Estimation of DEM Uncertainty Using Clustering Analysis

Laercio M. Namikawa

INPE - Instituto Nacional de Pesquisas Espaciais, C.P. 515, S.J.Campos, SP, 12201, Brazil
laercio@dpi.inpe.br

Abstract

This paper presents a method to estimate the uncertainty in a DEM using Cluster Analysis. The method considers that there are always more than one DEM available for a specific area, therefore, a statistical analysis can be performed and used to create a map with clusters of high and low uncertainty in elevation. The resulting map is particularly important for simulation applications, where the simulation process can apply the uncertainty information to select the best DEM for a region and to define the spatial uncertainty of the simulated result. The method is tested in a region of Sao Paulo State in Brazil, with heterogeneous terrain features. The results show that the method can be used not only in simulation, but also to define geographic regions where data collection can be improved.

Keywords: DEM, Uncertainty, Cluster Analysis, SRTM, ASTER GDEM.

1. Introduction

Environmental process models are useful for many different purposes, such as for simulation of scenarios and to fill gaps in data. These models use information about the environment to define values of the model variables. However, any piece of information has an uncertainty component, that is, any variable value is composed by the true value and the measurement error. Therefore, model results are affected by the uncertainty in the input data

If uncertainty information is available, then the model may create data with information about the reliability of the result. With spatial uncertainty information, map reliability at each location of the model output can be improved. For the hazard mapping example, a high vulnerable area may have uncertainty that is higher than the rest of the region, therefore, additional information with a better accuracy should be used to improve the vulnerability information at that location.

Therefore, the goal of this paper is to create a spatially distributed uncertainty map to be used by environmental process models. Since elevation data is used in many of these models, the uncertainty map to be created is about the spatial variability of uncertainty in elevation data set.

Traditionally, the accuracy of a data set is defined by taking samples of some locations with higher accuracy than the data set and comparing with the data set for the same locations. The drawback is that an uncertainty map cannot be created and the accuracy of the data set will be a global value instead of an uncertainty map. The approach here is to use publicly available data set to define the uncertainty map of a given data set. In this approach, locations are classified as belonging to regions with higher probability of low accuracy. The definition of regions with higher probability uses spatial statists to search for regions with clusters of high uncertainty values. Therefore, the aim is to extract the accuracy map with the support of freely available data and clustering analysis to define regions where the accuracy values are statically lower than in the rest of the area. In this paper, the study case is on elevation data, using SRTM and ASTER G-DEM to estimate the distribution of accuracy of topographic maps in 1:50000 scale of São José dos Campos, Brazil.

2. Uncertainty in Environmental Models

Environmental models are models of physical processes related to the Earth´s environment that occur in geographical scales. The model output accuracy is related to the accuracy of the model logic and to the accuracy of input data. Uncertainty is a known characteristic in all geographical data and one should use the most accurate data. However, the information about accuracy is not easily available or if available, it is not in a spatially distributed format. In this paper, data about the elevation is used as an example of input data for environmental models. Digital Elevation Models (DEMs) contain uncertainties (Hunter and Goodchild 1997; Canters et al. 2002). DEM represents the spatial distribution of elevation, which is a numeric value and accuracy values can be easily added to it. Unfortunately, current standards for DEM do not define a requirement for spatially distributed accuracy and require only a global value. For example, USGS specifies the desired accuracy standard of Level-1 DEM to be below 7 meters calculated using Root Mean Squared Error (RMSE) at few locations (USGS 2003). IBGE data should conform to the National Cartography Commission (CONCAR) standards, which defines the best standard to be half of the contour lines interval (Brazil, 1984).

3. Clustering Analysis

Since this paper proposes the generation of a map showing where the DEM has clusters of areas where the accuracy is statistically significant lower than the average using spatial statistics. This analysis considers elevation to be a random variable; therefore, each DEM is a sample of the "real" DEM and the probability density function can be defined from a set of different DEMs. When a global accuracy value is defined, the assumption is that errors are random, and every location has the same probability of being within the stated accuracy. If the 90% RMSE is 10 m, then the actual value at a location has 90% probability of being within 10 m from the stated one. Clusters are detected using the method proposed by Rogerson, 2001, which searches significant peaks on a surface representing a standardized measure that has been smoothed by a Gaussian kernel. A critical value is defined for a probability and the clusters are within the peaks with values higher than this value.

