Characterizing the error distribution of lidar elevation data for North Carolina

Paul A. Zandbergen

* Department of Geography, University of New Mexico, Bandelier West Room 111, MSC01 1110, 1 University of New Mexico, Albuquerque, NM, USA

Online publication date: 06 February 2011
Characterizing the error distribution of lidar elevation data for North Carolina

PAUL A. ZANDBERGEN*
Department of Geography, University of New Mexico, Bandelier West Room 111, MSC01 1110, 1 University of New Mexico, Albuquerque, NM 87131, USA

(Received 3 September 2008; in final form 24 August 2009)

Spatial data quality is a paramount concern in all geographical information systems (GIS) applications. Existing standards and guidelines for spatial data commonly assume the positional error is normally distributed. While non-normal behaviour of the error in digital elevation data has been observed in previous research, current guidelines for digital elevation data still assume that the errors for observations in open terrain are normally distributed. This research employed an accuracy assessment dataset from a substantial lidar data collection effort, the North Carolina Floodplain Mapping Program. Strong evidence was found that the vertical error of lidar elevation data is not normally distributed and that both major and minor outliers are very common. Of the five land cover types considered, only the distribution for urban areas approximated a normal distribution, even though these observations were generally much less accurate than those for open terrain. No influence of slope on the occurrence of non-normal behaviour in the distributions was found. The RMSEz (root mean square error) statistic used to characterize the fundamental accuracy of digital elevation data was found to be very sensitive to the occurrence of outliers, questioning its use in current guidelines.

1. Introduction

1.1 Accuracy standards for digital elevation data

Digital elevation models (DEMs) are used in a wide range of applications and need to be of sufficient quality to meet the needs of these applications. Concerns for DEM quality issues are clearly expressed in the development of standards for acquisition and dissemination of digital elevation data.

The most established map accuracy standard in the USA is the National Map Accuracy Standard (NMAS) (US Geological Survey (USGS) 1999). NMAS was developed in 1947 (US Bureau of the Budget 1947) specifically for printed maps. It has lost some of its meaning in recent years for digital spatial data but it remains a widely employed standard. NMAS contains both a horizontal and vertical component. Vertical accuracy in NMAS is specified in terms of the contour interval at the 90th per cent confidence interval as follows: ‘Vertical accuracy, as applied to contour maps on all publication scales, shall be such that not more than 10 per cent of the elevations tested shall be in error more than one-half the contour interval. In checking elevations taken from the map, the apparent vertical error may be decreased by

*Email: zandberg@unm.edu
assuming a horizontal displacement within the permissible horizontal error for a map of that scale’ (US Bureau of the Budget 1947). This one-half contour interval is referred to as the Vertical Map Accuracy Standard (VMAS). Since NMAS applies to contours only, the standard is only indirectly relevant for DEMs. With the advent of digital spatial data NMAS became technically obsolete, since within a computerized environment the display scale is independent of the scale for which the data was created.

In 1998 the Federal Geographic Data Committee (FGDC) published the National Standard for Spatial Data Accuracy (NSSDA) (FGDC 1998), which superseded NMAS for digital mapping products. NSSDA contains both a horizontal and vertical component. The NSSDA implements a testing methodology for determining the positional accuracy of locations on maps and in digital spatial data relative to clearly defined georeferenced positions of higher accuracy. NSSDA also provides specific guidelines for what type of reference data to use and the minimum number of points to be used, as well as their spatial distribution (FGDC 1998).

One key assumption of the NSSDA is that the data do not contain any systematic errors and that the positional errors follow a normal distribution. Based on this assumption, the NSSDA specifies a 95% confidence interval using the calculation of the root mean square error (RMSE) between test locations and reference locations. Vertical accuracy in the NSSDA is defined as ‘the linear uncertainty value, such that the true or theoretical location of the point falls within +/- of that linear uncertainty value 95 per cent of the time’ (FGDC 1998). The vertical accuracy statistic is determined as $1.96 \times \text{RMSE}$ on the basis of a normal distribution of the error values. Vertical accuracy is defined as a linear error in the Z direction and therefore the annotation $\text{RMSE}_z$ is used.

It is important to note that the NSSDA uses a 95% confidence interval, but that this is derived from the calculation of the RMSE. The 1.96 factor for vertical accuracy is directly derived from observations of the normal distribution as described by Greenwalt and Schultz (1968). This assumes there are no systematic errors and no major outliers, and that the distribution of vertical errors follows a normal distribution. However, these assumptions have not undergone much testing and are not elaborated upon in the original FGDC documents on NSSDA. For a large sample the 95th percentile can obviously be determined directly from the distribution of error values, but the NSSDA protocol specifies a minimum of only 20 control points. For such a small sample deriving the 95th percentile from the distribution is not very reliable and the use of the RMSE statistic was developed as a more robust alternative.

The NSSDA has been widely adopted in other spatial data accuracy standards and guidelines for elevation data. For example, the guidelines of the American Society for Photogrammetry and Remote Sensing (ASPRS) for vertical accuracy reporting for lidar data (ASPRS 2004) and the guidelines of the National Digital Elevation Program (NDEP) for handling digital elevation data (NDEP 2004) both incorporate the NSSDA, in addition to comparisons with the older NMAS standards. Both the ASPRS and NDEP guidelines, however, recognize that the error in elevation data may not be normally distributed and that the use of the $\text{RMSE}_z$ statistic is not appropriate for non-normal distributions. For relatively small sample sizes even a single major outlier can skew the RMSE statistic substantially. As a result, both the ASPRS and NDEP guidelines recommend the use of the 95th percentile as an alternative when the error distribution is not normal. It should be noted that in this case
the 95th percentile should be derived directly from the error distribution and not from the RMSE statistic.

