ESTIMATION OF INTERPOLATION ERROR IN DEMS USING STATISTICAL METHODS

Robert PAQUET, Australia

Key words: DEM, interpolation error, ALS, photogrammetry, Delaunay triangulation

SUMMARY

To work with DEMs requires a knowledge of their accuracy, to avoid having an error larger that the level of details to achieve the task. Such problems can lead to unplanned costs, resurveys, disputes, for example in flood studies (spurious wells), deformation detection (trying to detect mining subsidence effects smaller than the error), stockpile volume calculations (large discrepancy in estimation by buyer and seller).
The error can be divided into two components: the observation error (with its many different aspects resulting from the method of observation) and the interpolation error (where heights are estimated between observations). The total error is thus the estimate of the observation error propagated into the interpolation error.
The observation error is often difficult to determine, especially where the sensor is mounted on a platform in motion, such as an aircraft in airborne lidar. It is however more readily estimated for example with theodolites and terrestrial laser scanners.
The interpolation error in a DEM is a function of the roughness of the surface and the point density of the DEM. It does not depend on the platform chosen for the observations.
This article describes a statistical method to estimate the interpolation error in a DEM, regardless of which method was used for the observation. The error is then modelled on the density, using linear regression. The idea is to compare the interpolation error to the level of error required for the task. The accuracy/density model can be used to plan surveys at the observation phase to obtain a level of error adequate for the task.

ESTIMATION OF INTERPOLATION ERROR IN DEMS USING STATISTICAL METHODS

Robert PAQUET, Australia

1. INTRODUCTION

DEMs are a product used widely for engineering applications such as water flow modelling, infrastructure design, ground deformation studies or civil engineering works. Errors in the DEMs can be the source of contract disputes as for example the calculation of volumes, resulting in lengthy legal battles. Assessing the error for ground information where the points are interpolated between the observations is a difficult task, as the error in the observations is propagated to the points, which also have uncertainty due to interpolation. Moreover, some DEMs have their nodes themselves already interpolated from raw data, resulting in a second-generation interpolation. Third generation interpolations are also available: for example photogrammetric data may be used to generate contours, which are digitised in turn to create a DEM, from height information is sought at points interpolated between the nodes.
This article demonstrates two simple statistical techniques to assess the interpolation associated with a given DEM and to possibly model the interpolation, to provide a prediction tool.

2. ESTIMATION OF INTERPOLATION ERROR

2.1 Overview of the data

The data is a DEM generated directly from ALS (airborne laser scanning). The area covered is Islington Park, between the suburbs of Tighes Hill, Islington and Maryville in the inner west of the city of Newcastle upon Hunter, NSW. The data was provided to the author by AAMHATCH for research purposes. It is a typical urban ALS surface, where the observations have been cleared of buildings, trees and infrastructure to provide ground information. The data is a set of 27,729 points. Figure 1 shows a plot of the data set and an aerial photo (courtesy of Google Earth). The data set shown has its number of points reduced to 6,000 points for clarity. Note the coordinates system: the data originally on a MGA grid is shifted and centred on the origin. This is to minimise calculation error propagation due to operations with large numbers. The data provides ground information: the denser areas are in the street and in the park, the more patchy area represents areas where building and tree canopy information was removed to provide ground information.

