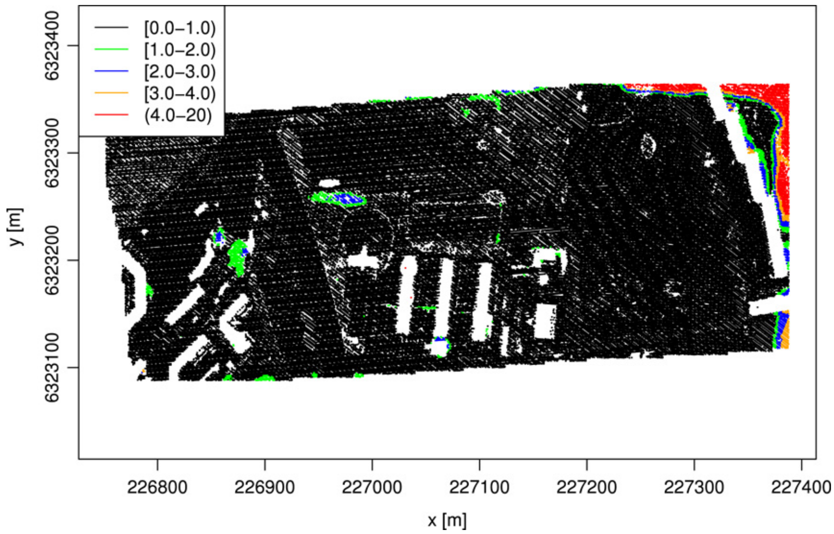

Fig. 7. Spatial distribution of differences between elevations from laser scanning and automated photogrammetry. The coloured areas have differences above 1 m (≈3 times standard deviation). The white areas are large buildings and elevations have been removed there by a filter program.

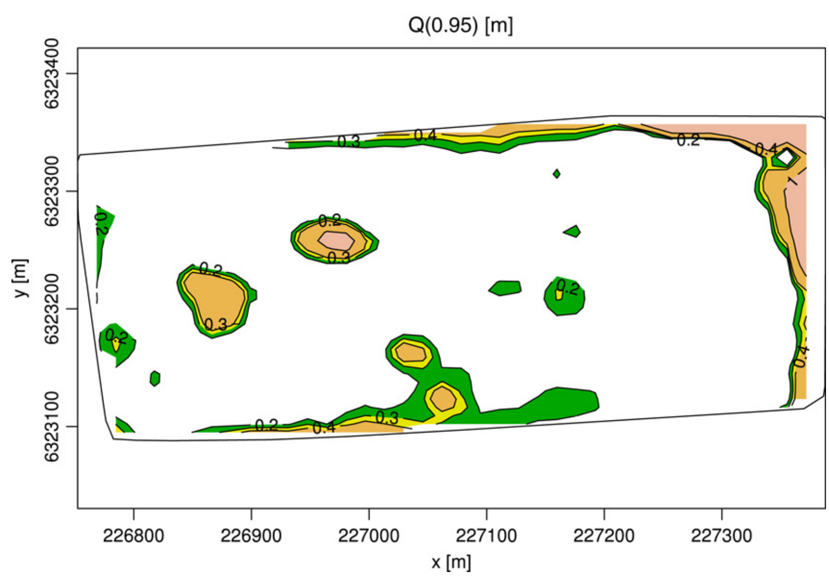

Fig. 8. Contour plot of the triogram surface describing the 95% quantile of the absolute error distribution. The iso-lines illustrate how the 95% quantile differs as a function of measurement location.

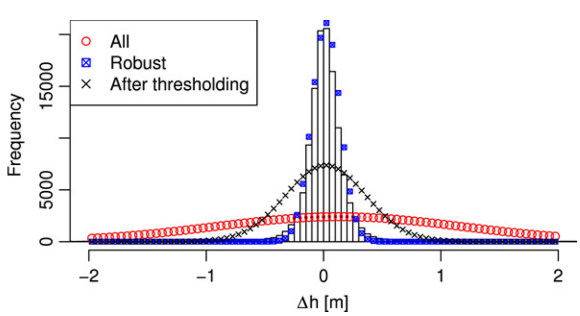

Fig. 9. Histogram of the differences Δh between two DEMs in metres truncated to the range $[-2\text{ m}, 2\text{ m}]$. Superimposed on the histogram are the corresponding normal distribution curves when estimating parameters 'mean' and 'variance' through one of the three approaches.

new quality measures (median, NMAD) is more adequate for DEMs derived by means of digital photogrammetry or laser scanning. A distribution-free and non-parametric alternative is the use of quantiles, therefore, we suggest computing the 68.3% and 95% quantile of the absolute errors additionally. The use of histogram and Q–Q plot provide a visual tool to help decide which accuracy measures are more appropriate for the tested DEM.

8. Discussion

The accuracy measures (systematic shift and standard deviation) should not be influenced from outliers and non-normality of the error distribution. Therefore, we suggest applying robust statistical methods in the assessment of DEM accuracy. Confidence intervals for the various quantiles are only small if the number of checkpoints is large. It is possible to treat the issue of sample size within a statistical context. Guidelines on how large the sample should be are given in Section 7.