The ASPRS and NDEP guidelines also recognize that the accuracy of high resolution elevation data varies with land cover type, and therefore recommend that accuracy testing consider several land cover types. This has given rise to the use of several types of vertical accuracy:

- **Fundamental vertical accuracy**: this is the vertical accuracy of elevation data as determined using checkpoints in open terrain. This is assumed to be the ‘best-case’ vertical accuracy since the potential influence of vegetation and buildings is minimal. The distribution of fundamental vertical accuracy is assumed to be normal, and it is recommended to use the RMSE\(_z\) statistic on 100% of the checkpoints to determine the 95th percentile according to the NSSDA protocol.
- **Supplemental vertical accuracy**: this is the vertical accuracy of elevation data in land cover categories other than open terrain, and may apply to a single land cover category or a combination of cover categories.
- **Consolidated vertical accuracy**: this is the vertical accuracy for all land cover categories combined, including open and other types.

The error distributions of supplemental and consolidated vertical accuracy are not expected to always follow a normal distribution and therefore a non-parametric statistic is recommended: the 95th percentile derived from the error distribution.

The National Digital Elevation Program (NDEP) has submitted a proposal to the FGDC to revise the NSSDA by incorporating the distinction between fundamental, supplemental and consolidated vertical accuracy, including the use of the 95th percentile derived from the error distribution as an alternative statistic for non-normal distributions (NDEP 2003). The use of the RMSE statistic, however, is still recommended for fundamental vertical accuracy.

While the occurrence of non-normal distributions in high-resolution elevation data is now widely recognized, most published elevation data accuracy assessments report the RMSE\(_z\) statistic without much consideration of the underlying distribution. One recent exception is the study by Oksanen and Sarjakoski (2006) which determined the vertical error in a high resolution DEM. Results indicate that the error distribution was non-normal in nature and could not be characterized with a single estimator of dispersion. Similar evidence of non-normality of the error distribution has been found by other studies (Fisher 1998, López 2000, Bonin and Rousseaux 2005).

The research on vertical error in DEMs suggests a number of explanations for the non-normality of the error distribution: (1) the frequent occurrence of gross errors (or blunders), in particular when data is interpolated from contours; (2) large positive spatial autocorrelation of the vertical error in DEMs and (3) non-stationary processes underlying the occurrence of vertical errors in DEMs. One critical source of non-stationary behaviour in DEMs obtained through Light Detection and Ranging (lidar) is the variability of vertical error with land cover (Hodgson and Bresnahan 2004, Hodgson et al. 2005).

### 1.2 Lidar elevation data

Light Detection and Ranging (lidar) was developed in the 1980s and has become a widely employed method for creating large-scale accurate DEMs. Several efforts are underway in the USA to collect lidar data statewide, primarily for floodplain
mapping. Overviews of the lidar system for terrain mapping may be found in Shan and Toth (2008) and Fowler (2001) and numerous research articles. Here, only some important elements in the lidar collection and processing approach are discussed as they relate to accuracy. In general, the accuracy of DEMs derived from lidar postings is affected by several factors: sensor, aircraft platform, navigation, lidar point processing and geographic environment.

The lidar sensor is mounted in a fixed-wing aircraft or helicopter. The Global Positioning System (GPS) and Inertial Navigation System (INS) are used to determine the position of the sensor and pointing direction of the laser. The \(x\), \(y\) and \(z\) position of the object the laser pulse is reflected from is determined by modelling the round-trip return time and position/pointing direction of the laser. The accuracy of the GPS, the INS, the pulse length, and the footprint size of the laser projected on the ground in combination set the technological limits of lidar data collection.

The emitted pulse interacts with numerous features encountered in the landscape. Some of the energy reflects from airborne objects but most of the energy reflects off vegetation, buildings, or the ground. The last identified return pulse may be the best estimate of the ground, often assumed to be bare earth. However, even the last return may not be the ground as vegetation canopy may totally obscure the ground. The last return may also be reflected from buildings and other large objects. These non-ground points can be identified using a combination of automated procedures and manual editing; this is referred to as point labelling.

The lidar point labelling process to remove vegetation and buildings remains an area of very active research (Lohmann et al. 2000, Cobby et al. 2001, Raber et al. 2002, Zhang et al. 2004, Shan and Sampath 2005, Sithole 2005, Han et al. 2007). Despite this progress in developing improved algorithms, many published studies on lidar elevation accuracy do not include documentation of the point labelling process used since most commercial vendors regard this as proprietary.