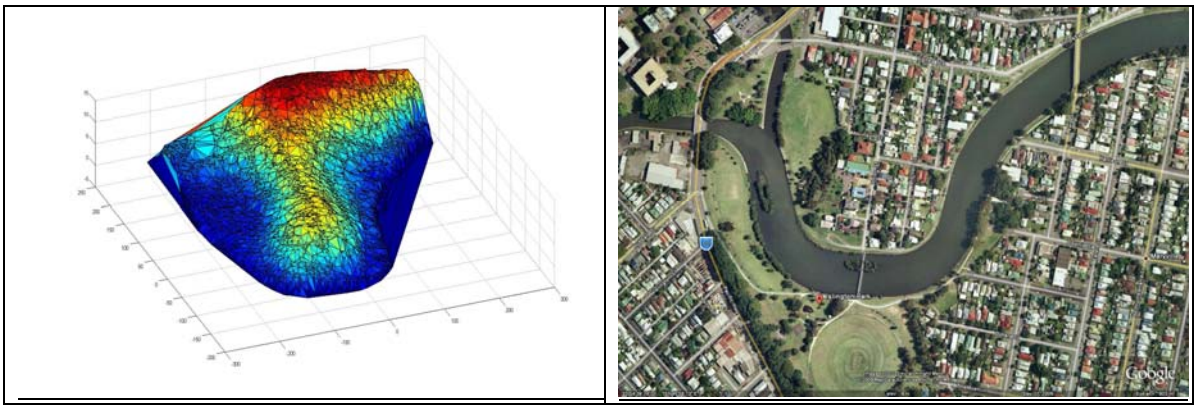

Figure 1: Data and Aerial Photo of Islington Park, NSW

2.2 Bootstrapping Method

2.2.1 Comments about Interpolation

The interpolated height of a point has increasing uncertainty with increasing distance from a measured point (of known height). In a TIN environment, the interpolated height of a mark is calculated using the known heights of the vertices of the triangle enclosing the said mark. The uncertainty is thus at a maximum for marks whose spatial position is at the geometric centre of the enclosing triangle. The interpolation error for several marks in the same triangle will therefore have different values depending of the position of the marks in the triangle. The potential magnitude of the error also varies with the size of the triangle. The interpolation error is also a function of the variation of relief within the triangle. For example the height of any point in a tennis court or a sports oval is not likely to be influenced by the size of the triangle, as the ground has no slope by design. Assessing the mean square root of the interpolation error of a surface is indeed a very complex task.

2.2.2 Bootstrapping

Bootstrapping methods are basically re-sampling methods to obtain complex statistical characteristics of samples which would prove difficult to obtain with traditional analytical methods (Efron and Tibshirani 1998). The processing power of modern personal computers is put to use to re-sample the data to be analysed many times around in order to create a stochastic model from which conclusions can be drawn.

2.2.3 Estimation of interpolation error

To evaluate the interpolation error at a point in the original set is a difficult task, as outlined in Section 2.2.1.

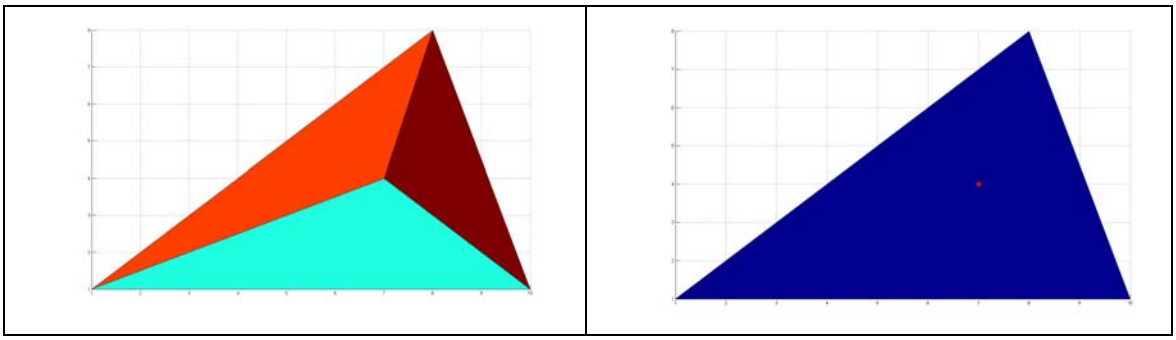

Figure 2: Interpolation Error Calculation

Consider now the triangulated group of 4 points in Figure 2. The re-sampling of this group may delete the mid-point and use the three outer points in the triangulation, replacing the three triangles with one larger triangle. Since we know the coordinates of the 4th point, the interpolation error is calculated as the difference between the known RL of the midpoint and its interpolated RL in the new larger triangle formed.
To calculate the interpolated RL of the point, we use the easting and northing coordinates of the point, and the equation of the enclosing triangle, defined by:

$$Ax + By + Cz - D = 0 \qquad \textbf{(Equation 1)}$$

Where the coefficients A, B, C and D are determined with the coordinates of the vertices P, Q and R of the enclosing triangle, as:

$$\begin{aligned} A &= y_p(z_r - z_q) + y_q(z_p - z_r) + y_r(z_q - z_p) \\ B &= x_p(z_q - z_r) + x_q(z_r - z_p) + x_r(z_p - z_q) \\ C &= x_p(y_r - y_q) + x_q(y_p - y_r) + x_r(y_q - y_p) \\ D &= Ax_r + By_r + Cz_r \end{aligned} \qquad \textbf{(Equation 2)}$$