3.1. Computing the Standardized Measure

The standardized measure is the *zscore* (*zs*), which is computed for normal distribution by:

$$zscore = \frac{x - \mu}{\sigma} \quad (1)$$

The *zs* is computed at each location of the DEM, using the global mean μ and the global σ. The measure that is used to provide the value at each point is the *coefficient of variation* (*CV*), which is the relative measure of the dispersion, that is, how relative *σ* is in relation to the mean *μ*. The *CV* is computed at each location of the DEM, using the local mean μ_{rc} and the local standard deviation σ_{rc}. Therefore, for a grid representing the DEM and with location defined in terms of row *r* and column *c* coordinates, the *coefficient of variation* (CV_{rc}) is computed by:

$$CV_{rc} = \frac{\sigma_{rc}}{\mu_{rc}} 100\% \qquad (2)$$

3.2. Gaussian Kernel

The *zs* values are smoothed since targets are clusters and not individual locations. The selection of the smoothing σ for the Gaussian kernel is based on the best one to filter random differences and to enhance clusters. The Gaussian kernel is created by a weighted sum of the grid cell neighbours, with weights given by:

$$w_{ij} = \frac{e^{\frac{-d_{ij}^2}{2\sigma^2}}}{\sqrt{\pi\sigma}} \qquad (3)$$

where *σ* is the standard deviation of the Gaussian kernel and d_{ij} is the distance from the center cell *i* to neighbour cell *j*. The weights are applied at cell *i* by:

$$y_i = \frac{\sum_j w_{ij} z_j}{\sqrt{\sum_j w_{ij}^2}} \qquad (4)$$

where y_i is the smoothed value of *zs* at the center cell *i*, w_{ij} is the weight for the cell at the distance from center cell *i* to neighbour cell *j*, and z_j is the *zs* of cell *j*. Note that the distance unit is number of cells.

3.2. Critical Value

Clusters of statistically significant high values of *zs* are defined based on a critical value M^*, selected based on the probability of finding a value greater than M^* at a selected significance level *α*. The M^* is computed by (Rogerson 2001):

$$p\left(\max z_i > M^*\right) = \frac{AM^*\varphi\left(M^*\right)}{4\pi\sigma^2} + \frac{D\varphi\left(M^*\right)}{\sqrt{\pi}\sigma} + \left[1 - \Phi\left(M^*\right)\right] \qquad (5)$$

where *A* is size of the DEM region in cell size units, *D* is the caliper diameter (*D* is half of the sum of the grid rectangle height and width), *σ* is the standard deviation of the Gaussian kernel, *φ* and *Φ* are the probability density function and the cumulative distribution function of the normal distribution, respectively.

The third term of Equation (5) is small enough to be discarded (Rogerson 2001), therefore, Equation (5) is simplified and approximated to:

$$M^* = \sqrt{-\sqrt{\pi}\ln\left(\frac{4\alpha\left(1+.81\sigma^2\right)}{A}\right)} \qquad (6)$$

Equation (6) is valid only if *A* is smaller than 10000 or if *σ* is not smaller than one. If *A* is greater than 10000, the approximation can be only used if:

$$\sigma_t \big/ \sqrt{A} > 0.01 \qquad (7)$$

where σ_t is the total smoothing given by $\sigma_t = \sqrt{\sigma_0^{\,2} + \sigma^2}$, with σ_0 equal to 10/9.

For most DEMs, *A* is greater than 10000, and *σ* is expected to be between 1 and 4. In these conditions, the restriction of Equation (7) is not satisfied, but the critical value computed by the approximation is only slightly smaller than the one from Equation (6) (Rogerson 2001).

4. Case Study

In this paper, clusters of high values of uncertainty in elevation data presented to define a spatially distributed accuracy map. The resulting accuracy map is qualitative and highlights regions in the DEM that should be checked carefully. The region around São José dos Campos, Brazil, was selected due to its diverse geomorphological features, with hilly areas, floodplain, escarpments and cuestas.

The analysed elevation is provided by IBGE, the Brazilian Institute for Geography and Statistics in 1:50000 scale. The topographic map, identified by the label São José dos Campos SF-23-Y-D-II-1, includes a vector file with contour lines representing altimetry, and covers the region between coordinates 23° south, 46° west and 23°15′ south, 45°45′ west. The contour lines were used to create two DEMs with 30 meter spatial resolution. The first one was created using the nearest neighbour interpolator and the second with the Triangular Irregular Network (TIN).