1.3 Accuracy of lidar-derived elevation

During the initial years of lidar mapping efforts most companies would quote accuracies of 15 cm RMSE, but most research suggests that this is only achievable under the most ideal conditions, for example, low altitude collections, flat terrain, minimal or no surface vegetation or obstructions and extensive human analysis in the point labelling process. In a controlled study of lidar performance over the Greenland ice sheet, Krabill et al. (2002) predicted an 8.6 cm RMSE theoretical error and documented a 5.4 cm RMSE error. More recent empirical studies suggest that accuracies of 25 cm or larger are more realistic for large-scale mapping applications in varied terrain. For example, Adams and Chandler (2002) measured lidar-derived elevation accuracy as 26 cm RMSE. They also found a slight bias in the lidar data, that is, a tendency to underpredict elevation. Bowen and Waltermine (2002) assessed the accuracy of lidar derived elevation data and how accuracy varied by topography. They found an overall accuracy of 43 cm RMSE for leaf-off conditions in the South Carolina piedmont. Hodgson and Bresnahan (2004) quantified the contribution of error from the lidar system, interpolation algorithm, terrain slope, land cover, and reference data and found RMSE values between 17.2 cm and 25.9 cm. Hodgson et al. (2003) analysed North Carolina lidar data collected during leaf-on conditions and compared the results to Interferometric Synthetic Aperture Radar (IFSAR)-derived DEMs and USGS Level 1 and 2 DEMs. Elevation error for the lidar-derived elevation data varied by land cover category and
ranged from 33 cm (low grass) to 153 cm (scrub/shrub). While this research has established variability in lidar accuracy by land cover, at present the body of research documenting accuracy estimates for specific land cover categories is limited.

In addition to land cover, a well-known characteristic of observed elevation error for terrain mapping is the relationship with terrain slope (Maling 1989). Even if the elevation of a surface observation is measured without error, the horizontal error in the observation may introduce error due to the displacement of the centre of the returned laser pulse. The maximum amount of elevation error introduced is a function of surface slope but the maximum error will only occur if the displacement is perpendicular to the contour line. However, additional error may occur in directions other than perpendicular, depending on the terrain.

1.4 Study objective

The non-normal distribution of positional error observed in many lidar elevation datasets presents a challenge to current map accuracy standards and error propagation modelling techniques which rely on assumptions of normality and utilize simple statistics to characterize its distribution, such as $\text{RMSE}_z$. The objective of this study, therefore, is threefold: (1) to develop a more rigorous characterization of the distribution of positional errors in lidar elevation data; (2) to examine the variability of this distribution with land cover type and terrain complexity; (3) to explore the sensitivity of the $\text{RMSE}_z$ statistic to variations in the underlying error distribution.

To allow for a proper characterization of the vertical error distribution of lidar elevation data, this research employs a dataset of a vertical accuracy assessment of a substantial lidar data collection effort, the North Carolina Flood Mapping Program.

2. Methods

2.1 Data collection

Lidar accuracy assessment data was obtained from the North Carolina Flood Mapping Program. Lidar data collection for the entire State of North Carolina has been undertaken since 2002. Data processing has been completed for the entire State while accuracy assessments have been completed for approximately 80% of the State, covering a total area of 109,000 km$^2$. Details on the collection, processing and accuracy assessment of the lidar data are provided in a series of Issue Papers produced by the North Carolina Flood Mapping Program (2006). A brief summary follows. The original lidar data were collected with a ground spacing of sampling points of approximately 3 m. To produce the bare earth elevation points a combination of manual and automated cleaning techniques were employed. These post-processing techniques included the use of automated procedures to detect elevation changes that appeared unnatural to remove buildings, as well as the use of last returns to remove vegetation canopy. The actual algorithms used in this stage of the post-processing are not disclosed since they are considered proprietary by the contractors who collected and processed the lidar data.

Initially two commercial vendors were contracted to obtain the lidar data. During the first phase of the data collection effort a number of study sites were identified where overlapping data occurred between contractors. This examination identified a number of inconsistencies between vendors and as a result only a single vendor was contracted for the remainder of the data collection. While the extent of the area
collected by the second vendor in the first phase is not known, the total area is relatively small compared to the total study area. The exact algorithms employed by the two vendors, and any changes in these algorithms during the course of the data collection, are not documented.

The final elevation data is distributed by the North Carolina Floodplain Mapping Program as bare earth lidar elevation points \((x, y, z)\) as well as a 6.1 m and 15.2 m bare earth DEM in raster format. Only the bare earth elevation points were used in the accuracy assessment in this study.

The original specifications for the collection of the lidar data state a vertical accuracy requirement of 25 cm for all the inland counties and 20 cm for all coastal counties. This accuracy requirement is based on the RMSE\(_z\) value. As part of the data collection, the North Carolina Flood Mapping Program also completed a county-by-county accuracy assessment of the lidar data. While the accuracy assessment was completed by independent survey contractors, the efforts were directed by the North Carolina Geodetic Survey (NGS). Specific requirements for the completion of the vertical accuracy assessment are documented in Issue Paper 37 from the North Carolina Flood Mapping Program (2004). A brief summary follows. NGS classified all land within the study area into one of five major land cover categories: (1) open terrain (e.g. bare-earth, sand, rock, ploughed fields, short grass, golf courses); (2) high grass weeds and crops (e.g. hay, corn, wheat, tobacco); (3) brush lands and low trees; (4) fully forested and (5) urban areas (vicinity of manmade structures, high density). Within each county, 20 checkpoints were selected within each of the five land cover categories, with the exception of forested areas where 40 checkpoints were selected. Each checkpoint was located on flat or uniformly sloped terrain within 5 m in all directions (i.e. no break-lines). At each checkpoint a survey was conducted using a combination of traditional surveying and real-time kinematic GPS to achieve a target vertical accuracy of 5 cm or better according to the National Oceanic and Atmospheric Administration (NOAA) Technical Memorandum NOS NGS-58.

The elevation obtained at each checkpoint was compared to the interpolated elevation from a triangulated irregular network (TIN) elevation model based on the bare-earth lidar elevation points. Vertical accuracy at each surveyed location was reported in feet or metres as a positive or negative value with three significant digits. The checkpoints, therefore, do not represent a direct accuracy assessment of the original bare earth elevation points, but instead the error in the TIN model derived from these elevation points. While the final elevation data are distributed as a gridded DEM, a TIN model was used in the accuracy assessment to limit the potential effect of the interpolation techniques employed to create the DEMs.