To obtain a stochastic model of the interpolation errors, the original surface of 27,729 points is treated in the following manner:
- At each iteration, 25 points are randomly selected from the original surface (the 25 points are thus different from iteration to iteration).
- The surface of 27,704 points still in the data set is thinned out randomly to hold 26,000 points. The interpolation error is then calculated for the 25 points.
- This operation described in the two items above is repeated 25 times. Recall that for each of the 25 iterations, the surface of 26,000 points is different, and the 25 points are different. We obtain therefore a sample of 625 interpolation error measurements for a random density of 26,000 points, all generated from a real surface of 27,729 points.
- The next 25 iterations are undertaken in the same manner, with different points, different surfaces, both picked randomly, but with a surface made of 25,750 points, that is, a decrement of 250 points from the previous (and initial) 25 iterations. The original surface has therefore 25 points randomly chosen, then is thinned out off 1,954 points instead of 1,704 points to generate a surface of 25,750 points density.

- The next 25 iterations (i.e. Iterations No.51 to No.75) generate a surface of 25,500 points, thus another decrement of 250 points.
- The number of decrements (of 250 points) is set to 102. The stochastic model is therefore constructed with 625 interpolation errors over 103 different density of the same surface, using only observed data.

The programme did the 64,375 calculations in 2,466.4 seconds.

Table 1: (Partial) Results of Bootstrapping Method

DEM Density (No. of Points)	No of Calculations (25 pts by 25 Iterations)	Mean of RL Differences (m)	Standard Deviation of means (m)
26,000	625	0.101	0.100
25,750	624	0.106	0.100
25,500	624	0.100	0.087
...	...	...	...
20,000	623	0.102	0.092
...	...	...	...
15,000	624	0.109	0.113
...	...	...	...
10,000	624	0.123	0.124
...	...	...	...
5,000	622	0.137	0.134
...	...	...	...
1,000	614	0.193	0.208
750	605	0.216	0.253
500	604	0.221	0.242

Note that the number of calculations does not always add to 625, as the algorithm:
- stops duplication of any of the 25 random points picked for the calculation of interpolation, although it is not likely to happen when the pool surface is large but note that it happens when the surface is small.
- evidently does not return a calculation when the point is outside the perimeter of the surface and an enclosing triangle cannot be found.

A plot of the model density Vs mean interpolation error is shown in Figure 3 below.