The comparison data sets are the SRTM and the G-DEM elevation data. Since elevation is considered a random variable, each one of these three data sets represents independent realizations of the elevation probability density function. Therefore, the accuracy of the DEM is not relevant to define the clusters.

SRTM elevation data is generated from data captured by the Shuttle Radar Topography Mission and is available globally at 3 arc-second resolution from ftp://edcsgs9.cr.usgs.gov/pub/data/srtm. The stated accuracy standard for this data is 16 m vertical 90% linear error. ASTER G-DEM is also available globally, but at 1 arc-second spatial resolution, from http://www.gdem.aster.ersdac.or.jp. The accuracy of G-DEM is estimated to be better than SRTM data; however, since elevation is extracted from stereo pairs of images, the accuracy is variable and dependent on the quality of the control points.

4.1. Clustering Analysis of IBGE Elevation

The three DEMs (IBGE, SRTM and G-DEM) were used to define the mean and the σ of the elevation probability density function at each point of the SRTM grid cell. Note that since the SRTM has the lowest spatial resolution, the other two DEMs were re-sampled and re-projected using the nearest neighbour interpolator to coincide with the original SRTM 3 arc-second locations.

The *cv* of the IBGE elevation was computed at each location using Equation (2). Next, the global mean and standard variation of the *cv* were computed in order to compute the *zs* using Equation (1). The Gaussian kernel was applied to the *cv* grid using two different σ, 1 and 2. The objective of using two different smoothing factors is to evaluate their influence in the detection of the clusters. The weights for σ 1 and 2 were computed using Equation (3). The critical value M^* was computed for the two *σs* of the Gaussian kernel ($\sigma = 1$ and $\sigma = 2$), region size A equal to 90000 (grid size is 300 rows by 300 columns) and significance level α equal to 0.05, using Equation (6). The critical value M^* are 4.692 for $\sigma = 1$ and 4.529 for $\sigma = 2$.

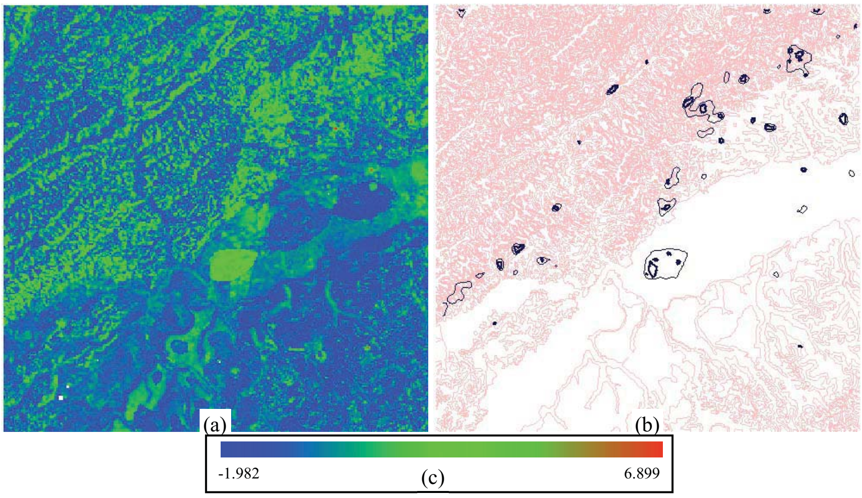

Figure 1: Clustering analysis of the IBGE DEM interpolated by nearest neighbour. (a) *zs* distribution. (b) Contour lines in pink colour, clusters of low accuracy for σ one in thick dark lines, and for σ two in thin dark lines. (c) Colour legend for the *zs* represented in (a).