For the purpose of this study, all available individual survey reports were obtained, including the land cover type and vertical accuracy assessment. The locations of all the check points were plotted in the North Carolina State Plane coordinate system using the easting and northing reported in the survey reports. Figure 1 shows the locations of all the check points \((n = 10\,952)\). The check points cover most counties in North Carolina, with the exception of the westernmost counties, for which the accuracy evaluation has not yet been released. The check points are distributed over a total area of 109 000 km\(^2\) for which lidar elevation data was collected. Within a single county some degree of local clustering can be observed. This reflects the nature of the collection of the survey data: within a county a number of square study areas of several square miles were delineated by the North Carolina Geodetic Survey, and independent survey contractors were tasked to select their check points within these predefined areas.
Individual survey locations are typically not closer together than several hundred metres, but still appear as highly clustered at the scale shown in figure 1.

The original vertical accuracy requirements of the lidar data collected have changed since the start of the North Carolina Flood Mapping Program. The original specifications required a vertical accuracy of 25 cm for all the inland counties and 20 cm for all coastal counties. This requirement was initially based on the RMSEz value for all data points within a single county. However, initial reports confirmed the occurrence of a small number of major outliers which skewed the RMSEz values, and as a result the accuracy requirements were revised to using the RMSEz statistic after removal of 5% outliers. When NDEP released their guidelines in 2004 for reporting the vertical accuracy of digital elevation data, these guidelines were adopted. This includes a fundamental vertical accuracy based on checkpoints in open terrain using the NSSDA protocol (RMSEz for all checkpoints × 1.96), a supplemental vertical accuracy based on the 95th percentile for all other land cover types, and a consolidated vertical accuracy based on the 95th percentile for all check points combined. The new vertical accuracy requirement under these guidelines is that the RMSEz for fundamental accuracy does not exceed 18.5 cm for all counties. The revisions of the vertical accuracy requirements of the North Carolina Flood Mapping Program are interesting in the context of the changing standards for digital elevation data. However, this study has employed the original data for the individual checkpoints and the results are therefore not influenced by the changes in reporting requirements for the lidar data.

2.2 Characterization of the vertical error distribution

Since the vertical accuracy of the lidar data is expected to vary with land cover type, all analysis of the original check points was completed separately for each of the five land cover types: (1) open terrain (e.g. bare-earth, sand, rock, ploughed fields, short grass, golf courses); (2) high grass weeds and crops (e.g. hay, corn, wheat, tobacco); (3) brush
lands and low trees; (4) fully forested and (5) urban areas (vicinity of manmade structures, high density).

The positional error distributions were characterized in a number of ways to determine the degree to which they follow a normal distribution. First, basic descriptive parameters of the distributions were derived, including mean, median, minimum, maximum, standard deviation, inter-quartile range and percentiles (5, 10, 90 and 95). Second, the degree of normality of the distribution was determined using skewness and kurtosis, in addition to standard normality tests, including the Kolmogorov–Smirnov test using Lilliefors correction (Lilliefors 1967) and the Shapiro–Wilk test (Shapiro and Wilk 1965). Third, normal Q–Q plots were created and compared to the normal curve. A normal Q–Q plot is a graphical method for diagnosing differences between the probability distribution of a statistical population and a theoretical normal distribution. The normal distribution is represented by a straight line and deviations from linearity indicate non-normal behaviour. Normal Q–Q plots provide a visual exploratory tool to examine normality of the distribution and to identify the nature of any deviations.

In addition to land cover, a second variable of potential influence on the vertical accuracy of the lidar data is local slope variability. While limitations were placed on the location of survey locations (flat or uniformly sloped within 5 m of the location), the slope of the immediately surrounding may have introduced some error in the lidar data. Figure 2 shows the slope grid for all of North Carolina. For display purposes a resampled DEM with a resolution of 46 m was used, which also combines the original lidar data with USGS 30 m data for the westernmost counties for which lidar data is not yet available. As illustrated in figure 2, coastal North Carolina is dominated by
very little relief, and the average slope for coastal counties is typically less than 3°. For inland counties, average slopes are typically between 5° and 10°, with steeper average slopes encountered in the most western counties.

For each surveyed location, the corresponding 6.1 m lidar elevation grid was obtained from the North Carolina Flood Mapping Program online data distribution website. The original American Standard Code for Information Interchange (ASCII) files were converted to an Environmental Systems Research Institute (ESRI) grid format using ArcGIS (ESRI, Redlands, CA, USA). Slope grids were derived using Horne’s (1981) method. Slope at each survey location was determined through a point-in-raster overlay between the check points and the slope grid. The algorithm used to derive the slope grid uses a 3 by 3 neighbourhood around each cell to determine the slope at the cell’s location. As a result, the derived slope is slightly different than the slope derived from a TIN based on the original bare-earth points. Nevertheless, the obtained values for slope are a meaningful estimate of the local slope variability.

For each land cover type, the distribution of the slope values at all check points was determined. A quintile method was used to split the sample for each land cover type into five categories from lowest to highest slope. Normality testing was performed on the 25 resulting distributions and normal Q–Q plots were created to determine the effect of slope on the distributions.