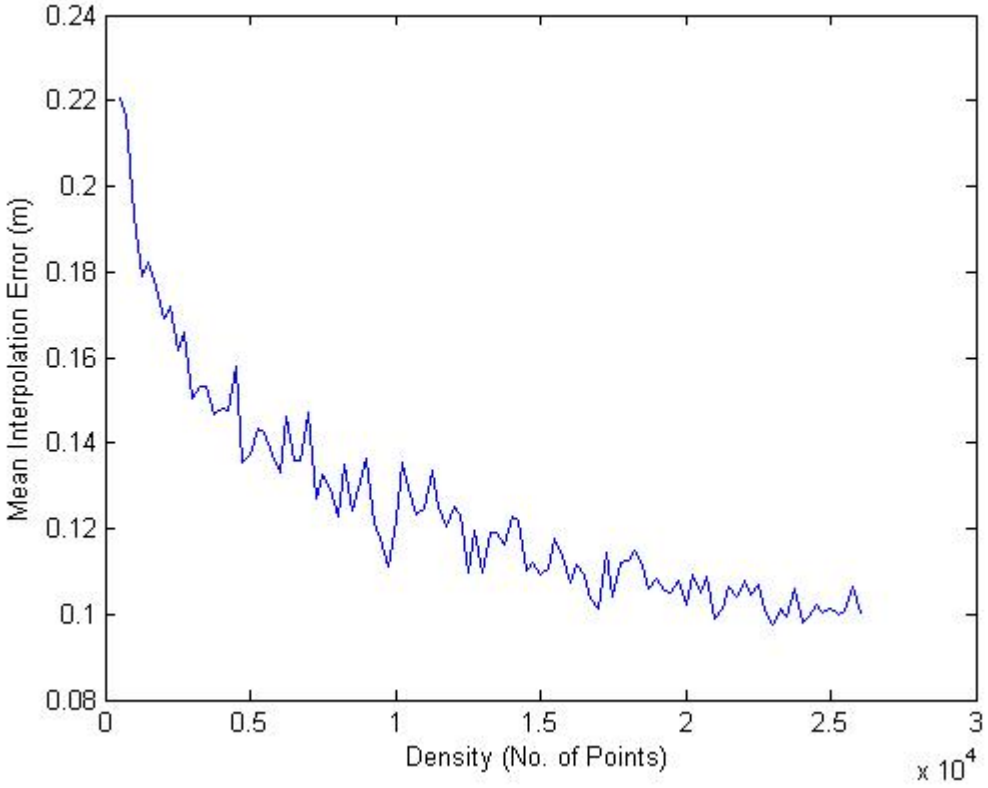

Figure 3: Plot Density Vs Mean Interpolation Error

Figure 3 shows that, as expected, the interpolation error value is decreasing as the density of the DEM is increasing. The interpolation error can then be used in conjunction with the other sources of error to assess the error budget of the DEM.
A generic error model is:

$$RMSE_{DEM} = \sqrt{RMSE^2_{Interpolation} + RMSE^2_{EverythingElse}} \qquad \textbf{(Equation 3)}$$

where 'Everything Else' includes the survey errors, the equipment and method errors, etc (see (Hodgson and Bresnahan 2004) for a detailed example).
The error model is an essential decision tool. A dense DEM may not be warranted for the task, but is very demanding on computing time and may require excessive computer resources for the manipulation of the data, in which case the DEM can be thinned out. On the other hand, the error model may reveal that the error is too large and not fit for purpose.

3. Prediction Model

The aim of this experiment is to create a density-interpolation error model, which could be used to predict one from the other. The statistical used is linear regression. In order to achieve this experiment and validate it, we use a thinned-out DEM to construct the model, which is verified with the original dense data.

3.1.1 Experiment

For this experiment, the original data set of 27,729 points is thinned out once and randomly to a set of 10,500 points. The algorithm presented in Section 2 is used to analyse the resulting surface with 95 decrements of 100 points of 25 iterations each, starting from a surface of density 10,000 points.
The variation of the means and standard deviation of the means is shown in Figure 4 below.