Quality control by means of visual inspection and photogrammetric measurements has to detect as many outliers as possible and to remove them. Stereo measurements and other editing by an operator may have to be added. The DTM is then cleaner, but some of the outliers may remain undetected. Use of robust methods is, therefore, highly recommended for DEMs derived by digital photogrammetry or laser scanning.

Our method for accuracy assessment can be summarized as follows: Compute vertical errors with all points in the sample. Then generate histograms and Q–Q plots to visualize the error distribution and to assess non-normality. Thereafter, compute mean error and standard deviation as well as median and NMAD together with confidence intervals. In case of big discrepancies assess whether outliers in the data are an issue. Also compute the 68.3% quantile and compare it with the NMAD value. In case of discrepancy decide (based on the histogram and Q–Q plot) if non-normality is an issue. If non-normality is an issue use the more robust and conservative quantile measures supplemented by a quantile surface plot. In case compliance with a specification has to be proven, use the appropriate statistical test procedure as described in Section 6.

The proposed methodology adapts to the specialities of laser scanning and digital photogrammetry, where blunders and non-normal distribution are often present, especially in non-open terrain. DEM standards have to take this into account and non-parametric and distribution free methods should be calculated in the assessment of accuracy.

Acknowledgements

We thank the Danish company COWI A/S for providing image and laser data and the students M. Smith and S. Borkmann of Aalborg University for reference data determined by ground surveying. Three anonymous reviewers have provided useful comments and suggestions improving the article. Birthe Nørskov, Aalborg University, is thanked for improving the English language.

References

ASPRS Lidar Committee. 2004. ASPRS Guidelines Vertical Accuracy Reporting for Lidar Data, http://www.asprs.org/society/committees/lidar/Downloads/Vertical_Accuracy_Reporting_for_Lidar_Data.pdf (accessed 28.01.2009) p. 20.

Aguilar, F., Agüera, F., Aguilar, A., 2007a. A theoretical approach to modeling the accuracy assessment of digital elevation models. Photogrammetric Engineering & Remote Sensing 73 (12), 1367–1379.

Aguilar, F., Aguilar, M., Agüera, F., 2007b. Accuracy assessment of digital elevation models using a non-parametric approach. International Journal of Geographical Information Science 21 (6), 667–686.

Atkinson, A.D.J., Ariza López, F.J., García-Balboa. 2005. Positional accuracy control using robust estimators. In: Proceedings of the 21st International Cartographic Conference, 09–16 July, Acoruña, Spain. Available from: http://www.cartesia.org/articulo206.html (accessed 28.01.2009).

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. International Archives of Photogrammetry and Remote Sensing 33 (Part B4/1), 110–117.

Carlisle, B.H., 2005. Modelling the spatial distribution of DEM error. Transactions in GIS 9 (4), 521–540.

Daniel, C., Tennant, K., 2001. DEM quality assessment. In: Maune, D.F. (Ed.), Digital Elevation Model Technologies and Applications: The DEM User Manual, 1st ed. ISBN: 1-57083-064-9, pp. 395–440.

D'Agostino, R.B., Belanger, A., D'Agostino Jr., R.B., 1990. A suggestion for using powerful and informative tests of normality. The American Statistician 44 (4), 316–321.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press.

Desu, M.M., Raghavarao, D., 1990. Sample Size Methodology. Academic Press, Inc.

Desu, M.M., Raghavarao, D., 2003. Nonparametric Statistical Methods for Complete and Censored Data. Chapman Hall/CRC.

Fisher, P.F., Tate, N.J., 2007. Causes and consequences of error in digital elevation models. Progress in Physical Geography 30 (4), 467–489.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1983. Understanding Robust and Exploratory Data Analysis. John Wiley & Sons, Inc.

Höhle, J., Potuckova, M., 2006. The EuroSDR test checking and improving of digital terrain models. In: EuroSDR Official Publication no. 51. ISBN: 9789051794915, pp. 10–55.

Hyndman, R.J., Fan, Y., 1996. Sample quantiles in statistical packages. The American Statistician 50 (4), 361–365.

Koenker, R., 2005. Quantile Regression. Economic Society Monographs. Cambridge University Press.

Koenker, R., Mizera, I., 2004. Penalized triograms: Total variation regularization for bivariate smoothing. Journal of the Royal Statistical Society. Series B 66 (1), 145–163.

Lahiri, S.N., 2003. Resampling Methods for Dependent Data. Springer.

Li, Z., Zhu, Q., Gold, C., 2005. Digital Terrain Modeling – Principles and Methodology. CRC Press, ISBN: 0-415-32462-9.

Maune, D.F. (Ed.), 2007. Digital Elevation Model Technologies and Applications: The DEM User Manual, 2nd ed. ISBN: 1-57083-082-7.

Mikhail, E., Ackermann, F., 1976. Observations and Least Squares. IEP — A Dun-Donnelley Publisher, New York, USA, ISBN: 0-7002-2481-5.

Moors, J.J.A., 1988. A quantile alternative for kurtosis. The Statistician 37 (1), 25–32.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical computing, Vienna, Austria. ISBN: 3-900051-07-0. http://www.R-project.org (accessed 28.01.2009).

Zandbergen, P.A., 2008. Positional accuracy of spatial data: Non-normal distributions and a critique of the national standard for spatial data accuracy. Transactions in GIS 12 (1), 103–130.