2.3 Spatial distribution of RMSE<sub>z</sub> value and outliers

Spatial variability in the accuracy of the lidar data were determined in two ways: (1) determining the RMSE<sub>z</sub> values by county; and (2) determining the percentage of outliers by county. For the first approach all the survey locations were pooled by county irrespective of land cover type. The RMSE<sub>z</sub> statistic for each county was determined after 5% outlier removal. This reflects the accuracy reporting employed in the first phase of the North Carolina Flood Mapping Program. The spatial pattern in the RMSE<sub>z</sub> values was investigated to determine the presence of specific counties with unusually high RMSE<sub>z</sub> values. For the second approach all the survey locations were pooled by land cover type. Outliers within each land cover type were identified as falling below the 5th percentile or above the 95th percentile of the error distribution for that land cover type. Subsequently for each county the number of outliers was determined as a percentage of all observations within that county.

2.4 Effect of outliers on RMSE<sub>z</sub> statistic

Finally, the sensitivity of the RMSE<sub>z</sub> statistics to the presence of outliers was determined. For each land cover type, all observations were ranked based on the absolute value of the vertical error. Then the RMSE<sub>z</sub> statistics was determined for all data points. After removal of the worst outlier, the RMSE<sub>z</sub> statistics of the remaining data points was determined again. This was repeated until 25% of the points were removed. Then for each land cover type the resulting RMSE<sub>z</sub> values were plotted against the percentile of the distribution used in the calculation.

3. Results

3.1 Description of vertical error distributions

Table 1 provides summary descriptive statistics for the error distribution of lidar elevation data by land cover type. The sample size for the fully forested category is
approximately twice as large as the other ones, based on the sampling design implemented for the accuracy assessment. The mean value is slightly negative for all distributions, suggesting a small bias towards an underestimate of the true elevation. The largest underestimate is 5.1 cm for the urban areas. The values for the median are very similar to the mean, providing no early indication the distributions are not normal. The values for the minimum and maximum of the distribution are very high, suggesting the occurrence of both positive and negative outliers. The range is largest for fully forested (8.197 m), followed by bare earth (6.205 m), while the ranges for the other three categories do not exceed 3 m.

Table 2 provides the percentiles for the error distributions. The 5th and 95th percentiles are of particular interest, since they provide a first measure to compare the accuracy of the lidar data by land cover type. The 5th and 95th percentiles for the bare earth and high grass categories are similar and lowest of the five land cover types, suggesting the lidar data for these types is most accurate. Percentiles for the other three types are considerably higher, with the exception of the 95th percentile for urban areas. The absolute value of the 5th percentile is also consistently larger than the 95th percentile for all land cover types, confirming the slight bias towards an underestimation of elevation already observed previously.

### 3.2 Normality testing

The first test of normality is provided by the values for skewness and kurtosis in table 3. The standard error for these two parameters varies slightly by land cover type.
due to differences in sample size. Values for skewness are strongly negative for bare earth and fully forested and positive for high grass and brush lands, suggesting asymmetrical behaviour for these land cover categories. For urban areas the value for skewness falls below the standard error, suggesting normal symmetrical behaviour. Values for kurtosis are positive and much larger than the standard error for all land cover types, suggesting long tails. The largest value for kurtosis is for bare earth, followed by fully forested. The smallest value is for urban areas.

Based on the observed values for skewness and kurtosis, the distributions for bare earth and fully forested deviate the most from a normal distribution, while the distribution for urban areas most resembles a normal distribution.

The next approach to characterize any deviation from the normal distribution is presented in figure 3 in the form of normal Q–Q plots. For each land cover type the empirical distribution is compared to a normal curve with the same mean and standard deviation. The first general observation is that a relatively small number of both positive and negative major outliers can be observed for all five land cover types. The second observation is that the normal Q–Q plots reveal a slightly sigmoid shape of the empirical distribution compared to the straight line of the normal curve. For urban areas, the empirical distribution is much closer to the straight line. This second observation indicates that the empirical error distributions not only have a number of major outliers, but also a substantial number of minor outliers, resulting in strong non-normal behaviour.

Statistical normality testing presents the third and final approach to characterize any non-normal behaviour. The results of the Kolmogorov–Smirnov and Shapiro–Wilk tests are shown in table 4. The only distribution for which the null hypothesis that the distribution is normal is not rejected is for the error distribution of the urban areas based on the KS test alone. This confirms the deviation from non-normal behaviour for the error distributions for all land cover types except urban areas.

3.3 Effect of slope

Slope is known to be a factor in the error of lidar elevation data (e.g. Peng and Shi 2006, Su and Bork 2006). Since the horizontal error in lidar is often much larger than the vertical error, the effect of any horizontal displacement on the vertical accuracy assessment will be much greater for areas of steeper slopes. It is therefore expected that the vertical error of lidar elevation data and the occurrence of major outliers will increase with steeper slopes.
Table 5 reports the 5th percentile, inter-quartile range (IQR) and 95th percentile for the error distribution for each of the five slope categories considered based on a quintile distribution. A strong effect of slope would result in increasing values for the (absolute values of) the percentiles and IQR. Table 5 reveals no such pattern for any of the land cover types. When considering each land cover type individually, there is no trend towards increasing or decreasing values for percentiles and IQR with increasing slope. When considering only the lowest and highest slope quintiles, the values for the percentiles and IQR do not reveal any consistent differences. This indicates there is no

Figure 3. Normal Q–Q plots of error in lidar elevation data by land cover type: (a) bare earth and low grass; (b) high grass, weeds and crops; (c) brush lands and low trees; (d) fully forested; (e) urban areas.
effect of slope on the error distributions. This can be explained by the relative modest slopes within the study area and the selection of field survey locations in areas of moderate slope: the break values between the 4th and 5th quintile are around 2.5–3° and only a small number of observations have higher slope values.