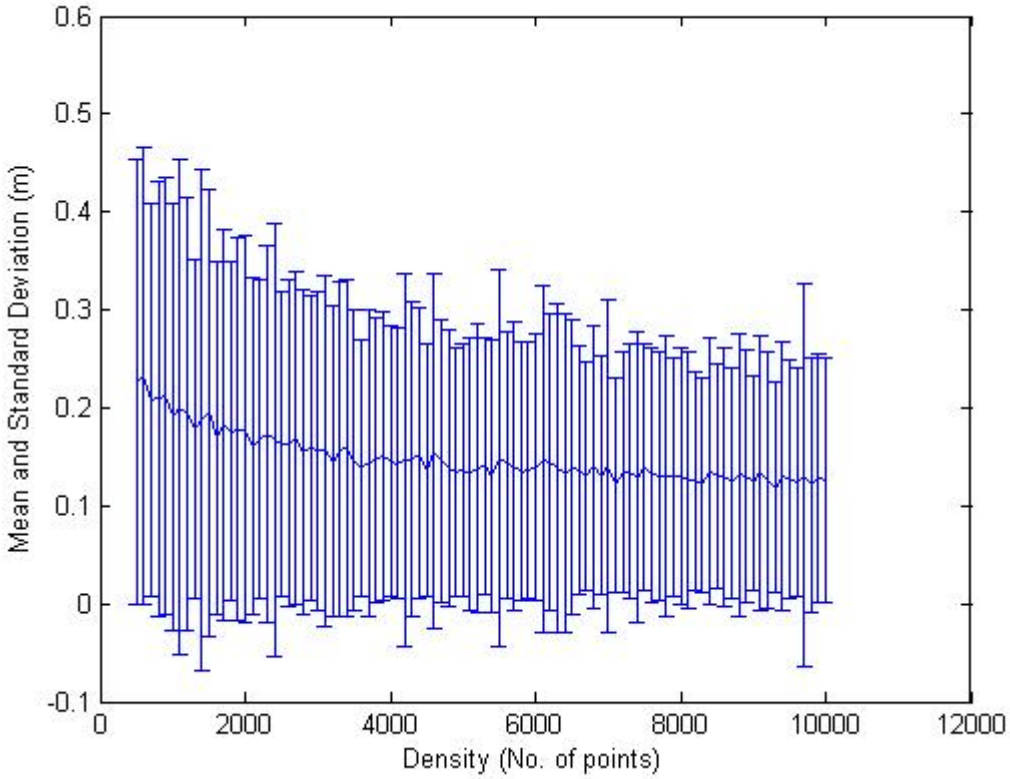

Figure 4: Means and SD of Interpolation Error Vs Density

3.1.2 Linear Model

In order to use linear regression, several data transformations are tried and a transformation using the log of the density and the log of the means is adopted (Kutner, Nachtsheim et al. 2004). The main reason for that choice is the straightness of the mean, as shown in Figure 5.

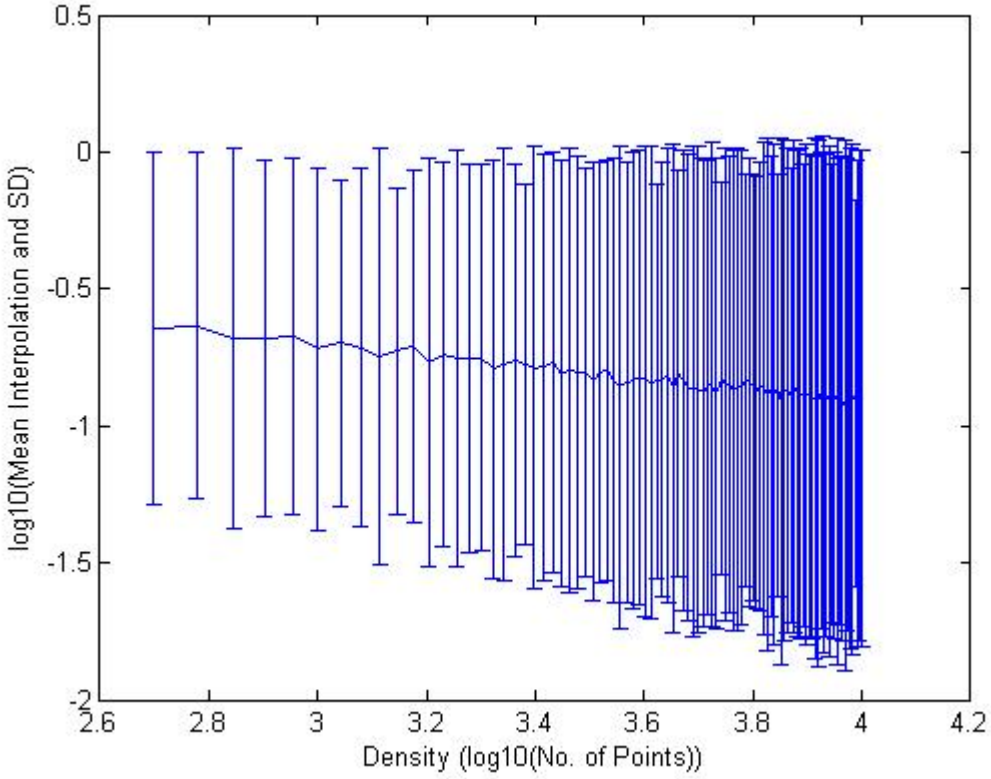

Figure 5: Transformed Data

The line is fitted on the means using linear regression techniques. The equation of the line is:

$$Y = -0.2032X - 0.0944 \qquad \textbf{(Equation 4)}$$

The goodness of the fit of the model and the residual plot are shown in Figure 6.