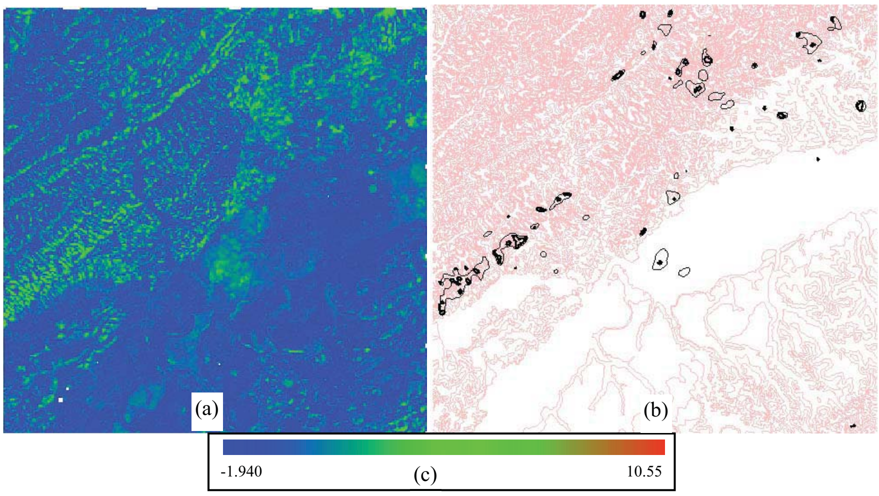

Figure 2: Clustering analysis of the IBGE DEM interpolated using TIN. (a) *zs* distribution. (b) Contour lines in pink colour, clusters of low accuracy for σ one in thick dark lines, and for σ two in thin dark lines. (c) Colour legend for the *zs* represented in (a).

4.2. Results Analysis

Figure 1 shows the clustering analysis of the IBGE DEM interpolated by the nearest neighbour method. The *zs* statistics is shown in Figure 1.a, with the colour code shown by Figure 1.c, and it indicates that there must be regions of significant low accuracy. Using the critical value M^* equal to 4.692 for σ one, the regions of significant low accuracy ($\alpha = 0.05$) are 1.8176 Km^2 in size, and are inside the thick dark lines in Figure 1.b. The thin lines in Figure 1.b indicates the regions of significant low accuracy ($\alpha = 0.05$) for σ two (critical value $M^* = 4.529$), with 14,158 Km^2 in size. Note the particular geomorphology of the case study region indicated by the pink colour lines representing the contour lines from the IBGE topographic map.

The clustering analysis of the IBGE DEM interpolated using TIN is shown in Figure 2. The spatial distribution of the zs is shown in Figure 2.a, with the colour code shown by Figure 2.c. Using M* = 4.692 for σ one, the regions of significant low accuracy (α = 0.05) are 1.710 Km^2 in size, and are inside the thick dark lines. The thin lines indicates regions of significant low accuracy (α = 0.05) for σ two, with 12,227 Km^2 in size. Note that the difference in the size between the interpolators is not large indicating that results are not influenced by the interpolators.

5. *Conclusion*

This paper proposes a method to create spatially distributed uncertainty information for elevation data from any source, in order to complement the information about the existing accuracy. Traditionally, accuracy for a data set is stated in terms of a global measure. Since there must be areas with lower accuracy in the whole region, the spatially distributed uncertainty map created by the proposed method can be used to verify if the DEM is suitable for the application or to direct data collection to improve data where it is more important.

The method uses the cluster analysis for data in a regular grid proposed by Rogerson, 2001. The clusters of low accuracy are detected on a grid of standardized measure, the *zscore* statistics generated from the map of the *coefficient of variation* for the DEM to be analyzed. This map of a *coefficient of variation* is created from the local statistics extracted using two another DEMs. In this paper, DEMs from SRTM and the G-DEM from ASTER were used to compute the local statistics.

In the study case, the method showed that can be used to create the uncertainty map. These maps can be set by the smoothing defined by the standard deviation. Therefore, if the user wants to have smaller clusters, small values of the Gaussian kernel standard deviation can be used. In the study case, the regions size decreased 7 times when the standard deviation changed from two to one (from 12 Km^2 to 1.7 Km^2 for the clusters extracted using the TIN interpolation). The numeric results indicate that the regions to be carefully analyzed decreased from 12412 Km^2 to 12 Km^2 (for the TIN interpolation case with standard deviation two); therefore, the costs in additional data collection could be one thousand times smaller (if costs are linearly proportional to region size.

References

Brazil (1984). Instruções Reguladoras das Normas Técnicas da Cartografia Nacional. *Decreto Número 89817*.

Canters, F., W. D. Genst, et al. (2002). "Assessing effects of input uncertainty in structural landscape classification." *International Journal of Geographical Information Science* 16(2): 129-149.

IBGE (1973). *São José dos Campos SF-23-Y-D-II-1*, IBGE.

Hunter, G. J. and M. F. Goodchild (1997). "Modeling the uncertainty of slope and aspect estimates derived from spatial databases." *Geographical Analysis* 29(1): 35-49.

Rogerson, P. A. (2001). "A Statistical Method for the Detection of Geographic Clustering." *Geographical Analysis* 33(2): 215-227.

USGS. (2003). "*USGS Digital Elevation Model Data*" Retrieved 03/12/2005, 2003, from http://edc.usgs.gov/glis/hyper/guide/usgs_dem.