Further analysis of the error distributions by slope category is provided in the form of normal Q–Q plots in figure 4. A small number of major outliers is observed in nearly all distributions and there are no apparent trends by slope category in this regard. Some of the most notable distributions are those for the bare earth observations. These distributions are expected to be normal since they represent

<table>
<thead>
<tr>
<th>Land cover</th>
<th>Slope interval</th>
<th>5th percentile</th>
<th>IQR</th>
<th>95th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bare earth and low grass</td>
<td>0.00–2.7</td>
<td>−0.274</td>
<td>0.180</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>0.27–0.63</td>
<td>−0.298</td>
<td>0.207</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>0.63–1.34</td>
<td>−0.245</td>
<td>0.158</td>
<td>0.178</td>
</tr>
<tr>
<td></td>
<td>1.34–2.49</td>
<td>−0.266</td>
<td>0.150</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td>2.49–21.72</td>
<td>−0.215</td>
<td>0.155</td>
<td>0.167</td>
</tr>
<tr>
<td>High grass, weeds and crops</td>
<td>0.00–0.27</td>
<td>−0.236</td>
<td>0.185</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>0.27–0.73</td>
<td>−0.281</td>
<td>0.177</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td>0.73–1.54</td>
<td>−0.266</td>
<td>0.161</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td>1.54–2.85</td>
<td>−0.243</td>
<td>0.142</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td>2.85–23.83</td>
<td>−0.203</td>
<td>0.145</td>
<td>0.252</td>
</tr>
<tr>
<td>Brush lands and low trees</td>
<td>0.00–0.27</td>
<td>−0.318</td>
<td>0.226</td>
<td>0.272</td>
</tr>
<tr>
<td></td>
<td>0.27–0.77</td>
<td>−0.352</td>
<td>0.248</td>
<td>0.370</td>
</tr>
<tr>
<td></td>
<td>0.77–1.67</td>
<td>−0.387</td>
<td>0.209</td>
<td>0.345</td>
</tr>
<tr>
<td></td>
<td>1.67–2.85</td>
<td>−0.323</td>
<td>0.209</td>
<td>0.268</td>
</tr>
<tr>
<td></td>
<td>2.85–21.56</td>
<td>−0.265</td>
<td>0.212</td>
<td>0.315</td>
</tr>
<tr>
<td>Fully forested</td>
<td>0.00–0.27</td>
<td>−0.371</td>
<td>0.252</td>
<td>0.364</td>
</tr>
<tr>
<td></td>
<td>0.27–0.75</td>
<td>−0.347</td>
<td>0.277</td>
<td>0.370</td>
</tr>
<tr>
<td></td>
<td>0.75–1.71</td>
<td>−0.325</td>
<td>0.226</td>
<td>0.292</td>
</tr>
<tr>
<td></td>
<td>1.71–3.24</td>
<td>−0.312</td>
<td>0.213</td>
<td>0.275</td>
</tr>
<tr>
<td></td>
<td>3.24–24.45</td>
<td>−0.357</td>
<td>0.246</td>
<td>0.296</td>
</tr>
<tr>
<td>Urban areas</td>
<td>0.00–0.27</td>
<td>−0.237</td>
<td>0.186</td>
<td>0.271</td>
</tr>
<tr>
<td></td>
<td>0.27–0.67</td>
<td>−0.279</td>
<td>0.202</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>0.67–1.29</td>
<td>−0.287</td>
<td>0.235</td>
<td>0.216</td>
</tr>
<tr>
<td></td>
<td>1.29–2.40</td>
<td>−0.362</td>
<td>0.243</td>
<td>0.223</td>
</tr>
<tr>
<td></td>
<td>2.40–20.97</td>
<td>−0.368</td>
<td>0.196</td>
<td>0.126</td>
</tr>
</tbody>
</table>

\(^{v}\)Lilliefors significance correction.

Table 4. Normality testing of the error distribution of lidar elevation data by land cover type.

<table>
<thead>
<tr>
<th>Land cover</th>
<th>Kolmogorov–Smirnov(^v)</th>
<th>Shapiro–Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bare earth and low grass</td>
<td>0.090 0.000</td>
<td>0.716 0.000</td>
</tr>
<tr>
<td>High grass, weeds and crops</td>
<td>0.069 0.000</td>
<td>0.897 0.000</td>
</tr>
<tr>
<td>Brush lands and low trees</td>
<td>0.063 0.000</td>
<td>0.944 0.000</td>
</tr>
<tr>
<td>Fully forested</td>
<td>0.101 0.000</td>
<td>0.737 0.000</td>
</tr>
<tr>
<td>Urban areas</td>
<td>0.024 0.021</td>
<td>0.976 0.000</td>
</tr>
</tbody>
</table>

Table 5. Effect of slope on the error distribution of lidar elevation data by land cover type.
Figure 4. Normal Q–Q plots of error in LIDAR elevation data by land cover type and slope category.
the fundamental accuracy as defined by the NDEP and ASPRS guidelines. Surprisingly, the distributions for the lower slope categories deviate most from the normal curve, while those for the 4th and 5th slope quintiles approximate the straight line of the normal curve very closely. The distributions for the urban areas are also of note, since they consistently approximate the normal curve across all slope categories. Very few of the other distributions approximate the normal curve.