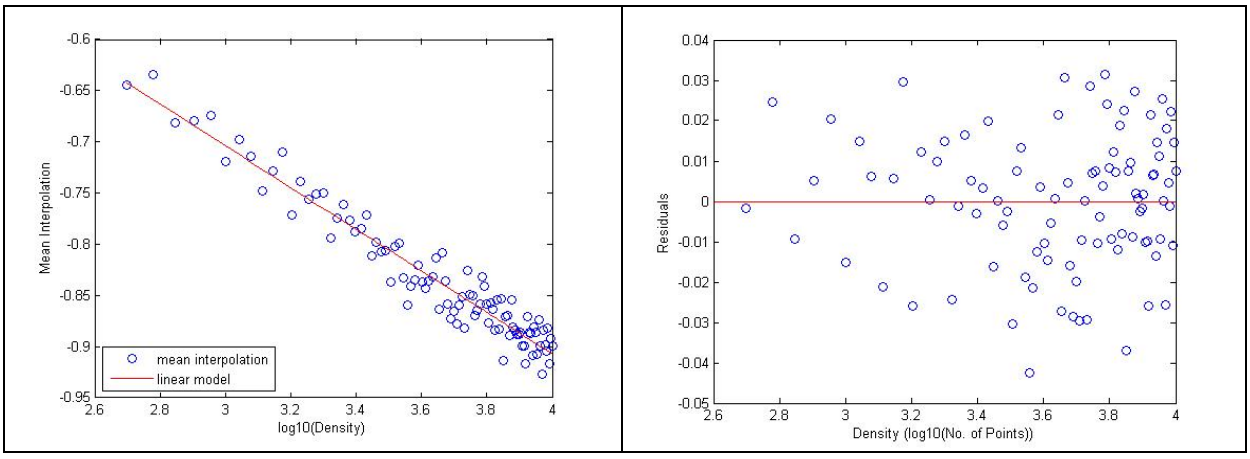

Figure 6: Linear Model Fit and Residual Plot

Figure 5 shows that the transformation adopted straightened the data in such a way that a linear model can be considered. A visual inspection of the residual plot indicates a slight departure constancy of the error variance. The plot is of the 'megaphone' type, with an increasing error variance as the density increase. Despite this, Figure 6 shows an excellent fit of the linear model with well-distributed residuals and no apparent trends.

3.1.3 Prediction et validation

To validate the model, which was created from a unique DEM of 10,500 points, we simply extend the linear model to density of 26,000 points and compare it with the data obtained in Section 3. The validation is done through plots in Figure 7.

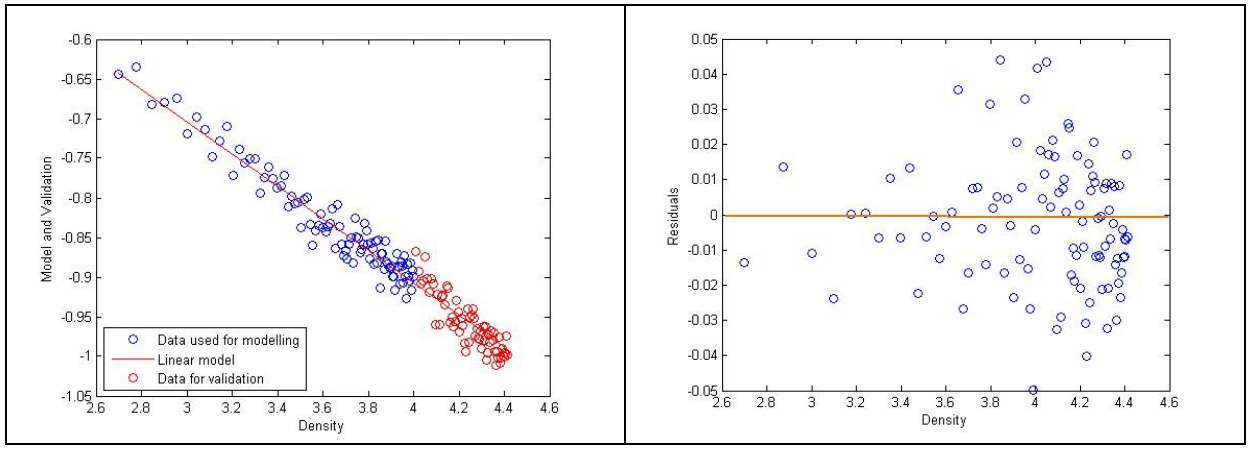

Figure 7: Prediction and Validation

The validation of the prediction shows that the model is over-predicting for the denser DEMs. This is verified by the residual plot which shows more negative residuals at higher density. The data and the model is transformed back and plotted in Figure 8. It confirmed that the model is over-predicting the interpolation error. The prediction is quite accurate if taking into consideration that the prediction is for a DEM almost three times as dense as the DEM used for modelling.