The general patterns observed in figure 4 are confirmed with formal normality testing using the Kolmogorov–Smirnov and Shapiro–Wilk test, the results of which are shown in table 6. Shapiro–Wilk is the more powerful of the two tests, even with Lilliefors modification of the Kolmogorov–Smirnov test. Of the 25 distributions tested, the null hypothesis that the distribution is normal is rejected 19 times using the Kolmogorov–Smirnov test and 22 times using the Shapiro–Wilk test. Based on the more powerful Shapiro–Wilk test the distributions that appear normal include the 5th slope quintiles of the bare earth and the 4th and 5th slope quintiles of the urban areas.

The results of the analysis of the normal Q–Q plots and the normality testing provide strong evidence of the lack of the influence of slope on the error distribution.

Table 6. Normality testing of the error distribution of lidar elevation data by land cover type and slope category.

<table>
<thead>
<tr>
<th>Land cover</th>
<th>Kolmogorov–Smirnov</th>
<th>Shapiro–Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>df</td>
</tr>
<tr>
<td>Bare earth and low grass</td>
<td>0.00–2.7</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>0.27–0.63</td>
<td>0.144</td>
</tr>
<tr>
<td></td>
<td>0.63–1.34</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>1.34–2.49</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>2.49–21.72</td>
<td>0.036</td>
</tr>
<tr>
<td>High grass, weeds and crops</td>
<td>0.00–0.27</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>0.27–0.73</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.73–1.54</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>1.54–2.85</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td>2.85–23.83</td>
<td>0.088</td>
</tr>
<tr>
<td>Brush lands and low trees</td>
<td>0.00–0.27</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>0.27–0.77</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>0.77–1.67</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>1.67–2.85</td>
<td>0.091</td>
</tr>
<tr>
<td></td>
<td>2.85–21.56</td>
<td>0.075</td>
</tr>
<tr>
<td>Fully forested</td>
<td>0.00–0.27</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>0.27–0.75</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td>0.75–1.71</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>1.71–3.24</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>3.24–24.45</td>
<td>0.102</td>
</tr>
<tr>
<td>Urban areas</td>
<td>0.00–0.27</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>0.27–0.67</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>0.67–1.29</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>1.29–2.40</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>2.40–20.97</td>
<td>0.043</td>
</tr>
</tbody>
</table>

*This is a lower bound of the true significance.

*Lilliefors significance correction.
The distributions for the lower slope categories do not contain fewer major outliers and do not approximate normal behaviour more than the distributions for higher slope categories.

3.4 Spatial patterns in \( \text{RMSE}_z \) and outliers

Figure 5 shows the spatial pattern in the 95% \( \text{RMSE}_z \) values by county for North Carolina. There is substantial variation in these \( \text{RMSE}_z \) values. A cluster of three counties with high \( \text{RMSE}_z \) values (>20 cm) in the western portion of the State corresponds to areas of relatively steep slopes, but four other counties with similarly high values are found in the much flatter central and coastal areas of the State. The counties with the lowest \( \text{RMSE}_z \) values (<12.5 cm) are found in the central portion of the State, and not near the coast as might be expected based on the initial accuracy requirements which were more stringent for coastal areas. Despite the observed variability, \( \text{RMSE}_z \) values for all counties are below 25 cm, suggesting the overall accuracy of the lidar data for a data collection effort at this scale is very good.

Figure 6 shows the percentage outliers based on the 5th and 95th percentiles of the error distributions by land cover types. The results reveal substantial variability among counties. For example, there are four counties with more than 20% of the observations considered outliers based on the 5th and 95th percentiles of the complete error distributions by land cover type for all counties combined. A review of the original accuracy reports for these four counties revealed no reference to any systematic errors, and the lidar data for each of these counties met both the initial
accuracy requirements (based on 95% RMSEz) and the revised accuracy requirements (based on fundamental accuracy) of the North Carolina Flood Mapping Program.

Despite the variability in the percentage of outliers among counties, there is no evidence that most of the outliers can be attributed to a small number of accuracy assessments with systematic errors. For example, of the 84 counties with accuracy assessment reports, only one county had no outliers and only 13 had fewer than five outliers. And while only four counties have more than 20% outliers, there are 17 counties with more than 15% outliers. So while figure 6 reveals substantial county-to-county differences in the percentage of outliers, the occurrence of outliers is not limited to only a handful of counties but is instead common throughout the entire study area.

3.5 Effect of outliers on RMSEz

The non-normal distribution of the error in lidar elevation data has implications for the statistics employed to characterize the error distribution. The most commonly used statistics in the current guidelines are the RMSEz value based on 100% of the data points for distributions that are assumed to be normal and the 95th percentile for those that are not. Another commonly employed technique is to apply some degree of ‘data trimming’, by removing outliers prior to determining the RMSEz value. Figure 7 shows the effect of this data trimming on the estimates for RMSEz.
Figure 7. Variation of RMSE$_z$ statistic by land cover type with the percentile of the error distribution used in the RMSE calculation.
The first major observation from figure 7 is the general difference in accuracy by land cover type. RMSE\(_z\) values for bare earth and high grass are very similar, and much lower than for the other three categories. Of the other three categories, fully forested is the least accurate. The comparison between urban areas and shrubs, however, reveals an interesting pattern since the two RMSE\(_z\) curves cross at approximately the 92nd percentile; below this percentile the lidar data for shrubs is most accurate while above this percentile lidar data for urban areas is most accurate.