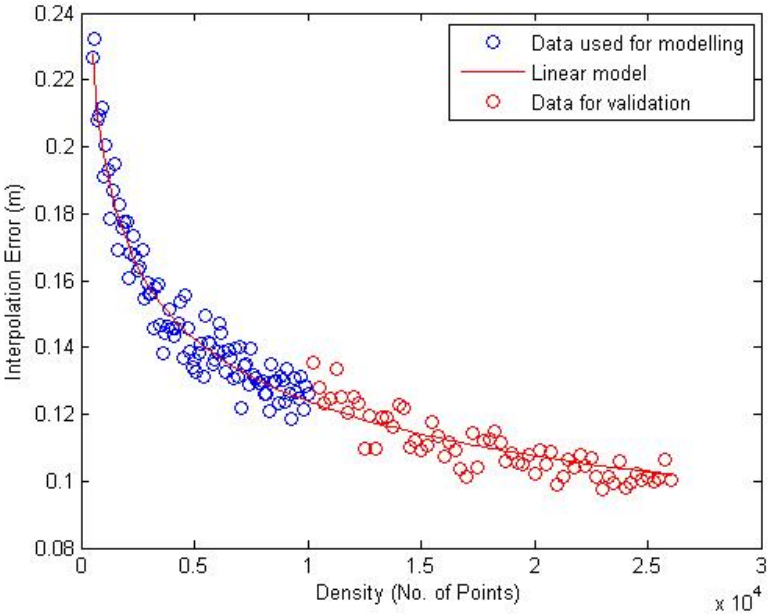

Figure 8: Data and Model (un-transformed)

4. Conclusions

End users of DEMs rarely query the accuracy of this product. Problems can occur where the DEM are used for example for engineering design or volume calculations and the error is larger than the requirements of the intended usage. On the other hand, the DEMs can also be very large files requiring abundant computer resources, in which case it would be beneficial to thin out the DEM before manipulation.
This article describes simple statistical tools to:
- Assess the interpolation error in the DEM, and
- Generate a Linear Regression model of the Density Vs Interpolation Error in order to predict the error outside the available DEM density. These models are useful if the estimated error is larger than the target error, and if a reduction of the error is necessary, for survey planning purposes, although extrapolation has to be used with caution.

The assessment of the interpolation error is using a cross validation (or bootstrapping) statistical technique. These techniques are empirical and helped to obtain statistical information difficult to find with theoretical techniques.

ACKNOWLEDGMENT

I would like to thank AAMHATCH for providing the ALS data used in this article.

REFERENCES

Efron, B. and R. J. Tibshirani, 1998, An Introduction to the Bootstrap, Ed. Chapman & Hall/CRC, 436p., London.

Hodgson, M. E. and P. Bresnahan, 2004, "Accuracy of Airborne Lidar-Derived Elevation: Empirical Assessment and Error Budget", Photogrammetric Engineering and Remote Sensing **70**(3): 331-339.

Kutner, M. H., C. J. Nachtsheim, et al., 2004, Applied Linear Statistical Models, Ed. McGraw-Hill Irwin, 1396p., fifth edt., Boston

BIOGRAPHICAL NOTES

Robert Pâquet has degrees in both civil engineering and surveying, and a doctorate in photogrammetry. His current works and research includes analysis of subsidence data.

CONTACTS

Dr Robert Pâquet
Mine Safety Operations
Industry & Investment NSW
PO Box 344
Hunter Region Mail Centre
NSW 2310
Australia
Tel. +61 2 4931 6647
Fax + 61 2 4931 6790
Email: robert.paquet@industry.nsw.gov.au
Web site: http://www.dpi.nsw.gov.au/