The second observation is the general shape of the curves: gradually increasing for larger percentiles, followed by a dramatic increase at very high percentiles. For a normal distribution there would not be such a dramatic increase, and it therefore does not come as a major surprise that the curve for urban areas shows the least amount of increase for the highest percentiles. The shape of the curves for RMSE\(_z\) at the highest percentiles is a strong illustration of the effect of non-normal behaviour in the form of minor and major outliers. For example, the 95\% RMSE\(_z\) for fully forested is 14.9 cm, while the 99\% RMSE\(_z\) is 20.4 cm and the 100\% RMSE\(_z\) is 28.1 cm. This illustrates the lack of reliability of the RMSE\(_z\) statistic. For urban areas these values are 15.5 cm, 17.1 cm and 18.6 cm, respectively, suggesting that for (close to) normal distributions the effect on RMSE\(_z\) is much less. Another dramatic illustration of the effect on non-normal behaviour is presented by the comparison of the bare earth and high grass categories. Observations for bare earth are assumed to be most accurate due to the lack of influence of buildings and vegetation in the collection and processing of the lidar data. Empirical results confirm that the RMSE\(_z\) values for bare earth are indeed slightly lower than for high grass for almost the entire range of percentiles, but that this is reversed for the complete sample of data points. For example, the 95\% RMSE\(_z\) values for bare earth and high grass are 11.7 and 12.0 cm, respectively, the 99\% RMSE\(_z\) values are 13.5 and 13.7 cm, respectively, and the 100\% RMSE\(_z\) values are 18.6 and 16.1 cm, respectively. Based on the entire sample, therefore, lidar data for high grass would be considered the most accurate. Inspection of the original data reveals that this is the result of only 18 observations (out of 2176 total) for bare earth with an absolute error larger than 50 cm.

4. Discussion and conclusions

This research has presented strong evidence that the distribution of vertical error of lidar elevation data is not normally distributed. While this has been observed by previous research, the current guidelines for elevation data by NDEP and ASPRS still assume that the errors for observations for bare earth (open terrain, no vegetation or buildings nearby) are normally distributed. This provides the justification for the use of the RMSE\(_z\) statistic based on all datapoints for bare earth to derive an assessment of fundamental accuracy. Using a dataset from one of the largest lidar data collection efforts in the world today, this research found no evidence that the errors for bare earth observations are normally distributed. On the contrary, strong evidence of the occurrence of both major and minor outliers was found.

Of the five land cover types considered, only the distribution for urban areas approximated a normal distribution, even though these observations were generally much less accurate than those for bare earth. An examination of the error distribution by slope category revealed no influence of slope on the occurrence of normal behaviour in the distributions. Investigation of the spatial pattern in RMSE\(_z\) values and outliers by county did not reveal any systematic errors associated with selected
accuracy assessment studies, and the occurrence of outliers was instead found to be distributed across the entire study area.

An examination of the influence of outliers on the estimates of RMSE$_z$ values confirmed the lack of robustness of the RMSE$_z$ statistic, with a dramatic increase resulting from a very small number of major outliers. The significance of this research lies in the fact that this was also observed for the bare earth observations, which brings assessment of fundamental accuracy as described in current guidelines into question. A more rigorous and complete characterization of the error distribution is recommended.

Among the limitations of this study is that the data are limited to a single geographic area. While the area is large in comparison to other studies on lidar accuracy, the diversity of landforms encountered is limited and in general the slopes are low to moderate. Furthermore, the method employed to select field survey locations resulted in relatively few locations with high slopes, limiting the analysis of slope as a factor. A second limitation is that only a single DEM resolution was considered (6.1 m) and the error distribution for higher resolution lidar might be somewhat different. A third limitation is that the data accuracy surveys were completed on processed lidar data provided by two commercial vendors. Since these vendors do not disclose their processing methods in great detail, any effects of data processing methods on the occurrence of lidar elevation errors could not be established. Differences in processing methods between the two vendors and any changes over time in these processing methods are not documented, and this remains one of several factors which may explain the observed patterns in lidar data accuracy. Finally, while this study has documented the occurrences of outliers in lidar elevation data, it provides limited insight into the factors which may cause this non-normal behaviour. Factors such as differences between lidar vendors, variations by acquisition dates/flight, reliability of the GPS observations during lidar collection could all play a role, but have not been examined in this study.

One of the major implications of the study is that it questions the use of the traditional RMSE statistic to characterize the accuracy of lidar elevation data, even for bare earth areas which have traditionally been assumed to result in normal error distributions. For small sample sizes in particular a single outlier can greatly skew the RMSE values. It also questions the use of the RMSE statistic in the NSSDA protocol to determine the 95th percentile of the error distribution, similar to earlier work on several different types of spatial data (Zandbergen 2008). Two recommendations emerge from the research findings: (1) employ a certain amount of data trimming (1–5%) or other method of outlier removal prior to determining RMSE$_z$ values; and (2) do not use the RMSE statistic in the NSSDA protocol but derive the 95th percentile directly from the error distribution, provided the sample size is large enough to do this reliably.

The non-normal distribution of vertical error in lidar data has implications beyond spatial data accuracy standards since most error propagation techniques for spatial data are also based on an assumption of normality. Given the occurrence of major errors in DEMs, alternative approaches to error propagation modelling will need to be developed.

In addition to the practical implications for spatial data accuracy standards and error propagation modelling, a better understanding of the distributions of vertical error in digital elevation data can provide insights into the underlying processes which explain the occurrence of errors in spatial data. While the occurrence of vertical errors
in lidar data has been the subject of recent investigations, the ongoing widespread adoption of lidar data as the new standard for collecting digital elevation data requires a more in-depth investigation of the causes behind the observed non-stationary behaviour.

References


